Mathematical and Numerical Analysis of Flow in Deformable Porous Media

Jakub Wiktor Both

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2019

UNIVERSITY OF BERGEN

Mathematical and Numerical Analysis of Flow in Deformable Porous Media

Jakub Wiktor Both

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 27.11.2019

Year: 2019

Title: Mathematical and Numerical Analysis of Flow in Deformable Porous Media

Name: Jakub Wiktor Both

Print: Skipnes Kommunikasjon / University of Bergen

Preface

This dissertation is submitted as a partial fulfillment of the requirements for the degree of Philosophiae Doctor (PhD) at the University of Bergen. The advisory committee has consisted of Florin Adrian Radu (University of Bergen), Kundan Kumar (Karlstad University, University of Bergen), and Jan Martin Nordbotten (University of Bergen).

Acknowledgments

Doing a PhD has been a journey – not only through exciting mathematics but also to a new country, to new friendships, to a new home. There is several people worth a special mention who have accompanied me on this journey and who I owe my sincere gratitude and deepest appreciation.

Foremost, this PhD thesis would not have been possible without my advisers, Florin Adrian Radu, Kundan Kumar, and Jan Martin Nordbotten. I want to express my deepest gratitude to them for their excellent guidance and continuous support. Thanks for giving me the freedom to pursue own ideas, being patient when longer works take their time, allowing me to fail, and yet helping me with constructive advice. Florin, special thanks go to you. Your enthusiasm and expertise was invaluable, ultimately letting me become a better mathematician. I particularly acknowledge your encouragement, constantly giving me opportunities to shape my career. Thank you Kundan for your enthusiasm and meticulous comments in all our works. Thanks to both of you, Kundan and Florin, for always sharing your guests with us PhD students and giving us the opportunity to discuss our work with many excellent international researchers – when not sharing your badminton skills with us. Last but not least, thank you Jan for sharing your knowledge, experience, and enthusiasm with me. That has always been an immense inspiration. I am deeply grateful for your support and the constant challenges – both in English and Norwegian – always resulting in significant improvements.

I am very grateful to have been part of the Porous Media Group. The last four years have been an amazing time, and I would like to show my greatest appreciation to all of my colleagues! Thank you Eirik for opening the door to your office, Uni CIPR, and finally the PMG by introducing me to Florin. Your help was essential for me ending up in this great environment. Besides discussing mathematics resulting in collaborations, I truly enjoyed the many distractions, great dinners, ski trips, and everyday lunch breaks. Special thanks go to Alessio, David, Erlend, Manuel, Michael, and Wietse. I also thank all friends of PMG, in particular, Uwe Köcher from Hamburg for our collaboration, and Evgueni and Stefano for sharing office with me for the last years.

During my PhD, I had the great opportunity to visit several interesting places. Particularly two longer exchange visits have shaped this thesis. I would like to show my greatest appreciation to Sorin Pop for hosting me several times at Hasselt University and continuously supporting my career. Likewise, I am deeply grateful to Ivan Yotov for being a great host during my stay at the University of Pittsburgh. Thanks to both of you for broadening my horizon, the fruitful collaboration, and helping me solving one of the most cumbersome riddles of this thesis!

Some of the background in my thesis I learned from previous mentors at RWTH Aachen University. I would like to express my gratitude to Lutz Pauli and Philipp Knechtges who were the first to introduce me to the world of research, shaped my coding skills during my time at CATS, and arouse my interest in Schur complements. Likewise, I am indebted to Oliver Sander for introducing me to DUNE and letting me be a part of dune-functions despite my little knowledge.

I would like thank my friends and family in Bergen and Aachen. Special thanks go to my parents for their exclusive commitment to offer everything possible for their children.

Last but not least, I owe my deepest gratitude to my most beloved ones. Anna-Lena, thank you so much for your continuous support, your love, your patience, and for being an endless source of energy. Isak, thank you for being the absolutely best distraction of the last two years. Your positivity, happiness, and daily increasing interest in life are truly the greatest inspiration. Without the two of you I wouldn't be who I am today. Thank you both for being THE essential part of this journey!

Abstract

This dissertation focuses on the mathematical analysis and numerical solution of coupled deformation- and flow-related processes in porous media, modeled by the theory of poro-elasticity. It comprises contributions within three topics: the gradient flow structures of quasi-static thermo-poro-visco-elastic processes in porous media; the numerical analysis and fine-tuning of the fixed-stress split for Biot's consolidation model; and the well-posedness analysis and robust solution of unsaturated poro-elasticity.

The fluid-structure interaction of fluids and porous solid materials is of great relevance for various applications within geotechnical, reservoir, structural, and biomedical engineering. Depending on the application different characteristics of a poro-elastic system may be relevant. In order to correctly predict such complex systems, accurate well-posed models are needed – as well as their robust numerical solution. As part of this thesis, a unified *gradient flow framework* is established for the modelling of coupled hydro-mechanical processes in porous media driven by the dissipation of energy. By involving thermodynamic knowledge, it allows for instance for the modelling of non-linearly compressible single-phase flow or non-Darcy flow within a non-linearly elastic (but geometrically linear) or visco-elastic, solid matrix – besides classical, linear poro-elasticity. The framework lays a foundation for the unified mathematical and numerical analysis of these models. In particular, well-posedness is established based on the abstract theory of doubly non-linear evolution equations. Furthermore, block-partitioned iterative solvers, exploiting the inherent block structure of coupled problems, are naturally developed by means of block-coordinate descent methods (with optional line search) for block-separable minimization problems, arising from the discretization of gradient flow formulations. Those are equipped with strong robustness properties under mild conditions, and allow for exploiting tailored, individual solver technology for the physical subproblems. On the one hand, the framework covers well-established theory fortifying its capability. For instance, applied to Biot's consolidation model, the widely-used fixed-stress split arises naturally, and guaranteed convergence rates consistent with the literature are derived. On the other hand, novel splitting schemes with guaranteed convergence rates are determined for more involved models as, e.g., linear poro-visco-elasticity. After all, the methodology applies equally to linear as to non-linear problems.

Biot's quasi-static consolidation model constitutes the simplest coupled model accounting for single-phase flow in a linearly elastic porous medium. At the same time it represents the prototype for any coupled poro-elasticity model. Consequently, a thorough understanding of this simple model is required in order to optimally solve more involved models. In this thesis, we gain deeper theoretical and practical understanding of the popular *fixed-stress split* for Biot's quasi-static consolidation model. More precisely, we consider a fixed-stress split with variable stabilization allowing for tuning the performance. Parameter-robust convergence of the splitting scheme is established for fully heterogeneous media. Furthermore, influences on the practically optimal stabilization parameter, leading to a minimal number of iterations, are practically and theoretically identified. The studies suggest that the optimal stabilization parameter does not only depend on mechanical and coupling material parameters as found in the literature. Instead, it also depends on fluid flow parameters, the stability of the spatial discretization, the computational domain, and boundary conditions. Alternatively to tuning the performance (or in addition), relaxation can be applied as, e.g., line search, based on the gradient flow structure of poro-elasticity.

The model for *unsaturated poro-elasticity* is a non-linear extension of Biot's quasi-static consolidation model. It results from the reduction of the general model for two-phase flow in deformable porous media to the unsaturated zone, i.e., effectively one fluid phase is simply neglected. Typical examples, in which this simplification is acceptable, origin from geotechnical engineering and building engineering, e.g., the stability analysis of dikes, or the drying shrinkage and cracking of cement. In this dissertation, unsaturated poro-elasticity is studied from two perspectives, constituting a step towards a better mathematical and numerical understanding of multi-phase flow in deformable porous media. First, the existence of weak solutions is established. Second, a robust block-partitioned linearization is provided, based in the fixed-stress split, and convergence is theoretically proved. Further acceleration is proposed using the Anderson acceleration, complying with the block-partitioned character of the solver. Theoretical justification is also provided for the Anderson acceleration to both accelerate contractive fixed point iterations, and to increase the robustness of non-contractive fixed point iterations. Thereby, the Anderson acceleration constitutes an adequate alternative to line search techniques – in particular in the context of non-linear, block-structured problems.

Outline

This dissertation consists of two parts. The scientific background is introduced in Part I, followed by the scientific results in Part II.

Part I consists of five chapters. Chapter 1 serves as an introduction to the subject of the mathematical modelling and the numerical approximation of coupled hydro-mechanical processes in porous media – in short, poroelasticity. In Chapter 2, the mathematical model for two-phase flow in deformable porous media is derived by means of the theory of porous media. It serves as basis for the reduction to the model of unsaturated poroelasticity and Biot's quasi-static consolidation model. In addition, connection between the frameworks of generalized and classical gradient flows is presented. Chapter 3 summarizes relevant tools from functional analysis and convex analysis useful for establishing the well-posedness of partial differential equations in continuous and discretized form. Chapter 4 is concerned with the numerical solution of coupled problems. After some general words on the numerical approximation of linear and non-linear problems by discretizations and iterative solvers, general ideas of block-partitioned solvers for linear saddle point problems and block-separable, convex minimization problems are presented. Finally, Chapter 5 summarizes the scientific contributions of the articles included in Part II and presents an outlook on future research.

Part II contains the scientific papers, which are grouped as main and related works. The main results consist of the following six scientific articles:

Paper A Both, J.W., Kumar, K., Nordbotten, J.M., and Radu, F.A. (2019), *The gradient flow structures of thermo-poro-visco-elastic processes in porous media*, In review.
arXiv:1907.03134 [math.NA]

Paper B 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.
doi:10.1016/j.aml.2016.12.019

Paper C Both, J.W., and Köcher, U. (2019), *Numerical Investigation on the Fixed-Stress Splitting Scheme for Biot's Equations: Optimality of the Tuning Parameter*. In Numerical Mathematics and Advanced Applications ENUMATH 2017, Lecture Notes in Computational Science and Engineering 126, pg. 789–797, Springer.
doi:10.1007/978-3-319-96415-7_74

Paper D Both, J.W., Pop, I.S., and Yotov, I. (2019), *Global existence of a weak solution to unsaturated poroelasticity.*
arXiv:1909.06679 [math.NA]

Paper E Both, J.W., Kumar, K., Nordbotten, J.M., and Radu, F.A. (2019), *Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media*. Computers & Mathematics with Applications 77(6), 1479–1502.
doi:10.1016/j.camwa.2018.07.033

Paper F Both, J.W., Kumar, K., Nordbotten, J.M., Pop, I.S., and Radu, F.A. (2019), *Iterative Linearisation Schemes for Doubly Degenerate Parabolic Equations*. In Numerical Mathematics and Advanced Applications ENUMATH 2017, Lecture Notes in Computational Science and Engineering 126, pg. 49–63, Springer.
doi:10.1007/978-3-319-96415-7_3

Additionally, the following two supplementary articles on related work are included:

Paper G 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, 179–194.
doi:10.1002/nme.6130

Paper H Both, J.W., Kumar, K., Nordbotten, J.M., and Radu, F.A. (2017), *Iterative Methods for Coupled Flow and Geomechanics in Unsaturated Porous Media*, In Poromechanics VI: Proceedings of the Sixth Biot Conference on Poromechanics, pg. 411–418, ASCE.
doi:10.1061/9780784480779.050

Contents

Part I

Scientific Background

Chapter 1

Introduction

Poroelasticity comprises a continuum theory combining the fields of structural mechanics and fluid mechanics in the context of porous media. It is is concerned with the macroscopic description of the interaction between fluid flow and the mechanical deformation of porous solids. These processes are two-way coupled. A macroscopic material deformation enforces fluid flow, and vice versa an increase in fluid pressure leads to the compression of solid grains on the microscale. Such a coupling is relevant in a variety of disciplines, from reservoir engineering to biomedical applications, giving rise for the need of a thorough understanding and accurate description.

1.1 Mathematical modeling of poroelasticity

Already early on, the relevance of the fluid-structure interaction of fluid and solid materials in porous media has been recognized by the engineering community. The first discoveries of fundamental mechanical effects of liquid-saturated porous solids have been made by Fillunger in the early 1910s [63]. Twenty years later, Fillunger founded the concept of the modern theory of liquid-saturated porous solids [77]; Terzaghi developed the one dimensional consolidation theory and introduced the concept of effective stress [149]; and Biot extended Terzaghi's theory to three dimensions [22]. All together founded the modern era for the effective modeling of fluid flow in deformable porous media in the language of continuum mechanics [63].

In the mid-twentieth century, especially the works by Biot governed the theory of poroelasticity. A central contribution has been a linear model accounting for single-phase flow in linearly elastic porous solids [22] – now called *Biot's (quasi-static linear) consolidation model*. It is based on fundamental physical laws including mass and momentum balance, combined with constitutive laws as Darcy's law and the effective stress concept. The model serves as prototype for poroelasticity models.

Today, the need for accurate models is still immediate. With the theory of porous media being applied to various fields and applications of societal relevance within, e.g., geotechnical, structural, and biomechanical engineering, multiple complex processes next to a hydro-mechanical coupling are of increased interest. This involves, for instance, thermal, hyper-elastic, visco-elastic, plastic, and chemical effects [57]. Of particular interest are strongly coupled hydro-mechanical processes in porous media saturated by multiple fluids. Relevant applications include CO_2 storage in depleted reservoirs [25, 88, 118], geothermal energy storage in naturally fractured reservoirs [134], compaction of reservoirs in the course of hydrocarbon production [89], swelling and drying shrinkage of building materials as concrete, potentially resulting in desiccation cracks [59], soil subsidence due to groundwater withdrawal for drinking water supply or industrial and agricultural purposes [147], and coupled stress and seepage analysis of dams [141] – just to mention a few. Recent models accounting for multi-phase flow in deformable porous media extend Biot's consolidation model to a set of highly non-linear and potentially strongly coupled partial differential equations [57, 103]. Motivated by geotechnical applications, *unsaturated poroelasticity* considers the simplified setup of a porous medium filled with two fluids (e.g., water and air) of which one phase (air) can approximately be neglected.

From a mathematical point of view, Biot's consolidation model has been studied particularly well. The existence, uniqueness and regularity of solutions has been established under various sorts of boundary conditions [14, 136, 161]. Recently, an increased interest has been showed in the analysis of linear and non-linear extensions of Biot's consolidation model, involving, e.g., dynamic and viscoelastic effects, deformation-dependent permeabilities, and fractured media, see the references within [36]. However, the analysis of various problems is still open, e.g., unsaturated poroelasticity as modeled by the theory of porous media [103].

1.2 Numerical solution of poroelasticity

Typically, analytical solutions of coupled poroelasticity models are not at hand, and large-scale experimental studies are impractical. On the other hand, computer simulations allow for the investigation of various scenarios in short time and under reduced cost.

Due to limited computer power, the numerical approximation of fully coupled poroelasticity models had not been feasible for decades. Instead, tailored numerical technology has independently been developed within the two branches that constitute poroelasticity: structural mechanics and fluid mechanics. Due to different needs, different numerical methods, as discretization methods and solver technologies, have been established.

Nowadays, computers are powerful enough. But compared to the advances within the two subbranches, the development of the accurate and efficient numerical approximation of poroelasticity models still slightly lags behind. For instance, the fundamental introductions of the effective stress concept by Terzaghi [149] and the generalization of Darcy's law to multiphase flow by Muskat [112] date back to the same year, 1936, demonstrating simultaneous interest; yet, the technology for numerically approximating flow in non-deformable porous media seems more mature than for flow in deformable media.

One of the major computational bottlenecks of the accurate numerical approximation of coupled, potentially non-linear models is typically the numerical solution of large algebraic (non-linear) systems of equations. These systems naturally inherit a block structure. Thus, block-partitioned solvers – either as iterative solver or preconditioner for a monolithic Krylov subspace method – have been of increased interest. One primary advantage is that this approach allows for flexible code design, enabling the well-developed solver technologies established in the fields of structural and fluid mechanics.

A particularly frequently used block-partitioned solver – also continuously recurring in this thesis – is the *fixed-stress split* [135]. In this method, one iterates back and forth between updating displacement and pressure unknowns separately by solving the momentum equations and stabilized mass balance equations, respectively, until convergence to the solution to the coupled problem is reached. The popularity of the method originates in its simplicity and unconditional robustness in the context of Biot's consolidation model [98, 111], and its allowance for flexible code design.

Despite the increased interest in complex (non-linear) poroelasticity models, the development and numerical analysis of robust block-partitioned solvers for such seem not yet fully understood. The classical fixed-stress split often constitutes a prototype for the development of block-partitioned solvers for more involved poroelasticity problems, cf., e.g., [34, 157]. By that, a thorough understanding of the plain fixed-stress split is vital, and insights may be transferred to more involved problems.

1.3 Main contributions

This dissertation contributes towards the mathematical analysis of coupled poroelasticity models, and the development and numerical analysis of iterative solvers, in particular block-partitioned solvers. The main contributions of this thesis are as follows:

1. **Identifying inherent gradient flow structures of thermo-poro-visco-elasticity.**
   In Paper A, a generalized gradient flow framework is established for the modeling of quasi-static coupled thermo-poro-visco-elasticity. Specific models are derived by identifying appropriate free energy and dissipation mechanisms driving the evolution. The gradient flow formulation is exploited for both the analysis of the well-posedness by employing the theory of doubly non-linear evolution equations, and the systematic development of robust block-partitioned solvers by utilizing the theory of convex minimization. In view of Biot's consolidation model, the general concepts allow for identifying the undrained and fixed-stress splits as natural choices among block-partitioned solvers, and deriving theoretical convergence rates consistent with the literature. Likewise, novel results are derived by the application to less-studied coupled systems, e.g., linear poro-visco-elasticity and non-linear extension of Biot's consolidation model, involving non-linear constitutive relations and non-Darcy flows.

2. **Establishing robustness of the fixed-stress split for the linear Biot equations for heterogeneous media.**
   By employing a problem-specific analysis in Paper B, global linear convergence is established for the fixed-stress split solving Biot's consolidation model for fully heterogeneous media. The derived theoretical convergence rate naturally extends previous results in the literature for homogeneous media.

3. **Providing analytical and practical tools for improving the performance of iterative solvers (for poroelasticity).**
   Four systematic approaches are provided for increasing the speed and robustness of iterative solvers:

   First, in Paper A, optimal relaxation is deduced for iterative solvers for poroelasticity models, which exhibit a gradient flow structure. Based on the fact that discretized gradient flows inherit a minimization structure, line search strategies are applicable.

   Second, in Paper B, Paper G, and Section 4.3.2, theoretical convergence rates are derived for the fixed-stress split with variable stabilization, giving rise to the *a priori* optimization with respect to the stabilization parameter. This technique is applicable to general block-partitioned solvers based on the L-scheme, cf. Section 4.3.1.

   Third, based on the mesh independent behavior of the fixed-stress split predicted by the previous convergence analysis, a sampling-based optimization is proposed. An optimal stabilization parameter is determined for a coarse mesh and later applied on the finer mesh of interest. This approach effectively accounts for including the influence of material parameters, domain, and boundary conditions, observed in Paper C.

   Fourth, in Paper E, Anderson acceleration is numerically demonstrated to effectively increase the speed and robustness of splitting schemes as, e.g., the fixed-stress split for unsaturated poroelasticity. In addition, theoretical justification is provided for this observation by studying a variant of Anderson acceleration applied to a simple linear problem. Ultimately, Anderson acceleration is identified as a suitable acceleration technique (in particular for non-linear problems) due to the minimal need of communication between physical subproblems.

4. **Establishing existence of weak solutions to unsaturated poroelasticity.**
   In Paper D, the existence of weak solutions to unsaturated poroelasticity (as modeled by the theory of porous media, cf. Section 2.1.2) is established under mild physical assumptions. In order to treat the highly non-linear model, regularization techniques are combined with the Galerkin method, utilizing a finite element-finite volume discretization. Compactness arguments yield the final result.

5. **Development and numerical analysis of robust iterative linearization schemes for possibly degenerate and coupled problems.**

   Similar concepts as for the fixed-stress split can be applied to non-linear problems – now, resulting in iterative linearization schemes. In Paper E and Paper H, an extension of the classical fixed-stress split is proposed for the block-partitioned linearization of unsaturated poroelasticity. By utilizing the close connection to the L-scheme linearization, it is proved to converge linearly under mild physical conditions. This extends the results from Paper B to the unsaturated regime. Furthermore, the block-partitioned solver is proposed to be coupled with Anderson acceleration for faster and more robust computations.

   The L-scheme has successfully been applied to problems involving Lipschitz continuous non-linearities. In Paper F, the convergence of an L-scheme linearization of a doubly degenerate non-linear parabolic-elliptic problem is established, by involving information of the stopping criterion. Moreover, in practice, the resulting scheme does not involve any tuning parameters and provides robust behavior, compared to standard linearization schemes applied to regularized models.

Chapter 2

Mathematical modeling

Macroscopic, effective descriptions of coupled processes in porous media are typically described in the form of (evolutionary) partial differential equations (PDE). In this chapter, the main mathematical models, employed in this thesis, are presented.

First, multi-phase flow in infinitesimally deformable porous media is modeled based on fundamental conservation principles combined with constitutive laws. This formulation is then restricted both to the partially saturated and fully saturated regimes, ultimately resulting in models utilized in several papers of this work: *unsaturated poroelasticity* and *Biot's quasi-static consolidation model*.

Secondly, the abstract modeling framework of generalized gradient flows is presented. Combined with thermodynamic considerations, it is utilized in this thesis in order to formulate poroelasticity models from a gradient flow perspective.

2.1 Flow in deformable porous media

Natural porous materials as clays, rocks, and sands are microscopically highly complex. So far, no technology is able to provide the exact geometry for field scale applications. And even if, computational power would be quickly exceeded by resolving the geometry exactly for the purpose of a computer simulation. For the sake of computable approximations, flow in porous media is commonly modeled using a continuum approach, i.e., the fluid-filled porous material is considered as a homogenized medium and an effective, macroscopic description is used for approximating microscopic processes. The same philosophy is employed in the context of deformable, fluid-filled porous materials.

In this section, we present the derivation of a continuum mechanical model for isothermal, immiscible two-phase flow in infinitesimally deformable porous media. It is based on fundamental conservation principles coupled with constitutive laws. The presentation is mainly based on [57, 103].

The resulting model yields a foundation for several models employed in this thesis, including the models for unsaturated poroelasticity, and linear poroelasticity. For the deduction, linearizing assumptions will be required. Although stated later in fuller detail, cf.

Section 2.1.1.6, we already mention two main hypotheses: (i) inertia is neglected; and (ii) the skeleton experiences solely infinitesimal deformations. Those allow for a simplified presentation of the derivation.

2.1.1 Immiscible two-phase flow in deformable porous media

In the theory of porous media, the fluid-saturated porous material is considered as a homogenized continuum. In this sense, at each location, both solid and fluid phases are present. In the following, we consider two immiscible fluid phases, a wetting and non-wetting fluid. We denote the solid, wetting fluid, and non-wetting fluid phases as s, w, and nw, respectively.

2.1.1.1 Basics of the Theory of Porous Media

In order to enable a continuous macroscopic description of a microscopically heterogeneous medium, the concept of averaging over representative elementary volumes (REV) is utilized, cf., e.g., [63]. The REV concept allows for the definition of essential effective quantities as the porosity, saturation as well as phase-averaged versions of physical variables as densities etc. For instance, the (Eulerian) porosity, locally measuring the amount of the volume occupied by pores (in the deformed configuration), is given by

$$\phi(\boldsymbol{x},t) := \frac{1}{|d\Omega_t(\boldsymbol{x})|} \int_{d\Omega_t(\boldsymbol{x})} \mathbb{1}_\mathrm{p}(\boldsymbol{r},t)\, d\boldsymbol{r},$$

where $d\Omega_t(\boldsymbol{x})$ denotes the REV corresponding to $\boldsymbol{x} \in \Omega_t$, $\Omega_t(\boldsymbol{x})$ denotes the deformed medium at time t, and $\mathbb{1}_\mathrm{p}$ denotes the characteristic function with respect to the pore space, defined by

$$\mathbb{1}_\mathrm{p}(\boldsymbol{x},t) := \begin{cases} 1 & \text{for } \boldsymbol{x} \text{ in the pore space of } \Omega_t \\ 0 & \text{else.} \end{cases}$$

Analogously, characteristic functions for the single fluid phases can be defined, allowing for defining saturations $s_\alpha \in [0,1]$, $\alpha \in \{\mathrm{w},\mathrm{nw}\}$, of the wetting and non-wetting fluid phases, quantifying the amount of the fluid occupied by the α phase at each location. Finally, we can define relevant volume fractions

$$\eta_\mathrm{s} := 1-\phi, \qquad \eta_\mathrm{w} := \phi s_\mathrm{w}, \qquad \eta_\mathrm{nw} := \phi s_\mathrm{nw}.$$

The same characteristic functions are utilized for defining phase-averaged quantities as densities ρ_π, $\pi \in \{\mathrm{s},\mathrm{w},\mathrm{nw}\}$, or fluid pressures p_α, $\alpha \in \{\mathrm{w},\mathrm{nw}\}$, based on the corresponding microscopic quantities, cf. Fig 2.1. Also, quantities referring to the homogenized medium may be defined, e.g., the bulk density $\rho = \sum_\pi \eta_\pi \rho_\pi$. An overview of the notation, introduced so far, as well as further on in this section, is summarize for reference in Table 2.1.

In the context of deformable media, REVs deform in time. Thus, Eulerian and Lagrangian variants exist for most of the relevant variables. In the following, we aim at working in Eulerian coordinates. However, eventually, we apply the hypotheses of small perturbations, cf. Section 2.1.1.6, to restrict ourselves to infinitesimal deformations.

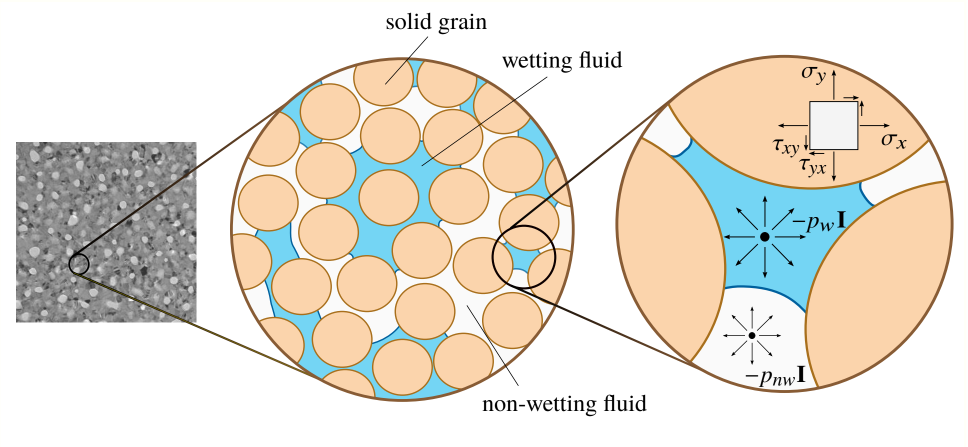


Figure 2.1: Schematic illustration: *(left)* a porous medium; *(center)* an REV occupied with solid grains, a wetting, and a non-wetting fluid phase; *(right)* micro-mechanical stress states within the solid and fluid materials (only hydrostatic for the latter), which are then averaged over the REV.

$\boldsymbol{x}$	spatial coordinate	ρ_π	material density
t	time	ρ	bulk density
Ω_0	reference configuration	$\rho_{\alpha,\mathrm{ref}}$	reference fluid density
Ω_t	deformed medium at time t	$\boldsymbol{\sigma}$	Cauchy stress tensor
π	material phase (s, w, nw)	$\boldsymbol{\sigma}_{\mathrm{eff}}$	effective stress
α	fluid phase (w, s)	p_{pore}	pore pressure
$\boldsymbol{X}^\pi$	material particle	α	Biot-Willis constant
$\boldsymbol{x}^\pi$	particle coordinate	N	Biot modulus
$\boldsymbol{v}^\pi$	particle velocity	μ	shear modulus
$\boldsymbol{v}^{\alpha\mathrm{s}}$	relative particle velocity	λ	Lamé parameter
	wrt. solid phase	K_{dr}	drained bulk modulus
$\frac{d^\pi}{dt}$	material derivative	p_{c}	capillary pressure
	wrt. phase π	K_π	bulk modulus
ϕ	porosity	μ_α	fluid viscosity
η_π	volume fraction	$\boldsymbol{\kappa}$	(absolute) permeability
s_α	fluid saturation	$k_{r\alpha}$	relative permeability
$\boldsymbol{u}$	solid displacement	$\boldsymbol{f}$	external body force
$\boldsymbol{\varepsilon}(\boldsymbol{u})$	linearized strain	$\boldsymbol{g}$	gravitational acceleration
p_α	fluid pressure	h_π	mass source

Table 2.1: Nomenclature of relevant physical variables used in Section 2.1.

2.1.1.2 Kinematics

In the theory of porous media, it is established to describe the motion of fluid phases with respect to the actual configuration of the moving solid skeleton. For that, we describe the motion of particles and introduce the material derivative for multi-phasic media.

Motion of particles. We recall, the fluid saturated deformable medium is modeled as a homogenized multi-phasic continuum. This implies, that each spatial point $\boldsymbol{x}$ in the current configuration Ω_t at time t is simultaneously occupied by three material particles $\boldsymbol{X}^\pi$, $\pi \in \{\mathrm{s}, \mathrm{w}, \mathrm{nw}\}$, cf. Figure 2.2. Their individual motion is tracked by individual deformation maps $\boldsymbol{x}^\pi = \boldsymbol{x}^\pi(\boldsymbol{X}^\pi, t)$, $\pi \in \{\mathrm{s}, \mathrm{w}, \mathrm{nw}\}$. The unique association $\boldsymbol{x} = \boldsymbol{x}^\pi(\boldsymbol{X}^\pi, t)$ defines the inverse maps $\boldsymbol{X}^\pi = \boldsymbol{X}^\pi(\boldsymbol{x}, t)$, $\pi \in \{\mathrm{s}, \mathrm{w}, \mathrm{nw}\}$. Individual particle velocities are then defined by

$$\boldsymbol{v}^\pi(\boldsymbol{x}, t) = \left. \frac{\partial \boldsymbol{x}^\pi(\boldsymbol{X}^\pi, t)}{\partial t} \right|_{\boldsymbol{X}^\pi = \boldsymbol{X}^\pi(\boldsymbol{x}, t)}, \quad \boldsymbol{x} \in \Omega_t,\ \pi \in \{\mathrm{s}, \mathrm{w}, \mathrm{nw}\}.$$

Furthermore, $\boldsymbol{v}^{\pi\alpha} := \boldsymbol{v}^\pi - \boldsymbol{v}^\alpha$ denotes the relative velocity of phase π with respect to phase α.

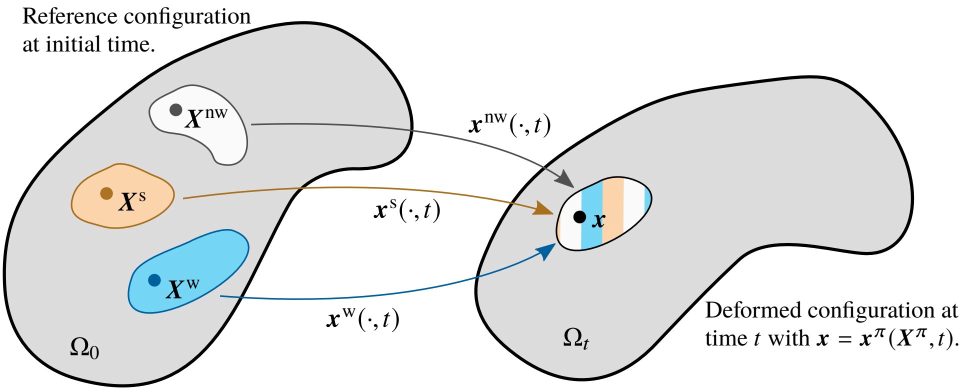

Reference configuration at initial time.
$\boldsymbol{X}^{\mathrm{nw}}$
$\boldsymbol{X}^{\mathrm{s}}$
$\boldsymbol{X}^{\mathrm{w}}$
Ω_0
$\boldsymbol{x}^{\mathrm{nw}}(\cdot, t)$
$\boldsymbol{x}^{\mathrm{s}}(\cdot, t)$
$\boldsymbol{x}^{\mathrm{w}}(\cdot, t)$
$\boldsymbol{x}$
Ω_t
Deformed configuration at time t with $\boldsymbol{x} = \boldsymbol{x}^\pi(\boldsymbol{X}^\pi, t)$.

Figure 2.2: Illustration of the independent movement of solid and fluid particles.

Naturally, it is assumed that two skeleton particles, juxtaposed at a given time, were always so and will remain so. Hence, the overall deformation of the medium is described directly by $\Omega_t = \boldsymbol{x}^{\mathrm{s}}(\Omega_0, t)$, where $\Omega_0 \subset \mathbb{R}^d$ denotes the initial, reference configuration. Consequently, the structural displacement is then given by $\boldsymbol{u}(\boldsymbol{x}, t) = \boldsymbol{x} - \boldsymbol{X}^{\mathrm{s}}(\boldsymbol{x}, t)$, $\boldsymbol{x} \in \Omega_t$.

Changes in distances and angles due to deformation are suitable measured by the Green-Lagrange tensor $\boldsymbol{E}(\boldsymbol{u}) := \frac{1}{2}\left(\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\nabla}\boldsymbol{u}^\top + \boldsymbol{\nabla}\boldsymbol{u}^\top\boldsymbol{\nabla}\boldsymbol{u}\right)$. Under infinitesimal deformations, a geometric linearization is sufficient, and $\boldsymbol{E}(\boldsymbol{u})$ can be approximated by the linearized strain $\boldsymbol{\varepsilon}(\boldsymbol{u}) := \frac{1}{2}\left(\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\nabla}\boldsymbol{u}^\top\right)$.

Material derivative of a field. The material derivative of a differentiable field $f(\boldsymbol{x}, t)$, given in its spatial description and referring to a moving particle of the phase $\pi \in \{\mathrm{s}, \mathrm{w}, \mathrm{nw}\}$, is

$$\frac{d^\pi f}{dt} := \frac{\partial f}{\partial t} + \boldsymbol{\nabla} f \cdot \boldsymbol{v}^\pi. \tag{2.1.1}$$

It is the time derivative of f that an observer attached to the particle $\boldsymbol{X}^\pi(\boldsymbol{x}, t)$ would derive.

Material derivative of an integral quantity. The material derivative also applies to integral quantities, now referring to all particles of a single phase $\pi \in \{\mathrm{s}, \mathrm{w}, \mathrm{nw}\}$ within a given volume. Let the field f be as above, and let $\omega_t^\pi = \boldsymbol{x}^\pi(\omega_0, t) \subset \Omega_t$ for some $\omega_0 \subset \Omega_0$ be an arbitrary control volume, moving with phase π. By virtue of the definition of the material derivative (2.1.1) and the Reynolds transport theorem, cf., e.g., [86], it holds that

$$\frac{d^\pi}{dt} \int_{\omega_t^\pi} f \, d\boldsymbol{x} = \int_{\omega_t^\pi} \left(\frac{d^\pi f}{dt} + f \boldsymbol{\nabla} \cdot \boldsymbol{v}^\pi \right) d\boldsymbol{x} = \int_{\omega_t^\pi} \left(\frac{\partial f}{\partial t} + \boldsymbol{\nabla} \cdot (f \boldsymbol{v}^\pi) \right) d\boldsymbol{x}. \tag{2.1.2}$$

2.1.1.3 Conservation laws

Based on fundamental principles, the linear and angular momenta of the homogenized medium, and the mass of the single phases are conserved under the deformation of a porous medium.

Momentum balance for the homogenized medium. Momentum balance comes in two versions: The balance of linear momentum and the balance of angular momentum. In the context of porous media, the first stipulates that the rate of change of the linear momentum of each single phase is equal to the creation rate of linear momentum due to external forces acting on the medium – similar for the angular momentum. Ultimately, the balance of linear momentum of phase $\pi \in \{\mathrm{s}, \mathrm{w}, \mathrm{nw}\}$ reads as follows. For any control volume $\omega_t \subset \Omega_t$ there exists $\omega_0^\pi \subset \Omega_0$ such that $\omega_t = \omega_t^\pi := \boldsymbol{x}^\pi(\omega_0^\pi, t)$ moves with phase π, and it holds that

$$\frac{d^\pi}{dt} \int_{\omega_t^\pi} \eta_\pi \rho_\pi \boldsymbol{v}^\pi \, d\boldsymbol{x} = \int_{\omega_t^\pi} \eta_\pi \rho_\pi \boldsymbol{f} \, d\boldsymbol{x} + \int_{\partial \omega_t^\pi} \boldsymbol{T}^\pi(\boldsymbol{x}, t, \boldsymbol{n}) \, ds,$$

where $\int_{\omega_t^\pi} \eta_\pi \rho_\pi \boldsymbol{v}^\pi \, d\boldsymbol{x}$ denotes the linear momentum relatively to the particles of phase π, $\boldsymbol{f}$ is an external, local body force, and $\boldsymbol{T}^\pi$ is a surface force resulting from local contact forces.

Summing over all phases, yields the linear momentum balance of the homogenized material. Under the assumption of negligible inertia, we obtain the quasi-static linear momentum balance of the homogenized medium

$$\int_{\omega_t} \rho \boldsymbol{f} \, d\boldsymbol{x} + \int_{\partial \omega_t} \boldsymbol{T}(\boldsymbol{x}, t, \boldsymbol{n}) \, ds = 0, \tag{2.1.3}$$

where $\rho := \sum_\pi \eta_\pi \rho_\pi$ is the bulk density, and $\boldsymbol{T} := \sum_\pi \boldsymbol{T}^\pi$ denotes the total surface force. Similarly, based on the angular momentum balance of each phase, the quasi-static angular momentum balance of the homogenized medium is obtained

$$\int_{\omega_t} \boldsymbol{x} \times \rho \boldsymbol{f} \, d\boldsymbol{x} + \int_{\partial \omega_t} \boldsymbol{x} \times \boldsymbol{T}(\boldsymbol{x}, t, \boldsymbol{n}) \, ds = 0. \tag{2.1.4}$$

Based on (2.1.3) and (2.1.4), classical elasticity theory [54] concludes the existence of a symmetric-tensor-valued field $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{x}, t)$ on Ω_t, *the Cauchy stress*, such that for the arbitrary control volume ω_t it holds that

$$\int_{\omega_t} \boldsymbol{\nabla} \cdot \boldsymbol{\sigma} + \rho \boldsymbol{f} \, d\boldsymbol{x} = 0.$$

A careful discussion utilizing averaging theory enables the interpretation of $\boldsymbol{\sigma}$ being the total stress acting on any unit area of the homogenized medium [57]. Ultimately, in differential form, the momentum balance reads $\boldsymbol{\sigma} = \boldsymbol{\sigma}^\top$ in Ω_t and

$$\boldsymbol{\nabla} \cdot \boldsymbol{\sigma} + \rho \boldsymbol{f} = 0 \quad \text{in } \Omega_t. \tag{2.1.5}$$

We remark, in general, it is customary to phrase the linear momentum balance in a Lagrangian framework. Since the focus of this thesis is on infinitesimally deforming porous media, a careful discussion allows for equating Ω_t with Ω_0.

Conservation of mass of single phases. In integral form the conservation of mass of phase $\pi \in \{\mathrm{s}, \mathrm{w}, \mathrm{nw}\}$ reads as follows. For any control volume $\omega_t^\pi \subset \Omega_t$ moving with phase π, the change in fluid mass is balanced by the amounts of fluid produced within and flowing into the control volume, i.e., it holds that

$$\frac{d^\pi}{dt} \int_{\omega_t^\pi} \eta_\pi \rho_\pi d\boldsymbol{x} = \int_{\omega_t^\pi} h_\pi d\boldsymbol{x} - \int_{\partial\omega_t^\pi \cap \partial\Omega_t} \boldsymbol{q}_\pi \cdot \boldsymbol{n}\, ds, \quad \text{for all } \omega_t \subset \Omega_t, \tag{2.1.6}$$

where $\eta_\pi \rho_\pi$ denotes the mass density of phase π, h_π is a prescribed mass source in Ω_t, $\boldsymbol{q}_\pi$ is a prescribed flux on the boundary $\partial_t \Omega_t$, and $\boldsymbol{n}$ is the outward normal. We note, since ω_t^π moves with phase π, no mass change occurs across parts of $\partial\omega_t$ inside Ω_t. By utilizing the definition of the material derivative (2.1.2), the mass balance of phase π can be formulated in differential form: For all times t it holds that

$$\frac{\partial}{\partial t}(\eta_\pi \rho_\pi) + \boldsymbol{\nabla} \cdot (\eta_\pi \rho_\pi \boldsymbol{v}^\pi) = h_\pi \quad \text{in } \Omega_t.$$

We note, it is common to formulate fluid mass balances in reference to the solid skeleton. In this regard, by utilizing (2.1.1), it holds that

$$\frac{d^{\mathrm{s}}}{dt}(\eta_\pi \rho_\pi) + \eta_\pi \rho_\pi \boldsymbol{\nabla} \cdot \boldsymbol{v}^{\mathrm{s}} + \boldsymbol{\nabla} \cdot (\eta_\pi \rho_\pi \boldsymbol{v}^{\pi s}) = h_\pi \quad \text{in } \Omega_t. \tag{2.1.7}$$

2.1.1.4 Constitutive equations

In order to complete the description of the mechanical behavior of fluid-saturated deformable porous media, constitutive relationships are required. For systematically choosing thermodynamically consistent relationships, entropy inequalities resulting from the second law of thermodynamics may generally be utilized, cf., e.g., [57, 84]. In the following, we mostly employ established, simple laws which depend only on quantities currently measurable in laboratory or field experiments. These mainly correspond to linearizations or simplifications of more general, complex laws.

Equation of state for fluids. The equation of state for fluids is idealistically assumed to be barotropic, i.e., depend solely on the phase-averaged fluid pressure such that

$$\rho_\alpha = \rho_\alpha(p_\alpha) = \rho_{\alpha,\mathrm{ref}} \exp\left(\frac{1}{K_\alpha}\left(p_\alpha - p_{\alpha,\mathrm{ref}}\right)\right), \qquad \alpha \in \{\mathrm{w}, \mathrm{nw}\}, \tag{2.1.8}$$

where K_α is the bulk modulus of fluid phase α.

Capillary pressure. Microscopically, fluid-fluid interfaces between wetting and non-wetting fluids are concave minisci caused by surface tension. The curvature of a meniscus is a result of the *capillary pressure* p_c, which is affected by the surface tension and pore size distribution. At equilibrium, p_c is given by the pressure difference

$$p_\mathrm{c} = p_\mathrm{nw} - p_\mathrm{w}. \tag{2.1.9}$$

It is common practice to assume that one can define a macroscopic capillary pressure, also satisfying (2.1.9). Moreover, it is assumed to be directly related to the fluid saturations, i.e., $s_\mathrm{w} = s_\mathrm{w}(p_\mathrm{c})$. Disregarding hysteresis, it is often further assumed that the relation can be inverted in the positive pressure regime, i.e., $p_\mathrm{c} = p_\mathrm{c}(s_\mathrm{w})$ for $s_\mathrm{w} < 1$.

As common in the poroelasticity literature, in this thesis it is assumed that saturation-pressure relations are independent of the deformation. However, it should be noted that it is evident that the capillary pressure depends on the actual pore size distribution, which changes under deformation. We mention two widely-used simple analytical saturation-pressure relations (neglecting residual saturations for simplicity): *Brooks-Corey* [42]

$$s_\mathrm{w}(p_\mathrm{c}) = \begin{cases} \left(\frac{p_\mathrm{c}}{p_\mathrm{e}}\right)^{-\lambda} & p_\mathrm{c} \geq p_\mathrm{e}, \\ 1 & \text{else}, \end{cases} \tag{2.1.10}$$

where $p_\mathrm{e} > 0$ is the entry pressure, and $\lambda > 0$ is the pore size distribution index; *van Genuchten* [152]

$$s_\mathrm{w}(p_\mathrm{c}) = \begin{cases} (1 + (a_\mathrm{vG} p_\mathrm{c})^{n_\mathrm{vG}})^{-m_\mathrm{vG}} & p_\mathrm{c} \geq 0, \\ 1 & \text{else}, \end{cases} \tag{2.1.11}$$

where $a_\mathrm{vG} > 0$ and $n_\mathrm{vG} > 1$ are model parameters, and $m_\mathrm{vG} := \frac{n_\mathrm{vG}-1}{n_\mathrm{vG}}$.

Darcy's law for multi-phase flow. Darcy's law, the basis of hydrology, relates fluid flow with forces caused by fluid pressure gradients and external (mostly gravitational) body forces acting on the fluid. Originally, it was empirically formulated for single-phase flow in porous media by Darcy [62]. Later, the concept of relative permeability was established by Muskat [112], allowing for a generalization to multi-phase flow. In its now most common form, the volumetric flux of fluid phase $\alpha \in \{\mathrm{w}, \mathrm{nw}\}$ relative to the skeleton, also called filtration vector, is described by

$$\boldsymbol{q}_\alpha := \eta_\alpha \boldsymbol{v}^{\alpha s} = -\frac{\boldsymbol{\kappa} k_{r\alpha}}{\mu_\alpha} (\boldsymbol{\nabla} p_\alpha - \rho_\alpha \boldsymbol{g}), \tag{2.1.12}$$

where $\boldsymbol{\kappa}$ is the intrinsic permeability of the skeleton, $k_{r\alpha} = k_{r\alpha}(s_\alpha)$ is the relative permeability, μ_α is the fluid viscosity, and $\boldsymbol{g}$ is the gravitational acceleration, assuming no other body forces are active.

Generalized Darcy's law (2.1.12) is only valid as a first approximation for slow laminar Newtonian fluid flow. For instance, inertial forces are neglected. Extensions to non-Newtonian fluids [122], non-laminar (Darcy-Forchheimer) flow [79] or transitional (Darcy-Brinkman) flow [41] require additional terms or non-linear dependence. In the following, we consider generalized Darcy's law of the form (2.1.12).

Two relative permeability models corresponding to the retention curves (2.1.10) and (2.1.11) are given by the Brooks-Corey relative permeability model [43]

$$k_{\mathrm{rw}}(s_{\mathrm{w}}) = s_{\mathrm{w}}^{\frac{2+3\lambda}{\lambda}}, \qquad k_{\mathrm{nw}}(s_{\mathrm{w}}) = (1 - s_{\mathrm{w}})^2 \left(1 - s_{\mathrm{w}}^{\frac{2+\lambda}{\lambda}}\right),$$

and the van Genuchten-Mualem relative permeability model [106]

$$k_{\mathrm{rw}}(s_{\mathrm{w}}) = \sqrt{s_{\mathrm{w}}} \left(1 - \left(1 - s_{\mathrm{w}}^{\frac{1}{m_{\mathrm{vG}}}}\right)^{m_{\mathrm{vG}}}\right)^2, \qquad k_{\mathrm{nw}}(s_{\mathrm{w}}) = \sqrt{1 - s_{\mathrm{w}}} \left(1 - s_{\mathrm{w}}^{\frac{1}{m_{\mathrm{vG}}}}\right)^{2m_{\mathrm{vG}}}.$$

Pore pressure. The *pore pressure* is the pressure of the homogenized, possibly multiphasic fluid acting on the skeleton as matric pressure or suction. For single-phasic fluids the pore pressure is equal to the fluid pressure. Yet, despite a 60 years ongoing discussion, no unified definition of the pore pressure has been established and fully accepted for unsaturated media. Several modeling attempts have been made in the literature – also of theoretical kind, cf. [114, 120] for reviews.

Bishop [24] stated one of the first generalizations for unsaturated media given by

$$p_{\mathrm{pore}} = p_{\mathrm{nw}} + \chi(p_{\mathrm{w}} - p_{\mathrm{nw}}).$$

placing emphasize on the matric suction $p_{\mathrm{nw}} - p_{\mathrm{w}}$ as a major influence. The weight χ, called the *Bishop's parameter*, quantifies the area of contact between solid and fluids. An often utilized choice is the *averaged pore pressure*, which is the volume-averaged fluid pressure

$$p_{\mathrm{avg}} = s_{\mathrm{w}} p_{\mathrm{w}} + s_{\mathrm{nw}} p_{\mathrm{nw}},$$

i.e., $\chi = s_{\mathrm{w}}$. However, it is criticized for not accounting for solid-fluid interfaces [58, 59].

In this thesis, we utilize the so-called *equivalent pore pressure*, cf. e.g. [57]. It is justified by a thermodynamic approach accounting also for solid-fluid interfaces via an interfacial energy. Ultimately, it is given by the corrected average pore pressure

$$p_{\mathrm{pore}} = s_{\mathrm{w}} p_{\mathrm{w}} + s_{\mathrm{nw}} p_{\mathrm{nw}} - \int_{s_{\mathrm{w}}}^{1} p_c(S)\, dS,$$

which is equivalent with the differential definition

$$dp_{\mathrm{pore}} = s_{\mathrm{w}} dp_{\mathrm{w}} + s_{\mathrm{nw}} dp_{\mathrm{nw}}.$$

In the single-phasic limit case or in the absence of capillarity, the definition of the pore pressure as the fluid pressure is recovered.

Cauchy stress tensor – The effective stress concept. Since the early stages of soil and rock mechanics, and by now established in geotechnical engineering, the *concept of effective stress* for describing the overall stress state of a homogenized porous medium has been introduced by Terzaghi [149] and Biot [22] for saturated porous media. Bishop [24] then naturally extended the concept to partially saturated media. All works assume that: (i) the effective stress $\boldsymbol{\sigma}_{\text{eff}}$ is assumed to be solely responsible for all major deformations of the (drained or undrained) skeleton; and (ii) the effective stress is a linear combination of the Cauchy stress $\boldsymbol{\sigma}$ and the pore pressure p_{pore}

$$\boldsymbol{\sigma}_{\text{eff}} = \boldsymbol{\sigma} + \alpha p_{\text{pore}} \mathbf{I}. \tag{2.1.13}$$

The weight α, called the *Biot-Willis constant*, has been determined for isotropic homogeneous solid materials under infinitesimal deformations among others by [23, 119] as

$$\alpha = 1 - \frac{K_{\text{dr}}}{K_{\text{s}}},$$

where K_{dr} and K_{s} are the bulk modulus of the drained skeleton and the bulk modulus of the solid grains, respectively.

Originally, phenomenologically stated on macroscopic level, (2.1.13) can be also theoretically justified utilizing, e.g., thermodynamic considerations [57], averaging theory [64], or homogenization [45, 133].

The effective stress for a St. Venant-Kirchhoff material. Provided the solid skeleton solely deforms under infinitesimal deformations, the hypotheses of linear elasticity allow for modeling the effective behavior by a generalized Hooke's laws as follows. The effective stress stress is proportional to the linearized strain tensor

$$\boldsymbol{\sigma}_{\text{eff}} = \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u})$$

where $\mathbb{C} \in \mathbb{R}^{d \times d \times d \times d}$ is a fourth-order Gassmann tensor corresponding to the drained skeleton. For isotropic materials, Hooke's law is equivalent with the description of a St. Venant-Kirchhoff material. There exist two parameters μ and λ, called shear modulus and Lamé's first parameter, respectively, such that

$$\boldsymbol{\sigma}_{\text{eff}} = 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \boldsymbol{\nabla} \cdot \boldsymbol{u}\, \mathbf{I}. \tag{2.1.14}$$

In this case, the bulk modulus of the drained solid skeleton as a whole is $K_{\text{dr}} = \frac{2\mu}{d} + \lambda$.

Solid density. The solid density is assumed to depend on the pore pressure and the first invariant of the effective stress $\sigma_{\text{eff,h}} := \frac{1}{d}\text{tr}\, \boldsymbol{\sigma}_{\text{eff}}$, i.e., its hydrostatic component, such that

$$\frac{1}{\rho_{\text{s}}} d\rho_s = \frac{1}{K_{\text{s}}} dp_{\text{pore}} + \frac{1}{\eta_{\text{s}} K_{\text{s}}} d\sigma_{\text{eff,h}}.$$

Moreover, the following constitutive relation is assumed between the hydrostatic effective stress and the overall volumetric strain rate, caused by structural deformation and uniform compression of solid particles as opposed to the skeleton,

$$\frac{d^s \sigma_{\text{eff,h}}}{dt} = K_{\text{dr}} \left(\boldsymbol{\nabla} \cdot \boldsymbol{v}^{\text{s}} - \frac{1}{K_{\text{s}}} \frac{d^s p_{\text{pore}}}{dt} \right).$$

For the solid density it follows that

$$\frac{1-\phi}{\rho_{\text{s}}} \frac{d^s \rho_{\text{s}}}{dt} = \frac{1}{K_{\text{s}}} (\alpha - \phi) \frac{d^s p_{\text{pore}}}{dt} + (\alpha - 1)\, \boldsymbol{\nabla} \cdot \boldsymbol{v}^{\text{s}}. \tag{2.1.15}$$

2.1.1.5 Resulting mass balance equations.

Combining the fundamental mass balance equations derived in Section 2.1.1.3 and the constitutive relations from Section 2.1.1.4 results in practical mass balance equations. Those yield the basis for the mathematical models for unsaturated poroelasticity and linear poroelasticity for the rest of the thesis.

Mass balance for the solid phase. The mass balance for the solid phase is solely a helping tool in order to derive an evolution equation for the Eulerian porosity. Assuming no solid mass production, (2.1.7) yields after expanding $\frac{d^s}{dt}$ and dividing by ρ_s

$$\frac{1-\phi}{\rho_\mathrm{s}} \frac{d^s \rho_\mathrm{s}}{dt} - \frac{d^s \phi}{dt} + (1-\phi)\, \boldsymbol{\nabla} \cdot \boldsymbol{v}^\mathrm{s} = 0.$$

Inserting the constitutive relation for the solid density (2.1.15) and rearranging, yields for the porosity

$$\frac{d^s \phi}{dt} = \frac{\alpha - \phi}{K_\mathrm{s}} \frac{d^s p_\mathrm{pore}}{dt} + (\alpha - \phi)\, \boldsymbol{\nabla} \cdot \boldsymbol{v}^\mathrm{s}. \tag{2.1.16}$$

Mass balance for the fluid phases. As for the solid phase, we consider (2.1.7), now for $\alpha \in \{\mathrm{w}, \mathrm{nw}\}$. After expanding $\frac{d^s}{dt}$ and dividing by ρ_α, we obtain

$$\phi \frac{d^s s_\alpha}{dt} + s_\alpha \frac{d^s \phi}{dt} + \frac{\phi s_\alpha}{\rho_\alpha} \frac{d^s \rho_\alpha}{dt} + \phi s_\alpha \boldsymbol{\nabla} \cdot \boldsymbol{v}^\mathrm{s} + \frac{1}{\rho_\alpha} \boldsymbol{\nabla} \cdot (\rho_\alpha \phi s_\alpha \boldsymbol{v}^{\alpha\mathrm{s}}) = \frac{h_\alpha}{\rho_\alpha}.$$

Inserting (2.1.16) and the equation of state for the fluid (2.1.8), yields in Ω_t

$$\phi \frac{d^s s_\alpha}{dt} + \frac{\phi s_\alpha}{K_\alpha} \frac{d^s p_\alpha}{dt} + \frac{\alpha - \phi}{K_\mathrm{s}} s_\alpha \frac{d^s p_\mathrm{pore}}{dt} + \alpha s_\alpha \boldsymbol{\nabla} \cdot \boldsymbol{v}^\mathrm{s} + \frac{1}{\rho_\alpha} \boldsymbol{\nabla} \cdot (\rho_\alpha \phi s_\alpha \boldsymbol{v}^{\alpha\mathrm{s}}) = \frac{h_\alpha}{\rho_\alpha}. \tag{2.1.17}$$

2.1.1.6 Hypotheses of small perturbations

The hypotheses of small perturbations are a set of hypotheses allowing for partial linearization of the otherwise highly non-linear problem. In this thesis, we impose:

- Inertia can be neglected, resulting in a quasi-static approximation of poroelasticity.
- The deformation of the solid skeleton and displacements of the solid particles are small, resulting in two particular implications: (i) infinitesimal displacements allow for the identification $\boldsymbol{x}^\mathrm{s} \approx \boldsymbol{X}^\mathrm{s}$, and hence $\Omega := \Omega_0 \approx \Omega_t$; and (ii) infinitesimal strains allow for the linearization of the Green-Lagrange strain $\boldsymbol{E}(\boldsymbol{u}) \approx \boldsymbol{\varepsilon}(\boldsymbol{u})$.
- The porosity is close to a reference porosity $\phi \approx \phi_0$.
- Fluid mass densities vary insignificantly, allowing for the approximations $\rho_\alpha \approx \rho_{\alpha,\mathrm{ref}}$ and $\boldsymbol{\nabla}\rho_\alpha \approx \boldsymbol{0}$.

A direct consequence of these hypotheses is that we can define the constant Biot modulus N

$$\frac{1}{N} := \frac{\alpha - \phi_0}{K_\mathrm{s}} \approx \frac{\alpha - \phi}{K_\mathrm{s}}.$$

Moreover, the material derivative with respect to solid particles can be approximated by the partial temporal derivative $\frac{d^s}{dt} \approx \frac{\partial}{\partial t}$, and one can identify $\boldsymbol{v}^\mathrm{s} = \frac{\partial \boldsymbol{u}}{\partial t}$.

2.1.1.7 Summary – Governing equations under linearizing hypotheses

Applying the hypotheses of small perturbations to the equations derived in this section, results in a partially non-linear mathematical model for multi-phase flow in deformable porous media under infinitesimal deformations. A minimal set of primary variables is chosen: the structural displacement $\boldsymbol{u}$ and the fluid pressures p_α, $\alpha \in \{\mathrm{w}, \mathrm{nw}\}$. Quantities as the fluid saturations s_α, the relative permeabilities $k_{\mathrm{r}\alpha}$ and the pore pressure p_pore are assumed to be given by constitutive relations, cf. Section 2.1.1.4.

Governing equations. Combining the fundamental momentum balance (2.1.5) with the effective stress concept (2.1.13) and Hooke's law for isotropic materials (2.1.14) results in the linear momentum balance

$$-\boldsymbol{\nabla} \cdot (2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \boldsymbol{\nabla} \cdot \boldsymbol{u}\,\mathbf{I}) + \alpha \boldsymbol{\nabla} p_\mathrm{pore} = \rho \boldsymbol{g} \qquad \text{in } \Omega. \tag{2.1.18}$$

The mass balance (2.1.17) for fluid phase $\alpha \in \{\mathrm{w}, \mathrm{nw}\}$ under the hypotheses of small perturbations becomes

$$\phi_0 \frac{\partial s_\alpha}{\partial t} + s_\alpha \frac{\partial}{\partial t}\left(\frac{\phi_0}{K_\alpha} p_\alpha + \frac{1}{N} p_\mathrm{pore} + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}\right) - \boldsymbol{\nabla} \cdot \left(\frac{\boldsymbol{\kappa} k_{\mathrm{r}\alpha}}{\mu_\alpha}\left(\boldsymbol{\nabla} p_\alpha - \rho_{\alpha,\mathrm{ref}} \boldsymbol{g}\right)\right) = \frac{h_\alpha}{\rho_{\alpha,\mathrm{ref}}} \quad \text{in } \Omega. \tag{2.1.19}$$

Boundary conditions and initial conditions. In order to close the system, boundary conditions and initial conditions need to be imposed. Given three partitions $(\Gamma_\mathrm{D}^\mathrm{m}, \Gamma_\mathrm{N}^\mathrm{m})$, $(\Gamma_\mathrm{D}^{\mathrm{f}\alpha}, \Gamma_\mathrm{N}^{\mathrm{f}\alpha})$, $\alpha \in \{\mathrm{w}, \mathrm{nw}\}$, of $\partial\Omega$, and prescribed displacement $\boldsymbol{u}_\mathrm{D}$, surface stress $\boldsymbol{\sigma}_\mathrm{N}$, pressure $p_{\alpha,\mathrm{D}}$, and normal fluxes $q_{\alpha,\mathrm{N}}$, we consider mixed essential and natural boundary conditions for $\alpha \in \{\mathrm{w}, \mathrm{nw}\}$

$$\begin{aligned}
\boldsymbol{u} &= \boldsymbol{u}_\mathrm{D} && \text{on } \Gamma_\mathrm{D}^\mathrm{m} \times (0,T),\\
\left(2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \boldsymbol{\nabla} \cdot \boldsymbol{u}\mathbf{I} - \alpha p_\mathrm{pore}\mathbf{I}\right) \boldsymbol{n} &= \boldsymbol{\sigma}_\mathrm{N} && \text{on } \Gamma_\mathrm{N}^\mathrm{m} \times (0,T),\\
p_\alpha &= p_{\alpha,\mathrm{D}} && \text{on } \Gamma_\mathrm{D}^{\mathrm{f}\alpha} \times (0,T),\\
-\boldsymbol{\kappa} k_{\mathrm{r}\alpha}\left(\boldsymbol{\nabla} p_\alpha - \rho_{\alpha,\mathrm{ref}} \boldsymbol{g}\right) \cdot \boldsymbol{n} &= q_{\alpha,\mathrm{N}} && \text{on } \Gamma_\mathrm{N}^{\mathrm{f}\alpha} \times (0,T).
\end{aligned}$$

Additionally, for prescribed displacement $\boldsymbol{u}_0$ and fluid pressures $p_{\alpha,0}$, $\alpha \in \{\mathrm{w}, \mathrm{nw}\}$, we consider the following initial conditions

$$\begin{aligned}
\boldsymbol{u} &= \boldsymbol{u}_0 && \text{in } \Omega \times \{0\},\\
p_\alpha &= p_{\alpha,0} && \text{in } \Omega \times \{0\}.
\end{aligned}$$

Comparison to the thermodynamic approach by Coussy. The papers in Part II of the thesis also refer to [57] for a thermodynamic derivation of the mathematical model. In the aforementioned work, the existence of an energy potential for the homogenized medium is assumed, allowing to utilize an entropy inequality and thereby resulting in constitutive equations for the Cauchy stress and the Lagrangian porosity. Balance equations are derived from fundamental principles translated to a Lagrangian framework. Ultimately, under linearizing hypotheses, cf. Section 2.1.1.6, the author arrives at the same governing equations as derived here with the equivalent pore pressure as pore pressure.

2.1.2 Unsaturated poroelasticity

Modelling immiscible two-phase flow in porous media can be considerably simplified under specific conditions occurring, e.g., for air and water in the vadose zone. For instance, assuming the viscosity of the non-wetting fluid phase is several orders smaller than the one of the wetting phase, it has a much greater mobility. Thereby, it can be expected that pressure differences in the non-wetting phase are much faster equilibrated than in the wetting phase. In addition, if the non-wetting fluid phase is assumed to be continuously connected to the atmosphere, variations in the non-wetting fluid pressure can be neglected. Consequently, the non-wetting fluid pressure can be assumed to be equal the atmospheric pressure, which for convenience is often assumed to be zero, i.e., $p_{\mathrm{nw}} = 0$. Those assumptions are commonly utilized for reducing two-phase flow models to Richards' equation [144].

The same assumptions can be applied in the context of deformable porous media. Consequently, the mass balance equation (2.1.19) can be neglected for the non-wetting fluid phase ($\alpha = \mathrm{nw}$), and kept unchanged for the wetting fluid phase ($\alpha = \mathrm{w}$). Other than that, the linear momentum balance (2.1.18) remains the same. Thereby, the model, presented in Section 2.1.1.7, essentially reduces to Richards' equation non-linearly coupled to the linear elasticity equations with the mechanical displacement $\boldsymbol{u}$ and wetting fluid pressure p_{w} as primary variables – from now on called model for *unsaturated poroelasticity*.

The formal reduction also simplifies the expressions for the capillary pressure and pore pressure. The negative capillary pressure p_{c} is identical with the inverse of $s_{\mathrm{w}} = s_{\mathrm{w}}(p_{\mathrm{w}})$ in the negative pressure regime, since

$$p_{\mathrm{c}}(p_{\mathrm{w}}) = -p_{\mathrm{w}},$$

whereas the equivalent pore pressure can be interpreted as a function of the wetting fluid pressure

$$p_{\mathrm{pore}}(p_{\mathrm{w}}) = s_{\mathrm{w}}(p_{\mathrm{w}})p_{\mathrm{w}} - \int_{s_{\mathrm{w}}}^{1} p_{\mathrm{c}}(S)\, dS$$

with $p'_{\mathrm{pore}}(p_{\mathrm{w}}) = s_{\mathrm{w}}(p_{\mathrm{w}})$.

It has to be stressed that the reduced model can lead to inaccurate results. As in the context of Richards' equation for non-deformable media, cf. [144] and the references within, the assumptions on the mobility difference of the two fluid phases and the connectivity of the non-wetting fluid phase are crucial. But even if those are fulfilled, the presence of obstacles in the medium as impermeable layers, or different highly heterogeneous structures, may disallow for neglecting the non-wetting fluid phase. In addition, in terms of deformable media, the pore pressure within a dry medium is equal the atmospheric pressure, according to the Biot theory. However, an increasing pore pressure as, e.g., the equivalent pore pressure, does not satisfy this relation. Thus, in the degenerate case of vanishing saturation, the coupled model is also questionable (independent of the validity of Richards' equation alone). In the attached papers on unsaturated poroelasticity, the case of vanishing saturation will be excluded.

2.1.3 Biot's quasi-static consolidation model

Considering the limit case of a single-phase flow in an infinitesimally deformable porous medium, the general model for two-phase flow as presented in Section 2.1.1.7, simplifies substantially. For instance, the non-wetting fluid phase can be entirely neglected, and the structural displacement $\boldsymbol{u}$ and wetting fluid pressure p_w remain as primary variables. Moreover, non-linearities in the fluid pressure become linear.

The resulting model is the classical two-field formulation of the famous and well-studied *Biot's (quasi-static) consolidation model*, originally introduced by Biot [22],

$$-\boldsymbol{\nabla} \cdot (2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \boldsymbol{\nabla} \cdot \boldsymbol{u}\, \mathbf{I}) + \alpha \boldsymbol{\nabla} p_\mathrm{w} = \rho \boldsymbol{g} \qquad \text{in } \Omega, \tag{2.1.20}$$

$$\frac{\partial}{\partial t}\left(\frac{1}{M} + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}\right) - \boldsymbol{\nabla} \cdot \left(\frac{\boldsymbol{\kappa}}{\mu_\mathrm{w}}\left(\boldsymbol{\nabla} p_\mathrm{w} - \rho_\mathrm{w,ref}\boldsymbol{g}\right)\right) = \frac{h_\mathrm{w}}{\rho_\mathrm{w,ref}} \qquad \text{in } \Omega, \tag{2.1.21}$$

with a constant bulk density $\rho = (1-\phi_0)\rho_\mathrm{s,ref} + \phi_0 \rho_\mathrm{w,ref}$ and a constant storage coefficient $\frac{1}{M} := \frac{\phi_0}{K_\alpha} + \frac{1}{N}$. It is also referred to as *linear Biot equations*. We mention, that the model can also be derived by means of homogenization [45, 133] or the framework of generalized gradient flows, cf. Paper A.

2.2 Variational modeling of dissipative systems – from classical gradient flows in Hilbert spaces to generalized gradient flows

The framework of *classical gradient flows* constitutes a modeling approach based on evolutionary PDEs for describing physical systems purely driven by dissipation of energy. The central idea is that within a physical state space, each state has an associated energy. Then given an initial datum, trajectories of states evolve along the negative gradient of that energy, until eventually ending in a stationary state corresponding to a local minimum of the energy landscape. This approach implicitly determines essential modeling assumptions. For instance, in the context of mechanical systems, viscous or friction forces have to dominate inertial forces, allowing for neglecting the latter. Many physical processes can be described as classical gradient flows, e.g., non-linear diffusion as the porous medium equation [121], the Cahn-Hilliard equation [146], quasi-stationary phase field models and the Stefan problem [131]. For a detailed, mathematical description of classical gradient flows, we refer to the seminal works by Komura [100], Crandall and Pazy [60], and Brezis [38, 39].

Classical gradient flows implicitly only allow for quadratic dissipation. Therefore (and for other purposes), the framework of *generalized gradient flows*, closely related to doubly non-linear evolution equations, has been introduced as, e.g., presented in [124]. Ultimately, various advanced applications can be modeled by generalized gradient flows. We mention incompressible immiscible two-phase flow in porous media [47], doubly non-linear Allen-Cahn equations [108], rate-independent finite elasticity [108], rate-dependent visco-plasticity at finite strain [109], and also thermo-poro-visco-elasticity including non-Darcy-type flow, cf. Paper A, just to mention a few.

The purpose of the remaining section is to depict the connection between the framework of classical gradient flows in Hilbert spaces and the framework of generalized gradient flows. With focus on modeling, technicalities as, e.g., non-smooth or $\mathbb{R} \cup \{\infty\}$-valued potentials are ignored in the following. For an introduction to Hilbert spaces and corresponding terminology, we refer to [55].

Classical gradient flows in Hilbert spaces. Let the *state space* H be a Hilbert space with an inner product $(\cdot, \cdot)$. Let $H^\star$ denote the dual space of H with the duality pairing $\langle \cdot, \cdot \rangle$. Furthermore, let the *energy* $\mathcal{E} : H \to \mathbb{R}$ be a functional, for which a Frechét derivative $\nabla\mathcal{E}(x) \in H^\star$ exists in the sense of functional derivatives, for all $x \in H$; let $\mathrm{grad}_H \mathcal{E}(x) \in H$ correspond to $\nabla\mathcal{E}(x) \in H^\star$ via Riesz's representation theorem.

Finally, for given initial datum $x_0 \in H$, the curve $t \mapsto x(t)$, $t \in [0, \infty)$ is called a *(classical) gradient flow* of $\mathcal{E}$ in H if it holds that

$$\dot{x} = -\mathrm{grad}_H \mathcal{E}(x) \quad \text{in } H, \tag{2.2.1}$$

$$x(0) = x_0. \tag{2.2.2}$$

We mention two limitations of the framework of classical gradient flows:

1. Classical gradient flows are restricted to quadratic dissipation of energy. For instance, for a trajectory of states defined by (2.2.1), the dissipation of energy is governed by

$$\partial_t \mathcal{E}(x) = \langle \nabla\mathcal{E}(x), \dot{x} \rangle = -\left(\mathrm{grad}_H \mathcal{E}(x), \dot{x}\right) = -\left(\dot{x}, \dot{x}\right).$$

2. It is often more convenient to describe changes of state via physical processes, which might not stand in a 1–1 correspondence to each other. Typical examples are fluxes associated to changes of concentration, mass, or temperature. The framework of classical gradient flows does not feature that.

An equivalent reformulation of (2.2.1)–(2.2.2) yields the foundation for extension to generalized gradient flows, which account for those two limitations. For instance, the change of state solves a convex minimization problem:

$$\dot{x} = \underset{s \in H}{\arg\min} \left\{ \frac{1}{2}(s, s) + \langle \nabla\mathcal{E}(x), s \rangle \right\}, \tag{2.2.3}$$

$$x(0) = x_0. \tag{2.2.4}$$

Generalized gradient flows in Banach spaces. Generalized gradient flows can be defined by extending the minimization formulation (2.2.3). Addressing the first limitation of classical gradient flows, mentioned above, the quadratic scalar product is replaced by a general convex dissipation potential $\mathcal{D}$. The second limitation is handled by the introduction of a process space P, consisting of feasible physical processes. Those are associated to changes of state via a transformation $T : P \longrightarrow H$. The curve $t \mapsto x(t)$, $t \in [0, T]$, is then called a *generalized gradient flow* if it holds that

$$\begin{aligned} \dot{x} &= T(x)q, && (2.2.5) \\ q &= \operatorname*{arg\,min}_{w \in P(x)} \left\{ \mathcal{D}(x; w) + \langle \nabla \mathcal{E}(x), T(x)w \rangle \right\}, && (2.2.6) \\ x(0) &= x_0. && (2.2.7) \end{aligned}$$

In principle, the dissipation potential and process space could be state-dependent, allowing for quite general scenarios. Choices for the dissipation potential in the literature range from vanishing, 1-homogeneous, non-quadratic, to non-finite potentials.

Finally, the generalization of (2.2.5)–(2.2.7) to Banach spaces is immediate.

Chapter 3

Establishing well-posedness of PDEs

The purpose of this section is to provide classical mathematical tools for establishing the well-posedness of (evolutionary) PDEs, which are used in Part II of this thesis. We mention four general approaches. We start with the most flexible and powerful approach: the combination of the *Galerkin method* and *compactness arguments*. It is applicable to a wide range of problems, and can be often utilized for analyzing the well-posedness of non-linear PDEs. Opposing to that, provided a problem has a specific structure, high-level techniques may be applicable instead, resulting in a more compact and elegant analysis. We mention tools from the theory of *doubly non-linear evolution equations*, *convex analysis*, and *saddle point problems*. The latter two are restricted to stationary partial differential equations.

We start with a brief overview of functional spaces used throughout this chapter.

3.1 Spaces

We make use of Hilbert, Banach, Lebesgue, Sobolev and Bochner spaces. For detailed introductions, we refer to the textbooks [1, 55, 73]. We employ the following notation:

- For *Hilbert spaces* H, let $(\cdot, \cdot)$ denote the associated *inner product*.
- For *Banach spaces* V, let $\|\cdot\|_V$ denote the associated norm. Furthermore, let $V^\star$ denote the *dual space* of V, consisting of all continuous linear functionals on V. The duality pairing is given by $\langle \cdot, \cdot \rangle$.
- For any open domain $\Omega \subset \mathbb{R}^N$, $N \in \mathbb{N}$, let $L^p(\Omega)$, $p \in [1, \infty]$, denote the standard *Lebesgue space*, consisting of measurable functions for which the p-th power of the absolute value is Lebesgue integrable, if $p < \infty$. The limit case $L^\infty(\Omega)$ consist of functions, which are essentially bounded on Ω, i.e., bounded up to a set of zero measure.
- Let $H^1(\Omega)$ denote the *Sobolev space*, consisting of functions in $L^2(\Omega)$, such that their weak derivatives of order 1 have finite $L^2(\Omega)$ norm.

- For any Banach space V, and $T > 0$, let $L^p(0,T;V)$, $p \in [1,\infty]$, denote the *Bochner space*, consisting of 'time'-depending functions f with values in V, for which $\|f\|_V$ lies in $L^p(0,T)$. Analogously, let $W^{1,p}(0,T;V)$ be the subspace of $L^p(0,T;V)$, whose elements have a weak derivative (in 'time') with finite $L^p(0,T;V)$ norm; for $p = 2$, we write $H^1(0,T;V) := W^{1,2}(0,T;V)$. And $C(0,T;V)$ denotes the space of functions, which are continuous in time.

3.2 Galerkin method and compactness arguments

The combination of the Galerkin method and compactness arguments is an often quite effective method for establishing the existence of solutions to stationary or evolutionary, non-linear PDEs, posed as a set of variational equations over an infinite dimensional, separable, reflexive Banach space [55]. In the following, we present the key steps illustrated for an evolutionary model problem reading as follows: *Find $u \in H^1(0,T;V^\star) \cap L^2(0,T;V)$ such that*

$$\int_0^T \langle \partial_t u, v\rangle \, dt + \int_0^T \langle \mathcal{A}(u), v\rangle \, dt = \int_0^T \langle f, v\rangle \, dt \qquad \textit{for all } v \in L^2(0,T;V), \tag{3.2.1}$$

with some initial data $u(0) = u_0$, where $T > 0$ denotes some final time; V is an infinite dimensional, separable, reflexive Banach space; ∂_t denotes the temporal derivative in the sense of distributions; $\mathcal{A} : V \to V^\star$ is some (non-linear) operator; $f \in L^2(0,T;V^\star)$; and $u_0 \in V$. The presentation is mainly based on [55].

1. Step – Approximation in finite dimensions. The first step is the actual Galerkin method (here we focus mainly on the conforming Galerkin method but stress that also non-conforming variants can be utilized). Since V is separable, there exists a countably infinite, independent family $\left(v^{(i)}\right)_{i=1}^\infty \subset V$, such that the union of the finite dimensional subspaces $V_m := \operatorname{span}\left\{v^{(i)}\right\}_{i=1}^m$ of V, $m \in \mathbb{N}$, is dense in V. A finite dimensional approximation of the problem (3.2.1) is then given by combining a time-discretization, e.g., the implicit Euler method, with the Galerkin method. For each $m, N \in \mathbb{N}$, we consider for $n \in \{1, \ldots, N\}$: *Given an approximation $u_m^{n-1} \in V_m$ for the previous time step, find $u_m^n \in V_m$ such that*

$$\left\langle \frac{u_m^n - u_m^{n-1}}{\Delta t}, v_m \right\rangle + \left\langle \mathcal{A}(u_m^n), v_m \right\rangle = \langle f^n, v_m\rangle \qquad \textit{for all } v_m \in V_m, \tag{3.2.2}$$

where $\Delta t := \frac{T}{N} > 0$ denotes a time step size corresponding to the time-discretization, and $f^n \in V^\star$ is an approximation of f on the time interval $(t_{n-1}, t_n]$ with suitable approximation quality. The initial conditions are approximated by some suitable $u_m^0 \in V_m$.

Next, for all $m, N \in \mathbb{N}$, the existence of at least one discrete solution of (3.2.2) has to be established for each time step. A helpful observation is that the discrete problem (3.2.2) is finite dimensional. Consequently, by utilizing a basis of V_m, it can be reformulated to an algebraic, non-linear problem: *Find $\mathbf{u}_m^n \in \mathbb{R}^m$ such that*

$$\mathbf{F}\left(\mathbf{u}_m^n\right) = \mathbf{0} \quad \text{in } \mathbb{R}^m, \tag{3.2.3}$$

for appropriate $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$. For proving the existence of a zero to (3.2.3), a corollary of Brouwer's fixed point theorem may be utilized.

Lemma 3.2.1 (Corollary of Brouwer's fixed point theorem [55])**.** *Let* $(\cdot,\cdot)$ *denote the (canonical) inner product on* $\mathbb{R}^m$*. Let* $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m$ *be a continuous function, satisfying*

$$(\mathbf{F}(\mathbf{x}), \mathbf{x}) \geq 0 \ \textit{for all} \ \mathbf{x} \in \mathbb{R}^m \ \textit{with} \ (\mathbf{x}, \mathbf{x}) \geq M, \tag{3.2.4}$$

for some fixed $M > 0$*. Then there exists an* $\mathbf{x}^\star \in \mathbb{R}^m$ *with* $\left(\mathbf{x}^\star, \mathbf{x}^\star\right) \leq M$ *and* $\mathbf{F}\left(\mathbf{x}^\star\right) = \mathbf{0}$*.*

Based on the time-discrete approximation, time-continuous approximations can be defined on the entire time interval $(0,T)$. For instance, piecewise constant or piecewise linear in time approximations $\bar{u}_{mN}$ and $\hat{u}_{mN}$ are defined by

$$\begin{aligned}
\bar{u}_{mN}(t) &:= u_m^n, && t \in (t_{n-1}, t_n],\\
\hat{u}_{mN}(t) &:= u_m^{n-1} + \frac{t - t_{n-1}}{\Delta t}(u_m^n - u_m^{n-1}), && t \in (t_{n-1}, t_n],
\end{aligned}$$

where $t_n := n\Delta t$; similarly, let $\bar{f}_N$ denote the piecewise constant function based on f^n. By construction and (3.2.2), it holds that

$$\int_0^T \langle \partial_t \hat{u}_{mN}, \bar{v}_m \rangle \, dt + \int_0^T \langle \mathcal{A}\left(\bar{u}_{mN}\right), \bar{v}_m \rangle \, dt = \int_0^T \left\langle \bar{f}_N, \bar{v}_m \right\rangle \, dt, \tag{3.2.5}$$

for all $\bar{v}_m$ such that $\bar{v}_m(t) = v_m^n$, $t \in (t_{n-1}, t_n]$ with arbitrary $v_m^n \in V_m$.

The main idea of the following steps will be: (i) to show that $\{\bar{u}_{mN}\}_{m,N}$ and $\{\hat{u}_{mN}\}_{m,N}$ converge to the same $u \in H^1(0,T;V^\star) \cap L^2(0,T;V)$; (ii) to conclude that the type of convergence also allows for considering the limit of (3.2.5) for $m, N \to \infty$; and finally (iii) deduce that u is a solution of the model problem (3.2.1).

2. Step – Uniform stability. Next, one has to show that the finite dimensional approximations are uniformly bounded in the correct spaces. This objective may be achieved as a byproduct of the corollary of Brouwer's fixed point theorem, cf. Lemma 3.2.1; this is by no means the only possibility, depending on the particular problem.

In the context of the model problem (3.2.1), it is natural to ask for uniform boundedness of $\{\bar{u}_{mN}\}_{m,N}$ in $L^2(0,T;V)$ and $\{\hat{u}_{mN}\}_{m,N}$ in $H^1(0,T;V^\star) \cap L^2(0,T;V)$, independently of $m, N \in \mathbb{N}$.

3. Step – Relative weak compactness. Resulting from uniform stability, compactness arguments allow for extracting subsequences which are weakly convergent. A fundamental result from functional analysis is the Eberlein-Šmulian theorem.

Lemma 3.2.2 (Eberlein-Šmulian theorem [55])**.** *Assume that* V *is a reflexive Banach space, and let* $\{u_m\}_m \subset V$ *be a bounded sequence in* V*. Then there exists a subsequence* $\{u_{m_k}\}_k$ *that converges weakly in* V*.*

In the context of the model problem (3.2.1), the Eberlein-Šmulian theorem yields the existence of a $u \in H^1(0,T;V^\star) \cap L^2(0,T;V)$ such that for $m,N \to \infty$ it holds that

$$\begin{aligned} \bar{u}_{mN} &\rightharpoonup u && \text{in } L^2(0,T;V), \\ \hat{u}_{mN} &\rightharpoonup u && \text{in } L^2(0,T;V), \\ \partial_t \hat{u}_{mN} &\rightharpoonup \partial_t u && \text{in } L^2(0,T;V^\star), \end{aligned}$$

(up to subsequences), where $\rightharpoonup$ denotes weak convergence.

If $\mathcal{A}$ in the model problem (3.2.1) allows for considering the limit of $\mathcal{A}(\bar{u}_{mN})$ towards $\mathcal{A}(u)$ (in the right sense) for $m,N \to \infty$, it is already possible to conclude that u is a solution to the model problem. However, for non-linear problems in general, strong convergence may often be required.

4. Step – Relative compactness and identification of a solution. For this step, we assume that for considering the limit of (3.2.5) for $m,N \to \infty$, it is sufficient to establish strong convergence $\bar{u}_{mN} \to u$ in $L^2(0,T;B)$ (up to a subsequence), $V \subset B$ is compactly embedded. We present three possible techniques for concluding that. The first, the famous *Aubin-Lions lemma*, is specific for evolutionary problems.

Lemma 3.2.3 (Aubin-Lions lemma [13])**.** *Let B be a Banach space. Let $\{v_m\}_m \subset L^p(0,T;B)$, $1 \leq p < \infty$. The sequence $\{v_m\}_m$ is relatively compact in $L^p(0,T;B)$ if the following two are fulfilled:*

- *$\{v_m\}_m$ is uniformly bounded in $L^p(0,T;V)$, for some $V \subset B$ with compact embedding.*
- *$\{\partial_t v_m\}_m$ is uniformly bounded in $L^p(0,T;W)$, for some $W \supset B$ with a continuous embedding.*

In the context of the modeling problem (3.2.1), assuming $V \subset B \subset V^\star (= W)$ satisfies the properties of Lemma 3.2.3, relative compactness of $\{\hat{u}_{mN}\}_{mN}$ in $L^2(0,T;B)$ can be concluded. Since $\bar{u}_{mN} - \hat{u}_{mN} \to 0$ in $L^2(0,T;B)$ for $N \to \infty$ by construction of the time interpolations, also relative compactness of $\{\bar{u}_{mN}\}_{mN}$ in $L^2(0,T;B)$ can be concluded. Ultimately, one can extract subsequences of $\{\bar{u}_{mN}\}_{m,N}$ and $\{\hat{u}_{mN}\}_{m,N}$, converging to a solution of the model problem (3.2.1).

For particular problems, the assumptions of the Aubin-Lions lemma might be too strong. A version, relaxing the condition on the time-derivatives, has been given by Simon [138].

Lemma 3.2.4 (A relaxed Aubin-Lions lemma [138])**.** *Let B be a Banach space. Let $\{v_m\}_m \subset L^p(0,T;B)$, $1 \leq p < \infty$. The sequence $\{v_m\}_m$ is relatively compact in $L^p(0,T;B)$ if the following two are fulfilled:*

- *$\{v_m\}_m$ is uniformly bounded in $L^p(0,T;V)$, for some $V \subset B$ with compact embedding.*
- *$\int_\tau^T \|v_m(t) - v_m(t-\tau)\|_B^p \, dt \leq \mathcal{O}(\tau)$, where $\mathcal{O}$ denotes a function such that $\mathcal{O}(\tau) \to 0$ as $\tau \to 0$.*

Furthermore, for problems, for which V is some subspace of $L^p(\Omega)$ for a domain $\Omega \subset \mathbb{R}^N$, $N \in \mathbb{N}$, one more relaxation may be helpful: the Riesz-Frechet-Kolmogorov compactness criterion. It is in particular often utilized when employing non-conforming, finite dimensional, dense subspace in the Galerkin method, based, e.g., on finite volume techniques [74].

Lemma 3.2.5 (Riesz-Frechet-Kolmogorov compactness criterion [40]). *Let G be a bounded set in $L^p\left(\mathbb{R}^N\right)$ with $1 \leq p < \infty$, $N \in \mathbb{N}$. Assume that*

$$\lim_{|h| \to 0} \|g(\cdot + h) - g(\cdot)\|_{L^p(\mathbb{R}^N)} = 0 \quad \textit{uniformly in } g \in G.$$

Then the closure of $G|_\Omega := \{g : \Omega \to \mathbb{R} \mid g \in G\}$ is compact for any measurable set $\Omega \subset \mathbb{R}^N$ with finite measure.

For instationary problems, the space $\mathbb{R}^N$ can be identified with $\mathbb{R}^d \times \mathbb{R}$, i.e., $N = d + 1$, with d the dimension of the physical space. The product space thereby covers the spatial and temporal spaces. For stationary problems, it is $N = d$.

3.3 Doubly non-linear evolution equations

For doubly non-linear evolution equations, which are closely related to generalized gradient flows, cf. Section 2.2, high-level abstract well-posedness results have been established in the literature. Those allow for a relatively simple analysis of the well-posedness provided a problem satisfies certain assumptions. We note that such high-level results are often derived by the Galerkin method combined with compactness arguments as introduced above.

In the following, we recall two classical results for simple setups. However, we emphasize that far more involved models can be analyzed in the context of doubly non-linear evolution equations, including state-dependent, non-smooth dissipation potentials, as well as non-autonomous, non-smooth energy potentials [56, 108].

We begin with a classical well-posedness result for gradient flows of convex functionals in Hilbert spaces, i.e., for the particular case of quadratic dissipation. For basic definitions and properties of convex functions, as the domain dom, and their subdifferentials ∂, we refer to the textbook [68].

Lemma 3.3.1 (Well-posedness for gradient flows of convex functionals [39]). *Let H be a real Hilbert space, and let $\mathcal{E} : H \to \mathbb{R} \cup \{\infty\}$ be convex and proper. Let $u_0 \in H$. Then there exists a unique function $u \in C\left([0, \infty); H\right)$ such that*

- *$u(0) = u_0$;*
- *$\partial_t u(t)$ exists in the classical sense for almost every $t > 0$;*
- *There exists a function $\zeta : [0, \infty) \to H^\star$ such that $\zeta(t) \in \partial\mathcal{E}(u(t))$ for every t, and*

$$(\partial_t u(t), w) = -\langle \zeta(t), w \rangle \quad \textit{for a.e. } t > 0 \textit{ and for all } w \in H;$$

- *At every $t \geq 0$, $\zeta(t)$ is the element of minimal norm in $\partial\mathcal{E}(u(t))$.*

Interestingly, regularity of solutions can be directly concluded from Lemma 3.3.1. From the existence of $\zeta(t) \in \partial\mathcal{E}(u(t))$ it immediately follows that $u(t) \in \operatorname{dom}\partial\mathcal{E}$ (for all times t), which might be stronger than $u(t) \in H$.

On the foundation of the fundamental works on classical gradient flows by Komura [100], Crandall and Pazy [60], and Brezis [38, 39], the theory of doubly non-linear evolution equations took off. We present one basic well-posedness result for a regular case, closely connected to generalized gradient flows as presented in Section 2.2.

Theorem 3.3.2 (Well-posedness for regular doubly evolution equations [56])**.** *Let B be a real, reflexive, and strictly convex Banach space. Let the subspace $V \subset B$ be a reflexive Banach space, dense and compactly embedded in B. Let $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Let the dissipation potential $\mathcal{D}$ and the energy potential $\mathcal{E}$ satisfy:*

- *$\mathcal{D} : \mathcal{B} \to [0, \infty)$ is differentiable, with $\mathcal{D}(0) = 0$, and $\nabla\mathcal{D}$ is coercive and continuous, i.e., there exist constants $C_1, C_2, C_3 > 0$ satisfying for all $u \in B$*

$$\begin{aligned} \langle \nabla\mathcal{D}(u), u\rangle &\geq C_1 \|u\|_B^p - C_2, \\ \|\nabla\mathcal{D}(u)\|_{B^\star}^q &\leq C_3 \left(\|u\|_B^p + 1\right). \end{aligned}$$

- *$\mathcal{E} : \mathcal{B} \to (-\infty, \infty]$ is proper, convex, and lower semicontinuous, such that* $\operatorname{dom}\mathcal{E} \subset V$ *and*

$$\|u\|_B^p + \mathcal{E}(u) \to \infty \textit{ whenever } u \in \operatorname{dom}\mathcal{E},\ \|u\|_V \to \infty.$$

Ultimately, for all $u_0 \in \operatorname{dom}\mathcal{E}$, there exists a $u \in W^{1,p}(0,T;B) \cap L^\infty(0,T;V)$ with $\nabla\mathcal{D}(\partial_t u), \nabla\mathcal{E}(u) \in L^q(0,T;B^\star)$ satisfying

$$\boldsymbol{\nabla}\mathcal{D}\left(\partial_t u(t)\right) + \boldsymbol{\nabla}\mathcal{E}\left(u(t)\right) = 0 \quad \textit{a.e. in } (0,T); \quad u(0) = u_0. \tag{3.3.1}$$

If $\nabla\mathcal{D}$ or $\nabla\mathcal{E}$ is linear and self-adjoint, the solution is unique.

A direct consequence of the characterization of doubly non-linear evolution equations as minimization problem for the time derivative, cf. Section 2.2, is that solutions u to (3.3.1) satisfy the regularity property

$$\int_0^t \mathcal{D}\left(\partial_t u(s)\right) ds + \mathcal{E}\left(u(t)\right) \leq \mathcal{E}\left(u(0)\right) \quad \text{for a.e. } t \in (0,T).$$

3.4 Convex minimization

Time-discrete problems or stationary problems that can be formulated as a convex minimization problem can be analyzed by means of convex analysis. We mention a fundamental result from calculus of variations in the context of convex energies. In some sense, it is the analog result to Lemma 3.3.2 for stationary (or time-discrete) problems.

Lemma 3.4.1 (Well-posedness for convex minimization [55, 68])**.** *Consider the problem*

$$\inf_{u \in C} \mathcal{E}(u), \tag{3.4.1}$$

where V is a reflexive Banach space, the objective function $\mathcal{E} : V \to \mathbb{R} \cup \{\infty\}$ is a proper, convex, lower semi-continuous function, and the feasible set $C \subset V$ is non-empty, closed, and convex. If C is bounded or $\mathcal{E}$ is coercive over C, i.e., $\mathcal{E}(u) \to \infty$ for $x \in C$ with $\|u\|_V \to \infty$, then (3.4.1) *has a solution. It is unique if $\mathcal{E}$ is strictly convex over C.*

3.5 Saddle point problems

Equality-constrained minimization problems are naturally related to saddle point problems via Lagrangian formulations. In the following, we present a classical well-posedness result for perturbed linear saddle-point formulations. It can be for instance applied for establishing the well-posedness of time-discretized linear Biot equations, cf. Section 2.1.3.

Lemma 3.5.1 (Well-posedness of perturbed saddle-point problems [27])**.** *Let V, Q be Hilbert spaces. Let $a : V \times V \to \mathbb{R}$, and $c : Q \times Q \to \mathbb{R}$ be continuous symmetric positive semi-definite bilinear forms. Moreover, let $b : V \times Q \to \mathbb{R}$ be a continuous bilinear form with associated canonical linear operators $B : V \to Q^\star$ and $B^\top : Q \to V^\star$. Assume the image of B is closed. Abbreviate their kernels $K := \ker B$, and $H := \ker B^\top$. Finally, assume a, b, and c satisfy:*

- *a is coercive on K, i.e., $a(v,v) \geq \alpha \|v\|_V^2$ for all $v \in K$, for some $\alpha > 0$;*
- *b satisfies an inf-sup condition, i.e., $\inf_{q \in H^\perp} \sup_{v \in V} \frac{b(v,q)}{\|q\|_Q \|v\|_V} = \inf_{v \in K^\perp} \sup_{q \in Q} \frac{b(v,q)}{\|q\|_Q \|v\|_V} = \beta$, for some $\beta > 0$, where $H^\perp$ denotes the orthogonal complement of H;*
- *c is coercive on H, i.e., $c(q,q) \geq \gamma \|q\|_Q^2$ for all $q \in H$, for some $\gamma > 0$.*

Then for all $f \in V^\star$, $g \in Q^\star$, there exists a unique $(u,p) \in V \times Q$ satisfying

$$a(u,v) - b(v,p) = \langle f, v \rangle \quad \forall v \in V, \tag{3.5.1}$$

$$b(u,q) + c(p,q) = \langle g, q \rangle \quad \forall q \in Q. \tag{3.5.2}$$

The following result from the theory of saddle point problems is a special case of the Banach Closed Range theorem – in certain cases also called Thomas' lemma [150]. It is a useful low-level tool for the analysis of both linear and non-linear coupled problems with a skew-symmetric coupling. It will be for instance employed in the analysis of Schur-complement based block-partitioned solvers for Biot equations.

Lemma 3.5.2 (Inf-sup argument [27])**.** *Let V and Q be Hilbert spaces, and let B be a linear continuous operator from V to $Q^\star$. Denote by $B^\top : Q \to V^\star$ the (canonical) transposed operator of B. Then, the following two statements are equivalent:*

- *$B^\top$ is bounding, i.e., $\exists \beta > 0$ such that $\left\|B^\top q\right\|_{V^\star} \geq \beta \|q\|_Q \; \forall q \in Q$.*
- *$\exists L_B : Q^\star \to V$, linear and bounded, such that $B(L_B(\xi)) = \xi \; \forall \xi \in Q^\star$ with $\|L_b\| = \beta^{-1}$.*

Chapter 4

Numerical solution of coupled problems

Given a (well-posed) potentially non-linear coupled PDE, closed-form solutions are in general not available. Instead computable numerical approximations may be employed. In this section, the conceptual procedure undertaken in this work is presented. It starts with the finite dimensional approximation of the problem by discretization in space and time, making it accessible to computer codes. The resulting potentially non-linear algebraic systems are solved using iterative solvers. This includes linearization schemes as Newton's method or L-scheme linearizations accounting for non-linearities, Schur-complement-based iterative splitting schemes tailored for saddle-point problems, or iterative splitting schemes based on block-coordinate descent methods for convex minimization problems. Finally, we comment on DUNE, the numerics environment used for the implementation of the above numerical method in the course of this thesis.

4.1 Discretization in space and time

In order to solve an infinitely dimensional problem utilizing a computer simulation, a finite dimensional approximation is required. Limited to the coupling of flow and deformation, poroelasticity models conceptually consist of two main blocks – equations from structural mechanics and fluid dynamics. In practice, discretization schemes tailored for the separate subproblems are combined for discretizing the coupled problem, mostly involving the classical *Finite Element Method* (FEM), the *Finite Volume Method* (FVM), or the *Mixed Finite Element Method* (MFEM), combined with an *Implicit time-stepping technique*, e.g., the Implicit Euler method. In the following, we comment on central properties of the spatial discretization methods, and refer to the textbooks [27, 71, 74] for detailed introductions.

The FEM is traditionally used for structural analyses and elliptic problems. It is based on the conforming Galerkin method, applied to a primal problem formulation, e.g., the displacement-formulation of linear elasticity, or the pressure formulation of single-phase flow in porous media. Major strengths of the FEM are a sound, mathematical foundation including the stability and convergence analysis, and a natural extension to higher order accurate FEMs. However, the FEM is less favored in the field of fluid dynamics due to a lack of conservation properties. In the context of poroelasticity, it is often utilized for discretizing the mechanics equation, combined with a conservative scheme for the fluid flow problem, cf., e.g., [75, 155].

The FVM is constructed to reflect the nature of conservation and balance laws. Therefore it is traditionally used in fluid dynamics and for hyperbolic problems. Convergence can be established; however, compared to the FEM the extension to higher order accurate FVMs is not trivial. Guaranteeing local conservation of mass and linear momentum, the FVM is widely-used for the discretization of flow in porous media [46, 74, 91, 92], linear elasticity [95, 148], and poroelasticity [117].

The MFEM joins advantages of both the FEM and the FVM, being a conforming finite element method based on mixed formulations. For this, separate equations for conservation laws and constitutive relations, as Darcy's law, are utilized. Thereby, it allows for a locally conservative discretization, but simultaneously enables classical FEM theory for the analysis, and additionally allows for a natural extension to higher order accurate MFEMs. The major drawback is the introduction of Lagrange multipliers, resulting in a saddle-point formulation. Discrete function spaces therefore have to be carefully chosen fulfilling an inf-sup stability criterion. By that the MFEM may become relatively expensive, depending on the problem. However, we mention that hybridization or the use of inexact quadrature allows for reducing the computational complexity without loss of accuracy. By the latter in fact, low-order MFEMs may translate to known cell-centered FVMs [7, 8, 15, 156]. The MFEM is widely-used for the discretization of flow in porous media [127, 160], linear elasticity [11, 12], and poroelasticity [44, 75, 155].

Finally, a range of alternative techniques has been established in the literature for the discretization of geomechanics and flow in porous media. We mention finite difference schemes [81], the discontinuous Galerkin method [125], the virtual element method [9, 80], the multiscale FEM [50], the adaptive FEM [4], and stabilized FEM [130].

4.1.1 Example: FEM-FEM discretization of Biot's consolidation model

In this section, we introduce a conforming discretization of the two-field formulation of the linear Biot equations (2.1.20)–(2.1.21). It will provide the foundation for later discussion of the fixed-stress split in Section 4.3. Based on the above discussion, we stress that we do not advocate the use of a classical FEM discretization of the fluid flow equations. Nevertheless, we choose it for a compact presentation. For simplicity, homogeneous boundary conditions are assumed on $\partial\Omega$ for both the displacement and fluid pressure.

For the temporal discretization, a time interval of interest $(0,T)$ is partitioned in intervals $(t_{n-1}, t_n]$ of (for simplicity) fixed time step size Δt. Given a tessellation $\mathcal{T}_h = \{T\}_T$ of Ω, conforming, discrete (finite element) spaces $\boldsymbol{V}_h = \boldsymbol{V}_h(\mathcal{T}_h) \subset H_0^1(\Omega)^d$ and $Q_h(\mathcal{T}_h) \subset H_0^1(\Omega)$ are assumed to be given for the approximation of the structural displacement and the fluid pressure; $H_0^1(\Omega)$ denotes the subspace of $H^1(\Omega)$ with zero trace on the boundary $\partial\Omega$.

A fully-discrete approximation of the two-field formulation of Biot's consolidation model (2.1.20)–(2.1.21) is then obtained by applying the (conforming) Galerkin method together with the Implicit Euler method. Given some discrete, initial data $(\boldsymbol{u}_h^0, p_h^0) \in \boldsymbol{V}_h \times Q_h$, the approximation at the n-th time step, $n \geq 1$, reads: *Given* $(\boldsymbol{u}_h^{n-1}, p_h^{n-1}) \in \boldsymbol{V}_h \times Q_h$, *define the fluid content* $\theta^{n-1} := \frac{1}{M} p_h^{n-1} + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}_h^{n-1}$, *and find* $(\boldsymbol{u}_h^n, p_h^n) \in \boldsymbol{V}_h \times Q_h$ *satisfying for all* $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h \times Q_h$

$$2\mu \left\langle \boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^n\right), \boldsymbol{\varepsilon}(\boldsymbol{v}_h) \right\rangle + \lambda \left\langle \boldsymbol{\nabla} \cdot \boldsymbol{u}_h^n, \boldsymbol{\nabla} \cdot \boldsymbol{v}_h \right\rangle - \alpha \left\langle p_h^n, \boldsymbol{\nabla} \cdot \boldsymbol{v}_h \right\rangle = \left\langle \rho \boldsymbol{g}, \boldsymbol{v}_h \right\rangle, \tag{4.1.1}$$

$$\frac{1}{M} \left\langle p_h^n, q_h \right\rangle + \alpha \left\langle \boldsymbol{\nabla} \cdot \boldsymbol{u}_h^n, q_h \right\rangle + \Delta t \left\langle \frac{\boldsymbol{\kappa}}{\mu_\mathrm{w}} \left(\boldsymbol{\nabla} p_h^n - \rho_{\mathrm{w,ref}} \boldsymbol{g} \right), \boldsymbol{\nabla} q_h \right\rangle = \left\langle \theta^{n-1} + \Delta t \frac{h_\mathrm{w}(t_n)}{\rho_{\mathrm{w,ref}}}, q_h \right\rangle. \tag{4.1.2}$$

We remark, that by virtue of the theory of saddle-point problems, $\boldsymbol{V}_h \times Q_h$ are required to be inf-sup stable with respect to the divergence operator, in order to guarantee parameter-robust stability, cf. Lemma 3.5.1. For instance, quadratic elements or the MINI element for the displacements, and linear elements for the pressure fulfill this requirement [27].

4.2 Iterative solution of algebraic systems

Next after discretization, an algebraic system of equations has to be solved. Provided there exists a solution, the main focus of this section is on the description of methods for approximating these solutions. For this, we consider the following algebraic model problem:

$$\textit{For given } \mathbf{F} : \mathbb{R}^N \to \mathbb{R}^N, \textit{ find } \mathbf{x}^\star \in \mathbb{R}^N \textit{ satisfying: } \mathbf{F}(\mathbf{x}^\star) = \mathbf{0}. \tag{4.2.1}$$

With (4.2.1) arising from discretizing PDEs, N may be very large, e.g., for large-scale simulations or highly accurate discretization schemes, or $\mathbf{F}$ may be non-linear. Both cases in general do not allow for the use of *direct methods*, which attempt to solve (4.2.1) within machine precision under fixed, computational cost. Instead, *iterative solvers* may be considered. Starting with an initial guess, approximations of solutions are successively improved until the error is acceptable, where the metric and tolerance for the error is controlled by the user.

In the following, we restrict the discussion to methods relevant to this thesis, and briefly comment on the special case that $\mathbf{F}$ is actually linear, as well as on Newton's method and L-scheme linearizations for the general case.

4.2.1 Special case of a linear problem

If $\mathbf{F}$ is linear, i.e., $\mathbf{F}(\mathbf{x}) = \mathbf{M}\mathbf{x} - \mathbf{b}$, for some $\mathbf{M} \in \mathbb{R}^{N \times N}$ and $\mathbf{b} \in \mathbb{R}^N$, three general approaches are most popular: *direct methods*, *iterative solvers* (for problems related to PDEs), and *preconditioned Krylov subspace methods*.

Direct methods utilize a suitable factorization of the matrix $\mathbf{M}$ and ultimately use forward and backward substitutions to solve (4.2.1). Those have the advantage of converging in a 'single iteration'. Furthermore, once a factorization of the matrix $\mathbf{M}$ has been constructed, low online cost per iteration allow for a cheap solution for changing right hand sides. This is in particular relevant for instationary problems. However, one has to note that the offline cost for constructing factorizations is relatively high in terms of computational operations and physical memory. For larger problems, the latter might prohibit the use of direct methods.

Iterative solvers instead approximate the solution to (4.2.1) in a sequence of iterations. Starting with an initial guess, the approximation is updated successively – favorably with lower computational cost per iteration than required for direct methods. Thus, compared to those, iterative solvers may either allow for making computations possible at all or lead to improved performance. This especially holds true for larger problems if efficient parallelization is available. Typical examples of iterative solvers for linear systems related to PDEs are multigrid methods [87], domain decomposition methods [140], and – as a special case of these – iterative splitting schemes for coupled problems, cf. Section 4.3. A slight drawback of these methods is the strict need for the methods to be contractive in order to ensure convergence, which follows directly from the fact that these methods eventually can be identified as preconditioned Richardson iterations, cf., e.g., [158]. This substantiates the need for theoretically studying the convergence of iterative solvers.

Krylov subspace methods are a particular class of iterative solvers specifically designed for linear, algebraic systems. Their general idea lies in minimizing the residual over a sequence of finite dimensional subpaces, the Krylov subspaces. For a detailed introduction and analysis, we refer to the textbook [132]. Most importantly, compared to the Richardson iteration, Krylov subspace methods are much more robust, being in principle always convergent. For practical convergence however, preconditioning is required, favorably resulting in a clustered spectrum of the preconditioned system, not necessary contained in the unit ball. The choice of suitable preconditioners is often problem-dependent and a large area of research. For linear systems related to PDEs one can often utilize iterative solvers as those mentioned above.

4.2.2 Newton's method

For the numerical solution of non-linear problems (4.2.1), iterative linearization schemes are usually required. The certainly most popular linearization method is *Newton's method* (or due to historical reasons sometimes also called the Newton-Raphson-Simpson method) [67]. It successively determines approximations of solutions to (4.2.1) by utilizing first-order Taylor approximations of $\mathbf{F}$. Starting with an initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^N$, the i-th iteration reads, $i \geq 1$: *Given an approximation* $\mathbf{x}^{(i-1)} \in \mathbb{R}^N$, *find* $\mathbf{x}^{(i)} \in \mathbb{R}^N$ *satisfying*

$$\mathbf{F}\left(\mathbf{x}^{(i-1)}\right) + \boldsymbol{\nabla}\mathbf{F}\left(\mathbf{x}^{(i-1)}\right)\left(\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}\right) = \mathbf{0}. \tag{4.2.2}$$

Evidently, $\mathbf{x}^{(i)}$ is only well-defined, if $\mathbf{F}$ is differentiable and its derivative $\boldsymbol{\nabla}\mathbf{F}$ is invertible.

By the Newton-Kantorovich theorem [67, 94], Newton's method converges locally and quadratically if $\boldsymbol{\nabla}\mathbf{F}$ is locally Lipschitz continuous and its inverse is bounded. Possible drawbacks arise immediately from that result and may be indeed also observed in practice: global convergence is not guaranteed if the initial guess is not well chosen. Damping strategies allow for the recovery of robustness. Furthermore, problems involving, e.g., Hölder continuous non-linearities may give rise to ill-conditioned Jacobians as, e.g., for Richards' equation [53, 105] or unsaturated poroelasticity, cf. Paper E, such that the regularity assumptions may not be satisfied. Efforts to remedy difficulties in the context of porous media applications include the re-parametrization of constitutive relations [37], trust-region Newton methods [104, 154], and reordering techniques [101, 113], just to mention a few.

4.2.3 L-scheme linearization

The *L-scheme* [126, 139] is an inexact Newton method. In its essence, the idea is to replace the exact Jacobian $\nabla\mathbf{F}\left(\mathbf{x}^{(i-1)}\right)$ in Newton's method (4.2.2) by a constant approximation $\mathbf{L} \in \mathbb{R}^{N\times N}$. Thus, starting with an initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^N$, the i-th iteration of the resulting scheme reads, $i \geq 1$: *Given an approximation* $\mathbf{x}^{(i-1)} \in \mathbb{R}^N$*, find* $\mathbf{x}^{(i)} \in \mathbb{R}^N$ *such that*

$$\mathbf{F}\left(\mathbf{x}^{(i-1)}\right) + \mathbf{L}\left(\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}\right) = \mathbf{0}. \tag{4.2.3}$$

The reason for applying the L-scheme is a possible increase of robustness compared to Newton's method – in particular in cases in which Newton's method fails to converge. By construction, the L-scheme can be applied to non-smooth problems as no evaluation of derivatives is required. By that the conditioning of the problem potentially improves, and the cost per iteration compared to Newton's method is lowered [105]. On the other hand, at most linear convergence can be expected. We finally remark that for linear $\mathbf{F}$, the L-scheme linearization (4.2.3) is identical with a preconditioned Richardson iteration [132].

In order to allow for better performance, the L-scheme linearization may be only applied to (non-smooth) parts of non-linearities. For instance, assuming $\mathbf{F}$ decomposes as

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}_{\text{lin}}\mathbf{x} + \mathbf{F}_{\text{qlin}}(\mathbf{x})\mathbf{x} + \mathbf{F}_{\text{L}}(\mathbf{x}) + \mathbf{F}_{\text{smooth}}(\mathbf{x})$$

with $\mathbf{F}_{\text{lin}} \in \mathbb{R}^{N\times N}$, $\mathbf{F}_{\text{qlin}} : \mathbb{R}^N \to \mathbb{R}^{N\times N}$, some non-smooth $\mathbf{F}_{\text{L}} : \mathbb{R}^N \to \mathbb{R}^N$, and smooth $\mathbf{F}_{\text{smooth}} : \mathbb{R}^N \to \mathbb{R}^N$, a possible approximation of the Jacobian $\nabla\mathbf{F}(\mathbf{x})$ might be given by

$$\nabla\mathbf{F}(\mathbf{x}) \approx \mathbf{F}_{\text{lin}} + \mathbf{F}_{\text{qlin}}(\mathbf{x}) + \mathbf{L} + \mathbf{F}'_{\text{smooth}}(\mathbf{x})$$

for some constant $\mathbf{L} \in \mathbb{R}^{N\times N}$. The resulting inexact Newton method, also sometimes regarded as *modified Picard method*, would involve iteration-dependent Jacobians, increasing the computational cost per iteration. On the other hand, same properties as for the plain L-scheme can be concluded regarding non-smooth problems.

A key question is, how $\mathbf{L}$ should be chosen. First of all, it is important to note that in general constant linearizations can only be appropriate for non-decreasing Lipschitz continuous non-linearities. Ultimately, in algebraic terms $0 \leq \langle \mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \leq \langle \mathbf{L}(\mathbf{x} - \mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ is sufficient, cf. Paper E. Finally, suitable explicit choices for $\mathbf{L}$ are problem-dependent. L-scheme linearizations have been shown to be robust and linearly convergent with mesh-independent rates for several porous media applications as, e.g., Richards' equation [105], two-phase flow in porous media [129], non-linear single-phase poroelasticity [28, 29], and unsaturated poroelasticity, cf. Paper E. Moreover, extensions can be made to Hölder continuous non-linearities by including information on stopping criteria [128], see also Paper F.

4.2.4 Relaxation strategies for iterative solvers

In order to increase the robustness or speed of iterative solvers, relaxation techniques can be employed. The basic concept is to enhance each iteration by a correction step, cf. Algorithm 1.

Algorithm 1: Relaxation of iterative solvers illustrated for Newton's method

```
Prediction: Find $\hat{\mathbf{x}}^{(i)}$ satisfying, e.g., (4.2.2).
Relaxation: Choose $\omega \in \mathbb{R}$ based on some heuristic or mathematical foundation.
Correction: Set $\mathbf{x}^{(i)} := \hat{\mathbf{x}}^{(i)} + \omega \left(\hat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i-1)}\right)$.
```

We mention three relaxation concepts frequently used in the literature:

- *Line search for first-order optimality conditions* [115]: Typically line search is applied for problems which correspond to a minimization problem. For instance, let $\mathbf{F} = \boldsymbol{\nabla}\mathcal{E}$ for some $\mathcal{E} : \mathbb{R}^N \to \mathbb{R}$. Then (4.2.1) defines the first-order optimality conditions of minimizing $\mathcal{E}$. Ultimately, ω is chosen by minimization along the search direction

$$\omega := \arg\inf_{w \in \mathbb{R}} \mathcal{E}\left(\hat{\mathbf{x}}^{(i)} + w\left(\hat{\mathbf{x}}^{(i)} - \mathbf{x}^{(i-1)}\right)\right),$$

  where most often inexact minimization is sufficient.

- *Residual based descent* [67]: Zeros of $\mathbf{F}$ solve the minimization problem

$$\inf_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{F}(\mathbf{x})\| . \tag{4.2.4}$$

  Thereby, the relaxation parameter ω can be chosen as (inexact) minimizer of the residual (4.2.4) along the search direction, analogous to the above line search approach.

- *Anderson acceleration* [10]: Anderson acceleration is a multi-secant method for vector-valued functions. Thereby, it can be utilized for correcting predictions made by iterative solvers for finding zeros. The main idea is that an arbitrary amount of previous search directions $\hat{\mathbf{x}}^{(j)} - \mathbf{x}^{(j-1)}$ is utilized for approximating the Jacobian of $\mathbf{F}$ in low dimensions [153]. For depth $m \geq 1$, the following minimization problem has to be solved:

$$\inf_{\boldsymbol{\alpha} \in \mathbb{R}^{m+1}} \left\| \sum_{j=0}^{m} \alpha_j \left(\hat{\mathbf{x}}^{(i-j)} - \mathbf{x}^{(i-j-1)}\right) \right\|_2 \quad \text{such that } \sum_{j=0}^{m} \alpha_j = 1.$$

  Opposing to the relaxation as in Algorithm 1, the corrected approximation is defined by

$$\mathbf{x}^{(i)} := \sum_{j=0}^{m} \alpha_j \hat{\mathbf{x}}^{(i-j)}.$$

  Anderson acceleration can be interpreted as quasi-Newton method and preconditioned non-linear GMRES [153], and it has recently been showed to locally accelerate any contractive fixed point iteration [72].

A major difference of Anderson acceleration to line search and a residual based descent is that no assembly of any objective function or Jacobian is required for setting up the minimization problem. This is in particular beneficial for splitting schemes for coupled problems, cf. Paper E. On the other hand, that lack of global information prevents guaranteed robustness, especially for non-contractive fixed point iterations. Nevertheless, it may be applied for damping Newton's method as also the residual based descent.

4.3 Block-partitioned solvers for saddle-point problems

Discretized coupled PDEs naturally inherit a block structure of the corresponding continuous problems. Naturally, a block structure motivates the design of block-partitioned iterative solvers. Such may be beneficial opposing to fully-implicit *monolithic solvers* as smaller and better conditioned systems are solved instead. In addition, a major advantage of a staggered approach is that it allows for a flexible design of computer codes. Separate simulators with solver technologies tailored to independent subproblems may be utilized, which for instance has been a driving force in the context of industrial poroelasticity applications. On the other hand, block-partitioned solvers are not inherently unconditionally stable, whereas monolithic solvers are per sé robust. Thus, the numerical analysis is an essential component of the development processes of block-partitioned solvers.

Linear saddle point problems are a particular class of block structured problems. In the context of block-partitioned solvers, they naturally give rise to Schur-complement-based *splitting schemes*.

In this section, the L-scheme, cf. Section 4.2.3, is utilized for formulating iterative splitting schemes for linear saddle point problems involving a feasible approximation of the exact Schur complement. Conditions for the convergence of the resulting scheme with variable stabilization are established. This allows for deducing feasible choices of the stabilization introduced by the L-scheme.

With Biot's consolidation model, cf. Section 2.1.3, falling into the class of linear (block-structured) saddle point problems, the general, algebraic L-scheme-based framework is applied to it. That results in an L-scheme perspective of the widely-used *fixed-stress split* for the linear Biot equations, an often recurring method in this work and chosen here as representative for iterative splitting schemes for poroelasticity problems. By that, novel convergence results are deduced for the fixed-stress split.

Despite the focus on linear problems, we emphasize that similar concepts can be employed for non-linear problems. A short comment is also provided in Section 4.3.3.

4.3.1 Schur-complement-based splitting scheme via the L-scheme

We illustrate the essentials of Schur-complement-based iterative splitting schemes, utilizing the following linear model problem. It is in particular representative for a wide range of poroelasticity problems as for instance discretized linear poroelasticity, cf. Section 4.1.1.

$$\textit{Find } (\mathbf{u},\mathbf{p}) \in \mathbb{R}^{n_\mathrm{u}} \times \mathbb{R}^{n_\mathrm{p}} \textit{ satisfying} \quad \begin{bmatrix} \mathbf{A} & -\mathbf{B}^\top \\ \mathbf{B} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{g} \\ \mathbf{h} \end{bmatrix}, \tag{4.3.1}$$

where $n_\mathrm{u}, n_\mathrm{p} \in \mathbb{N}$, $\mathbf{A} \in \mathbb{R}^{n_\mathrm{u}\times n_\mathrm{u}}$, $\mathbf{B} \in \mathbb{R}^{n_\mathrm{p}\times n_\mathrm{u}}$, $\mathbf{C}^{n_\mathrm{p}\times n_\mathrm{p}}$, $\mathbf{g} \in \mathbb{R}^{n_\mathrm{u}}$, $\mathbf{h} \in \mathbb{R}^{n_\mathrm{p}}$. Assume the matrix $\mathbf{A}$ is symmetric positive definite, $\mathbf{B}$ has full rank, and $\mathbf{C}$ is symmetric positive semi-definite. By the theory of saddle-point problems, cf. Lemma 3.5.1, (4.3.1) has a unique solution.

The block structure of (4.3.1) allows for a systematic decoupling of solving for $\mathbf{u}$ and $\mathbf{p}$. For instance, (4.3.1) is equivalent with

$$\begin{bmatrix} \mathbf{A} & -\mathbf{B}^\top \\ \mathbf{0} & \mathbf{S}+\mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{g} \\ \mathbf{h} - \mathbf{B}\mathbf{A}^{-1}\mathbf{g} \end{bmatrix}, \tag{4.3.2}$$

where $\mathbf{S} := \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$ denotes the (exact) Schur complement. The formulation (4.3.2) effectively allows for consecutively solving for $\mathbf{p}$ and $\mathbf{u}$. However, the construction of $\mathbf{S}$ requires the inversion of $\mathbf{A}$, which is in general infeasible.

Despite its impractical and rather theoretical character, (4.3.2) yields the foundation for the construction of efficient, iterative splitting schemes, which do not require $\mathbf{A}^{-1}$. In the following, this is illustrated by utilizing the L-scheme, cf. Section 4.2.3. Although linear, we treat the pressure problem as non-linear and define

$$\mathbf{F} : \mathbb{R}^{n_\mathrm{p}} \to \mathbb{R}^{n_\mathrm{p}}, \qquad \mathbf{F}(\mathbf{p}) := (\mathbf{S}+\mathbf{C})\,\mathbf{p} - \left(\mathbf{h} - \mathbf{B}\mathbf{A}^{-1}\mathbf{g}\right).$$

Let for now $\mathbf{L} \in \mathbb{R}^{n_\mathrm{p}\times n_\mathrm{p}}$ be some symmetric positive definite matrix. Provided some $\mathbf{p}^{(0)} \in \mathbb{R}^{n_\mathrm{p}}$, the L-scheme defines an iterative scheme with the i-th iteration reading as follows: *For $i \geq 1$, given $\mathbf{p}^{(i-1)} \in \mathbb{R}^{n_\mathrm{p}}$, find $\mathbf{p}^{(i)} \in \mathbb{R}^{n_\mathrm{p}}$ such that*

$$\mathbf{F}\left(\mathbf{p}^{(i-1)}\right) + (\mathbf{L}+\mathbf{C})\left(\mathbf{p}^{(i)} - \mathbf{p}^{(i-1)}\right) = \mathbf{0}. \tag{4.3.3}$$

The definition of $\mathbf{F}$ contains the exact Schur complement $\mathbf{S}$. In order to get rid of it, an auxiliary problem is utilized. Motivated by (4.3.2), introduce $\mathbf{u}^{(i)}$ for all $i \geq 0$ as the solution of the problem: *Given $\mathbf{p}^{(i)} \in \mathbb{R}^{n_\mathrm{p}}$, find $\mathbf{u}^{(i)} \in \mathbf{R}^{n_\mathrm{u}}$ such that*

$$\mathbf{A}\mathbf{u}^{(i)} = \mathbf{g} + \mathbf{B}^\top\mathbf{p}^{(i)}. \tag{4.3.4}$$

Then (4.3.3) is equivalent with

$$\mathbf{B}\mathbf{u}^{(i-1)} + \mathbf{L}\left(\mathbf{p}^{(i)} - \mathbf{p}^{(i-1)}\right) + \mathbf{C}\mathbf{p}^{(i)} = \mathbf{h}. \tag{4.3.5}$$

Finally, (4.3.4)–(4.3.5) (in reverse order) define a feasible two-step splitting scheme, assuming $\mathbf{A}$ and $\mathbf{L}+\mathbf{C}$ can be efficiently 'inverted'. In fact, various prominent splitting schemes in the field of poroelasticity as, e.g. the fixed-stress split, cf. Section 4.3.2, are of that form.

As always for the L-scheme, a central question is, how should $\mathbf{L}$ be chosen. We answer this question by a brief convergence analysis, closely related to the standard analysis of the modified Richardson iteration [132].

Lemma 4.3.1 (Convergence of the algebraic splitting scheme)**.** *Let* $(\mathbf{u}, \mathbf{p}) \in \mathbb{R}^{n_\mathrm{u}} \times \mathbb{R}^{n_\mathrm{p}}$ *denote the solution to* (4.3.1)*. Let* $\mathbf{p}^{(i)} \in \mathbb{R}^{n_\mathrm{p}}$*,* $i \geq 1$*, be defined by the iterative scheme* (4.3.3)*. Let* $\mathbf{e}_\mathbf{p}^{(i)} := \mathbf{p}^{(i)} - \mathbf{p}$ *denote the error at iteration* $i \geq 0$*. For all* $i \geq 1$*, it holds that*

$$\sqrt{\left\langle \mathbf{L}\mathbf{e}_\mathbf{p}^{(i)}, \mathbf{e}_\mathbf{p}^{(i)} \right\rangle} \leq \frac{\max\left\{\left|1 - \lambda_{\max}\left(\mathbf{L}^{-1/2}\mathbf{S}\mathbf{L}^{-1/2}\right)\right|, \left|1 - \lambda_{\min}\left(\mathbf{L}^{-1/2}\mathbf{S}\mathbf{L}^{-1/2}\right)\right|\right\}}{1 + \lambda_{\min}\left(\mathbf{L}^{-1/2}\mathbf{C}\mathbf{L}^{-1/2}\right)} \sqrt{\left\langle \mathbf{L}\mathbf{e}_\mathbf{p}^{(i-1)}, \mathbf{e}_\mathbf{p}^{(i-1)} \right\rangle}, \tag{4.3.6}$$

where $\lambda_{\max}(\cdot)$ *and* $\lambda_{\min}(\cdot)$ *denote the maximum and minimum absolute eigenvalues.*

Proof. Let $i \geq 1$. By subtraction of (4.3.3) and the second row of (4.3.2), we obtain the error propagation relation

$$(\mathbf{L} + \mathbf{C})\, \mathbf{e}_\mathbf{p}^{(i)} = (\mathbf{L} - \mathbf{S})\, \mathbf{e}_\mathbf{p}^{(i-1)}.$$

By rearranging terms it follows that

$$\mathbf{L}^{1/2}\mathbf{e}_\mathbf{p}^{(i)} = \left(\mathbf{I} + \mathbf{L}^{-1/2}\mathbf{C}\mathbf{L}^{-1/2}\right)^{-1} \left(\mathbf{I} - \mathbf{L}^{-1/2}\mathbf{S}\mathbf{L}^{-1/2}\right) \mathbf{L}^{1/2}\mathbf{e}_\mathbf{p}^{(i-1)}.$$

By multiplying $\mathbf{L}^{1/2}\mathbf{e}_\mathbf{p}^{(i)}$ to both sides, using the Cauchy-Schwarz inequality, and employ properties of matrix and vector norms, we obtain

$$\begin{aligned} \sqrt{\left\langle \mathbf{L}\mathbf{e}_\mathbf{p}^{(i)}, \mathbf{e}_\mathbf{p}^{(i)} \right\rangle} &\leq \left\| \left(\mathbf{I} + \mathbf{L}^{-1/2}\mathbf{C}\mathbf{L}^{-1/2}\right)^{-1} \left(\mathbf{I} - \mathbf{L}^{-1/2}\mathbf{S}\mathbf{L}^{-1/2}\right) \mathbf{L}^{1/2}\mathbf{e}_\mathbf{p}^{(i-1)} \right\|_2 \qquad (4.3.7) \\ &\leq \left\| \left(\mathbf{I} + \mathbf{L}^{-1/2}\mathbf{C}\mathbf{L}^{-1/2}\right)^{-1} \right\| \left\| \left(\mathbf{I} - \mathbf{L}^{-1/2}\mathbf{S}\mathbf{L}^{-1/2}\right) \right\|_2 \sqrt{\left\langle \mathbf{L}\mathbf{e}_\mathbf{p}^{(i-1)}, \mathbf{e}_\mathbf{p}^{(i-1)} \right\rangle}. \end{aligned}$$

The final result follows using standard tools from linear algebra. □

Revisiting the question for a suitable choice of **L**, the objective is the that iterative splitting scheme (4.3.3) – equivalently (4.3.4)–(4.3.5) – is contractive. By Lemma 4.3.1, the optimal choice for **L** is $\mathbf{L} = \mathbf{S}$. But again, this is infeasible. However, it becomes evident, that a close relation, i.e., $\mathbf{L} \approx \mathbf{S}$, e.g., in the sense of equivalent spectra, yields a contractive splitting scheme, similar to the concepts of norm-equivalent preconditioners [2, 107]. After all, a suitable choice is problem-dependent, involving $\mathbf{A}, \mathbf{B}$, and **C**. An example discussion in the context of the linear Biot equations is given in the subsequent section.

Remark 4.3.2 (Optimality of the convergence result)**.** *If* $\mathbf{C} \neq \omega\mathbf{L}$ *for some* $\omega \in \mathbb{R}$*, the final convergence result* (4.3.6) *is not sharp. In the view of the proof, in particular* (4.3.7)*, the use of the sub-multiplicativity of matrix norms causes loss of information on the interaction of the matrices. On the other hand, for the general case,* (4.3.6) *allows for a simple discussion of a suitable* **L** *opposing to* (4.3.7)*.*

4.3.2 Example: Fixed-stress split for Biot's consolidation model

The abstract considerations in the previous section allow for a slightly novel perspective onto the widely-used *fixed-stress split* for Biot's consolidation model for homogeneous media. The physically motivated scheme sequentially solves the fluid flow equations under fixed (volumetric) stress conditions, and the mechanics equations with updated fluid fields, iterating until convergence. When first introduced in [135], the main motivation lied in the possible use of independent structural mechanics and reservoir simulators. Since its introduction, the robustness of the fixed-stress split has been confirmed mathematically, including unconditional stability in the sense of a von Neumann analysis [98], guaranteed convergence for various discretizations [19, 30, 111, 143], multi-rate settings [5], two-grid settings [61], adaptive setting [3], and an equivalence to alternating minimization for a strongly convex energy functional, cf. Paper A. It is furthermore used as smoother for multigrid methods [82], preconditioner for Krylov subspace methods [51, 52, 158], and yields also the conceptual foundation for more involved poroelasticity models [28, 34, 83, 99, 157]. We emphasize that also alternative schemes are utilized in the literature including the undrained split [97], algebraically motivated, block-partitioned splitting schemes [76, 158], and norm-equivalent preconditioners [2, 90, 102], to only mention a few.

The scheme. In terms of the fully-discretized two-field formulation of the linear Biot equations (4.1.1)–(4.1.2), the fixed-stress split iteratively defines displacements $\boldsymbol{u}_h^{n,i}$ and pressures $p_h^{n,i}$ as approximations of the solution $\left(\boldsymbol{u}_h^n, p_h^n\right)$, where i denotes the iteration index. The fixed volumetric stress condition allows for eliminating the unknown volumetric deformation $\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n,i}$ in the flow equation. For instance, utilizing the effective stress (2.1.13) with the pore pressure being the fluid pressure, and the generalized Hooke's law (2.1.14), the condition translates to

$$K_{\mathrm{dr}}\boldsymbol{\nabla}\cdot\left(\boldsymbol{u}_h^{n,i}-\boldsymbol{u}_h^{n,i-1}\right)\overset{!}{=}\alpha\left(p_h^{n,i}-p_h^{n,i-1}\right). \tag{4.3.8}$$

We recall it is $K_{\mathrm{dr}}=\frac{2\mu}{d}+\lambda$. Consequently, the fixed volumetric stress condition translates to an $L^2(\Omega)$-type pressure stabilization in the fluid flow equations.

As observed in [111], the idea of fixing an artificial volumetric stress introduces a tuning parameter β_{FS}. The relation (4.3.8) is then replaced by

$$\alpha\boldsymbol{\nabla}\cdot\left(\boldsymbol{u}_h^{n,i}-\boldsymbol{u}_h^{n,i-1}\right)\overset{!}{=}\beta_{\mathrm{FS}}\left(p_h^{n,i}-p_h^{n,i-1}\right).$$

Each iteration of the fixed-stress split is then divided into a predictor and a corrector step. Assuming a homogeneous medium, the predictor step of iteration $i\in\mathbb{N}$ reads: *Given* $\left(\boldsymbol{u}_h^{n,i-1}, p_h^{n,i-1}\right)\in\boldsymbol{V}_h\times Q_h$, *find* $p_h^{n,i}\in Q_h$ *such that for all* $q_h\in Q_h$

$$\begin{aligned}\beta_{\mathrm{FS}}\left\langle p_h^{n,i}-p_h^{n,i-1},q_h\right\rangle+\frac{1}{M}\left\langle p_h^{n,i},q_h\right\rangle+\alpha\left\langle\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n,i-1},q_h\right\rangle\\+\Delta t\left\langle\frac{\boldsymbol{\kappa}}{\mu_{\mathrm{w}}}\left(\boldsymbol{\nabla}p_h^{n,i}-\rho_{\mathrm{w,ref}}\boldsymbol{g}\right),\boldsymbol{\nabla}q_h\right\rangle=\left\langle\theta^{n-1}+\Delta t\frac{h_{\mathrm{w}}(t_n)}{\rho_{\mathrm{w,ref}}},q_h\right\rangle.\end{aligned} \tag{4.3.9}$$

The corrector step then reads: *Given* $p_h^{n,i}\in Q_h$, *find* $\boldsymbol{u}_h^{n,i}\in\boldsymbol{V}_h$ *such that for all* $\boldsymbol{v}_h\in\boldsymbol{V}_h$

$$2\mu\left\langle\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right),\boldsymbol{\varepsilon}(\boldsymbol{v}_h)\right\rangle+\lambda\left\langle\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n,i},\boldsymbol{\nabla}\cdot\boldsymbol{v}_h\right\rangle-\alpha\left\langle p_h^{n,i},\boldsymbol{\nabla}\cdot\boldsymbol{v}_h\right\rangle=\langle\rho\boldsymbol{g},\boldsymbol{v}_h\rangle. \tag{4.3.10}$$

As, e.g., observed in Paper C, the performance of the fixed-stress split may significantly depend on the specific choice of the tuning parameter β_{FS}. Thus, in particular for large scale applications, a good choice of the stabilization is crucial for good performance of the splitting scheme. Different choices for β_{FS} resulted from optimization attempts, e.g., $\beta_{\mathrm{FS}} = \frac{\alpha^2}{2K_{\mathrm{dr}}}$ [30, 111], $\beta_{\mathrm{FS}} = \frac{\alpha^2}{2\lambda}$ [19, 83], and a more involved choice also depending on fluid flow parameters [143]. Also the ad-hoc proposed choice of $\beta_{\mathrm{FS}} = \frac{\alpha^2}{K_{\mathrm{dr,d}}}$ with a dimension-based estimation of the 'exact local bulk modulus' [98] is noteworthy. However, none of the above choices for β_{FS} has proved to be generally optimal in practice in the sense of yielding the minimal amount of iterations. On the other hand, it also has to be stressed, that the specific value of the stabilization becomes less important when treating the splitting scheme as preconditioner to a monolithic Krylov subspace method (not considered here).

Algebraic interpretation and analysis. Using the canonical isomorphism between finite element functions and coefficient vectors, the fully-discretized two-field formulation of the linear Biot equations (4.1.1)–(4.1.2) translates to an algebraic saddle point problem of the form (4.3.1). Similarly, the fixed-stress split (4.3.9)–(4.3.10) translates to an algebraic splitting scheme (4.3.4)–(4.3.5). In particular, $\mathbf{u}$ and $\mathbf{p}$ are the coefficient vectors corresponding to $\boldsymbol{u}_h^{n,i}$ and $p_h^{n,i}$, respectively; $\mathbf{A}$ and $\mathbf{C}$ correspond to the bilinear forms of linear elasticity and single-phase flow, respectively; $\mathbf{B}$ is associated to the divergence operator weighted by the Biot coefficient α; and $\mathbf{L} = \beta_{\mathrm{FS}}\mathbf{M}$ with $\mathbf{M}$ being the pressure mass matrix. This enables Lemma 4.3.1 for optimizing β_{FS} and thereby the performance of the fixed-stress split. For simplicity, the material is assumed to be homogeneous and isotropic.

Corollary 4.3.3 (Optimized convergence of the fixed-stress split)**.** *Assume a homogeneous and isotropic material. Let $e_{p_h}^{(i)} := p_h^{n,i} - p_h$ denote the pressure error for the fixed-stress split. Then it holds*

$$\left\|e_{p_h}^{(i)}\right\|^2 \leq \left(\frac{\max\left\{\left|\beta_{\mathrm{FS}} - \frac{\alpha^2}{K_{\mathrm{dr}}^\star}\right|, \left|\beta_{\mathrm{FS}} - \alpha^2\beta_{\mathrm{is}}\right|\right\}}{\beta_{\mathrm{FS}} + \frac{1}{M} + \frac{\Delta t \kappa}{\mu_{\mathrm{w}} C_\Omega^2}} \right)^2 \left\|e_{p_h}^{(i-1)}\right\|^2 \tag{4.3.11}$$

where $K_{\mathrm{dr}}^\star := \alpha^2/\lambda_{\max}\left(\mathbf{M}^{-1}\mathbf{S}\right) \in [K_{\mathrm{dr}}, 2\mu + \lambda]$, $\beta_{\mathrm{is}} \geq 0$ is the inf-sup constant corresponding to the divergence operator $\mathbf{B}$, cf. Lemma 3.5.2, and $C_\Omega > 0$ is a Poincaré constant. Consequently, the optimized choice is given by $\beta_{\mathrm{FS}} = \frac{\alpha^2}{2}\left(\frac{1}{K_{\mathrm{dr}}^\star} + \beta_{\mathrm{is}}\right)$ resulting in the convergence result

$$\left\|e_{p_h}^{(i)}\right\|^2 \leq \left(\frac{\frac{\alpha^2}{K_{\mathrm{dr}}^\star} - \alpha^2\beta_{\mathrm{is}}}{\frac{\alpha^2}{K_{\mathrm{dr}}^\star} + \alpha^2\beta_{\mathrm{is}} + 2\left(\frac{1}{M} + \frac{\Delta t \kappa}{\mu_{\mathrm{w}} C_\Omega^2}\right)} \right)^2 \left\|e_{p_h}^{(i-1)}\right\|^2. \tag{4.3.12}$$

Proof. By Lemma 4.3.1, it holds

$$\left\|e_{p_h}^{(i)}\right\| \leq \frac{\max\left\{\left|\beta_{\mathrm{FS}} - \lambda_{\max}\left(\mathbf{M}^{-1/2}\mathbf{S}\mathbf{M}^{-1/2}\right)\right|, \left|\beta_{\mathrm{FS}} - \lambda_{\min}\left(\mathbf{M}^{-1/2}\mathbf{S}\mathbf{M}^{-1/2}\right)\right|\right\}}{\beta_{\mathrm{FS}} + \lambda_{\min}\left(\mathbf{M}^{-1/2}\mathbf{C}\mathbf{M}^{-1/2}\right)} \left\|e_{p_h}^{(i-1)}\right\|$$

Using some linear algebra, inf-sup stability, cf. Lemma 3.5.2, and the Poincaré inequality (introducing the Poincaré constant C_Ω), one can show that

$$\begin{aligned}\lambda_{\max}\left(\mathbf{M}^{-1/2}\mathbf{S}\mathbf{M}^{-1/2}\right) &= \frac{\alpha^2}{K_{\mathrm{dr}}^\star} \leq \frac{\alpha^2}{K_{\mathrm{dr}}},\\ \lambda_{\min}\left(\mathbf{M}^{-1/2}\mathbf{S}\mathbf{M}^{-1/2}\right) &= \alpha^2\beta_{\mathrm{is}},\\ \lambda_{\min}\left(\mathbf{M}^{-1/2}\mathbf{C}\mathbf{M}^{-1/2}\right) &= \frac{1}{M} + \frac{\Delta t\kappa}{\mu_{\mathrm{w}} C_\Omega^2},\end{aligned}$$

which proves (4.3.11). Optimization of the rate with respect to β_{FS} results directly in the remaining assertion. □

By Remark 4.3.2, if $\mathbf{C}$ is not a multiplicative of the pressure mass matrix, the final optimized choice is not over all optimal. On the other hand, if the medium is low-permeable and the fluid is slightly compressible, the resulting stabilization is expected to yield good results – in particular since $K_{\mathrm{dr}}^\star$ involves information on the mechanical boundary conditions and the effective dimension of the problem, which has been identified to have a significant impact on the optimal choice for β_{FS} [31, 143]. The exact values for $K_{\mathrm{dr}}^\star$ and β_{is} may be, e.g., approximated by utilizing the Power method. The additional offline cost should be feasible if relatively many coupled problems have to be solved, e.g., due to many time steps.

After all, the convergence result reveals potential shortcomings in previous optimization studies, also part of this thesis. This will be discussed in more detail in Section 5.

Finally, we note that by directly involving the origin of the problem we have ultimately been able to include relevant problem information for deriving a theoretical convergence rate of the fixed-stress split. On the other hand, a corresponding analysis for heterogeneous media can be more intuitively performed by a problem-close analysis compared to an algebraic approach, see in Paper A, Paper B and Paper G.

Remark 4.3.4 (The undrained split in relation to Lemma (4.3.1))**.** *We briefly comment, that in an analogous way the undrained split [97] may be interpreted as L-scheme linearization of the displacement reduced problem. However, the relevant Schur complement* $\mathbf{B}^\top\mathbf{C}^{-1}\mathbf{B}$ *is not invertible since* $\mathbf{B}^\top$ *does not have full rank. Hence, essentially it holds* $\beta_{\mathrm{is}} = 0$ *(with abuse of notation), fortifying the potential of the fixed-stress split compared to the undrained split.*

4.3.3 Comment on non-linear coupled problems

As for linear problems, block-partitioned solvers can be beneficial for non-linear coupled problems. Compared to the monolithic Newton method, only linear convergence can in general be expected – however, possibly in exchange for increased robustness. In addition, block-partitioned solvers allow for breaking up the problem in smaller subproblems, potentially minimizing the overall cost, comparable to the discussion of direct solvers vs. splitting schemes for linear coupled problems. Furthermore, since block-partitioned solvers simultaneously account for the linearization and decoupling, the performance of the overall solver is essentially governed by the dominating complexity of the problem, be it strong non-linearities or a strong coupling strength. Consequently, for strongly coupled, non-linear problems, block-partitioned solvers account for non-linearities almost for free in terms of number of iterations, compared to a linear analog.

This idea has been pursued in the literature, e.g., for non-linear poroelasticity [28, 29].

4.4 Block-partitioned solvers for minimization problems

Next to saddle point problems, another important class of block-structured problems is given by block-separable, constrained minimization problems:

$$\textit{Find } \mathbf{x}^\star \in \mathbf{X} \textit{ such that } \mathcal{E}\left(\mathbf{x}^\star\right) = \inf_{\mathbf{x}\in\mathbf{X}} \mathcal{E}\left(\mathbf{x}\right), \tag{4.4.1}$$

where $\mathcal{E} : \mathbb{R}^N \to \mathbb{R}$ denotes the objective function, and the feasible set $\mathbf{X}$ is a Cartesian product of non-empty, closed sets $\mathbf{X}_j \subset \mathbb{R}^{N_j}$, such that $\sum_j N_j = N$.

Such problems naturally arise, e.g., after discretization of generalized gradient flows by the Minimizing Movement Scheme [65], closely related to the Implicit Euler method and a classical technique in the analysis of gradient flows. Relevant to this thesis, as explored in Paper A, thermo-poro-visco-elasticity problems can be formulated as generalized gradient flows and thereby as block-structured minimization problem after suitable discretization.

A natural approach to exploit the block-structured character of (4.4.1) for the development of block-partitioned solvers is to apply a cyclic block-non-linear Gauss-Seidel method, which frequently is also called *block coordinate descent method* (BCD), cf. Algorithm 2. The method may be practically valuable, if the block-component-wise minimization is fairly cheap. In the special case of two blocks, i.e., $m = 2$, BCD reduces to the fundamental *alternating minimization*, which becomes relevant for two-way coupled problems.

Algorithm 2: Block Coordinate Descent (BCD) for general number of blocks m

```
Input: Initial guess $\mathbf{x}^{(0)} = \left(\mathbf{x}_1^{(0)}, \dots, \mathbf{x}_m^{(0)}\right) \in \mathbf{X}$
For all iteration indices $i \geq 1$
    For all components $j = 1, \dots, m$
        Determine $\mathbf{x}_j^{(i)} \in \arg\inf_{\mathbf{x}\in\mathbf{X}_j} \mathcal{E}\left(\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{j-1}^{(i)}, \mathbf{x}, \mathbf{x}_{j+1}^{(i-1)}, \dots, \mathbf{x}_m^{(i-1)}\right)$
```

A major strength of the BCD is guaranteed global convergence under very loose conditions. For continuously differentiable objective functions $\mathcal{E}$, it is sufficient that the block-partitioned feasible set $\mathbf{X}$ is convex, and the block-component-wise minimization of $\mathcal{E}$ is well-defined [21], i.e., '∈' can be replaced with '=' in Step 4 of Algorithm 2. Various advances have been made to in fact further relax the convergence result by replacing smoothness with convexity assumptions for $\mathcal{E}$, e.g., [85, 151]. For the special case of alternating minimization even neither smoothness nor convexity is required [85]. Guaranteed convergence rates can be established under stronger convexity assumptions as showed for alternating minimization [20], see also Paper A; similarly for the BCD viewed as an orthogonal, successive subspace correction method for solving non-linear elliptic PDEs [145].

It becomes evident that the BCD yields mathematically founded, reliable partitioned solvers for non-linear coupled PDEs. Thus, a formulation of the type (4.4.1) seems very beneficial for the development and analysis of numerical solvers. This is explored in view of thermo-poro-visco-elasticity in Paper A.

4.5 The numerics environment DUNE

For the numerical studies in this dissertation the numerics environment DUNE (Distributed and Unified Numerics Environment) was used [16, 17, 26]. DUNE is a modular C++ library for the solution of PDEs using mesh-based methods including the FEM, the FVM, and the MFEM, cf. Section 4.1. Various modern C++ techniques as template and generic programming are heavily used ensuring efficiency in scientific computations.

DUNE itself comes with six elementary core modules providing basic infrastructure, grid interfaces, geometry classes, linear algebra classes, and local finite element basis functions on reference elements. An additional discretization module is required for the actual simulation of PDEs, taking care of managing global discrete functions, local and global assembly of (physical) algebraic problems, enforcing essential boundary conditions, and finally solving the resulting algebraic problems using, e.g., linearization, etc. There exist highly-developed general discretization modules as PDELab [18] and DUNE-FEM [66], as well as problem-specific discretization modules, e.g., DuMux [78], a module targeting various porous media applications.

Recently, a DUNE module for defining and managing global discrete functions has been developed: `dune-functions` [69, 70]. Together with the core modules, it provides all essential tools for developing a light-weight discretization module.

In this sense, as part of this thesis, the new discretization module `dune-biot` has been developed, tailored for the simulation of coupled multi-physics problems. The main purpose of that module has been the verification of the theoretical results obtained in this work, cf. Section 5. For that reason, it provides tools for the finite element simulation of, e.g., linear poroelasticity, unsaturated poroelasticity, and linear poro-visco-elasticity, solved either monolithically or sequentially. Various conforming (mixed) finite element discretizations for both linear elasticity, single-phase flow, and Richards' equation for heterogeneous media are available, utilizing primal and mixed formulations. The solver technology makes use of methods and concepts presented in this chapter: direct solvers, Newton's method, the L-scheme, and fixed-stress-type block-partitioned solvers. In addition, relaxation of iterative solvers by problem-specific line search strategies and general Anderson acceleration is available.

Chapter 5

Summary and Outlook

This chapter provides a summary and discussion of the scientific results presented in form of eight scientific articles in Part II.

Section 5.1.1 contains six papers comprising the main results of this thesis. In Paper A, the gradient flow structures of thermo-poro-visco-elasticity are revealed, and implications are presented for the well-posedness analysis, and the natural development and analysis of block-partitioned solvers. In Paper B, guaranteed convergence of the fixed-stress split for Biot's consolidation model for fully heterogeneous media is established. Paper C addresses the issue of the practical optimality of the stabilization parameter within the fixed-stress split and can be read as motivation for Paper G. Paper D and Paper E constitute the contributions to the mathematical discussion of unsaturated poroelasticity, as modeled in Section 2.1.2. In particular, the existence of weak solutions is established in Paper D. Paper E addresses the robust linearization using block-partitioned iterative solvers. Finally in Paper F, the convergence of the L-scheme linearization for doubly degenerate parabolic problems is established.

Section 5.1.2 contains two papers on related work. In Paper G, the convergence analysis of the fixed-stress split from Paper B is improved under the assumption of an inf-sup stable discretization, and a simple sampling strategy is proposed for the numerical optimization of the performance of the solver. Paper H constitutes a proceeding work to Paper E on robust block-partitioned iterative solvers for unsaturated poroelasticity, containing a numerical study.

This chapter ends with concluding remarks and an outlook.

5.1 Summary of the papers

5.1.1 Main results

Paper A [35]

Title: *The gradient flow structures of thermo-poro-visco-elastic processes in porous media*

Authors: Both, J.W., Kumar, K., Nordbotten, J.M., Radu, F.A.

Journal: In review (2019).

Preprint: arXiv:1907.03134 [math.NA]

Coupled thermo-hydro-mechanical processes in porous media are commonly modeled as quasi-static and dissipative. Motivated by that, this paper studies such processes from a gradient flow perspective, which ultimately results in unified approaches for the modeling, analysis, and robust numerical solution of thermo-poro-visco-elasticity.

The main contribution of this work is fivefold. First, an abstract framework for the modeling of thermo-poro-visco-elasticity is provided, found on the notion of generalized gradient flows, cf. Section 2.2. By this, the gradient flow structures of thermo-poro-visco-elasticity are explicitly revealed for the first time in the literature, to our best knowledge. Specific models are obtained by choosing free energy and dissipation potentials. In particular, gradient flow formulations are derived for previously employed PDE-based models [57], as Biot's consolidation model, non-linear poroelasticity in the infinitesimal strain regime, linear poro-visco-elasticity, non-Newtonian Darcy and non-Darcy flows in poro-elastic media, and thermo-poroelasticity without thermal convection. These identifications are in itself interesting as they determine the driving forces of the evolution of those systems.

Second, an high-level abstract well-posedness result for models arising from the above modeling framework is established. It combines tools from the theory of doubly non-linear evolution equations as well as convex analysis, cf. Sections 3.3–3.4. The application of the final well-posedness result boils down to checking simple convexity and continuity properties of the energy and dissipation potentials. Along these lines, a new concise proof for the well-posedness of the linear Biot equations is provided. With same complexity, well-posedness is also obtained for models not yet studied in the literature, as general poro-visco-elasticity.

Third, a unified methodology is established for the natural development of robust block-partitioned iterative solvers for thermo-poro-visco-elasticity. Due to the coupled character and gradient flow structure, discretized thermo-poro-visco-elasticity models inherit a convex, block-separable minimization structure. Motivated by that, block-coordinate descent methods are applied for the robust numerical solution, cf. Section 4.4; alternatively, a dual problem is derived first. Optionally, founded on the minimization structure, (inexact) line search is applied for relaxation of the inexact minimization. In view of two-block-structured problems, abstract convergence theory for alternating minimization is established, providing simple tools for deriving guaranteed convergence rates.

Fourth, the capabilities of the methodology are demonstrated by application to specific models. For instance, as proof of concept, various well-established block-partitioned iterative solvers used in the literature are derived including the fixed-stress and undrained splits for Biot's consolidation model [97, 98], and the undrained-adiabatic and extended fixed-stress splits recently introduced for thermo-poroelasticity [96]. Moreover, novel extensions are provided for linear poro-visco-elasticity and non-linear poroelasticity under infinitesimal strains. Tensorial and heterogeneous stabilizations are automatically deduced whenever suitable. In addition, the abstract convergence theory for alternating minimization yields theoretical convergence rates that are consistent with previous problem-dependent analyses (if existing), e.g., for the fixed-stress and undrained splits [111].

Finally, a numerical study is provided considering a three-dimensional footing problem governed by differing linear and non-linear physics. Two observations are made: (i) the robustness of block-partitioned solvers derived by the above methodology is confirmed; and (ii) a potentially significant impact on the acceleration of block-partitioned iterative solvers by line search strategies is demonstrated. In particular, applying line search has been identified as valid alternative to the tuning of solvers, cf. Paper B, Paper C, and Paper G – especially, since no *a priori* knowledge or user-interaction is required.

The approaches in this paper have some limitations. Most prominently, physical models describing complex coupled thermo-poro-visco-elastic processes in porous media do not necessarily exhibit the generalized gradient flow structure considered in this work, e.g., general convective-dominated processes, or materials with limit behavior as incompressible fluids or solids. However, in certain cases, non-monotone perturbations of gradient flows, as convection, may be neglected in the construction of block-partitioned solvers. By utilizing operator splitting techniques as, e.g., a Strang splitting, those perturbations may effectively be accounted for. Furthermore, for materials with limit behavior, utilizing dual problems is often more promising.

In addition, the methodology for the development of block-partitioned solvers, does not consider non-physical predictor-corrector methods, as the optimized fixed-stress. Those are treated separately in Paper B, Paper C, and Paper G.

Paper B [30]

Title:	*Robust fixed stress splitting for Biot's equations in heterogeneous media*
Authors:	Both, J.W., Borregales, M., Nordbotten, J.M., Kumar, K., Radu, F.A.
Journal:	Applied Mathematics Letters 68, 101–108 (2017).
DOI:	10.1016/j.aml.2016.12.019

The fixed-stress split constitutes one of the most used block-partitioned iterative solvers for Biot's consolidation model, cf. Section 4.3.2. One of the reasons for its popularity is its unconditional stability [98, 111], which algorithmically is the consequence of a well-chosen stabilization of the flow problem, explicitly approximating the Schur complement of the linear elasticity equations. To our best knowledge, previous convergence analyses, justifying appropriate stabilizations, solely considered homogeneous media [111] (apart from a heterogeneous permeability). A robust choice in the case of locally varying mechanical and coupling parameters, had not been discussed.

In this paper, previous convergence results for the fixed-stress split are extended to general fully heterogeneous media, i.e., media with locally varying mechanical and fluid material parameters. For this, a problem-specific convergence analysis is performed.

Two conclusions are drawn. First, unconditional stability holds for the classical fixed-stress split, i.e., iterating back and forth, solving the fluid flow problem under – we emphasize – local fixed volumetric stress, and the mechanics problem with updated fluid fields. Thus, algorithmically, this leads to locally varying stabilization, which solely depends on local mechanics parameters, despite the elliptic character of the linear elasticity equation. Numerical examples confirm the robustness of the fixed-stress split with respect to variations in the material parameters of different orders of magnitude (see also the supplementary material of this paper).

Second, the theoretical convergence result naturally extends the homogeneous case. For this, the use of weighted norms has been crucial, not requiring any worst-case-type bounds before the end of the proof. In addition, the final result is stated in energy norms for the fluid pressure, opposing to [111], in which the authors utilize a problem-dependent metric.

The problem-specific analysis in this paper allows for fixing an artificial volumetric stress in the first half-step of the algorithm. This approach has been inspired by [111], see also Section 4.3.2. By this, the introduced stabilization acts as tuning parameter (or more precisely tuning vector). The above theoretical convergence rate becomes dependent on that tuning vector, giving rise to minimization with respect to the tuning vector. The resulting, optimized stabilization corresponds to half the stabilization dictated by the classical fixed-stress ansatz. This is consistent with the optimized stabilization determined in [110, 111], despite differing approaches. Numerical examples demonstrate superior performance for the optimized stabilization compared to the classical choice. However, as it turns out, the theoretically optimized stabilization is in general not practically optimal, i.e., not leading to the minimal amount of iterations, cf. Paper C, which ultimately has given rise to the follow-up Paper G.

To our best knowledge, this work has been the first to theoretically investigate the fixed-stress split for fully heterogeneous media. Extending the approach in [111], similar results to ours have been derived in the later preprint [6], but the proposed stabilization deteriorates for vanishing Poisson's ratio. In that view, the result is not parameter-robust. Nevertheless, the overall drawn conclusions are consistent with our work. In hindsight, the later developed framework provided in Paper A allows for a slightly simpler construction and improved analysis of the fixed-stress split for heterogeneous media compared to this work (but without optimization).

Paper C [31]

Title: *Numerical Investigation on the Fixed-Stress Splitting Scheme for Biot's Equations: Optimality of the Tuning Parameter*

Authors: Both, J.W., Köcher, U.

Book: Numerical Mathematics and Advanced Applications ENUMATH 2017, Lecture Notes in Computational Science and Engineering 126, pg. 789–797 (2019).

DOI: 10.1007/978-3-319-96415-7_74

The performance of the fixed-stress split (for Biot's consolidation model) is known to may significantly depend on the particular choice of the stabilization parameter [98]. Theoretical investigations, see e.g., Paper B, commonly suggest that the practically optimal choice, resulting in a minimal amount of iterations, only depends on the Lamé parameters and the Biot coefficient – parameters associated to the Schur complement of the linear elasticity equations. However, additional properties may have an impact as, e.g. follows from the theoretical discussion in Section 4.3.2 and [98].

In this conference proceeding, the practical optimality of the stabilization parameter is numerically assessed, and its sensitivity with respect to further problem properties is investigated. A particular focus is on the boundary conditions for the mechanics problem and flow material parameters. The numerical study consists of several test cases, all based on the same geometry: a two-dimensional L-shaped domain under oscillating compression. We consider various material parameters and two sets of boundary conditions, generating either (A) an almost uniaxial compression, and (B) a true two dimensional deformation. The practically optimal tuning parameter is simply determined by sampling.

Three major conclusions are drawn. First, the boundary conditions associated to the mechanics equation primarily govern a real-valued effective dimension $d^\star \in [1, d]$ of the problem. For instance, for boundary conditions (A) and (B), the effective dimension is approximately $d^\star = 1$ and $d^\star = 2$, respectively, corresponding to the overall physical behavior. Ultimately, using the notation from Section 4.3.2, the practically optimal stabilization parameter is close to $\beta_\mathrm{FS} = \frac{\alpha^2}{K_\mathrm{dr}^\star}$, where $K_\mathrm{dr}^\star = \frac{2\mu}{d^\star} + \lambda$ denotes an effective bulk modulus of the matrix. Thereby, a similar conclusion is made as in [98], in which the authors introduce the 'exact local bulk modulus', mainly governed by boundary conditions. In hindsight, the effective bulk modulus $K_\mathrm{dr}^\star$ is expected to depend on the largest generalized eigenvalue of the exact Schur complement with respect to the pressure matrix, consistent with Corollary 4.3.3 and Paper G.

Second, fluid flow parameters affect the optimality of the tuning parameter. Considering a soft, strongly compressible bulk, just the permeability values are varied. A significant sensitivity of the optimal tuning parameter can be observed despite fixed mechanical and coupling parameters, and boundary conditions. In this part of the study, a deterioration of the performance of the fixed-stress split is observed for low permeability values. This is consistent with the known theoretical convergence rates. However, we note that finite elements have been used which are not inf-sup stable across the physical subproblems. This issue will be again addressed in Paper G.

Third, the optimized tuning parameters based on *a priori* analyses, as presented in Paper B or [19, 111], are generally not practically optimal. This has been recognized by considering an almost incompressible matrix. The corresponding practically optimal tuning parameter is $\beta_{\text{FS}} \approx \alpha^2/K_{\text{dr}} \approx \alpha^2/\lambda$. A factor of $\frac{1}{2}$, as suggested by the analyses, is lacking. This is inconsistent with the interpretation that the lower the stabilization (as long as robust) the faster the convergence, which is suggested by the theoretical investigations.

Altogether, there is clear evidence that the *a priori* analyses published so far are missing the dependence of many significant factors and thereby do not provide sharp theoretical convergence rates. Consequently, those have to be revisited, provided the goal is to optimally tune the stabilization. Such an attempt is made in Paper G.

Paper D [36]

Title: *Global existence of a weak solution to unsaturated poroelasticity*

Authors: Both, J.W., Pop, I.S., Yotov, I.

Preprint: arXiv:1909.06679 [math.NA]

This paper focuses on the existence analysis of weak solutions to unsaturated poroelasticity as modeled by the theory of porous media, cf. Section 2.1.2. It thereby constitutes a step towards the analysis of the general model considering multi-phase flow, cf. Section 2.1.1.

The mathematical model for unsaturated poroelasticity is a non-linear extension of the linear Biot equations. Non-linearities arise in various places: the coupling terms; the fluid compressibility; and the mobility. Those make an analysis particularly difficult. Therefore, a transformed problem is considered. By introducing a new pressure-like variable, defined by the Kirchhoff transformation, the diffusion term is linearized. This trick is often applied in the analysis of non-linear elliptic-parabolic PDEs, cf., e.g., [159]. Ultimately, the existence of weak solutions for the transformed problem is established under continuity assumptions on the non-linearities, and a non-degeneracy condition, which essentially requires a positive minimal residual saturation. The assumptions are demonstrated to typically hold true for geotechnical applications.

The proof utilizes the combination of regularization techniques, the Galerkin method, and compactness arguments, cf. Section 3.2. It can be summarized in three steps. First, the transformed problem is doubly regularized, enhancing the problem with a uniform parabolic character. The regularization can also be physically interpreted. A viscoelastic effect is included in the mechanics equations, and solid grains are assumed to be compressible. The latter is only required for the case of both incompressible solid grains and fluids.

Second, in the sense of the Galerkin method, the regularized problem is discretized in time and space. A combination of a conforming finite element method and a cell-centered finite volume method is employed for approximating the mechanics and flow problems, respectively. The latter is crucial for handling the non-linear coupling terms. Based on uniform stability of the discretized solution, classical compactness arguments lead to the existence of a weak solution to the doubly regularized, continuous problem.

Third, the limit case of vanishing regularization is discussed under the non-degeneracy assumption. Finally, by showing uniform stability and employing compactness arguments, the existence of a weak solution to the original, continuous problem is established.

This paper constitutes the first advance in the literature to establish well-posedness of the model of unsaturated poroelasticity, as modeled by the theory of porous media, cf. Section 2.1.2. However, we stress that a decidedly simplified model has been analyzed before [137]. That model involves non-physical linearizations of coupling terms and strict correlations between the non-linearities, which finally allow for an elegant mathematical proof based on the theory of maximal monotone operators.

Paper E [34]

Title: *Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media*

Authors: Both, J.W., Kumar, K, Nordbotten, J.M., Radu, F.A.

Journal: Computers & Mathematics with Applications 77(6), 1479–1502 (2019).

DOI: 10.1016/j.camwa.2018.07.033

In this paper, we discuss the robust linearization of the model for unsaturated poroelasticity, cf. Section 2.1.2. Similar to the linear Biot equations, the inherent block structure of the model motivates the development of block-partitioned solvers – here acting as simultaneous linearization and decoupling, cf. Section 4.3.3. In this sense, based on the similarities to the linear Biot equations, the extension of the fixed-stress split to unsaturated poroelasticity is investigated.

The Richards' equation constitutes a part of the model for unsaturated poroelasticity. Frequently employed constitutive relations for the hydraulic properties are vanishing for low saturations and may be not Lipschitz continuous, which may result in ill-conditioned Jacobians arising from Newton's method. Therefore, simple (more) robust alternatives to Newton's method have been developed in the literature as the modified Picard method [53] and the L-scheme [105], cf. Section 4.2.

In this work, three different block-partitioned solvers are proposed, combining the fixed-stress split and one-time applications of the three linearization schemes for Richards' equation: Newton's method, the modified Picard method, and the L-scheme. One iteration of each resulting block-partitioned linearization scheme is designed as follows. First, stabilized fluid flow equations are solved utilizing a single linearization step; afterwards, the mechanics equations are solved with updated fluid fields. The three resulting schemes are referred to as Fixed-Stress-Newton, Fixed-Stress-Modified-Picard, and Fixed-Stress-L-scheme, in the order of increasing expected robustness.

Theoretical convergence of the Fixed-Stress-L-scheme is established under mild physical assumptions and the use of an inf-sup stable discretization with respect to the pressure-displacement coupling. The first is required to fully disregard the degenerate case of a completely dried material, similar to Paper D, whereas the latter is required for parameter-robust convergence in the fully saturated regime, similar to Paper G. The convergence proof is founded on the close connection between the Fixed-Stress-L-scheme and the L-scheme applied to an equivalent pressure formulation, which is obtained by exactly inverting the mechanics problem similarly as in Section 4.3.1. The proof in particular suggests a choice for the stabilization parameters associated with the fixed-stress split and the L-scheme.

Despite theoretical robustness, the performance of the Fixed-Stress-L-scheme may deteriorate in unfavorable situations. For instance, if extensive stabilization is required by the theory, this might result in stagnation. The less robust quasi-Newton variants may even diverge. Consequently, we are concerned with two issues regarding the above block-partitioned linearization schemes: slow convergence and robustness. To remedy this situation, we propose to post-process each iteration by applying Anderson acceleration, cf. Section 4.2.4. Compared to common line search techniques, Anderson acceleration does not require any information on the coupled problem, and thus is conforming with the decoupling character of the linearization schemes. Furthermore, it has been previously observed to accelerate contractive fixed point iterations.

In the course of this work, novel theoretical insights are gained for Anderson acceleration. For a simple linear problem solved by the Richardson iteration, the error propagation of a restarted Anderson acceleration with depth 1 is quantified. By this, the effective acceleration and the potential recovery of convergence can be concluded for respectively contractive and non-contractive fixed-point iterations.

Finally, numerical investigations demonstrate a potentially significant improvement of the performance of all proposed block-partitioned linearizations under the use of Anderson acceleration, confirming the theoretical considerations. Three particular observations are made: (i) robustness is observed, also in challenging situations, in which each non-accelerated method including the monolithic Newton method fails to converge; (ii) Anderson acceleration with a low depth shows potential to increase the robustness of the monolithic Newton method; (iii) Anderson acceleration relaxes the need for the tuning of the stability parameter of block-partitioned solvers. After all, in practice, the Fixed-Stress-Newton method coupled with Anderson acceleration has showed the best performance among the splitting schemes.

Paper F [32]

Title:	*Iterative Linearisation Schemes for Doubly Degenerate Parabolic Equations*
Authors:	Both, J.W., Kumar, K., Nordbotten, J.M., Pop, I.S., and Radu, F.A.
Book:	Numerical Mathematics and Advanced Applications ENUMATH 2017, Lecture Notes in Computational Science and Engineering 126, pg. 49–63 (2019).
DOI:	10.1007/978-3-319-96415-7_3

This conference proceeding deals with the robust linearization of non-linear doubly degenerate parabolic problems with linear diffusion. Such a problem arises, e.g., when applying the Kirchhoff transformation to Richards' equation in the absence of gravity. We consider the general case in which the non-linear parabolic term is only non-decreasing and potentially Hölder continuous. In other words, we allow for the parabolic character to locally turn into a hyperbolic or an elliptic one. This creates the main difficulty in the development of robust iterative solvers.

One possible fix is to solve a regularized problem. However, if accurate solutions are required, the regularization has to be chosen sufficiently small, which may worsens the conditioning of the problem. Consequently, linearization schemes as Newton's and Picard's methods may exhibit convergence problems.

In this work, we consider the original problem linearized by the L-scheme adapted to Hölder continuous non-linearities. For the choice of the L-scheme stabilization, the central idea is to include not only continuity properties of the non-linearity but also the desired error tolerance. Convergence is theoretically established for the proposed linearization. Thus, a robust numerical solution is viable. On the other hand, the theoretical convergence rate also predicts the following: the finer the error tolerance is chosen, the larger the stabilization has to be chosen, and thereby the slower the convergence.

A numerical investigation confirms the theoretical robustness of the proposed L-scheme. In comparison, a regularized problem is considered, solved by Newton's method and the (standard) L-scheme. For both (and especially Newton's method) robustness issues are observed. Both require some fine-tuning of discretization parameters in order to successfully converge. Opposing to that, the proposed L-scheme (without regularization) does not require any fine-tuning. In addition, reassembling Jacobians is not required. Yet, the number of iterations required for convergence is much higher compared to the case when Newton's method converges.

5.1.2 Supplementary results

Paper G [143]

Title: *On the optimization of the fixed-stress splitting for Biot's equations*

Authors: Storvik, E., Both, J.W., Kumar, K., Nordbotten, J.M., and Radu, F.A.

Journal: International Journal for Numerical Methods in Engineering 120, 179–194 (2019).

DOI: 10.1002/nme.6130

This paper constitutes a follow-up paper of Paper B and Paper C, aiming at the the optimization of the performance of the fixed-stress split for Biot's consolidation model. Despite several optimization attempts in the literature [19, 98, 110, 111], an accurate, cheaply applicable estimate of the practically optimal tuning parameter, resulting in a minimal amount of iterations, has been lacking. So far, theoretically justified estimates depend only on mechanical and coupling material parameters, which is known to be insufficient. This paper provides novel theoretical and numerical results on the optimization of the tuning parameter.

The main contribution of this work is twofold. First, we revisit the problem-specific convergence analysis of the fixed-stress split in Paper B under the additional assumption of an inf-sup stable discretization with respect to the displacement-pressure coupling. The previous result is improved in two aspects: (i) the fixed-stress split exhibits parameter-robust convergence, even for incompressible materials and impermeable media; (ii) the convergence rate again gives rise for the optimization with respect to the tuning parameter, resulting in a parameter depending on both mechanical, coupling, fluid flow, and discretization parameters. Both (i) and (ii) are theoretically concluded for the first time. Regarding (ii), in comparison to Corollary 4.3.3, similar ingredients are utilized, e.g., $K_{\mathrm{dr}}^{\star}$ and β_{is}, including implicit information on the boundary conditions. However, the two predicted convergence rates and theoretically optimized stabilization parameters differ – also for the limit case of incompressible materials and impermeable media. Thus, based on Remark 4.3.2, the theoretically optimized parameter is in general not practically optimal. Same is concluded by numerical studies (see in particular in the associated preprint [142]). A compelling reason for the discrepancy is the use of too coarse estimates in the theoretical convergence analysis, similarly as in Paper B. This has resulted in a non-sharp theoretical convergence rate.

Second, both the convergence theory and the numerical observations demonstrate essentially mesh-independent convergence rates for the fixed-stress split. Motivated by that, a simple strategy is proposed: the practically optimal stabilization parameter is numerically estimated by cheap brute force optimization on a coarse mesh. Numerical examples demonstrate that the resulting parameter is in good agreement with the practically optimal convergence also on finer meshes. Clearly, the approach requires the access to a coarse mesh, which might not be available as, e.g., in industrial applications.

Paper H [33]

Title: *Iterative Methods for Coupled Flow and Geomechanics in Unsaturated Porous Media*

Authors: J.W. Both, K. Kumar, J.M. Nordbotten, F.A. Radu

Book: Poromechanics VI: Proceedings of the Sixth Biot Conference on Poromechanics, pg. 411–418, ASCE (2017).

DOI: 10.1061/9780784480779.050

This conference proceeding is a predecessor of Paper E and thereby also focuses on the robust linearization of unsaturated poroelasticity, as modeled in Section 2.1.2. Complementing Paper E, a combination of the two block-partitioned linearizations, the Fixed-Stress-L-scheme and the Fixed-Stress-Newton method, is proposed. Inspired by [105, 123], the first method is used as a warm-up for the latter for a given number of iterations. The main idea behind this is to exploit the robustness of the L-scheme-based linearization, and to obtain a sufficiently accurate approximation in order to trigger the faster (but local) convergence of the Newton-type linearization.

Employing a simple numerical study, different monolithic and block-partitioned solvers are compared. The test case involves discontinuous initial data, resulting in failing convergence of the plain monolithic Newton method. The slightly more robust Fixed-Stress-Newton method converges only for certain mesh sizes. On the other hand, the Fixed-Stress-L-scheme converges robustly. This is in agreement with the theoretical convergence analysis from Paper E. Yet, as expected, the Fixed-Stress-Newton method is significantly faster if it converges successfully.

In the same numerical study, the combined Fixed-Stress-L-scheme/Fixed-Stress-Newton method is considered. It shows superior performance over the remaining methods. A small user-defined number of iterations of the slower L-scheme-based linearization suffices for making the combined variant converging for all considered scenarios. In addition, the total number of iterations is basically the same as for the plain Fixed-Stress-Newton method, if the latter converges.

All in all, the combined variant joins the advantages of the single Fixed-Stress-L-scheme and the Fixed-Stress-Newton method as predicted. It thereby provides an alternative to the Fixed-Stress-Newton method accelerated by Anderson acceleration as suggested by Paper E. However, one drawback is the need of deciding when to switch in between the two methods. It can be resolved, e.g., by choosing a fixed number of iterations or a coarser stopping criterion. After all, a tuning parameter is introduced.

5.2 Outlook

This thesis concerns the mathematical analysis and numerical solution of coupled deformation- and flow-related processes in porous media. Those involve linearized single-phase flow, unsaturated flow, or non-Darcy flow within a linearly elastic or visco-elastic, solid matrix. Contributions are made in both the well-posedness analysis as well as the development and numerical analysis of block-partitioned iterative solvers for coupled models within the overarching topic of poroelasticity. In the following, we comment on possible future efforts arising from the results of this thesis.

As the presented unified gradient flow framework for the modeling, analysis, and development of block-partitioned solvers has led to promising results, continued investigation on the application to more involved systems deserves future attention. So far, fairly simple models have been considered, constituting a first step. However, it would be interesting to investigate the applicability of the framework to, e.g., multi-phase flow systems or hyper-elasto-plastic solids. Different results in this view have recently been published for just single components [47, 109] – not in the context of coupled poroelasticity.

Moreover, it would be interesting to investigate to what extent the gradient flow structures can be exploited in the development of robust or structure-preserving numerical methods. Possible directions could be *a posteriori*-type methods, time and space discretization, or improved numerical solvers, as partially already investigated for particular problems [48, 49, 93, 116].

Aiming at developing robust block-partitioned solvers for linear and non-linear coupled problems in view of this thesis, Anderson acceleration has been considered as post-processing acceleration technique. It has practically showed to be a very promising tool for the acceleration of especially non-linear block-partitioned (potentially non-contractive) solvers, not requiring any global Jacobian. In the course of that, first theoretical evidence has been provided confirming previous observations in the literature: Anderson acceleration may effectively accelerate the convergence of fixed point iterations, and it may increase the robustness of those. On the other hand, the theoretical result has been given for a fairly simple example. A more general investigation, complementing recent results on contractive, smooth fixed point iterations [72], should be made to better understand the potential of Anderson acceleration.

As the modeling assumptions for reducing the model for multi-phase flow in deformable porous media to unsaturated poroelasticity are violated in various real-life applications, a deeper study of the more general model would be of great interest. In particular, despite additional technicalities, previous difficulties and shortcomings of our results on unsaturated poroelasticity in the zone of vanishing fluid saturation may possibly be canceled by taking into account all fluid phases – in particular those which are predominant. Thereby, a natural extension of our results on unsaturated poroelasticity may be possible.

Bibliography

[1] Adams, R. (1975). Sobolev Spaces. Pure and applied mathematics. Academic Press.

[2] Adler, J. H., Gaspar, F. J., Hu, X., Rodrigo, C., and Zikatanov, L. T. (2017). Robust block preconditioners for Biot's model. In *International Conference on Domain Decomposition Methods*, pages 3–16. Springer. doi: 10.1007/978-3-319-93873-8.

[3] Ahmed, E., Nordbotten, J. M., and Radu, F. A. (2019). Adaptive asynchronous time-stepping, stopping criteria, and a posteriori error estimates for fixed-stress iterative schemes for coupled poromechanics problems. *arXiv e-prints*. arXiv:1901.01206 [math.NA].

[4] Ahmed, E., Radu, F. A., and Nordbotten, J. M. (2019). Adaptive poromechanics computations based on a posteriori error estimates for fully mixed formulations of Biot's consolidation model. *Computer Methods in Applied Mechanics and Engineering 347*, 264–294. doi: 10.1016/j.cma.2018.12.016.

[5] Almani, T., Kumar, K., Dogru, A., Singh, G., and Wheeler, M. (2016). Convergence analysis of multirate fixed-stress split iterative schemes for coupling flow with geomechanics. *Computer Methods in Applied Mechanics and Engineering 311*, 180–207. doi: 10.1016/j.cma.2016.07.036.

[6] Almani, T., Kumar, K., and Wheeler, M. (2017). Convergence analysis of single rate and multirate fixed stress split iterative coupling schemes in heterogeneous poroelastic media. *ICES REPORT 17–23* .

[7] Ambartsumyan, I., Khattatov, E., Nordbotten, J. M., and Yotov, I. (2018). A multipoint stress mixed finite element method for elasticity I: Simplicial grids. *arXiv e-prints*. arXiv:1805.09920 [math.NA].

[8] Ambartsumyan, I., Khattatov, E., Nordbotten, J. M., and Yotov, I. (2018). A multipoint stress mixed finite element method for elasticity II: Quadrilateral grids. *arXiv e-prints*. arXiv:1811.01928 [math.NA].

[9] Andersen, O., Nilsen, H. M., and Raynaud, X. (2017). Virtual element method for geomechanical simulations of reservoir models. *Computational Geosciences 21*(5), 877–893. doi: 10.1007/s10596-017-9636-1.

[10] Anderson, D. G. (1965). Iterative procedures for nonlinear integral equations. *Journal of the ACM 12*(4), 547–560. doi: 10.1145/321296.321305.

[11] Arnold, D., Falk, R., and Winther, R. (2007). Mixed finite element methods for linear elasticity with weakly imposed symmetry. *Mathematics of Computation 76*(260), 1699–1723. doi: 10.1090/S0025-5718-07-01998-9.

[12] Arnold, D. N. and Winther, R. (2002). Mixed finite elements for elasticity. *Numerische Mathematik 92*(3), 401–419. doi: 10.1007/s002110100348.

[13] Aubin, J.-P. (1963). Un théoreme de compacité. *Comptes rendus de l'Académie des science 256*(24), 5042–5044.

[14] Auriault, J.-L. and Sanchez-Palencia, E. (1977). Etude de comportment macroscopique d'un milieu poreux sature deformable. *Journal de Mécanique 16*, 575–603.

[15] Baranger, Jacques, Maitre, Jean-François, and Oudin, Fabienne (1996). Connection between finite volume and mixed finite element methods. *ESAIM: Mathematical Modelling and Numerical Analysis 30*(4), 445–465. doi: 10.1051/m2an/1996300404451.

[16] Bastian, P., Blatt, M., Dedner, A., Engwer, C., Klöfkorn, R., Kornhuber, R., Ohlberger, M., and Sander, O. (2008). A generic grid interface for parallel and adaptive scientific computing. Part II: implementation and tests in DUNE. *Computing 82*(2-3), 121–138. doi: 10.1007/s00607-008-0004-9.

[17] Bastian, P., Blatt, M., Dedner, A., Engwer, C., Klöfkorn, R., Ohlberger, M., and Sander, O. (2008). A generic grid interface for parallel and adaptive scientific computing. Part I: abstract framework. *Computing 82*(2-3), 103–119. doi: 10.1007/s00607-008-0003-x.

[18] Bastian, P., Heimann, F., and Marnach, S. (2010). Generic implementation of finite element methods in the distributed and unified numerics environment (DUNE). *Kybernetika 46*(2), 294–315.

[19] Bause, M., Radu, F., and Köcher, U. (2017). Space–time finite element approximation of the Biot poroelasticity system with iterative coupling. *Computer Methods in Applied Mechanics and Engineering 320*, 745–768. doi: 10.1016/j.cma.2017.03.017.

[20] Beck, A. and Tetruashvili, L. (2013). On the Convergence of Block Coordinate Descent Type Methods. *SIAM Journal on Optimization 23*(4), 2037–2060. doi: 10.1137/120887679.

[21] Bertsekas, D. P. (1997). Nonlinear programming, volume 48. Taylor & Francis.

[22] Biot, M. (1941). General theory of three-dimensional consolidation. *Journal of Applied Physics 12*(2), 155–164. doi: 10.1063/1.1712886.

[23] Biot, M. and Willis, D. (1957). The elastic coefficients of the theory of consolidation. *Journal of Applied Mechanics 24*, 594–601.

[24] Bishop, A. W. (1959). The principle of effective stress. *Teknisk ukeblad 39*, 859–863.

[25] Bjørnarå, T. I., Nordbotten, J. M., and Park, J. (2016). Vertically integrated models for coupled two-phase flow and geomechanics in porous media. *Water Resources Research 52*(2), 1398–1417. doi: 10.1002/2015WR017290.

[26] Blatt, M., Burchardt, A., Dedner, A., Engwer, C., Fahlke, J., Flemisch, B., Gersbacher, C., Gräser, C., Gruber, F., Grüninger, C., et al. (2016). The distributed and unified numerics environment, version 2.4. *Archive of Numerical Software 4*(100), 13–29. doi: 10.11588/ans.2016.100.26526.

[27] Boffi, D., Brezzi, F., and Fortin, M. (2013). Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics. Springer.

[28] Borregales, M., Radu, F. A., Kumar, K., and Nordbotten, J. M. (2018). Robust iterative schemes for non-linear poromechanics. *Computational Geosciences 22*(4), 1021–1038. doi: 10.1007/s10596-018-9736-6.

[29] Borregales, M. A., Kumar, K., Nordbotten, J. M., and Radu, F. A. (2019). Iterative solvers for Biot model under small and large deformation. *arXiv e-prints*. arXiv:1905.12996 [math.NA].

[30] 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. doi: 10.1016/j.aml.2016.12.019.

[31] Both, J. W. and Köcher, U. (2019). Numerical investigation on the fixed-stress splitting scheme for Biot's equations: Optimality of the tuning parameter. In *Numerical Mathematics and Advanced Applications ENUMATH 2017, Lecture Notes in Computational Science and Engineering 126*, pages 789–797. Springer. doi: 10.1007/978-3-319-96415-7_74.

[32] Both, J. W., Kumar, K., Nordbotten, J. M., Pop, I. S., and Radu, F. A. (2019). Iterative Linearisation Schemes for Doubly Degenerate Parabolic Equations. In *Numerical Mathematics and Advanced Applications ENUMATH 2017, Lecture Notes in Computational Science and Engineering 126*, pages 49–63. Springer. doi: 10.1007/978-3-319-96415-7_3.

[33] Both, J. W., Kumar, K., Nordbotten, J. M., and Radu, F. A. (2017). Iterative Methods for Coupled Flow and Geomechanics in Unsaturated Porous Media. In *Poromechanics VI: Proceedings of the Sixth Biot Conference on Poromechanics*, pages 411–418. ASCE. doi: 10.1061/9780784480779.050.

[34] Both, J. W., Kumar, K., Nordbotten, J. M., and Radu, F. A. (2019). Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media. *Computers & Mathematics with Applications 77*(6), 1479–1502. doi: 10.1016/j.camwa.2018.07.033.

[35] Both, J. W., Kumar, K., Nordbotten, J. M., and Radu, F. A. (2019). The gradient flow structures of thermo-poro-visco-elastic processes in porous media. *arXiv e-prints*. arXiv:1907.03134 [math.NA].

[36] Both, J. W., Pop, I. S., and Yotov, I. (2019). Global existence of a weak solution to unsaturated poroelasticity. *arXiv e-prints*. arXiv:1909.06679 [math.NA].

[37] Brenner, K. and Cancès, C. (2017). Improving Newton's Method Performance by Parametrization: The Case of the Richards Equation. *SIAM Journal on Numerical Analysis 55*(4), 1760–1785. doi: 10.1137/16M1083414.

[38] Brezis, H. (1971). Monotonicity Methods in Hilbert Spaces and Some Applications to Nonlinear Partial Differential Equations. In Zarantonello, E. H. (editor), *Contributions to Nonlinear Functional Analysis*, pages 101–156. Academic Press. doi: 10.1016/B978-0-12-775850-3.50009-1.

[39] Brezis, H. (1973). Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert. North-Holland Mathematics Studies. Elsevier Science.

[40] Brezis, H. (2010). Functional analysis, Sobolev spaces and partial differential equations. Springer Science & Business Media.

[41] Brinkman, H. C. (1949). A calculation of the viscous force exerted by a flowing fluid on a dense swarm of particles. *Flow, Turbulence and Combustion 1*(1), 27. doi: 10.1007/BF02120313.

[42] Brooks, R. (1964). Hydraulic properties of porous media. *Hydrology Papers 3*.

[43] Brooks, R. H. and Corey, A. T. (1966). Properties of porous media affecting fluid flow. *Journal of the irrigation and drainage division 92*(2), 61–90.

[44] Brun, M. K., Ahmed, E., Berre, I., Nordbotten, J. M., and Radu, F. A. (2019). Monolithic and splitting based solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport. *arXiv e-prints*. arXiv:1902.05783 [math.NA].

[45] Burridge, R. and Keller, J. B. (1981). Poroelasticity equations derived from microstructure. *The Journal of the Acoustical Society of America 70*(4), 1140–1146. doi: 10.1121/1.386945.

[46] Cancès, C. (2009). Finite volume scheme for two-phase flows in heterogeneous porous media involving capillary pressure discontinuities. *ESAIM: Mathematical Modelling and Numerical Analysis 43*(5), 973–1001. doi: 10.1051/m2an/2009032.

[47] Cancès, C., Gallouët, T. O., and Monsaingeon, L. (2015). The gradient flow structure for incompressible immiscible two-phase flows in porous media. *Comptes Rendus Mathematique 353*(11), 985–989. doi: 10.1016/j.crma.2015.09.021.

[48] Cancès, C., Gallouët, T. O., and Todeschi, G. (2019). A variational finite volume scheme for Wasserstein gradient flows. *arXiv e-prints*. arXiv:1907.08305 [math.NA].

[49] Cancès, C. and Guichard, C. (2017). Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Foundations of Computational Mathematics 17*(6), 1525–1584. doi: 10.1007/s10208-016-9328-6.

[50] Castelletto, N., Hajibeygi, H., and Tchelepi, H. A. (2017). Multiscale finite-element method for linear elastic geomechanics. *Journal of Computational Physics 331*, 337–356. doi: 10.1016/j.jcp.2016.11.044.

[51] Castelletto, N., White, J. A., and Ferronato, M. (2016). Scalable algorithms for three-field mixed finite element coupled poromechanics. *Journal of Computational Physics 327*, 894–918. doi: 10.1016/j.jcp.2016.09.063.

[52] Castelletto, N., White, J. A., and Tchelepi, H. A. (2015). Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics. *International Journal for Numerical and Analytical Methods in Geomechanics 39*(14), 1593–1618. doi: 10.1002/nag.2400.

[53] Celia, M. A., Bouloutas, E. T., and Zarba, R. L. (1990). A general mass-conservative numerical solution for the unsaturated flow equation. *Water resources research 26*(7), 1483–1496. doi: 10.1029/WR026i007p01483.

[54] Ciarlet, P. G. (1988). Mathematical Elasticity: Volume I: three-dimensional elasticity. North-Holland.

[55] Ciarlet, P. G. (2013). Linear and Nonlinear Functional Analysis with Applications. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[56] Colli, P. (1992). On some doubly nonlinear evolution equations in Banach spaces. *Japan Journal of Industrial and Applied Mathematics 9*(2), 181. doi: 10.1007/BF03167565.

[57] Coussy, O. (2004). Poromechanics. Wiley.

[58] Coussy, O. (2007). Revisiting the constitutive equations of unsaturated porous solids using a Lagrangian saturation concept. *International Journal for Numerical and Analytical Methods in Geomechanics 31*(15), 1675–1694. doi: 10.1002/nag.613.

[59] Coussy, O., Dangla, P., Lassabatère, T., and Baroghel-Bouny, V. (2004). The equivalent pore pressure and the swelling and shrinkage of cement-based materials. *Materials and Structures 37*(1), 15–20. doi: 10.1007/BF02481623.

[60] Crandall, M. G. and Pazy, A. (1969). Semi-groups of nonlinear contractions and dissipative sets. *Journal of Functional Analysis 3*(3), 376–418. doi: 10.1016/0022-1236(69)90032-9.

[61] Dana, S. and Wheeler, M. F. (2018). Convergence analysis of two-grid fixed stress split iterative scheme for coupled flow and deformation in heterogeneous poroelastic media. *Computer Methods in Applied Mechanics and Engineering 341*, 788–806. doi: 10.1016/j.cma.2018.07.018.

[62] Darcy, H. P. G. (1856). Les Fontaines publiques de la ville de Dijon. Exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau, etc. V. Dalamont.

[63] De Boer, R. (2000). Theory of porous media: highlights in historical development and current state. Springer Science & Business Media.

[64] de Boer, R. and Ehlers, W. (1990). The development of the concept of effective stresses. *Acta Mechanica 83*(1), 77–92. doi: 10.1007/BF01174734.

[65] De Giorgi, E. (1993). New problems on minimizing movements. *Ennio de Giorgi: Selected Papers* pages 699–713.

[66] Dedner, A., Klöfkorn, R., Nolte, M., and Ohlberger, M. (2010). A generic interface for parallel and adaptive discretization schemes: abstraction principles and the DUNE-FEM module. *Computing 90*(3-4), 165–196. doi: 10.1007/s00607-010-0110-3.

[67] Deuflhard, P. (2011). Newton methods for nonlinear problems: affine invariance and adaptive algorithms, volume 35. Springer Science & Business Media.

[68] Ekeland, I. and Temam, R. (1999). Convex analysis and variational problems, volume 28. SIAM.

[69] Engwer, C., Gräser, C., Müthing, S., and Sander, O. (2015). The interface for functions in the dune-functions module. *arXiv e-prints.* arXiv:1512.06136 [math.NA].

[70] Engwer, C., Gräser, C., Müthing, S., and Sander, O. (2018). Function space bases in the dune-functions module. *arXiv e-prints.* arXiv:1806.09545 [math.NA].

[71] Ern, A. and Guermond, J.-L. (2013). Theory and practice of finite elements, volume 159. Springer Science & Business Media.

[72] Evans, C., Pollock, S., Rebholz, L. G., and Xiao, M. (2018). A proof that Anderson acceleration improves the convergence rate in linearly converging fixed point methods (but not in those converging quadratically). *arXiv e-prints.* arXiv:1810.08455 [math.NA].

[73] Evans, L. (1998). Partial Differential Equations. Graduate studies in mathematics. American Mathematical Society.

[74] Eymard, R., Gallouët, T., and Herbin, R. (2000). Finite volume methods *7*, 713–1018. doi: 10.1016/S1570-8659(00)07005-8.

[75] 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. doi: 10.1016/j.jcp.2010.03.018.

[76] Ferronato, M., Franceschini, A., Janna, C., Castelletto, N., and Tchelepi, H. A. (2019). A general preconditioning framework for coupled multiphysics problems with application to contact- and poro-mechanics. *Journal of Computational Physics 398*, 108887. doi: 10.1016/j.jcp.2019.108887.

[77] Fillunger, P. (1936). Erdbaumechanik? Selbstverl. d. Verf. Wien.

[78] Flemisch, B., Darcis, M., Erbertseder, K., Faigle, B., Lauser, A., Mosthaf, K., Müthing, S., Nuske, P., Tatomir, A., Wolff, M., et al. (2011). DuMux: DUNE for multi-{phase, component, scale, physics,...} flow and transport in porous media. *Advances in Water Resources 34*(9), 1102–1112. doi: 10.1016/j.advwatres.2011.03.007.

[79] Forchheimer, P. (1901). Wasserbewegung durch Boden. *Zeitschrift des Vereins deutscher Ingenieure 45*, 1782–1788.

[80] Fumagalli, A. and Keilegavlen, E. (2018). Dual Virtual Element Method for Discrete Fractures Networks. *SIAM Journal on Scientific Computing 40*(1), B228–B258. doi: 10.1137/16M1098231.

[81] Gaspar, F., Lisbona, F., and Vabishchevich, P. (2003). A finite difference analysis of Biot's consolidation model. *Applied numerical mathematics 44*(4), 487–506. doi: 10.1016/S0168-9274(02)00190-3.

[82] Gaspar, F. J. and Rodrigo, C. (2017). On the fixed-stress split scheme as smoother in multigrid methods for coupling flow and geomechanics. *Computer Methods in Applied Mechanics and Engineering 326*, 526–540. doi: 10.1016/j.cma.2017.08.025.

[83] Girault, V., Kumar, K., and Wheeler, M. F. (2016). Convergence of iterative coupling of geomechanics with flow in a fractured poroelastic medium. *Computational Geosciences 20*(5), 997–1011. doi: 10.1007/s10596-016-9573-4.

[84] Gray, W. G. and Hassanizadeh, S. M. (1991). Unsaturated Flow Theory Including Interfacial Phenomena. *Water Resources Research 27*(8), 1855–1863. doi: 10.1029/91WR01260.

[85] Grippo, L. and Sciandrone, M. (2000). On the Convergence of the Block Nonlinear Gauss-Seidel Method Under Convex Constraints. *Operations Research Letters 26*(3), 127–136. doi: 10.1016/S0167-6377(99)00074-7.

[86] Gurtin, M. E. (1982). An introduction to continuum mechanics, volume 158. Academic press.

[87] Hackbusch, W. (2013). Multi-grid methods and applications, volume 4. Springer Science & Business Media.

[88] Hawkes, C., McLellan, P., Zimmer, U., Bachu, S., et al. (2004). Geomechanical factors affecting geological storage of CO_2 in depleted oil and gas reservoirs. In *Canadian International Petroleum Conference*. Petroleum Society of Canada. doi: 10.2118/05-10-05.

[89] Hettema, M., Schutjens, P., Verboom, B., Gussinklo, H., et al. (2000). Production-induced compaction of a sandstone reservoir: the strong influence of stress path. *SPE Reservoir Evaluation & Engineering 3*(04), 342–347. doi: 10.2118/65410-PA.

[90] Hong, Q., Kraus, J., Lymbery, M., and Philo, F. (2019). Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models. *Numerical Linear Algebra with Applications 26*(4), e2242. doi: 10.1002/nla.2242.

[91] Huber, R. and Helmig, R. (2000). Node-centered finite volume discretizations for the numerical simulation of multiphase flow in heterogeneous porous media. *Computational Geosciences 4*(2), 141–164. doi: 10.1023/A:1011559916309.

[92] Jenny, P., Lee, S., and Tchelepi, H. A. (2003). Multi-scale finite-volume method for elliptic problems in subsurface flow simulation. *Journal of Computational Physics 187*(1), 47–67. doi: 10.1016/S0021-9991(03)00075-5.

[93] Jüngel, A., Stefanelli, U., and Trussardi, L. (2018). Two time discretizations for gradient flows exactly replicating energy dissipation. *arXiv e-prints.* arXiv:1811.06033 [math.NA].

[94] Kantorovich, L. V. (1948). Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk 3*(6), 89–185.

[95] Keilegavlen, E. and Nordbotten, J. M. (2017). Finite volume methods for elasticity with weak symmetry. *International Journal for Numerical Methods in Engineering 112*(8), 939–962. doi: 10.1002/nme.5538.

[96] Kim, J. (2018). Unconditionally stable sequential schemes for all-way coupled thermoporomechanics: Undrained-adiabatic and extended fixed-stress splits. *Computer Methods in Applied Mechanics and Engineering 341*, 93–112. doi: 10.1016/j.cma.2018.06.030.

[97] Kim, J., Tchelepi, H., and Juanes, R. (2011). Stability and convergence of sequential methods for coupled flow and geomechanics: Drained and undrained splits. *Computer Methods in Applied Mechanics and Engineering 200*(23), 2094–2116. doi: 10.1016/j.cma.2011.02.011.

[98] Kim, J., Tchelepi, H., and Juanes, R. (2011). Stability and convergence of sequential methods for coupled flow and geomechanics: Fixed-stress and fixed-strain splits. *Computer Methods in Applied Mechanics and Engineering 200*(13), 1591–1606. doi: 10.1016/j.cma.2010.12.022.

[99] Kim, J., Tchelepi, H. A., and Juanes, R. (2013). Rigorous Coupling of Geomechanics and Multiphase Flow with Strong Capillarity. *Society of Petroleum Engineers* doi: doi:10.2118/141268-PA.

[100] Komura, Y. (1967). Nonlinear semi-groups in Hilbert space. *Journal of the Mathematical Society of Japan 19*(4), 493–507. doi: 10.2969/jmsj/01940493.

[101] Kwok, F. and Tchelepi, H. (2007). Potential-based reduced Newton algorithm for nonlinear multiphase flow in porous media. *Journal of Computational Physics 227*(1), 706–727. doi: 10.1016/j.jcp.2007.08.012.

[102] Lee, J., Mardal, K., and Winther, R. (2017). Parameter-Robust Discretization and Preconditioning of Biot's Consolidation Model. *SIAM Journal on Scientific Computing 39*(1), A1–A24. doi: 10.1137/15M1029473.

[103] Lewis, R. and Schrefler, B. (1998). The finite element method in the static and dynamic deformation and consolidation of porous media. Numerical methods in engineering. John Wiley.

[104] Li, B. and Tchelepi, H. A. (2014). Unconditionally Convergent Nonlinear Solver for Multiphase Flow in Porous Media under Viscous Force, Buoyancy, and Capillarity. *Energy Procedia 59*, 404–411. doi: 10.1016/j.egypro.2014.10.395. European Geosciences Union General Assembly 2014, EGU Division Energy, Resources & the Environment (ERE).

[105] List, F. and Radu, F. A. (2016). A study on iterative methods for solving Richards' equation. *Computational Geosciences 20*(2), 341–353. doi: 10.1007/s10596-016-9566-3.

[106] Luckner, L., Van Genuchten, M. T., and Nielsen, D. (1989). A consistent set of parametric models for the two-phase flow of immiscible fluids in the subsurface. *Water Resources Research 25*(10), 2187–2193. doi: 10.1029/WR025i010p02187.

[107] Mardal, K.-A. and Winther, R. (2011). Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications 18*(1), 1–40. doi: 10.1002/nla.716.

[108] Mielke, A., Rossi, R., and Savaré, G. (2013). Nonsmooth analysis of doubly nonlinear evolution equations. *Calculus of Variations and Partial Differential Equations 46*(1), 253–310. doi: 10.1007/s00526-011-0482-z.

[109] Mielke, A., Rossi, R., and Savaré, G. (2018). Global existence results for viscoplasticity at finite strain. *Archive for Rational Mechanics and Analysis 227*(1), 423–475. doi: 10.1007/s00205-017-1164-6.

[110] Mikelić, A., Wang, B., and Wheeler, M. F. (2014). Numerical convergence study of iterative coupling for coupled flow and geomechanics. *Computational Geosciences 18*(3), 325–341. doi: 10.1007/s10596-013-9393-8.

[111] Mikelić, A. and Wheeler, M. F. (2013). Convergence of iterative coupling for coupled flow and geomechanics. *Computational Geosciences 17*(3), 455–461. doi: 10.1007/s10596-012-9318-y.

[112] Muskat, M. and Meres, M. W. (1936). The Flow of Heterogeneous Fluids Through Porous Media. *Physics 7*(9), 346–363. doi: 10.1063/1.1745403.

[113] Natvig, J. R. and Lie, K.-A. (2008). Fast computation of multiphase flow in porous media by implicit discontinuous Galerkin schemes with optimal ordering of elements. *Journal of Computational Physics 227*(24), 10108–10124. doi: 10.1016/j.jcp.2008.08.024.

[114] Nikooee, E., Habibagahi, G., Hassanizadeh, S. M., and Ghahramani, A. (2013). Effective Stress in Unsaturated Soils: A Thermodynamic Approach Based on the Interfacial Energy and Hydromechanical Coupling. *Transport in Porous Media 96*(2), 369–396. doi: 10.1007/s11242-012-0093-y.

[115] Nocedal, J. and Wright, S. (2006). Numerical optimization. Springer Science & Business Media.

[116] Nochetto, R. H., Savaré, G., and Verdi, C. (2000). A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations. *Communications on Pure and Applied Mathematics 53*(5), 525–589. doi: 10.1002/(SICI)1097-0312(200005)53:5<525::AID-CPA1>3.0.CO;2-M.

[117] Nordbotten, J. M. (2016). Stable cell-centered finite volume discretization for Biot equations. *SIAM Journal on Numerical Analysis 54*(2), 942–968. doi: 10.1137/15M1014280.

[118] Nordbotten, J. M. and Celia, M. A. (2011). Geological storage of CO2: modeling approaches for large-scale simulation. John Wiley & Sons.

[119] Nur, A. and Byerlee, J. D. (1971). An exact effective stress law for elastic deformation of rock with fluids. *Journal of Geophysical Research 76*(26), 6414–6419. doi: 10.1029/JB076i026p06414.

[120] Nuth, M. and Laloui, L. (2008). Effective stress concept in unsaturated soils: Clarification and validation of a unified framework. *International Journal for Numerical and Analytical Methods in Geomechanics 32*(7), 771–801. doi: 10.1002/nag.645.

[121] Otto, F. (2001). The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations 26*(1-2), 101–174. doi: 10.1081/PDE-100002243.

[122] Owens, R. G. and Phillips, T. N. (2002). Computational Rheology. Imperial College Press. doi: 10.1142/p160.

[123] Paniconi, C. and Putti, M. (1994). A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems. *Water Resources Research 30*(12), 3357–3374. doi: 10.1029/94WR02046.

[124] Peletier, M. A. (2014). Variational Modelling: Energies, gradient flows, and large deviations. *arXiv e-prints*. arXiv:1402.1990 [math.NA].

[125] 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. doi: 10.1007/s10596-008-9082-1.

[126] Pop, I. S., Radu, F. A., and Knabner, P. (2004). Mixed finite elements for the Richards' equation: linearization procedure. *Journal of Computational and Applied Mathematics 168*(1), 365–373. doi: 10.1016/j.cam.2003.04.008.

[127] Radu, F., Pop, I. S., and Knabner, P. (2004). Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation. *SIAM Journal on Numerical Analysis 42*(4), 1452–1478. doi: 10.1137/S0036142902405229.

[128] Radu, F. A., Kumar, K., Nordbotten, J. M., and Pop, I. S. (2017). A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities. *IMA Journal of Numerical Analysis 38*(2), 884–920. doi: 10.1093/imanum/drx032.

[129] Radu, F. A., Nordbotten, J. M., Pop, I. S., and Kumar, K. (2015). A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media. *Journal of Computational and Applied Mathematics 289*, 134–141. doi: http://dx.doi.org/10.1016/j.cam.2015.02.051.

[130] Rodrigo, C., Gaspar, F., Hu, X., and Zikatanov, L. (2016). Stability and monotonicity for some discretizations of the Biot's consolidation model. *Computer Methods in Applied Mechanics and Engineering 298*, 183–204. doi: 10.1016/j.cma.2015.09.019.

[131] Rossi, R. and Savaré, G. (2006). Gradient flows of non convex functionals in Hilbert spaces and applications. *ESAIM: Control, Optimisation and Calculus of Variations 12*(3), 564–614. doi: 10.1051/cocv2006013.

[132] Saad, Y. (2003). Iterative methods for sparse linear systems, volume 82. Siam.

[133] Sánchez-Palencia, E. (1980). Non-homogeneous media and vibration theory. *Lecture notes in physics 127*.

[134] Segall, P. and Fitzgerald, S. D. (1998). A note on induced stress changes in hydrocarbon and geothermal reservoirs. *Tectonophysics 289*(1), 117–128. doi: 10.1016/S0040-1951(97)00311-9.

[135] Settari, A., Mourits, F., et al. (1998). A coupled reservoir and geomechanical simulation system. *SPE Journal 3*(03), 219–226. doi: 10.2118/50939-PA.

[136] Showalter, R. (2000). Diffusion in Poro-Elastic Media. *Journal of Mathematical Analysis and Applications 251*(1), 310–340. doi: 10.1006/jmaa.2000.7048.

[137] Showalter, R. and Su, N. (2001). Partially saturated flow in a poroelastic medium. *Discrete and Continuous Dynamical Systems–Series B 1*(4), 403–420. doi: 10.3934/dcdsb.2001.1.403.

[138] Simon, J. (1986). Compact sets in the space $L^p(0,T;B)$. *Annali di Matematica Pura ed Applicata 146*(1), 65–96. doi: 10.1007/BF01762360.

[139] Slodicka, M. (2002). A Robust and Efficient Linearization Scheme for Doubly Nonlinear and Degenerate Parabolic Problems Arising in Flow in Porous Media. *SIAM Journal on Scientific Computing 23*(5), 1593–1614. doi: 10.1137/S1064827500381860.

[140] Smith, B., Bjorstad, P., and Gropp, W. (2004). Domain decomposition: parallel multilevel methods for elliptic partial differential equations. Cambridge university press.

[141] Steffen, G. S., Candelaria, S. M., Stapledon, D., Bell, G., and Foster, M. (2014). Geotechnical engineering of dams. CRC press.

[142] Storvik, E., Both, J. W., Kumar, K., Nordbotten, J. M., and Radu, F. A. (2018). On the optimization of the fixed-stress splitting for Biot's equations. *arXiv e-prints.* arXiv:1811.06242 [math.NA].

[143] 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. doi: 10.1002/nme.6130.

[144] Szymkiewicz, A. (2012). Modelling water flow in unsaturated porous media: accounting for nonlinear permeability and material heterogeneity. Springer Science & Business Media.

[145] Tai, X.-C. and Espedal, M. (1998). Rate Of Convergence Of Some Space Decomposition Methods For Linear And Nonlinear Problems. *SIAM Journal on Numerical Analysis 35*, 1558–1570. doi: 10.1137/S0036142996297461.

[146] Taylor, J. E. and Cahn, J. W. (1994). Linking anisotropic sharp and diffuse surface motion laws via gradient flows. *Journal of Statistical Physics 77*(1), 183–197. doi: 10.1007/BF02186838.

[147] Teatini, P., Ferronato, M., Gambolati, G., Bertoni, W., and Gonella, M. (2005). A century of land subsidence in Ravenna, Italy. *Environmental Geology 47*(6), 831–846. doi: 10.1007/s00254-004-1215-9.

[148] Terekhov, K. M. and Tchelepi, H. A. (2020). Cell-centered finite-volume method for elastic deformation of heterogeneous media with full-tensor properties. *Journal of Computational and Applied Mathematics 364*, 112331. doi: 10.1016/j.cam.2019.06.047.

[149] Terzaghi, K. v. (1936). The shearing resistance of saturated soils and the angle between the planes of shear. In *First international conference on soil Mechanics, 1936*, volume 1, pages 54–59.

[150] Thomas, J.-M. (1977). Sur l'analyse numérique des méthodes d'éléments finis hybrides et mixtes. Ph.D. thesis.

[151] Tseng, P. (2001). Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications 109*(3), 475–494. doi: 10.1023/A:1017501703105.

[152] van Genuchten (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal 44(5)*, 892–898. doi: 10.2136/sssaj1980.03615995004400050002x.

[153] Walker, H. F. and Ni, P. (2011). Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis 49*(4), 1715–1735. doi: 10.1137/10078356X.

[154] Wang, X. and Tchelepi, H. A. (2013). Trust-region based solver for nonlinear transport in heterogeneous porous media. *Journal of Computational Physics 253*, 114–137. doi: 10.1016/j.jcp.2013.06.041.

[155] Wheeler, M., Xue, G., and Yotov, I. (2014). Coupling multipoint flux mixed finite element methods with continuous Galerkin methods for poroelasticity. *Computational Geosciences 18*(1), 57–75. doi: 10.1007/s10596-013-9382-y.

[156] Wheeler, M. F. and Yotov, I. (2006). A multipoint flux mixed finite element method. *SIAM Journal on Numerical Analysis 44*(5), 2082–2106. doi: 10.1007/s00211-011-0427-7.

[157] White, J. A., Castelletto, N., Klevtsov, S., Bui, Q. M., Osei-Kuffuor, D., and Tchelepi, H. A. (2019). A two-stage preconditioner for multiphase poromechanics in reservoir simulation. *Computer Methods in Applied Mechanics and Engineering 357*, 112575. doi: 10.1016/j.cma.2019.112575.

[158] White, J. A., Castelletto, N., and Tchelepi, H. A. (2016). Block-partitioned solvers for coupled poromechanics: A unified framework. *Computer Methods in Applied Mechanics and Engineering 303*, 55–74. doi: 10.1016/j.cma.2016.01.008.

[159] Wilhelm Alt, H. and Luckhaus, S. (1983). Quasilinear elliptic-parabolic differential equations. *Mathematische Zeitschrift 183*(3), 311–341. doi: 10.1007/BF01176474.

[160] Yotov, I. P. (1996). Mixed finite element methods for flow in porous media. Ph.D. thesis, Rice University.

[161] Zenisek, A. (1984). The existence and uniquencess theorem in Biot's consolidation theory. *Aplikace matematiky 29*(3), 194–211.

Part II

Included papers

Paper A

The gradient flow structures of thermo-poro-visco-elastic processes in porous media

Both, J.W., Kumar, K., Nordbotten, J.M., and Radu, F.A.

In review.

arXiv:1907.03134 [math.NA]

The gradient flow structures of thermo-poro-visco-elastic processes in porous media

Jakub W. Both[*] Kundan Kumar[†] Jan M. Nordbotten[*] Florin A. Radu[*]

Abstract

In this paper, the inherent gradient flow structures of thermo-poro-visco-elastic processes in porous media are examined for the first time. In the first part, a modelling framework is introduced aiming for describing such processes as generalized gradient flows requiring choices of physical states, corresponding energies, dissipation potentials and external work rates. It is demonstrated that various existing models can be in fact written within this framework. Ultimately, the particular structure allows for a unified well-posedness analysis performed for different classes of linear and non-linear models. In the second part, the gradient flow structures are utilized for constructing efficient discrete approximation schemes for thermo-poro-visco-elasticity – in particular robust, physical splitting schemes. Applying alternating minimization to naturally arising minimization formulations of (semi-)discrete models is proposed. For such, the energy decrease per iteration is quantified by applying abstract convergence theory only utilizing convexity and Lipschitz continuity properties of the problem – a fairly simple but flexible machinery. By this approach, e.g., the widely used undrained and fixed-stress splits for the linear Biot equations are derived and analyzed. By application of the framework to more advanced models, novel splitting schemes with guaranteed theoretical convergence rates are naturally derived. Moreover, based on the minimization character of the (semi-)discrete equations, relaxation of splitting schemes by line search is proposed; numerical results show a potentially great impact on the acceleration of splitting schemes for both linear and nonlinear problems.

1 Introduction

Gradient flows describe the evolution of purely dissipative systems. Given an initial state x_0, a state x evolves along the negative gradient of an energy $\mathcal{E}$ under the influence of an external force f_{ext}, i.e.,

$$\dot{x} + \boldsymbol{\nabla}\mathcal{E}(x) = f_{\mathrm{ext}}, \quad \text{a.e. in } (0,T), \quad x(0) = x_0, \tag{1.1}$$

where $\dot{x}$ denotes the temporal derivative of x and $\boldsymbol{\nabla}\mathcal{E}$ denotes the Gâteux-derivative of $\mathcal{E}$ wrt. x.

The formal gradient flow structure (1.1) is ubiquitous in a broad set of applications and has been therefore of great research interest since the fundamental works by Komura [1], Crandall and Pazy [2] and Brezis [3, 4]. Meanwhile, gradient flows have been studied in Hilbert spaces and metric spaces [5]; in particular, since the seminal work by Otto [6], much attraction has been paid to gradient flows in probability spaces endowed with the Wasserstein metric. It is not our intention to review the vast literature on the topic; we mention a small fragment of the long list of applications with an inherent gradient structure: heat conduction, the Stefan

[*] Department of Mathematics, University of Bergen, Bergen, Norway; {erlend.storvik@uib.no, jakub.both@uib.no, jan.nordbotten@uib.no, florin.radu@uib.no}

[†] Department of Mathematics and Computer Science, Karlstad University, Karlstad, Sweden; {kundan.kumar@kau.se}

problem, Hele-Shaw cell, flow in porous media, parabolic variational inequalities, degenerate and quasi-linear parabolic PDEs, and transport.

Classical gradient flows are limited to dissipation mechanisms induced by a quadratic potential, which is quite restrictive for many practical situations. Far more systems can be modelled using the notion of *generalized gradient flows* as, e.g., described by Peletier [7]. Those allow in particular for non-quadratic dissipation potentials, including those which are vanishing, not finite, positively homogeneous of degree 1 or state-dependent. Additionally, generalized gradient flows allow for relating the tangent space of the state space with a process space. In this perspective, generalized gradient flows are formally defined by five components:

1. A state space $\mathcal{X}$.

2. A process space $\mathcal{P}_{\dot{\mathcal{X}}}$ together with an instruction how states change $\dot{x} = \mathcal{T}(x)p$, where $x \in \mathcal{X}$, $p \in \mathcal{P}_{\dot{\mathcal{X}}}$, and $\mathcal{T}(x)$ a transformation operator.

3. An (internal) free energy $\mathcal{E}(x)$ for states $x \in \mathcal{X}$.

4. An (external) work rate $\mathcal{P}_{\mathrm{ext}}(x;p)$ in terms of the process vectors.

5. A dissipation potential $\mathcal{D}(x;p)$ in terms of process vectors inducing the cost of the change of state.

Then for given state $x \in \mathcal{X}$, the current change of state $\dot{x}$, in terms of the corresponding process vector $p \in \mathcal{P}_{\dot{\mathcal{X}}}$, is defined by

$$\begin{aligned} \dot{x} &= \mathcal{T}(x)p \\ p &= \operatorname*{arg\,min}_{q \in \mathcal{P}_{\dot{\mathcal{X}}}} \Big\{ \mathcal{D}(x;q) + \langle \boldsymbol{\nabla}\mathcal{E}(x), \mathcal{T}(x)q \rangle - \mathcal{P}_{\mathrm{ext}}(x;q) \Big\}, \end{aligned} \tag{1.2}$$

i.e., the loss of energy is maximized along the steepest descent of the energy under minimum cost. Again, many applications can be modelled by generalized gradient flows. We mention incompressible, immiscible two-phase flow in porous media [8], doubly non-linear Allen-Cahn equations [9], rate-independent finite elasticity [9], rate-dependent visco-plasticity at finite strain [10].

Apart from the structure itself, a (generalized) gradient flow interpretation may be beneficial in many ways. A wide range of abstract theory for gradient systems has been established dealing, e.g., with the well-posedness analysis [3, 4, 11], *a posteriori* error analysis for time discretizations [12], or *a priori* error analysis for numerical discretizations in time and space [13]. Furthermore, energy preserving time discretizations can be constructed [14], and optimization algorithms can be utilized for the construction of robust numerical solvers.

In this work, for the first time, we explore the gradient structure in the consolidation of fluid-saturated porous media, also called *theory of poro-elasticity*, and provide a generalized gradient flow formulation (1.2) for various poro-elasticity models. Coupled thermo-hydro-mechanical-chemical processes in porous media have been of great research interest recently, due to the presence of many practical applications of societal and industrial relevance. We mention not only classical, geotechnical applications within soil and reservoir mechanics, but also geothermal reservoirs [15], $\mathrm{CO_2}$ storage [16], deformation of hydrogels [17] or biomechanical applications [18] among others.

The theory of poro-elasticity goes back to the early seminal contributions by Terzaghi [19] and Biot [20]; since then many mathematical models for thermo-hydro-mechanical-chemical processes in porous media have been established utilizing, e.g., averaging processes [21], thermodynamic arguments [22], or homogenization [23, 24, 25, 26]. Traditionally, corresponding models are formulated as partial differential equations (PDE). Based on those formulations, there exists a mature literature on both analytical and numerical, rigorous mathematical theory for specific poro-elasticity models. It is beyond the scope of this work to give a comprehensive review; we only point out results connected to this paper: For the linear Biot model well-posedness has been showed using semigroup theory [27]. Recent advances on extensions of the linear Biot equations include well-posedness for the dynamic poro-elasticity [24], thermo-poro-elasticity with non-linear, thermal convection [28], poro-visco-elasticity with a purely visco-elastic strain [27, 29], and linear poro-elasticity with a deformation-dependent, non-linear permeability [29]. We are not aware of any explicit result on the well-posedness of linear poro-visco-elasticity models, which consider strains composed of an elastic and a visco-elastic contribution as modelled by [22], or non-linear poro-elasticity under an infinitesimal strain assumption as studied in a fully discretized setting by [30]. In terms of numerical discretization and solution of the linear Biot model, stable, spatial discretizations for various choices of primary variables have been introduced [31, 32, 33, 34]. Furthermore, physically motivated, robust operator splittings have been of great, recent interest, allowing for either using independent, tailored simulators for different physics or developing good block preconditioners for monolithic Krylov subspace methods. Such have been developed and studied for in particular the linear Biot model [35, 36, 37, 38, 39, 40, 41, 42, 43, 44], for non-linear poro-elasticity under an infinitesimal strain assumption [30], thermo-poro-elasticity [45] and large strain poro-elasticity [46].

To our knowledge, the connection between gradient flows and poro-elasticity from a mathematical point of view has not yet been studied in the literature. However, we have to honor the work by Miehe [47], which has also been an inspiration for this paper. In the aforementioned work with a focus on general modelling, isothermal flow in fully-saturated poro-elastic media under large strains is formulated using minimization principles, which eventually can be identified as a generalized gradient flow (1.2). The authors have furthermore noted that the minimization structure allows arbitrary pairs of finite elements as spatial discretization of the coupled problem.

The aim of this paper is not only to reveal a natural gradient structure of thermo-poro-visco-elasticity, but also to discuss how to exploit this structure to study well-posedness and naturally develop numerical methods. Serving as proof of concept, we explore thoroughly the linear Biot equations: We highlight the gradient structure of the linear Biot equations; well-posedness results are deduced employing abstract theory for doubly non-linear evolution equations and convex analysis; additionally, we identify widely used splitting schemes [35, 48] as alternating minimization, which are *a priori* guaranteed to converge. Utilizing abstract convergence theory, we are able to prove the same convergence rates as previously reported in the literature [37], in which problem-specific proofs are performed. We further apply the same workflow to more advanced poro-elasticity models with increased complexity and derive novel robust splitting schemes with guaranteed theoretical convergence rates. The findings are presented in two parts.

Part I (Sec. 2–7) is concerned with two aspects: (i) The modelling of coupled processes in poro-elastic materials as generalized gradient flows, and (ii) a subsequent well-posedness analysis. By combining the abstract generalized gradient flow formulation (1.2) with conceptual considerations regarding poro-elasticity, an abstract modelling framework for thermo-poro-visco-elasticity is established in Sec. 2. In its most general form, it allows for non-isothermal, (non-)Darcy flow in a saturated, non-linearly poro-visco-elastic material governed by dissipation only. Specific models are then obtained by involving common thermodynamic knowledge on free energies and dissipation potentials: A gradient flow formulation is derived for linear poro-elasticity (Sec. 3), linear poro-visco-elasticity (Sec. 4), non-linear poro-elasticity in the infinitesimal strain regime (Sec. 5), non-Newtonian Darcy and non-Darcy flows in poro-elastic media (Sec. 6), and thermo-poro-elasticity without thermal convection (Sec. 7), all consistent with previously employed PDE-based models [22]. Regarding the well-posedness analysis for poro-elasticity models, the main difficulty is the characteristic fact that the dissipation potential does not depend on all process vectors, as e.g., the change in mechanical displacement; this is solved by combining an abstract decoupling approach [49] with classical convex analysis [50] and theory on doubly non-linear evolution equations [51, 9], tailored to our needs, cf. Appendix A. It is summarized in a unified well-posedness result, Thm. 2.1, the main theoretical result of Part I. Furthermore, it is applied to practically all models listed above; in particular it gives a new concise proof for the well-posedness of the linear Biot equations.

Part II (Sec. 8–14) deals with the robust, numerical solution of aforementioned thermo-poro-visco-elasticity models, by exploiting the generalized gradient flow structure discussed in Part I. More precisely, after a semi-implicit time discretization along the lines of the minimizing movement scheme for gradient flows [52], the generalized gradient flow formulation translates into a minimization problem. For models discussed in Sec. 3–7, the minimization problem is convex, enabling the vast literature on convex optimization for efficient numerical solution, i.e., a problem non-specific machinery. Motivated by the recent advances on splitting schemes in the community, we discuss in particular the application of the plain alternating minimization or cyclic block coordinate descent methods [53, 54]. They allow for natural decoupling of the entire problem into its physical subproblems. Additionally, guaranteed convergence follows directly from abstract optimization theory, adjusted to our needs, cf. Appendix B. By this, we provide a new perspective on widely used, physically motivated splitting schemes, as the undrained and fixed-stress splits [35, 48] for linear poro-elasticity (Sec. 9). In particular, we provide a simple, mathematical intuition why those schemes are natural choices among predictor-corrector methods for which physical variables are simply fixed in the predictor step – in contrast for example to the drained and fixed-strain splits which are only conditionally stable [35, 48]. In addition, by applying the unified approach, we derive novel, robust splitting schemes for linear poro-visco-elasticity (Sec. 10) and nonlinear poro-elasticity under infinitesimal strains (Sec. 11), and provide a theoretical basis for the undrained-adiabatic and extended fixed-stress splits [45] for thermo-poro-elasticity (Sec. 12). This annexes the mathematically intuitive interpretation of directional minimization to the physical motivation of the splitting schemes. Finally, the minimization formulation allows for acceleration of the previously discussed splitting schemes, using a line search relaxation strategy (Sec. 13). In the context of poro-elasticity, this has not yet been observed in the literature. In particular, for linear problems, exact line search can be performed cheaply using quadratic interpolation due to the quadratic nature of the time-discrete minimization problem; the same technique is proposed as inexact line search for semi-linear models. We close the second part with a succinct numerical study (Sec. 14) aiming for answering four questions: (i) what is the impact of the relaxation of splitting schemes by line search; (ii) how does it relate to the optimization of tuning parameters employed within splitting schemes; (iii) how do relaxed splitting schemes perform for poro-visco-elasticity and (iv) and non-linear poro-elasticity? We observe that applying line search is effectively identical with optimizing splitting schemes, but no *a priori* knowledge or user-interaction is required. Furthermore, splitting schemes for poro-visco-elasticity and non-linear poro-elasticity show similar performance as for linear poro-elasticity.

1.1 Notation

Throughout this work, let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be an open, connected domain, with Lipschitz boundary $\partial\Omega$ and outward normal $\boldsymbol{n}$; let $[0,T]$ denote a finite time interval with finite time $T > 0$.

We use the following notation for standard function spaces and their norms [55]: Let $L^p(\Omega)$ be the space of functions for which the p-th power of the absolute value is Lebesgue integrable. For $L^2(\Omega)$, let $\langle \cdot,\cdot \rangle = \langle \cdot,\cdot \rangle_{L^2(\Omega)}$ denote the standard $L^2(\Omega)$ scalar product, $\| \cdot \| = \| \cdot \|_{L^2(\Omega)}$ the associated norm. Let $\langle \cdot,\cdot \rangle_\Gamma := \langle \cdot,\cdot \rangle_{L^2(\Gamma)}$ for measurable boundary segments $\Gamma \subset \Omega$. Let $W^{1,p}(\Omega)$, $p \geq 1$, denote the usual Sobolev space, consisting of functions in $L^p(\Omega)$ with a weak derivative in $L^p(\Omega)$, $H^1(\Omega) = W^{1,2}(\Omega)$ and $H_0^1(\Omega)$ its subspace with zero trace on $\partial\Omega$. Furthermore, $H_{\mathrm{div}}^p(\Omega)$, $p \geq 1$, denotes vectorial functions with d components in $L^p(\Omega)$ with a weak divergence in $L^2(\Omega)$; and $H(\mathrm{div};\Omega) = H_{\mathrm{div}}^2(\Omega)$.

We use bold symbols for vectors and tensors. Similarly, we use bold symbols for vector valued function spaces, e.g., $\boldsymbol{H}^1(\Omega)$. For elements of $\boldsymbol{H}^1(\Omega)$, let $\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\left(\boldsymbol{\nabla u} + \boldsymbol{\nabla u}^\top\right)$ denote the symmetric gradient, also called linearized strain; $\boldsymbol{\nabla}$ denotes both the spatial gradient and the (partial) functional derivative given by the Gâteaux-derivative, depending on the context. For $\mathcal{V}$ a Banach space, let $L^p(0,T;\mathcal{V})$ and $H^1(0,T;\mathcal{V})$ denote standard Bochner spaces endowed with standard norms. Newton's notation is used for denoting temporal derivatives of variables, e.g., $\dot{x}$ for the temporal derivative of x, whereas partial temporal derivatives of functionals are denoted by ∂_t. For $\mathcal{V}$ a Banach space, we denote $\mathcal{V}^\star$ its dual space and $\langle\cdot,\cdot\rangle_{\mathcal{V}^\star\times\mathcal{V}}$ a duality pairing. If obvious, we omit the subscript.

Finally, let $|\cdot|$ denote the absolute value, the Euclidean distance and the Frobenius norm for scalars, vectors and second-order tensors, respectively. And let $\operatorname{tr}\mathbf{A} = \sum_i A_{ii}$ denote the trace of a quadratic second-order tensor $\mathbf{A}$. The inequality $a \lesssim b$ means there exists a generic constant $C > 0$ independent of a and b such that $a \leq Cb$.

Let $\otimes$ denote the Kronecker product, and for the special case of two vectors. Moreover, let $:$ denote the single, double or triple (depending on the context) inner product for tensors. For the double inner product of a fourth order and a second order tensor we often omit $:$ as often done in mathematical literature for linear elasticity. Finally, $\langle\cdot,\cdot\rangle$ with tensorial arguments of same order is equivalent to a Lebesgue integral over the double inner product of the arguments.

A nomenclature regarding notation for generalized gradient flows, physical fields, function spaces etc. is provided in Appendix C.

Part I – Modelling and analyzing thermo-poro-visco-elasticity as generalized gradient flow

The main objective of part I is to highlight the inherent gradient structure of various poro-elasticity models. Secondary, we prove well-posedness for such models. Sec. 2 lays a foundation for this, providing an abstract gradient flow modelling framework for poro-elasticity, and subsequently an abstract well-posedness result for degenerate, doubly non-linear evolution equations, which will allow for a unified well-posedness analysis of poro-elasticity models. Based on those tools, we discuss linear poro-elasticity (Sec. 3), linear poro-visco-elasticity (Sec. 4), non-linear poro-elasticity in the infinitesimal strain regime (Sec. 5), non-Darcy flows in poro-elastic media (Sec. 6), and linear thermo-poro-elasticity without thermal convection (Sec. 7).

2 Foundation for modelling and analyzing poro-elasticity as generalized gradient flow

In the following, tools are introduced which will be applied throughout Part I of the paper. First, in Sec. 2.1, a general framework for modelling poro-elasticity based on the formal definition of generalized gradient flows (1.2) is proposed. Additionally, in Sec. 2.3, an abstract well-posedness result is derived, which allows for a unified analysis of poro-elasticity in the subsequent sections.

2.1 Formal modelling framework for non-isothermal flow in poro-visco-elastic media

From a continuum mechanical perspective, it is fair to assume that fluid-saturated, deformable porous media are purely governed by dissipation. That remains true, when allowing for additional structural visco-elasticity or non-isothermal flow with negligible, thermal convection. Consequently, it is natural to expect that a wide class of poro-elasticity models have an inherent gradient flow structure. Indeed, by incorporating thermodynamic interpretation into the notion of generalized gradient flows (1.2), we introduce a general modelling framework for non-isothermal flow in poro-visco-elastic media.

To set modelling limits, we restrict the discussion to fully-saturated media which deform under an infinitesimal strain assumption. Visco-elastic and thermal effects are allowed. But it is implicitly assumed that the considered system can be formulated as a gradient flow. This cannot always be true, e.g., when thermal convection or non-quasi-static mechanical behavior are non-negligible.

In the following the single components of a generalized gradient flow are defined based on thermodynamic knowledge:

1. As state space, we choose

$$\mathcal{X} = \{(\boldsymbol{u}, \theta, \boldsymbol{\varepsilon}_{\mathrm{v}}, S)\}\,, \tag{2.1}$$

   where $\boldsymbol{u}$ is the displacement of the matrix with respect to a reference state Ω; θ is the change of the fluid mass on Ω with respect to some reference configuration scaled by the inverse of a reference fluid density; $\boldsymbol{\varepsilon}_{\mathrm{v}}$ is the visco-elastic strain such that $\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{\varepsilon}_{\mathrm{v}}$ denotes the elastic strain; and S is the total entropy. Depending on which processes are considered, we choose only a suitable subset of $\mathcal{X}$ as state space.

2. Structural displacements $\boldsymbol{u}$ and visco-elastic strains $\boldsymbol{\varepsilon}_{\mathrm{v}}$ change with rates $\dot{\boldsymbol{u}}$ and $\dot{\boldsymbol{\varepsilon}}_{\mathrm{v}}$, respectively. Instead of using the rates $\dot{\theta}$ and $\dot{S}$ directly, we associate those with a volumetric flux $\boldsymbol{q}$ and an entropy flux $\boldsymbol{j}$, respectively. Their relations are imposed by the conservation of mass and balance of entropy

$$\dot{\theta} + \boldsymbol{\nabla} \cdot \boldsymbol{q} = q_\theta \;\text{ on } \Omega, \tag{2.2}$$

$$\dot{S} + \boldsymbol{\nabla} \cdot \boldsymbol{j} = q_S \;\text{ on } \Omega, \tag{2.3}$$

   where q_θ and q_S denote given, time-dependent production terms.

   Gradient flows effectively define changes of states, and boundary conditions can be imposed for those on boundary segments $\Gamma_{\boldsymbol{u}}, \Gamma_{\boldsymbol{q}}, \Gamma_{\boldsymbol{j}} \subset \partial\Omega$. We define the function spaces for $t \in [0, T]$ (without specifying regularity for now)

$$\dot{\mathcal{V}}(t) = \left\{\boldsymbol{v} : \Omega \to \mathbb{R}^d \,|\, \boldsymbol{v} = \dot{\boldsymbol{u}}_\Gamma(t) \text{ on } \Gamma_{\boldsymbol{u}}\right\}, \tag{2.4}$$

$$\mathcal{Z}(t) = \left\{\boldsymbol{z} : \Omega \to \mathbb{R}^d \,|\, \boldsymbol{z} \cdot \boldsymbol{n} = q_{\Gamma,\mathrm{n}}(t) \text{ on } \Gamma_{\boldsymbol{q}}\right\}, \tag{2.5}$$

$$\dot{\mathcal{T}}(t) = \left\{\boldsymbol{t} : \Omega \to \mathbb{R}^{d\times d}\right\}, \tag{2.6}$$

$$\mathcal{W}(t) = \left\{\boldsymbol{w} : \Omega \to \mathbb{R}^d \,|\, \boldsymbol{w} \cdot \boldsymbol{n} = j_{\Gamma,\mathrm{n}}(t) \text{ on } \Gamma_{\boldsymbol{j}}\right\}. \tag{2.7}$$

   associated with the change of structural displacement, volumetric flux, the change of the visco-elastic strain and entropy flux. Function spaces associated with the states are implicitly defined. Due to its internal character, no boundary conditions are imposed for the change of the visco-elastic strain. We suppress the explicit time-dependence of function spaces and boundary data in the rest of the article; e.g. we write $\dot{\mathcal{V}}$ instead of $\dot{\mathcal{V}}(t)$.

3. For given state $(\boldsymbol{u}, \theta, \boldsymbol{\varepsilon}_{\mathrm{v}}, S)$, let the energy $\mathcal{E}$ be given by the Helmholtz free energy of the system. According to thermodynamic derivations [22], we can derive the total stress $\boldsymbol{\sigma}$, the fluid pressure p and the temperature T by

$$\boldsymbol{\sigma} := \partial_{\boldsymbol{\nabla u}}\mathcal{E}, \qquad p := \partial_\theta \mathcal{E}, \qquad T := \partial_S \mathcal{E}. \tag{2.8}$$

Those also act as dual variables to $(\boldsymbol{u}, \theta, S)$, for which complementary boundary conditions to (2.4)–(2.7) have to be prescribed

$$\boldsymbol{\sigma n} = \boldsymbol{\sigma}_{\Gamma,\mathrm{n}} \quad \text{on } \Gamma_{\boldsymbol{\sigma}} := \partial\Omega \setminus \Gamma_{\boldsymbol{u}}, \tag{2.9}$$

$$p = p_\Gamma \quad \text{on } \Gamma_p := \partial\Omega \setminus \Gamma_{\boldsymbol{q}}, \tag{2.10}$$

$$T = T_\Gamma \quad \text{on } \Gamma_T := \partial\Omega \setminus \Gamma_{\boldsymbol{j}}. \tag{2.11}$$

As common in poro-elasticity, we employ an effective stress approach. We assume therefore, the total energy $\mathcal{E}$ can be decomposed into three contributions

$$\mathcal{E}(\boldsymbol{u}, \theta, \varepsilon_\mathrm{v}, S) = \mathcal{E}_\mathrm{eff}(\boldsymbol{\nabla u}, \varepsilon_\mathrm{v}) + \mathcal{E}_\mathrm{v}(\varepsilon_\mathrm{v}) + \mathcal{E}_\mathrm{fluid}(\boldsymbol{\nabla u}, \theta, \varepsilon_\mathrm{v}, S), \tag{2.12}$$

where the first contribution is assigned to the solid and defines the effective stress and will finally depend only on the elastic strain; the second contribution is the energy stored (and potentially lost) due to inelastic effects; and the third contribution corresponds to the fluid, allowing for defining the fluid quantities. We obtain the effective stress $\boldsymbol{\sigma}_\mathrm{eff}$, p and T also from

$$\boldsymbol{\sigma}_\mathrm{eff} := \partial_{\boldsymbol{\nabla u}}\mathcal{E}_\mathrm{eff}, \qquad p = \partial_\theta \mathcal{E}_\mathrm{fluid}, \qquad T = \partial_S \mathcal{E}_\mathrm{fluid}. \tag{2.13}$$

4. The external work rate $\mathcal{P}_\mathrm{ext}$ acts as a negative potential for changes of state or associated process vectors. Throughout this work, we assume $\mathcal{P}_\mathrm{ext}$ is linear and state-independent, and we allow $\mathcal{P}_\mathrm{ext}$ to vary in time. Furthermore, it is natural to assume the total external work rate decomposes into separate, independent contributions

$$\mathcal{P}_\mathrm{ext}(t, \dot{\boldsymbol{u}}, \boldsymbol{q}, \dot{\varepsilon}_\mathrm{v}, \boldsymbol{j}) = \mathcal{P}_\mathrm{ext,mech}(t, \dot{\boldsymbol{u}}) + \mathcal{P}_\mathrm{ext,fluid}(t, \boldsymbol{q}) + \mathcal{P}_\mathrm{ext,temp}(t, \boldsymbol{j}).$$

Since the visco-elastic strain is interpreted as an internal variable, no external work rate is associated to $\dot{\varepsilon}_\mathrm{v}$. In the context of poro-elasticity, external work rates integrate external body and surface forces acting on the fluid and the matrix. In particular, surface forces can be identified as the boundary conditions imposed on the dual variables (2.9)–(2.11). All in all, we employ

$$\begin{aligned}\mathcal{P}_\mathrm{ext,mech}(t, \dot{\boldsymbol{u}}) &= \langle \boldsymbol{f}_\mathrm{ext}(t), \dot{\boldsymbol{u}}\rangle + \langle \boldsymbol{\sigma}_{\Gamma,\mathrm{n}}(t), \dot{\boldsymbol{u}}\rangle_{\Gamma_{\boldsymbol{\sigma}}},\\ \mathcal{P}_\mathrm{ext,fluid}(t, \boldsymbol{q}) &= \langle \boldsymbol{g}_\mathrm{ext}(t), \boldsymbol{q}\rangle + \langle p_\Gamma(t), \boldsymbol{q}\cdot\boldsymbol{n}\rangle_{\Gamma_p},\\ \mathcal{P}_\mathrm{ext,temp}(t, \boldsymbol{j}) &= \langle T_\Gamma(t), \boldsymbol{j}\cdot\boldsymbol{n}\rangle_{\Gamma_T}.\end{aligned}$$

Here, $\boldsymbol{f}_\mathrm{ext}$ and $\boldsymbol{g}_\mathrm{ext}$ denote external body forces applied to the matrix and the fluid, respectively. We stress that under the hypothesis of small perturbations of the Lagrangian porosity [22], often coming along with the assumptions of linear elasticity, it is indeed fair to assume that $\boldsymbol{f}_\mathrm{ext}$ and $\boldsymbol{g}_\mathrm{ext}$ are state-independent.

5. Accounting for viscous dissipation, changes of states come at cost governed by a dissipation potential. For the poro-elasticity models considered in this work, it is adequate to assume that the underlying dissipation mechanisms are state-independent; e.g., for large strain poro-elasticity, this is not the case. Furthermore, we presume independent cost for each process such that the total dissipation potential decomposes into

$$\mathcal{D}(\dot{\boldsymbol{u}}, \boldsymbol{q}, \dot{\varepsilon}_\mathrm{v}, \boldsymbol{j}) = \mathcal{D}_\mathrm{mech}(\dot{\boldsymbol{u}}) + \mathcal{D}_\mathrm{fluid}(\boldsymbol{q}) + \mathcal{D}_\mathrm{v}(\dot{\varepsilon}_\mathrm{v}) + \mathcal{D}_\mathrm{th}(\boldsymbol{j}). \tag{2.14}$$

A common feature for many poro-elasticity models is to assume that structural displacements react instantaneously, which corresponds to the choice $\mathcal{D}_\mathrm{mech} = 0$. The potentials $\mathcal{D}_\mathrm{fluid}$, $\mathcal{D}_\mathrm{v}$ and $\mathcal{D}_\mathrm{th}$ correspond to a (non-)Darcy-type law, some viscosity law for strain rates and a Fourier-type law, respectively. We will essentially consider quadratic dissipation potentials; besides in the context of non-Darcy flows in poro-elastic materials, cf. Sec. 6.

Finally, (1.2) yields an abstract model for describing the evolution of a fluid-saturated, deformable porous medium. The states $(\boldsymbol{u}, \theta, \varepsilon_{\mathrm{v}}, S)$ change in time $t \in (0,T)$ by

$$\dot{\theta} = q_\theta - \boldsymbol{\nabla} \cdot \boldsymbol{q}, \tag{2.15}$$

$$\dot{S} = q_S - \boldsymbol{\nabla} \cdot \boldsymbol{j}, \tag{2.16}$$

$$\begin{aligned}(\dot{\boldsymbol{u}}, \boldsymbol{q}, \dot{\varepsilon}_{\mathrm{v}}, \boldsymbol{j}) = \underset{(\boldsymbol{v},\boldsymbol{z},\boldsymbol{t},\boldsymbol{w}) \in \dot{\mathcal{V}} \times \mathcal{Z} \times \dot{\mathcal{T}} \times \mathcal{W}}{\arg\min} \Big\{ &\langle \partial_{\boldsymbol{\nabla u}} \mathcal{E}(\boldsymbol{\nabla u}, \theta, \varepsilon_{\mathrm{v}}, S), \boldsymbol{\nabla v} \rangle - \mathcal{P}_{\mathrm{ext,mech}}(t, \boldsymbol{v}) \\ &+ \mathcal{D}_{\mathrm{fluid}}(\boldsymbol{z}) - \langle \partial_\theta \mathcal{E}_{\mathrm{fluid}}(\boldsymbol{\nabla u}, \theta, \varepsilon_{\mathrm{v}}, T), \boldsymbol{\nabla} \cdot \boldsymbol{z} \rangle - \mathcal{P}_{\mathrm{ext,fluid}}(t, \boldsymbol{z}) \\ &+ \mathcal{D}_{\mathrm{v}}(\boldsymbol{t}) + \langle \partial_{\varepsilon_{\mathrm{v}}} \mathcal{E}(\boldsymbol{\nabla u}, \theta, \varepsilon_{\mathrm{v}}, S), \boldsymbol{t} \rangle \\ &+ \mathcal{D}_{\mathrm{th}}(\boldsymbol{w}) - \langle \partial_S \mathcal{E}_{\mathrm{fluid}}(\boldsymbol{\nabla u}, \theta, \varepsilon_{\mathrm{v}}, T), \boldsymbol{\nabla} \cdot \boldsymbol{w} \rangle - \mathcal{P}_{\mathrm{ext,temp}}(t, \boldsymbol{w}) \Big\}\end{aligned} \tag{2.17}$$

and are subject to initial conditions at time $t = 0$

$$\boldsymbol{u} = \boldsymbol{u}_0, \qquad \theta = \theta_0, \qquad \varepsilon_{\mathrm{v}} = \varepsilon_{\mathrm{v},0}, \qquad S = S_0 \qquad \text{on } \Omega. \tag{2.18}$$

Remark 2.1 (Primal and dual formulation)**.** *We distinguish between primal and dual variables. The gradient flow formulation* (2.15)–(2.17) *governs primal variables. Hence, we will call this the primal formulation. In certain situations, a dual formulation governing dual variables can be derived from the primal formulation. This is, e.g., discussed for linear poro-elasticity, cf. Sec. 3.2.*

2.2 Poro-elasticity formulated as doubly non-linear evolution equation

The framework as described in the previous section is suitable for modelling poro-elasticity. Yet, in the next section, we provide tools for a unified well-posedness analysis of models of type (2.15)–(2.18), which utilize the closely related reformulation of a generalized gradient flow as a doubly non-linear evolution equation. A natural reformulation of the general poro-elasticity model (2.15)–(2.18) is achieved by introducing accumulated fluxes

$$\boldsymbol{q}_{\int}(t) := \int_0^t \boldsymbol{q}(\tau)\, d\tau, \tag{2.19}$$

$$\boldsymbol{j}_{\int}(t) := \int_0^t \boldsymbol{j}(\tau)\, d\tau \tag{2.20}$$

as alternative states to θ and S, respectively. Corresponding function spaces $\mathcal{Z}_{\int}$ and $\mathcal{W}_{\int}$ are implicitly defined by $\dot{\mathcal{Z}}_{\int} = \mathcal{Z}$ and $\dot{\mathcal{W}}_{\int} = \mathcal{W}$, i.e., $\boldsymbol{q}_{\int} \in \mathcal{Z}_{\int}$ if and only if $\dot{\boldsymbol{q}}_{\int} \in \mathcal{Z}$. Analogously, we set $Q_\theta(t) := \int_0^t q_\theta(s)\, ds$ and $Q_S(t) := \int_0^t q_S(s)\, ds$. By (2.15)–(2.16), the accumulated fluxes are associated with θ and S by

$$\theta = \theta_0 + Q_\theta - \boldsymbol{\nabla} \cdot \boldsymbol{q}_{\int}, \tag{2.21}$$

$$S = S_0 + Q_S - \boldsymbol{\nabla} \cdot \boldsymbol{j}_{\int}. \tag{2.22}$$

By eliminating θ and S, the generalized gradient flow formulation (2.15)–(2.18) becomes a degenerate doubly non-linear evolution equation for $(\boldsymbol{u}, \boldsymbol{q}_{\int}, \varepsilon_{\mathrm{v}}, \boldsymbol{j}_{\int}) \in \mathcal{V} \times \mathcal{Z}_{\int} \times \mathcal{T} \times \mathcal{W}_{\int}$

$$\boldsymbol{\nabla} \mathcal{D}(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_{\int}, \dot{\varepsilon}_{\mathrm{v}}, \dot{\boldsymbol{j}}_{\int}) + \boldsymbol{\nabla} \tilde{\mathcal{E}}(t, \boldsymbol{u}, \boldsymbol{q}_{\int}, \varepsilon_{\mathrm{v}}, \boldsymbol{j}_{\int}) = \boldsymbol{\nabla} \mathcal{P}_{\mathrm{ext}}(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_{\int}, \dot{\varepsilon}_{\mathrm{v}}, \dot{\boldsymbol{j}}_{\int}), \tag{2.23}$$

with reinterpreted (potentially, explicitly time-dependent) energy

$$\tilde{\mathcal{E}}(t, \boldsymbol{u}, \boldsymbol{q}_{\int}, \varepsilon_{\mathrm{v}}, \boldsymbol{j}_{\int}) := \mathcal{E}(\boldsymbol{\nabla u}, \theta_0 + Q_\theta(t) - \boldsymbol{\nabla} \cdot \boldsymbol{q}_{\int}, \varepsilon_{\mathrm{v}}, S_0 + Q_S(t) - \boldsymbol{\nabla} \cdot \boldsymbol{j}_{\int}).$$

and initial conditions

$$\boldsymbol{u} = \boldsymbol{u}_0, \qquad \boldsymbol{q}_{\int} = \boldsymbol{0}, \qquad \varepsilon_{\mathrm{v}} = \varepsilon_{\mathrm{v},0}, \qquad \boldsymbol{j}_{\int} = \boldsymbol{0} \qquad \text{on } \Omega. \tag{2.24}$$

Remark 2.2 (Reformulation over linear spaces)**.** *The function spaces $\mathcal{V}\times\mathcal{Z}_{\int}\times\mathcal{T}\times\mathcal{W}_{\int}$ are given by linear spaces translated by some essential boundary conditions, cf.* (2.4)–(2.7)*. By explicitly incorporating the translation into the definitions of $\mathcal{D}$, $\tilde{\mathcal{E}}$ and $\mathcal{P}_{\mathrm{ext}}$, the problem* (2.23) *can be reformulated over time-independent, linear spaces; however, each functional becomes explicitly time-dependent.*

Remark 2.3 (Time-independent dissipation potential and energy functional and linear function spaces)**.** *From a modelling perspective, imposing time-dependent, essential boundary conditions is straightforward. However, the analysis of gradient systems under essential boundary conditions is known to be a delicate topic, cf., e.g., [7]. A model-specific discussion is most often required; e.g., for quadratic potentials and energies, boundary conditions or external sources can be equivalently reformulated as linear contributions of the external work rates, allowing for reducing the discussion to linear, time independent spaces and time-independent energy functionals. Non-homogeneous, time-independent boundary conditions are less of a problem, as the driving functional remains decreasing along solutions.*

2.3 Abstract well-posedness result for degenerate doubly non-linear evolution equations

In the following, we establish an abstract well-posedness result which allows for a unified discussion of poro-elasticity models arising from the gradient flow modelling framework introduced above, cf. Sec. 3–7. For this, we first note that the problem (2.23) falls into the category of degenerate, doubly non-linear evolution equations on Banach spaces. More specifically, the structural assumptions made in Sec. 2.1, and assuming solely external work rates are time-dependent, cf. Rem. 2.3, motivates to consider the abstract evolutionary system

$$(\dot{x}_1,\dot{x}_2) = \underset{(y_1,y_2)\in\mathcal{V}_1\times\mathcal{V}_2}{\arg\min} \Big\{ \Psi(y_2) + \langle \boldsymbol{\nabla}\mathcal{E}(x_1,x_2),(y_1,y_2)\rangle \\ - \langle \mathcal{P}_1(t),y_1\rangle - \langle \mathcal{P}_2(t),y_2\rangle \Big\}. \tag{2.25}$$

In particular, we assume:

(P1) The set of primary variables can be partitioned into two sets with either vanishing or non-vanishing dissipative character. Those can be respectively grouped in two (multi-valued) variables x_1, x_2. Let x_1 denote the variables that change without cost.

(P2) The function spaces $\mathcal{V}_1$ and $\mathcal{V}_2$ corresponding to x_1 and x_2, respectively, are assumed to be time-independent and to have a linear structure. Thereby, they can be identified as both state and tangent spaces. Furthermore, let $\mathcal{V}_i$ be Banach spaces with norms $\|\cdot\|_{\mathcal{V}_i}$, $i=1,2$. In particular, assume there exists a semi-norm $|\cdot|_{\mathcal{V}_2}$ on $\mathcal{V}_2$ such that

$$\|y_2\|_{\mathcal{V}_2}^p = \|y_2\|_{\mathcal{B}_2}^p + |y_2|_{\mathcal{V}_2}^p,$$

where $\mathcal{B}_2\supset\mathcal{V}_2$ is a larger Banach space with norm $\|\cdot\|_{\mathcal{B}_2}$, and $p:=\min\{p_\psi,p_\mathcal{E}\}\in(1,\infty)$ with p_ψ and $p_\mathcal{E}$ introduced in (P3) and (P4).

(P3) The dissipation potential $\Psi:\mathcal{B}_2\to[0,\infty)$ is convex, continuously differentiable and coercive wrt. to $\mathcal{B}_2$. In particular, there exists a constant $C>0$ and $p_\psi\in(1,\infty)$ satisfying

$$\Psi(y_2)\geq C\|y_2\|_{\mathcal{B}_2}^{p_\psi},\quad y_2\in\mathcal{B}_2.$$

(P4) The free energy of the system is convex, lower semi-continuous and continuously differentiable. Furthermore, it can be decomposed into a strictly convex part in the variable with vanishing dissipation, and a convex contribution in an affine combination of the primary variables

$$\mathcal{E}(x_1,x_2)=\mathcal{E}_1(x_1)+\mathcal{E}_2(\Lambda(x_1,x_2)),\ (x_1,x_2)\in\mathcal{V}_1\times\mathcal{V}_2, \tag{2.26}$$

i.e., $\mathcal{E}_1:\mathcal{V}_1\to[0,\infty)$ is strictly convex; $\mathcal{E}_2:\tilde{\mathcal{V}}\to[0,\infty)$ is convex with $\tilde{\mathcal{V}}$ a (separable) Banach space; and $\Lambda:\mathcal{V}_1\times\mathcal{V}_2\to\tilde{\mathcal{V}}$ is an affine operator, satisfying

$$\Lambda(x_1,x_2)-\Lambda(y_1,y_2)=\Lambda_1(x_1-y_1)+\Lambda_2(x_2-y_2),\quad \forall x_i,y_i\in\mathcal{V}_i,\ i=1,2,$$

for $\Lambda_i:\mathcal{V}_i\to\tilde{\mathcal{V}}$ two linear operators with adjoint operators $\Lambda_i^\star$, $i=1,2$. Furthermore, there exist constants C_1,C_2,C_3 and $p_1,p_\mathcal{E}\in(1,\infty)$ satisfying

$$\begin{aligned}\mathcal{E}_1(x_1)&\geq C_1\|x_1\|_{\mathcal{V}_1}^{p_1},\\ \mathcal{E}(x_1,x_2)&\geq C_2|x_2|_{\mathcal{V}_2}^{p_\mathcal{E}}-C_3.\end{aligned}$$

(P5) The external loads satisfy $\mathcal{P}_1\in C(0,T;\mathcal{V}_1^\star)\cap W^{1,p_1^\star}(0,T;\mathcal{V}_1^\star)$ and $\mathcal{P}_2\in C(0,T;\mathcal{V}_2^\star)\cap W^{1,p^\star}(0,T;\mathcal{V}_2^\star)$, where $\frac{1}{p_1}+\frac{1}{p_1^\star}=\frac{1}{p}+\frac{1}{p^\star}=1$.

(P6) The initial conditions $(x_1(0),x_2(0))\in\mathcal{V}_1\times\mathcal{V}_2$ have finite energy $\mathcal{E}(x_1(0),x_2(0))<\infty$ and satisfy the compatibility condition

$$x_1(0)=\underset{x_1\in\mathcal{V}_1}{\arg\min}\left\{\mathcal{E}(x_1,x_2(0))-\langle\mathcal{P}_1(0),x_1\rangle\right\}.$$

Quasi-static systems of type (2.25) have been studied in the literature before, cf., e.g., [49, 9]; however in the aforementioned works, energies with decompositions different than the poro-elasticity-specific choice (2.26) are treated. And in the theory on doubly non-linear evolution equations in general, the external loading $\mathcal{P}_2$ would usually be assumed to be in the dual of a larger space ($\mathcal{B}_2^\star$ in our context). Here, the weaker regularity assumption on $\mathcal{P}_2$ originates from the nature of external loadings applied in the context of flow in porous media. In order to handle the weak, spatial regularity within the theory on doubly non-linear evolution equations, stronger temporal regularity is required along with above growth conditions on the energy functional. All in all, using similar ideas as [49, 9] but tailored to the above problem structure, we prove well-posedness of (2.25) under (P1)–(P6).

Theorem 2.1 (Well-posedness for generalized gradient flow system (2.25))**.** *Assuming* (P1)–(P6)*, there exists a solution* (x_1,x_2) *to* (2.25) *satisfying*

$$\begin{aligned}x_1&\in L^\infty(0,T;\mathcal{V}_1),\\ x_2&\in W^{1,p}(0,T;\mathcal{B}_2)\cap L^\infty(0,T;\mathcal{V}_2).\end{aligned}$$

If $\boldsymbol{\nabla}\Psi$ *or* $\boldsymbol{\nabla}\mathcal{E}$ *is linear and self-adjoint, it is unique.*

Proof. We follow ideas by [49, 9] and decouple the system into a minimization problem and a gradient flow problem. The first is discussed using classical convex analysis (Thm. A.2, cf. Appendix A); the discussion of the gradient flow problem utilizes theory on doubly non-linear evolution equations (Thm. A.1, cf. Appendix A).

Decoupling. For fixed time $t\in[0,T]$, the optimality conditions for $(x_1(t),x_2(t))$, derived as first variation, corresponding to (2.25) read

$$\boldsymbol{\nabla}\mathcal{E}_1(x_1(t))+\Lambda_1^\star\boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1(t),x_2(t))\big)=\mathcal{P}_1(t)\ \text{ in }\mathcal{V}_1^\star, \tag{2.27}$$

$$\boldsymbol{\nabla}\Psi(\dot{x}_2(t))+\Lambda_2^\star\boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1(t),x_2(t))\big)=\mathcal{P}_2(t)\ \text{ in }\mathcal{V}_2^\star. \tag{2.28}$$

We conclude that x_1 is defined as solution to a minimization problem for given x_2. Hence, given $t \in [0,T]$ and $x_2 \in \mathcal{V}_2$, we denote $x_1^\star := x_1^\star(t,x_2) \in \mathcal{V}_1$ to be the solution of the problem

$$\inf_{y_1 \in \mathcal{V}_1} \mathcal{E}(y_1, x_2) - \langle \mathcal{P}_1(t), y_1 \rangle \,. \tag{2.29}$$

Since $\mathcal{E}_1$ is strictly convex, $x_1^\star$ is well-defined by Thm. A.2. We introduce the reduced energy

$$\mathcal{E}_{\text{red}}(t,x_2) := \mathcal{E}\big(x_1^\star(t,x_2), x_2\big) - \langle \mathcal{P}_1(t), x_1^\star(t,x_2)\rangle \,,\ t \in [0,T],\ x_2 \in \mathcal{V}_2.$$

We observe, that the optimality conditions (2.27)–(2.28) can be decoupled into

$$\boldsymbol{\nabla}\mathcal{E}_1(x_1^\star(t,x_2)) + \Lambda_1^\star \boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,x_2),x_2)\big) = \mathcal{P}_1(t) \quad \text{in } \mathcal{V}_1^\star, \tag{2.30}$$

$$\boldsymbol{\nabla}\Psi(\dot{x}_2(t)) + \boldsymbol{\nabla}\mathcal{E}_{\text{red}}(t, x_2(t)) = \mathcal{P}_2(t) \quad \text{in } \mathcal{V}_2^\star. \tag{2.31}$$

Existence and uniqueness for the gradient flow problem. Eq. (2.31) has the structure of a doubly non-linear evolution equation. The existence (and uniqueness) of a solution to (2.31) follows by employing Thm. A.1; under above assumptions, it is sufficient to check that $\mathcal{E}_{\text{red}}$ complies with the assumptions of Thm. A.1.

First, by (P4) and (P5), it is simple to show that there exist constants $C_1 > 0$, $C_2 \geq 0$, independent of t, satisfying

$$\mathcal{E}_{\text{red}}(t,x_2) \geq C_1 |x_2|_{\mathcal{V}_2}^{p_\mathcal{E}} - C_2.$$

Second, $\mathcal{E}_{\text{red}}(t,\cdot)$ is convex on $\mathcal{V}_2$: This follows from the fact that $\boldsymbol{\nabla}\mathcal{E}_{\text{red}}$ is monotone [56]. In order to see this, we derive an explicit expression for $\boldsymbol{\nabla}\mathcal{E}_{\text{red}}$. Let $x_2, y_2 \in \mathcal{V}_2$ be arbitrary, and let $Dx_1^\star(t,x_2)[y_2] := \frac{\mathrm{d}}{\mathrm{d}\delta}\big|_{\delta=0} x_1^\star(t, x_2 + \delta y_2)$. Using the chain rule, the optimality condition corresponding to (2.29), and the definitions of $x_1^\star$ and $\mathcal{E}$, we obtain

$$\begin{aligned}
&\langle \boldsymbol{\nabla}\mathcal{E}_{\text{red}}(t,x_2), y_2\rangle \\
&\quad = \big\langle \boldsymbol{\nabla}_1\mathcal{E}\big(x_1^\star(t,x_2),x_2\big) - \mathcal{P}_1(t), Dx_1^\star(t,x_2)[y_2]\big\rangle + \big\langle \boldsymbol{\nabla}_2\mathcal{E}\big(x_1^\star(t,x_2),x_2\big), y_2\big\rangle \\
&\quad = \big\langle \boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,x_2),x_2)\big), \Lambda_2 y_2\big\rangle \,.
\end{aligned}$$

Hence, from the definition of Λ, we obtain

$$\begin{aligned}
&\langle \boldsymbol{\nabla}\mathcal{E}_{\text{red}}(t,x_2) - \boldsymbol{\nabla}\mathcal{E}_{\text{red}}(t,y_2), x_2 - y_2\rangle \\
&\quad = \big\langle \boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,x_2),x_2)\big) - \boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,y_2),y_2)\big), \Lambda_2(x_2 - y_2)\big\rangle \\
&\quad = \big\langle \boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,x_2),x_2)\big) - \boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,y_2),y_2)\big), \\
&\qquad\quad \big(\Lambda(x_1^\star(t,x_2),x_2) - \Lambda(x_1^\star(t,y_2),y_2)\big) - \Lambda_1\big(x_1^\star(t,x_2) - x_1^\star(t,y_2)\big)\big\rangle \,.
\end{aligned}$$

Subtracting the optimality condition for arbitrary $x_2, y_2 \in \mathcal{V}_2$, yields

$$\begin{aligned}
&\boldsymbol{\nabla}\mathcal{E}_1(x_1^\star(t,x_2)) - \boldsymbol{\nabla}\mathcal{E}_1(x_1^\star(t,y_2)) \\
&\quad = -\Lambda_1^\star\left(\boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,x_2),x_2)\big) - \boldsymbol{\nabla}\mathcal{E}_2(\Lambda(x_1^\star(t,y_2),y_2))\right) \quad \text{in } \mathcal{V}_1^\star.
\end{aligned}$$

Hence, together, we obtain

$$\begin{aligned}
&\langle \boldsymbol{\nabla}\mathcal{E}_{\text{red}}(t,x_2) - \boldsymbol{\nabla}\mathcal{E}_{\text{red}}(t,y_2), x_2 - y_2\rangle \\
&\quad = \big\langle \boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,x_2),x_2)\big) - \boldsymbol{\nabla}\mathcal{E}_2\big(\Lambda(x_1^\star(t,y_2),y_2)\big), \Lambda(x_1^\star(t,x_2),x_2) - \Lambda(x_1^\star(t,y_2),y_2)\big\rangle \\
&\qquad + \langle \boldsymbol{\nabla}\mathcal{E}_1\left(x_1^\star(t,x_2)\right) - \boldsymbol{\nabla}\mathcal{E}_1(x_1^\star(t,x_2)), x_1^\star(t,x_2) - x_1^\star(t,y_2)\rangle \,.
\end{aligned}$$

From the convexity of $\mathcal{E}_1$ and $\mathcal{E}_2$, we obtain the convexity of $\mathcal{E}_{\text{red}}$.

By exploiting the definition of $\mathcal{E}_{\mathrm{red}}$, the optimality condition (2.30), (P4), and (P5), we obtain for almost every $t \in (0,T)$ and $p_1^\star$ as in (P5)

$$\begin{aligned}
|\partial_t \mathcal{E}_{\mathrm{red}}(t,x_2)| &= \big| \left\langle \boldsymbol{\nabla}\mathcal{E}_1(x_1^\star(t,x_2)), \dot{x}_1^\star(t,x_2)\right\rangle \\
&\qquad + \left\langle \boldsymbol{\nabla}\mathcal{E}_1\left(\Lambda\left(x_1^\star(t,x_2),x_2\right)\right), \Lambda_1 \dot{x}_1^\star(t,x_2)\right\rangle - \left\langle \mathcal{P}_1, \dot{x}_1^\star(t,x_2)\right\rangle \\
&\qquad - \left\langle \partial_t \mathcal{P}_1, x_1^\star(t,x_2)\right\rangle \big| \\
&= |\langle \partial_t \mathcal{P}_1, x_1^\star(t,x_2)\rangle| \\
&\lesssim \|\partial_t \mathcal{P}_1\|_{\mathcal{V}_1^\star}^{p_1^\star} + \|x_1^\star(t,x_2)\|_{\mathcal{V}_1}^{p_1}.
\end{aligned}$$

In addition, by employing (P4), and using that $\mathcal{E}_2$ is positive, it follows

$$\|x_1^\star(t,x_2)\|_{\mathcal{V}_1}^{p_1} \lesssim \mathcal{E}_1(x_1^\star(t,x_2)) \leq \mathcal{E}_{\mathrm{red}}(t,x_2) + \langle \mathcal{P}_1, x_1^\star(t,x_2)\rangle.$$

Employing (P5) and Young's inequality, yields

$$\|x_1^\star(t,x_2)\|_{\mathcal{V}_1}^{p_1} \lesssim \mathcal{E}_{\mathrm{red}}(t,x_2) + \langle \mathcal{P}_1, x_1^\star(t,x_2)\rangle.$$

Altogether, we obtain

$$|\partial_t \mathcal{E}_{\mathrm{red}}(t,x_2)| \lesssim \|\mathcal{P}_1\|_{\mathcal{V}_1^\star}^{p_1^\star} + \|\partial_t \mathcal{P}_1\|_{\mathcal{V}_1^\star}^{p_1^\star} + \mathcal{E}_{\mathrm{red}}(t,x_2). \tag{2.32}$$

Consequently, $\mathcal{E}_{\mathrm{red}}$ complies with Thm. A.1, and together with (P1)–(P6), there exists a solution $x_2 \in W^{1,p}(0,T;\mathcal{B}_2) \cap L^\infty(0,T;\mathcal{V}_2)$ to (2.30). It is unique in case $\boldsymbol{\nabla}\Psi$ or $\boldsymbol{\nabla}\mathcal{E}_{\mathrm{red}}$ are linear and self-adjoint, cf. Thm. A.1. The latter follows for linear and self-adjoint $\boldsymbol{\nabla}\mathcal{E}_i$, $i = 1,2$.

Finite energy. By Thm. A.1, x_2 satisfies the characteristic energy identity

$$\begin{aligned}
&\int_0^T \Psi(\dot{x}_2(t))\,dt + \mathcal{E}_{\mathrm{red}}(x_2(T)) - \langle \mathcal{P}_2(T), x_2(T)\rangle \\
&\quad = \mathcal{E}_{\mathrm{red}}\left(x_2(0)\right) - \langle \mathcal{P}_2(0), x_2(0)\rangle + \int_0^T \partial_t \mathcal{E}_{\mathrm{red}}(t,x_2(t))\,dt - \int_0^T \left\langle \dot{\mathcal{P}}_2(t), x_2(t)\right\rangle\,dt.
\end{aligned}$$

Using (P4) and (2.32), we obtain

$$\begin{aligned}
\int_0^T \Psi(\dot{x}_2(t))\,dt + \mathcal{E}_{\mathrm{red}}(x_2(T)) &\lesssim \langle \mathcal{P}_2(T), x_2(T)\rangle \\
&\leq \mathcal{E}_{\mathrm{red}}\left(x_2(0)\right) - \langle \mathcal{P}_2(0), x_2(0)\rangle + \|\mathcal{P}_1\|_{W^{1,p_1^\star}(0,T;\mathcal{V}_1^\star)}^{p_1^\star} \\
&\quad + \|\mathcal{P}_2(T)\|_{\mathcal{V}_2^\star}^{p_{\mathcal{E}}^\star} + \|\mathcal{P}_2\|_{W^{1,p_{\mathcal{E}}^\star}(0,T;\mathcal{V}_2^\star)}^{p_{\mathcal{E}}^\star} + C_3 + \int_0^T \mathcal{E}_{\mathrm{red}}(t,x_2)\,dt,
\end{aligned}$$

with $\frac{1}{p_{\mathcal{E}}} + \frac{1}{p_{\mathcal{E}}^\star} = 1$ and C_3 from (P4). The assumptions on the external data are chosen such that the right hand side is uniformly bounded in T up to the last term. By the Grönwall inequality it follows that $\mathcal{E}_{\mathrm{red}}(t,x_2(t))$ is uniformly bounded in time.

Existence and uniqueness for the minimization problem. Since $x_2 \in L^\infty(0,T;\mathcal{V}_2)$, (2.29) is well-defined for $x_2 = x_2(t)$ for a.e. $t \in (0,T)$; thereby also $x_1 = x_1^\star(t,x_2)$. Finally, by the definition of $\mathcal{E}_{\mathrm{red}}$ and (P4), it follows for a.e. $t \in (0,T)$

$$\|x_1(t)\|_{\mathcal{V}_1}^{p_1} \lesssim \mathcal{E}_{\mathrm{red}}(t,x_2) + \|\mathcal{P}_1(t)\|_{\mathcal{V}_1}^{p_1^\star}.$$

Hence, by the above paragraph and (P5), $x_1^\star \in L^\infty(0,T;\mathcal{V}_1)$. Altogether, we obtain existence (and uniqueness) of the coupled system (2.25), which concludes the proof.

□

Remark 2.4. *Detailed stability bounds can be derived using energy identities for gradient flows, cf., e.g.,* (A.3).

3 Linear Biot equations as generalized gradient flow

The theory of linear poro-elasticity describes the continuum mechanics of coupled flow and geomechanics in porous media under several simplifying hypotheses: in particular, the fundamental linearizing assumptions of linear elasticity; the hypothesis of small perturbations of the Lagrangian porosity; and an at most slightly compressible, Newtonian fluid. Together with first principles and Darcy's law, the *Biot's consolidation model*, also called *linear Biot equations*, can be deduced, coupling elliptic and parabolic equations. For a detailed introduction, we refer to the seminal work by Biot [20] and the comprehensive books [22, 21].

In this section, we provide a derivation of the linear Biot equations employing the modelling framework described in Sec. 2.1. Thereby we demonstrate the inherent gradient flow structure of the linear Biot equations. Acknowledging the fact that the linear Biot equations have been already studied quite thoroughly in the literature, the following discussion serves mostly as proof of concept and guide for subsequent discussions of more involved poro-elasticity models.

3.1 Generalized gradient flow formulation of linear poro-elasticity

Using the modelling approach described in Sec. 2.1, we derive Biot's consolidation model as a generalized gradient flow. It suffices to specify states, an associated Helmholtz free energy $\mathcal{E}$ and a dissipation potential $\mathcal{D}$.

As states, we choose the mechanical displacement $\boldsymbol{u}$ and the volume content θ with associated processes $\dot{\boldsymbol{u}}$ and the volumetric flux $\boldsymbol{q}$, respectively. Suitable function spaces for the latter, incorporating essential boundary conditions are given by

$$\mathcal{V} = \left\{ \boldsymbol{v} \in H^1(\Omega) \,|\, \boldsymbol{v} = \boldsymbol{u}_\Gamma \text{ on } \Gamma_{\boldsymbol{u}} \right\}, \tag{3.1}$$

$$\dot{\mathcal{V}} = \left\{ \boldsymbol{v} \in H^1(\Omega) \,|\, \boldsymbol{v} = \dot{\boldsymbol{u}}_\Gamma \text{ on } \Gamma_{\boldsymbol{u}} \right\}, \tag{3.2}$$

$$\mathcal{Z} = \left\{ \boldsymbol{z} \in H(\mathrm{div};\Omega) \,|\, \boldsymbol{z} \cdot \boldsymbol{n} = q_{\Gamma,\mathrm{n}} \text{ on } \Gamma_{\boldsymbol{q}} \right\}. \tag{3.3}$$

For their variations, we define correspondingly

$$\mathcal{V}_0 = \left\{ \boldsymbol{v} \in H^1(\Omega) \,|\, \boldsymbol{v} = \boldsymbol{0} \text{ on } \Gamma_{\boldsymbol{u}} \right\}, \tag{3.4}$$

$$\mathcal{Z}_0 = \left\{ \boldsymbol{z} \in H(\mathrm{div};\Omega) \,|\, \boldsymbol{z} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma_{\boldsymbol{q}} \right\}. \tag{3.5}$$

The energy is chosen to be the Helmholtz free energy for linearly deformable porous media, cf. Ch. 4.2.2, [22],

$$\begin{aligned} \mathcal{E}(\boldsymbol{u},\theta) &= \mathcal{E}_{\mathrm{eff}}(\boldsymbol{u}) + \mathcal{E}_{\mathrm{fluid}}(\boldsymbol{u},\theta), \\ \mathcal{E}_{\mathrm{eff}}(\boldsymbol{u}) &= \tfrac{1}{2} \left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u}) \right\rangle, \\ \mathcal{E}_{\mathrm{fluid}}(\boldsymbol{u},\theta) &= \tfrac{M}{2} \left\| \theta - \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u} \right\|^2, \end{aligned}$$

where $\mathbb{C}$ is a symmetric, uniformly positive definite, fourth-order stiffness tensor, M can be identified as the inverse of the compressibility of the bulk and α is the Biot coefficient. In this work, we assume isotropic materials, modelled as St. Venant Kirchhoff material, i.e., there exist constants $\mu > 0$ and $\lambda \geq 0$ satisfying

$$\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) = 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \boldsymbol{\nabla} \cdot \boldsymbol{u}\, \mathbf{I}.$$

From (2.8), we recover the classical relations

$$\theta = \tfrac{1}{M} p + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}, \tag{3.6}$$

$$\boldsymbol{\sigma} = \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) - \alpha p\, \mathbf{I}, \tag{3.7}$$

$$\boldsymbol{\sigma}_{\mathrm{eff}} = \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) = \boldsymbol{\sigma} + \alpha p\, \mathbf{I}. \tag{3.8}$$

A standard assumption in linear poro-elasticity is the quasi-static character of the mechanical problem. As consequence, mechanical deformation occur instantaneously and hence changes without any cost. Hence, allowing for viscous dissipation for the fluid, changes of displacements and volumetric fluxes come at costs based on the dissipation potentials

$$\begin{aligned}\mathcal{D}_{\text{mech}}(\dot{\boldsymbol{u}}) &= 0,\\ \mathcal{D}_{\text{fluid}}(\boldsymbol{q}) &= \tfrac{1}{2}\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{q}\right\rangle,\end{aligned}$$

where the conductivity $\boldsymbol{\kappa}$ is a symmetric, uniformly positive definite and uniformly bounded second-order tensor. It can be identified as the permeability, scaled by the inverse of the fluid viscosity.

Given the current state $(\boldsymbol{u}, \theta)$, its change is then described by (2.15)–(2.17):

$$\dot{\theta} = q_\theta - \boldsymbol{\nabla}\cdot\boldsymbol{q} \qquad \text{and} \tag{3.9}$$

$$\begin{aligned}(\dot{\boldsymbol{u}}, \boldsymbol{q}) = \underset{(\boldsymbol{v},\boldsymbol{z})\in\dot{\mathcal{V}}\times\mathcal{Z}}{\arg\min} \Big\{ &\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v})\rangle - \alpha\,\langle M(\theta - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}), \boldsymbol{\nabla}\cdot\boldsymbol{v}\rangle - \mathcal{P}_{\text{ext,mech}}(\boldsymbol{v})\\ &+ \tfrac{1}{2}\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{z}, \boldsymbol{z}\right\rangle - \langle M(\theta - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}), \boldsymbol{\nabla}\cdot\boldsymbol{z}\rangle - \mathcal{P}_{\text{ext,fluid}}(\boldsymbol{z})\Big\}.\end{aligned} \tag{3.10}$$

The system (3.9)–(3.10) can be reduced to a compact two-field formulation using ideas from Sec. 2.2. Recalling the definition of the accumulated flux $\boldsymbol{q}_{\int}$, cf. Eq. (2.19), living in

$$\mathcal{Z}_{\int}(t) = \left\{ \boldsymbol{z} \in H(\text{div};\Omega) \;\middle|\; \boldsymbol{z}\cdot\boldsymbol{n} = \int_0^t q_{\Gamma,\text{n}}\,dt \text{ on } \Gamma_{\boldsymbol{q}} \right\}, \quad t \in [0,T], \tag{3.11}$$

we introduce the generalized displacement $\boldsymbol{U} = (\boldsymbol{u}, \boldsymbol{q}_{\int})$ and its change $\dot{\boldsymbol{U}} = (\dot{\boldsymbol{u}}, \boldsymbol{q})$. Energies, external work rates and dissipation potentials can be naturally interpreted as functions of $\boldsymbol{U}$ and $\dot{\boldsymbol{U}}$, respectively. After all, the evolution of the generalized displacement $\boldsymbol{U}$ is governed by the generalized gradient flow

$$\dot{\boldsymbol{U}}(t) = \underset{\boldsymbol{V}\in\dot{\mathcal{V}}(t)\times\dot{\mathcal{Z}}_{\int}(t)}{\arg\min} \Big\{ \mathcal{D}(\boldsymbol{V}) + \langle \boldsymbol{\nabla}\mathcal{E}(t, \boldsymbol{U}(t)), \boldsymbol{V}\rangle - \mathcal{P}_{\text{ext}}(t, \boldsymbol{V}) \Big\}. \tag{3.12}$$

Formulations based on the generalized displacement are in the following referred to as the *primal formulation* of linear poro-elasticity.

In order to verify that (3.12) is indeed formally equivalent to the Biot equations, we derive the corresponding optimality conditions. Written in variational form, they read: Find $(\boldsymbol{u}, \boldsymbol{q}) \in \mathcal{V}\times\mathcal{Z}$ and θ with suitable regularity such that

$$\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v})\rangle - \alpha\,\langle M(\theta - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}), \boldsymbol{\nabla}\cdot\boldsymbol{v}\rangle = \mathcal{P}_{\text{ext,mech}}(\boldsymbol{v}) \qquad \forall \boldsymbol{v}\in\mathcal{V}_0, \tag{3.13}$$

$$\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{z}\right\rangle - \langle M(\theta - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}), \boldsymbol{\nabla}\cdot\boldsymbol{z}\rangle = \mathcal{P}_{\text{ext,fluid}}(\boldsymbol{z}), \qquad \forall \boldsymbol{z}\in\mathcal{Z}_0, \tag{3.14}$$

$$\dot{\theta} + \boldsymbol{\nabla}\cdot\boldsymbol{q} = q_\theta, \qquad \text{in } L^2(\Omega). \tag{3.15}$$

Identifying the fluid pressure from (3.6), we recover the three-field formulation of the classical quasi-static linear Biot equations.

3.2 Dual formulation of linear poro-elasticity

For the special case of quasi-static linear poro-elasticity, a natural *dual formulation* can be derived by applying the Legendre-Fenchel duality theory [50] to (3.12). The procedure is analogous to the discussion of primal and dual formulations of linear elastostatics in the context of convex analysis [57, 58]. We skip the derivation here and present directly the dual formulation. It naturally employs the dual generalized stress $\boldsymbol{\Sigma} = (\boldsymbol{\sigma}, p)$ as primary variable, pairing up the total mechanical stress and the fluid stress, i.e., fluid pressure. Suitable function spaces incorporating essential boundary conditions are given by

$$\dot{\mathcal{S}} := \left\{ \boldsymbol{\sigma} \in H(\mathrm{div};\Omega)^d \;\middle|\; \begin{array}{l} \boldsymbol{\sigma}\boldsymbol{n} = \dot{\boldsymbol{\sigma}}_{\Gamma,n} \text{ on } \Gamma_{\boldsymbol{\sigma}}, \\ \boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + \dot{\boldsymbol{f}}_{\mathrm{ext}} = \mathbf{0} \text{ in } L^2(\Omega), \\ \langle \boldsymbol{\sigma}, \boldsymbol{\gamma}\rangle = 0 \; \forall \boldsymbol{\gamma} \in \boldsymbol{Q}_{\mathrm{AS}} \end{array} \right\}, \tag{3.16}$$

$$\boldsymbol{Q}_{\mathrm{AS}} := \left\{ \boldsymbol{\gamma} \in L^2(\Omega)^{d\times d} \,|\, \boldsymbol{\gamma} \text{ skew-symmetric on } \Omega \right\} \tag{3.17}$$

$$\dot{\mathcal{Q}} := \left\{ q \in H^1(\Omega) \,|\, q = \dot{p}_\Gamma \text{ on } \Gamma_p \right\}, \tag{3.18}$$

$$\dot{\mathcal{H}}^\star := \dot{\mathcal{S}} \times \dot{\mathcal{Q}}. \tag{3.19}$$

We note, that the balance of linear momentum is incorporated intrinsically in $\dot{\mathcal{S}}$, which is characteristic for the dual formulation. Imposing only weak symmetry of stress tensors however is our choice, which is motivated by current advances in the robust discretization of the mixed formulation of elasticity and poro-elasticity, cf., e.g., [59, 33, 60, 61, 32]; imposing strong symmetry is also possible.

In between the primal and the dual formulation, the mathematical interpretation of dissipation and energy essentially swaps, similarly for essential and natural boundary conditions. Hence, utilizing (3.6)–(3.7) and Darcy's law, we define the dual energy, dissipation and external work rate by

$$\begin{aligned} \mathcal{E}^\star(\boldsymbol{\Sigma}) &= \mathcal{D}(\boldsymbol{\Sigma}) = \frac{1}{2}\left\langle \boldsymbol{\kappa}(\boldsymbol{\nabla}p - \boldsymbol{g}_{\mathrm{ext}}), \boldsymbol{\nabla}p - \boldsymbol{g}_{\mathrm{ext}} \right\rangle, \\ \mathcal{D}^\star(\dot{\boldsymbol{\Sigma}}) &= \mathcal{E}(\dot{\boldsymbol{\Sigma}}) = \tfrac{1}{2}\left\langle \mathbb{A}(\dot{\boldsymbol{\sigma}} + \alpha\dot{p}\,\mathbf{I}), \dot{\boldsymbol{\sigma}} + \alpha\dot{p}\,\mathbf{I}\right\rangle + \tfrac{1}{2M}\|\dot{p}\|^2, \\ \mathcal{P}^\star_{\mathrm{ext}}(\dot{\boldsymbol{\Sigma}}) &= \langle \dot{\boldsymbol{u}}_\Gamma, \dot{\boldsymbol{\sigma}}\boldsymbol{n}\rangle_{\Gamma_{\boldsymbol{\sigma}}} + \langle q_\theta, \dot{p}\rangle + \langle q_{\Gamma,\mathrm{n}}, \dot{p}\rangle_{\Gamma_{\boldsymbol{q}}}. \end{aligned}$$

Here, $\mathbb{A} = \mathbb{C}^{-1}$ denotes the compliance tensor; for homogeneous, isotropic materials, it satisfies for $\boldsymbol{\sigma} \in \mathbb{R}^{d\times d}$, with deviatoric and hydrostatic components $\boldsymbol{\sigma}^{\mathrm{d}} := \boldsymbol{\sigma} - \sigma^{\mathrm{h}}\,\mathbf{I}$ and $\sigma^{\mathrm{h}} := \frac{1}{d}\mathrm{tr}\,\boldsymbol{\sigma}$, respectively,

$$(\mathcal{A}\boldsymbol{\sigma}) : \boldsymbol{\sigma} = \frac{1}{2\mu}\left|\boldsymbol{\sigma}^{\mathrm{d}}\right|^2 + \frac{1}{K_{\mathrm{dr}}}|\sigma^{\mathrm{h}}|^2 \tag{3.20}$$

Finally, the evolution of the generalized stress $\boldsymbol{\Sigma}$ is prescribed by the generalized gradient flow

$$\dot{\boldsymbol{\Sigma}} = \underset{\boldsymbol{T}\in\dot{\mathcal{H}}^\star}{\arg\min} \left\{ \mathcal{D}^\star(\boldsymbol{T}) + \langle \boldsymbol{\nabla}\mathcal{E}^\star(\boldsymbol{\Sigma}), \boldsymbol{T}\rangle - \mathcal{P}^\star_{\mathrm{ext}}(\boldsymbol{T}) \right\}, \tag{3.21}$$

subject to compatible, initial data $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ at time $t = 0$. Evidently one major advantage of the dual formulation (3.21) compared to the primal formulation (3.12) is, it allows for incompressible solids and fluids.

The corresponding optimality conditions can be shown to be identical to the four field formulation of linear poro-elasticity [60], employing the total stress and the fluid pressure as primary variables, and mechanical displacement and rotation as Lagrange multipliers.

3.3 Well-posedness of linear poro-elasticity

In the following, we establish existence and uniqueness of a weak solution to the primal formulation of linear poro-elasticity and discuss its regularity. For this, we apply the abstract well-posedness result, Thm. 2.1.

Lemma 3.1 (Well-posedness and regularity for linear poro-elasticity)**.** *Let* $\mathcal{V}, \mathcal{V}_0, \mathcal{Z}_0$ *and* $\mathcal{Z}_\int$ *as defined in* (3.1)*–*(3.5) *and* (3.11) *with*

$$\begin{aligned}\boldsymbol{u}_\Gamma &\in C(0,T;H^{1/2}(\Gamma_{\boldsymbol{u}})^d) \cap H^1(0,T;H^{1/2}(\Gamma_{\boldsymbol{u}})^d),\\ q_{\Gamma,\mathrm{n}} &\in C(0,T;H^{-1/2}(\Gamma_{\boldsymbol{q}})^d).\end{aligned}$$

For the external loadings, and natural boundary and initial conditions assume

$$\begin{aligned}\boldsymbol{\sigma}_{\Gamma,\mathrm{n}} &\in C(0,T;H^{-1/2}(\Gamma_{\boldsymbol{\sigma}})^d) \cap H^1(0,T;H^{-1/2}(\Gamma_{\boldsymbol{\sigma}})^d),\\ p_\Gamma &\in C(0,T;H^{1/2}(\Gamma_{\boldsymbol{q}})^d) \cap H^1(0,T;H^{1/2}(\Gamma_{\boldsymbol{q}}),\\ \boldsymbol{f}_{\mathrm{ext}} &\in C(0,T;(\boldsymbol{H}^1(\Omega))^\star) \cap H^1(0,T;(\boldsymbol{H}^1(\Omega))^\star),\\ \boldsymbol{g}_{\mathrm{ext}} &\in C(0,T;H(\mathrm{div},\Omega)^\star) \cap H^1(0,T;H(\mathrm{div};\Omega)^\star),\\ q_\theta &\in L^2(0,T;\mathcal{V}_0^\star \cap \mathcal{Z}_0^\star),\\ \theta_0 &\in \mathcal{V}_0^\star \cap \mathcal{Z}_0^\star\\ \boldsymbol{u}_0 &\in \boldsymbol{H}^1(\Omega), \ \textit{such that}\ \boldsymbol{u}_{0|_{\Gamma_{\boldsymbol{u}}}} = \boldsymbol{u}_\Gamma(0),\end{aligned}$$

with the initial conditions satisfying the compatibility condition

$$\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_0), \boldsymbol{\varepsilon}(v)\rangle - \langle M\left(\theta_0 - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}_0\right), \boldsymbol{\nabla}\cdot\boldsymbol{v}\rangle = \langle \boldsymbol{f}_{\mathrm{ext}}(0), \boldsymbol{v}\rangle\,, \quad \forall \boldsymbol{v} \in \mathcal{V}_0.$$

Then there exists a unique solution $\boldsymbol{U} = (\boldsymbol{u}, \boldsymbol{q}_\int)$ *of* (3.12) *and equivalently* (3.9)*–*(3.10)*, satisfying*

$$\boldsymbol{u} \in L^\infty(0,T;H^1(\Omega)), \tag{3.22}$$

$$\boldsymbol{q}_\int \in H^1(0,T;L^2(\Omega)) \cap L^\infty(0,T;H(\mathrm{div};\Omega)), \tag{3.23}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{q}_\int \in H^1(0,T;H^{-1}(\Omega)) \cap L^\infty(0,T;L^2(\Omega)). \tag{3.24}$$

For the volumetric flux $\boldsymbol{q}$*, fluid content* θ*, fluid pressure* p *and stress* $\boldsymbol{\sigma}$*, associated with the states by* (2.19)*,* (2.21)*,* (3.6) *and* (3.7)*, respectively, it holds*

$$\boldsymbol{q} \in L^2(0,T;L^2(\Omega)),\ \boldsymbol{\nabla}\cdot\boldsymbol{q} \in L^2(0,T;H^{-1}(\Omega)), \tag{3.25}$$

$$\theta \in H^1(0,T;H^{-1}(\Omega)) \cap L^\infty(0,T;L^2(\Omega)), \tag{3.26}$$

$$p \in L^\infty(0,T;L^2(\Omega)), \tag{3.27}$$

$$\boldsymbol{\sigma} \in L^\infty(0,T;L^2(\Omega)). \tag{3.28}$$

Proof. The well-posedness result is a direct consequence of Thm. 2.1, applied to poro-elasticity written as doubly non-linear evolution equation, cf. Sec. 2.2. For this, we reformulate (3.12) having two objectives in mind: (i) time-dependent contributions due to essential boundary conditions and external sources are required to impact only the external work rates; (ii) the final structure is required to match the abstract setting of Thm. 2.1.

Due to objective (i), the problem has to be formulated for homogeneous contributions of the generalized displacement. It will be denoted by

$$\left(\boldsymbol{u}_{\mathrm{hom}}, \boldsymbol{q}_{\int,\mathrm{hom}}\right) := \left(\boldsymbol{u}, \boldsymbol{q}_\int\right) - \left(\tilde{\boldsymbol{u}}_\Gamma, \tilde{\boldsymbol{q}}_{\int,\Gamma}\right) \in \mathcal{V}_0 \times \mathcal{Z}_0,$$

where

$$\begin{aligned}\tilde{\boldsymbol{u}}_\Gamma &\in C(0,T;\boldsymbol{H}^1(\Omega)) \cap H^1(0,T;\boldsymbol{H}^1(\Omega)),\\ \tilde{\boldsymbol{q}}_{\int,\Gamma} &\in C^1(0,T;H(\mathrm{div};\Omega)).\end{aligned}$$

are extensions of the essential boundary conditions onto Ω, such that $\tilde{\boldsymbol{u}}(0) = \boldsymbol{u}_0$, $\tilde{\boldsymbol{u}}_\Gamma|_{\Gamma_{\boldsymbol{u}}}(t) = \boldsymbol{u}_\Gamma(t)$ and $\tilde{\boldsymbol{q}}_\Gamma \cdot \boldsymbol{n}|_{\Gamma_{\boldsymbol{q}}}(t) = \int_0^t q_{\Gamma,\mathrm{n}}\, dt$ for a.e. $t \in (0,T)$.

Using the notation of Thm. 2.1, we define for $(\boldsymbol{v}, \boldsymbol{z}) \in \mathcal{V}_0 \times \mathcal{Z}_0$

$$\begin{aligned} \Psi(\boldsymbol{z}) &:= \mathcal{D}_{\text{fluid}}(\boldsymbol{z}) & \mathcal{E}_1(\boldsymbol{v}) &:= \mathcal{E}_{\text{eff}}(\boldsymbol{v}), \\ \mathcal{E}_2(m) &:= \tfrac{M}{2}\|m\|^2, & \Lambda(\boldsymbol{v}, \boldsymbol{z}) &:= \theta_0 - \boldsymbol{\nabla}\cdot\boldsymbol{z} - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{v}, \\ && \mathcal{E}(\boldsymbol{v}, \boldsymbol{z}) &:= \mathcal{E}_1(\boldsymbol{v}) + \mathcal{E}_2(\Lambda(\boldsymbol{v}, \boldsymbol{z})), \end{aligned}$$

and in order to fulfill objective (i), we set

$$\begin{aligned} \mathcal{P}_1(t, \boldsymbol{v}) &:= \mathcal{P}_{\text{ext,mech}}(t, \boldsymbol{v}) - \langle \mathbb{C}\boldsymbol{\varepsilon}(\tilde{\boldsymbol{u}}_\Gamma(t)), \boldsymbol{\varepsilon}(\boldsymbol{v})\rangle \\ &\quad + M \left\langle Q_\theta(t) - \boldsymbol{\nabla}\cdot\tilde{\boldsymbol{q}}_{\int,\Gamma}(t) - \alpha\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{u}}_\Gamma(t), \alpha\boldsymbol{\nabla}\cdot\boldsymbol{v} \right\rangle, \\ \mathcal{P}_2(t, \boldsymbol{z}) &:= \mathcal{P}_{\text{ext,fluid}}(t, \boldsymbol{z}) - \left\langle \boldsymbol{\kappa}^{-1}\dot{\tilde{\boldsymbol{q}}}_{\int,\Gamma}(t), \boldsymbol{z} \right\rangle \\ &\quad + M \left\langle Q_\theta(t) - \boldsymbol{\nabla}\cdot\tilde{\boldsymbol{q}}_{\int,\Gamma}(t) - \alpha\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{u}}_\Gamma(t), \boldsymbol{\nabla}\cdot\boldsymbol{z} \right\rangle. \end{aligned}$$

One can simply verify that (3.12) is equivalent to

$$\begin{aligned} (\dot{\boldsymbol{u}}_{\text{hom}}, \dot{\boldsymbol{q}}_{\int,\text{hom}}) = \underset{(\boldsymbol{v},\boldsymbol{z})\in\mathcal{V}_0\times\mathcal{Z}_0}{\arg\min} \Big\{ &\Psi(\boldsymbol{z}) + \left\langle \boldsymbol{\nabla}\mathcal{E}\left(\boldsymbol{u}_{\text{hom}}, \boldsymbol{q}_{\int,\text{hom}}\right), (\boldsymbol{v}, \boldsymbol{z}) \right\rangle \\ &- \mathcal{P}_1(t, \boldsymbol{v}) - \mathcal{P}_2(t, \boldsymbol{z}) \Big\} \end{aligned} \tag{3.29}$$

together with zero initial conditions. Furthermore, it is simple to verify (P1)–(P6); we just note that $\mathcal{V}_1 = \mathcal{V}_0$, $\mathcal{V}_2 = \mathcal{Z}_0$ and $\mathcal{B}_2 = \boldsymbol{L}^2(\Omega)$ using the notation of Thm. 2.1. Consequently, we obtain existence and uniqueness of a solution to (3.29), and consequently of (3.12), satisfying (3.22)–(3.23). Since

$$\int_0^T \frac{\left\langle \boldsymbol{\nabla}\cdot\dot{\boldsymbol{q}}_{\int,\text{hom}}, \phi \right\rangle}{\|\boldsymbol{\nabla}\phi\|}\, dt \leq \|\dot{\boldsymbol{q}}_{\int,\text{hom}}\|^2_{L^2(0,T;L^2(\Omega))},$$

and the regularity of $\tilde{\boldsymbol{q}}_{\int,\Gamma}$, it follows (3.24). Finally, (3.25)–(3.28) follow directly using (2.19), (2.21) and (3.6)–(3.7). □

4 Linear poro-visco-elasticity as generalized gradient flow

Biot's consolidation model considers solely primary consolidation, which results in a characteristic, quasi-static, mechanical response of the poro-elastic system. Modelling-wise, this originates from neglecting viscous dissipation due to mechanical deformation. However, in physical, deformable porous media viscous dissipation always occurs, and it leads to partially non-instantaneous deformation, also called *secondary consolidation*. The theory of poro-visco-elasticity incorporates such visco-elastic effects.

Classical models for visco-elasticity consider a separation of the total strain into an elastic and a visco-elastic strain [22]; the elastic contribution is instantaneously recovered during an unloading process, whereas the visco-elastic contribution is not. In the extreme case, the elastic behavior can be neglected and the total strain can be assumed to be identical to the visco-elastic strain [27, 29]; the corresponding model is also referred to as linear, quasi-static poro-elasticity with secondary consolidation. Here, we treat the general case, including the elastic and visco-elastic strains.

Linear poro-visco-elasticity can be naturally formulated as generalized gradient flow by suitably enhancing the primal model for linear poro-elasticity. As state we choose the generalized displacement $(\boldsymbol{u}, \boldsymbol{q}_\int, \boldsymbol{\varepsilon}_\mathrm{v})$, incorporating now also the visco-elastic strain $\boldsymbol{\varepsilon}_\mathrm{v}$, living for a.e. $t \in (0,T)$ in

$$\mathcal{T} := \{\boldsymbol{t} \in L^2(\Omega)^{d\times d} \,|\, \boldsymbol{t} \text{ symmetric}\}.$$

We distinguish between the standard elastic strain energy and a stored, visco-elastic energy. Additionally, we allow for different impacts of the elastic and visco-elastic strains on the energy corresponding to the fluid. The associated energy functional is consistently defined as by [22]

$$\begin{aligned}\mathcal{E}_\mathrm{v}(\boldsymbol{u}, \boldsymbol{q}_\int, \boldsymbol{\varepsilon}_\mathrm{v}) =& \tfrac{1}{2}\langle \mathbb{C}(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{\varepsilon}_\mathrm{v}), \boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{\varepsilon}_\mathrm{v}\rangle + \tfrac{1}{2}\langle \mathbb{C}_\mathrm{v}\boldsymbol{\varepsilon}_\mathrm{v}, \boldsymbol{\varepsilon}_\mathrm{v}\rangle \\ &+ \tfrac{M}{2}\left\|\theta_0 + Q_\theta - \boldsymbol{\nabla}\cdot\boldsymbol{q}_\int - \alpha_\mathrm{v}\,\mathrm{tr}\,\boldsymbol{\varepsilon}_\mathrm{v} - \alpha(\boldsymbol{\nabla}\cdot\boldsymbol{u} - \mathrm{tr}\,\boldsymbol{\varepsilon}_\mathrm{v})\right\|^2,\end{aligned}$$

where $\mathbb{C}_\mathrm{v}$ is a symmetric, uniformly positive definite, fourth-order tensor.

Taking also into account the dissipation of the stored, visco-elastic energy, the dissipation potential is defined by

$$\mathcal{D}_\mathrm{v}(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_\int, \dot{\boldsymbol{\varepsilon}}_\mathrm{v}) = \tfrac{1}{2}\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}_\int, \boldsymbol{q}_\int\rangle + \tfrac{1}{2}\langle \mathbb{C}_\mathrm{v}'\dot{\boldsymbol{\varepsilon}}_\mathrm{v}, \dot{\boldsymbol{\varepsilon}}_\mathrm{v}\rangle.$$

For many materials a visco-elastic effect is only encountered, e.g., for the volumetric part [22, 29]. Consequently, the symmetric, positive semi-definite, fourth-order tensor $\mathbb{C}_\mathrm{v}'$ may be singular.

Again, restricting to isotropic materials, the fourth-order tensors $\mathbb{C}_\mathrm{v}$, $\mathbb{C}_\mathrm{v}'$ can be associated with Lamé parameters $\mu_\mathrm{v} > 0, \lambda_\mathrm{v} \geq 0, \mu_\mathrm{v}' \geq 0, \lambda_\mathrm{v}' > 0$, and corresponding bulk moduli $K_{\mathrm{dr,v}} = \frac{2\mu_\mathrm{v}}{d} + \lambda_\mathrm{v}$ and $K_{\mathrm{dr,v}}' = \frac{2\mu_\mathrm{v}'}{d} + \lambda_\mathrm{v}'$, via

$$\begin{aligned}\mathbb{C}_\mathrm{v}\boldsymbol{\varepsilon}_\mathrm{v} &= 2\mu_\mathrm{v}\boldsymbol{\varepsilon}_\mathrm{v} + \lambda_\mathrm{v}\mathrm{tr}\,\boldsymbol{\varepsilon}_\mathrm{v}, \\ \mathbb{C}_\mathrm{v}'\dot{\boldsymbol{\varepsilon}}_\mathrm{v} &= 2\mu_\mathrm{v}'\dot{\boldsymbol{\varepsilon}}_\mathrm{v} + \lambda_\mathrm{v}'\mathrm{tr}\,\dot{\boldsymbol{\varepsilon}}_\mathrm{v}.\end{aligned}$$

Since $\boldsymbol{\varepsilon}_\mathrm{v}$ is interpreted as internal variable, the external work rates can be chosen as in the context of linear poro-elasticity

$$\mathcal{P}_\mathrm{ext}(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_\int, \dot{\boldsymbol{\varepsilon}}_\mathrm{v}) = \mathcal{P}_\mathrm{ext,mech}(\dot{\boldsymbol{u}}) + \mathcal{P}_\mathrm{ext,fluid}(\dot{\boldsymbol{q}}_\int).$$

Finally, within the framework introduced in Sec. 2.1, the resulting evolution equation reads for current state $(\boldsymbol{u}, \boldsymbol{q}_\int, \boldsymbol{\varepsilon}_\mathrm{v})$

$$\begin{aligned}(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_\int, \dot{\boldsymbol{\varepsilon}}_\mathrm{v}) = \underset{(\boldsymbol{v},\boldsymbol{z},\boldsymbol{t})\in\dot{\mathcal{V}}\times\dot{\mathcal{Z}}_\int\times\mathcal{T}}{\arg\min} & \Big\{\mathcal{D}_\mathrm{v}(\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{t}) + \Big\langle \boldsymbol{\nabla}\mathcal{E}_\mathrm{v}(\boldsymbol{u}, \boldsymbol{q}_\int, \boldsymbol{\varepsilon}_\mathrm{v}), (\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{t})\Big\rangle \\ & -\mathcal{P}_\mathrm{ext}(\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{t})\Big\}.\end{aligned} \tag{4.1}$$

The corresponding optimality conditions yield the model for linear poro-visco-elasticity as discussed by [22]

$$\langle \mathbb{C}(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{\varepsilon}_\mathrm{v}), \boldsymbol{\varepsilon}(\boldsymbol{v})\rangle - \alpha\,\langle p, \boldsymbol{\nabla}\cdot\boldsymbol{v}\rangle = \mathcal{P}_\mathrm{ext,mech}(\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathcal{V}_0, \tag{4.2}$$

$$\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{z}\right\rangle - \langle p, \boldsymbol{\nabla}\cdot\boldsymbol{z}\rangle = \mathcal{P}_\mathrm{ext,fluid}(\boldsymbol{z}), \qquad \forall \boldsymbol{z} \in \mathcal{Z}_0, \tag{4.3}$$

$$\left\langle \mathbb{C}_\mathrm{v}'\dot{\boldsymbol{\varepsilon}}_\mathrm{v} + \mathbb{C}_\mathrm{v}\boldsymbol{\varepsilon}_\mathrm{v} - \alpha_\mathrm{v} p\,\mathbf{I}, \boldsymbol{t}\right\rangle = \langle \boldsymbol{\sigma}, \boldsymbol{t}\rangle \qquad \forall \boldsymbol{t} \in \mathcal{T}, \tag{4.4}$$

$$\theta - \theta_0 + \boldsymbol{\nabla}\cdot\boldsymbol{q}_\int = Q_\theta, \qquad \text{in } L^2(\Omega), \tag{4.5}$$

where we explicitly introduced the fluid pressure and total stress by

$$p = \partial_m\mathcal{E}_\mathrm{v} = M\left(\theta_0 + Q_\theta - \boldsymbol{\nabla}\cdot\boldsymbol{q}_\int - \alpha_\mathrm{v}\,\mathrm{tr}\,\boldsymbol{\varepsilon}_\mathrm{v} - \alpha(\boldsymbol{\nabla}\cdot\boldsymbol{u} - \mathrm{tr}\,\boldsymbol{\varepsilon}_\mathrm{v})\right), \tag{4.6}$$

$$\boldsymbol{\sigma} = \partial_{\boldsymbol{\nabla}\boldsymbol{u}}\mathcal{E}_\mathrm{v} = \mathbb{C}(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{\varepsilon}_\mathrm{v}) - \alpha p\,\mathbf{I}. \tag{4.7}$$

Well-posedness can be analyzed analogously to Lemma 3.1.

Lemma 4.1 (Well-posedness of linear poro-visco-elasticity)**.** *Let $\boldsymbol{\varepsilon}_{\mathrm{v},0} \in L^2(\Omega)^{d\times d}$, satisfying the compatibility condition for the deviatoric components*

$$2\mu_{\mathrm{v}} \left\langle \boldsymbol{\varepsilon}_{\mathrm{v},0}^{\mathrm{d}}, \boldsymbol{t} \right\rangle = 2\mu \left\langle \boldsymbol{\varepsilon}(\boldsymbol{u}_0)^{\mathrm{d}} - \boldsymbol{\varepsilon}_{\mathrm{v},0}^{\mathrm{d}}, \boldsymbol{t} \right\rangle, \quad \forall \boldsymbol{t} \in \mathcal{T},$$

in case $\mu_{\mathrm{v}}' = 0$, where the deviatoric components are defined according to (4.11) *and* (4.12)*. Then under the same regularity assumptions as in Lemma 3.1 there exists a unique solution $(\boldsymbol{u}, \boldsymbol{q}_{\int}, \boldsymbol{\varepsilon}_{\mathrm{v}})$ of* (4.2)*–*(4.5)*, satisfying* (3.22)*–*(3.24) *as for linear poro-elasticity. Additionally, for $\mu_{\mathrm{v}}' > 0$ it holds*

$$\boldsymbol{\varepsilon}_{\mathrm{v}} \in H^1(0,T;L^2(\Omega)), \tag{4.8}$$

whereas for $\mu_{\mathrm{v}}' = 0$, $\lambda_{\mathrm{v}}' > 0$ it holds

$$\boldsymbol{\varepsilon}_{\mathrm{v}} \in L^\infty(0,T;L^2(\Omega)), \tag{4.9}$$

$$\operatorname{tr} \boldsymbol{\varepsilon}_{\mathrm{v}} \in H^1(0,T;L^2(\Omega)). \tag{4.10}$$

For the flux $\boldsymbol{q}$, mass m, pressure p and stress $\boldsymbol{\sigma}$, associated with the states by (2.19)*,* (2.21)*,* (4.6) *and* (4.7)*, respectively, same regularity holds as for linear poro-elasticity, cf.* (3.25)*–*(3.28)*.*

Proof. Non-homogeneous boundary conditions $\boldsymbol{u}_\Gamma$, $q_{\Gamma,n}$ and a non-zero, external source term q_θ can be discussed as in the proof of Lemma 3.1. In the following, the focus is exclusively on the visco-elastic contribution, and only the case of zero boundary data and source terms is discussed.

The case $\mu_{\mathrm{v}}' > 0$ follows analogously to the proof of Lemma 3.1; employ Thm. 2.1 with the partition of state variables $\{\boldsymbol{u}\}$ and $\{\boldsymbol{q}_{\int}, \boldsymbol{\varepsilon}_{\mathrm{v}}\}$.

The second case, $\mu_{\mathrm{v}}' = 0$, $\lambda_{\mathrm{v}}' > 0$, requires a problem reformulation before applying Thm. 2.1. As $\mathbb{C}_{\mathrm{v}}'$ is singular, $\mathcal{D}_{\mathrm{v}}(\cdot)$ is not coercive on $\boldsymbol{L}^2(\Omega) \times \mathcal{T}$. This can be fixed by decomposing strains. We introduce an orthogonal decomposition of visco-elastic strains into their hydrostatic and deviatoric parts. Let

$$\begin{aligned} \varepsilon_{\mathrm{v}}^{\mathrm{h}} &:= \tfrac{1}{d} \operatorname{tr} \boldsymbol{\varepsilon}_{\mathrm{v}} \in \mathcal{T}^{\mathrm{h}}, \qquad \boldsymbol{\varepsilon}_{\mathrm{v}}^{\mathrm{d}} := \boldsymbol{\varepsilon}_{\mathrm{v}} - \varepsilon_{\mathrm{v}}^{\mathrm{h}} \mathbf{I} \in \mathcal{T}^{\mathrm{d}}, \\ \mathcal{T}^{\mathrm{h}} &:= \left\{ t\mathbf{I} \;\middle|\; t \in L^2(\Omega) \right\}, \\ \mathcal{T}^{\mathrm{d}} &:= \left\{ \boldsymbol{t} \in L^2(\Omega)^{d\times d} \;\middle|\; \operatorname{tr} \boldsymbol{t} = 0 \right\}. \end{aligned} \tag{4.11}$$

such that $\langle \mathbb{C}_{\mathrm{v}} \boldsymbol{\varepsilon}_{\mathrm{v}}, \boldsymbol{\varepsilon}_{\mathrm{v}} \rangle = 2\mu_{\mathrm{v}} \|\boldsymbol{\varepsilon}_{\mathrm{v}}^{\mathrm{d}}\|^2 + d^2 \left(\frac{2\mu_{\mathrm{v}}}{d} + \lambda_{\mathrm{v}} \right) \|\varepsilon_{\mathrm{v}}^{\mathrm{h}}\|^2$. Similarly, we introduce

$$\varepsilon^{\mathrm{h}}(\boldsymbol{u}) := \tfrac{1}{d} \operatorname{tr} \boldsymbol{\varepsilon}(\boldsymbol{u}), \qquad \boldsymbol{\varepsilon}^{\mathrm{d}}(\boldsymbol{u}) := \boldsymbol{\varepsilon}(\boldsymbol{u}) - \varepsilon^{\mathrm{h}}(\boldsymbol{u})\, \mathbf{I}. \tag{4.12}$$

We re-interpret the energy and dissipation potential as functions of $\boldsymbol{u}, \boldsymbol{q}_{\int}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{\mathrm{d}}, \varepsilon_{\mathrm{v}}^{\mathrm{h}}$ and their temporal changes, respectively,

$$\begin{aligned} \mathcal{E}_{\mathrm{v}}(\boldsymbol{u}, \boldsymbol{q}_{\int}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{\mathrm{d}}, \varepsilon_{\mathrm{v}}^{\mathrm{h}}) &= \tfrac{1}{2} \left(2\mu \|\boldsymbol{\varepsilon}^{\mathrm{d}}(\boldsymbol{u}) - \boldsymbol{\varepsilon}_{\mathrm{v}}^{\mathrm{d}}\|^2 + 2\mu_{\mathrm{v}} \|\boldsymbol{\varepsilon}_{\mathrm{v}}^{\mathrm{d}}\|^2 \right) \\ &\quad + \tfrac{d^2}{2} \left(K_{\mathrm{dr}} \|\varepsilon^{\mathrm{h}}(\boldsymbol{u}) - \varepsilon_{\mathrm{v}}^{\mathrm{h}}\|^2 + K_{\mathrm{dr,v}} \|\varepsilon_{\mathrm{v}}^{\mathrm{h}}\|^2 \right) \\ &\quad + \tfrac{M}{2} \left\| \theta_0 - \boldsymbol{\nabla} \cdot \boldsymbol{q}_{\int} - d\alpha_{\mathrm{v}} \varepsilon_{\mathrm{v}}^{\mathrm{h}} - d\alpha \left(\varepsilon^{\mathrm{h}}(\boldsymbol{u}) - \varepsilon_{\mathrm{v}}^{\mathrm{h}} \right) \right\|^2, \\ \mathcal{D}_{\mathrm{v}}(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_{\int}, \dot{\boldsymbol{\varepsilon}}_{\mathrm{v}}^{\mathrm{d}}, \dot{\varepsilon}_{\mathrm{v}}^{\mathrm{h}}) &= \tfrac{1}{2} \left\langle \boldsymbol{\kappa}^{-1} \dot{\boldsymbol{q}}_{\int}, \dot{\boldsymbol{q}}_{\int} \right\rangle + \tfrac{d^2 \lambda_{\mathrm{v}}'}{2} \|\dot{\varepsilon}_{\mathrm{v}}^{\mathrm{h}}\|^2. \end{aligned}$$

where $K_{\mathrm{dr}} = \frac{2\mu}{d} + \lambda$ and $K_{\mathrm{dr,v}} = \frac{2\mu_{\mathrm{v}}}{d} + \lambda_{\mathrm{v}}$. The external work rate $\mathcal{P}_{\mathrm{ext}}$ is independent of $\dot{\boldsymbol{\varepsilon}}_{\mathrm{v}}$, and hence, remains unaltered after re-interpretation.

The dissipation potential $\mathcal{D}_\mathrm{v}$ defines a norm for $(\dot{\boldsymbol{q}}_\int, \dot{\varepsilon}^\mathrm{h}_\mathrm{v})$ and hence is coercive on $\dot{\mathcal{Z}}_\int \times \mathcal{T}^\mathrm{h}$. In order to apply Thm. 2.1, we decompose $\mathcal{E}_\mathrm{v}$ into a strictly convex part, depending only on the complementary part, $(\boldsymbol{u}, \boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v})$, and a convex remainder. Let

$$\begin{aligned}
\Psi_\mathrm{v}(\dot{\boldsymbol{q}}_\int, \varepsilon^\mathrm{h}_\mathrm{v}) &:= \tfrac{1}{2}\left\langle \boldsymbol{\kappa}^{-1}\dot{\boldsymbol{q}}_\int, \dot{\boldsymbol{q}}_\int \right\rangle + \tfrac{d^2\lambda'_\mathrm{v}}{2}\|\dot{\varepsilon}^\mathrm{h}_\mathrm{v}\|^2, \\
\mathcal{E}_{\mathrm{v},1}(\boldsymbol{u}, \boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v}) &:= \tfrac{1}{2}\left(2\mu\|\boldsymbol{\varepsilon}^\mathrm{d}(\boldsymbol{u}) - \boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v}\|^2 + 2\mu_\mathrm{v}\|\boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v}\|^2\right), \\
\mathcal{E}_{\mathrm{v},2}(x_1, x_2, x_3) &:= \tfrac{d^2}{2}\left(K_\mathrm{dr}\|x_1\|^2 + K_\mathrm{dr,v}\|x_2\|^2\right) + \tfrac{M}{2}\|x_3\|^2, \\
\Lambda_\mathrm{v}(\boldsymbol{u}, \boldsymbol{q}_\int, \boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v}, \varepsilon^\mathrm{h}_\mathrm{v}) &:= \left[\varepsilon^\mathrm{h}(\boldsymbol{u}) - \varepsilon^\mathrm{h}_\mathrm{v},\, \varepsilon^\mathrm{h}_\mathrm{v},\, \theta_0 - \boldsymbol{\nabla}\cdot\boldsymbol{q}_\int - d\alpha_\mathrm{v}\varepsilon^\mathrm{h}_\mathrm{v} - d\alpha\left(\varepsilon^\mathrm{h}(\boldsymbol{u}) - \varepsilon^\mathrm{h}_\mathrm{v}\right)\right],
\end{aligned}$$

satisfying

$$\mathcal{E}_\mathrm{v}\left(\boldsymbol{u}, \boldsymbol{q}_\int, \boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v}, \varepsilon^\mathrm{h}_\mathrm{v}\right) = \mathcal{E}_{\mathrm{v},1}\left(\boldsymbol{u}, \boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v}\right) + \mathcal{E}_{\mathrm{v},2}\left(\Lambda_\mathrm{v}\left(\boldsymbol{u}, \boldsymbol{q}_\int, \boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v}, \varepsilon^\mathrm{h}_\mathrm{v}\right)\right).$$

Finally, (4.1) takes the form

$$\begin{aligned}
(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_\int, \dot{\boldsymbol{\varepsilon}}^\mathrm{d}_\mathrm{v}, \dot{\varepsilon}^\mathrm{h}_\mathrm{v}) \;=\; &\operatorname*{arg\,min}_{(\boldsymbol{v},\boldsymbol{z},\boldsymbol{t}^\mathrm{d},t^\mathrm{h})\in\mathcal{V}_0\times\mathcal{Z}_0\times\mathcal{T}^\mathrm{d}\times\mathcal{T}^\mathrm{h}} \Big\{\Psi_\mathrm{v}(\boldsymbol{v}, t^\mathrm{h}) \\
&+ \left\langle \boldsymbol{\nabla}\mathcal{E}_\mathrm{v}\left(\boldsymbol{u}, \boldsymbol{q}_\int, \boldsymbol{\varepsilon}^\mathrm{d}_\mathrm{v}, \varepsilon^\mathrm{h}_\mathrm{v}\right), (\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{t}^\mathrm{d}, t^\mathrm{h})\right\rangle - \mathcal{P}_\mathrm{ext,mech}(\boldsymbol{u}) - \mathcal{P}_\mathrm{ext,fluid}(\boldsymbol{z})\Big\}.
\end{aligned}$$

complying with the abstract structure in Thm. 2.1. Additionally, the properties (P1)–(P6) are satisfied. Hence, by Thm. 2.1, we obtain the existence of a unique solution to (4.1). Regularity follows along the lines of Lemma 3.1. □

Remark 4.1 (Purely visco-elastic deformation). *As mentioned above, the mechanical displacement may be assumed to be purely visco-elastic, i.e., $\boldsymbol{\varepsilon}_\mathrm{v} = \boldsymbol{\varepsilon}(\boldsymbol{u})$. This corresponds to $\mu, \lambda \to \infty$, while $\mathbb{C}_\mathrm{v}$ and $\mathbb{C}'_\mathrm{v}$ remain finite, i.e., $\mathbb{C}$ acts as penalty parameter. In the limit, following from Lemma 4.1, $\boldsymbol{\varepsilon}(\boldsymbol{u})$ inherits the regularity of $\boldsymbol{\varepsilon}_\mathrm{v}$. This yields an alternative approach to [27, 29] for the analysis of quasi-static, linear poro-elasticity with secondary consolidation.*

5 Non-linear poro-elasticity under infinitesimal strains as generalized gradient flow

In many applications linear constitutive laws are not sufficient in order to describe the physical behavior of a fluid-saturated, deformable porous medium – even when restricted to the infinitesimal strain regime [62, 63]; similar to soil mechanics [64] and solid mechanics [58]. The gradient flow modeling framework introduced in Sec. 2.1 allows for involving a variety of non-linear relationships. In the following, we consider a non-linear stress-strain relationship together with a non-linear fluid compressibility, assuming infinitesimal strains. Based on the gradient flow structure of the resulting models, we analyze their well-posedness along the lines of Sec. 3. This setting has been also studied numerically by [30].

We generalize the primal formulation of linear poro-elasticity, cf. Sec. 3.1: We choose the same state variables, $(\boldsymbol{u}, \boldsymbol{q}_\int)$, living in the same function spaces as before. In the spirit of hyperelasticity [58], we consider energies

$$\mathcal{E}_\mathrm{nl}(\boldsymbol{u}, \boldsymbol{q}_\int) := \mathcal{E}_\mathrm{nl,eff}(\boldsymbol{u}) + \mathcal{E}_\mathrm{nl,fluid}(\boldsymbol{u}, \boldsymbol{q}_\int), \tag{5.1}$$

$$\mathcal{E}_\mathrm{nl,eff}(\boldsymbol{u}) := \int_\Omega W(\boldsymbol{\varepsilon}(\boldsymbol{u}))\,dx, \tag{5.2}$$

$$\mathcal{E}_\mathrm{nl,fluid}(\boldsymbol{u}, \boldsymbol{q}_\int) := \int_\Omega \int_0^{\theta_0 + Q_\theta - \boldsymbol{\nabla}\cdot\boldsymbol{q}_\int - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}} b^{-1}(s)\,ds\,dx, \tag{5.3}$$

for some convex strain energy density W and an invertible function b. This choice results in the generalized, (implicit) definition of the fluid pressure and mechanical stress, cf. (3.6)–(3.7), via

$$m = b(p) + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}, \tag{5.4}$$

$$\boldsymbol{\sigma} = \frac{\partial W(\boldsymbol{\varepsilon}(\boldsymbol{u}))}{\partial \boldsymbol{\varepsilon}(\boldsymbol{u})} - \alpha p \, \mathbf{I}, \tag{5.5}$$

which follows directly from (2.8). Examples for strain energy densities W and the corresponding effective stress $\boldsymbol{\sigma}_{\text{eff}}$, cf. (2.13), are given in Tab. 1; in principle also non-convex potentials [58] could be employed, but are not considered here. For b we allow arbitrary invertible, increasing, Lipschitz continuous functions with $b(0) = 0$, generalizing the linear compressibility law $b(p) = \frac{1}{M}p$, employed in linear poro-elasticity. We refer to [30] for possible choices.

Linear elasticity (cf. Sec. 3)	$W(\boldsymbol{\varepsilon}(\boldsymbol{u})) = \frac{1}{2}\left(2\mu\lvert\boldsymbol{\varepsilon}(\boldsymbol{u})\rvert^2 + \lambda(\boldsymbol{\nabla}\cdot\boldsymbol{u})^2\right)$ $\boldsymbol{\sigma}_{\text{eff}} = 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda(\boldsymbol{\nabla}\cdot\boldsymbol{u})\,\mathbf{I}$
Non-linear compressibility (cf. [63, 30])	$W(\boldsymbol{\varepsilon}(\boldsymbol{u})) = \frac{1}{2}\left(2\mu\lvert\boldsymbol{\varepsilon}(\boldsymbol{u})\rvert^2 + \int_0^{\boldsymbol{\nabla}\cdot\boldsymbol{u}} l(s)\,ds\right)$, l increasing, $l(0) = 0$ $\boldsymbol{\sigma}_{\text{eff}} = 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + l(\boldsymbol{\nabla}\cdot\boldsymbol{u})\,\mathbf{I}$
Non-linear shear modulus (cf. [64, 65])	$W(\boldsymbol{\varepsilon}(\boldsymbol{u})) = \int_0^{\lvert\boldsymbol{\varepsilon}(\boldsymbol{u})\rvert} sf(s)\,ds + \frac{\lambda}{2}(\boldsymbol{\nabla}\cdot\boldsymbol{u})^2$, f unif. pos. and non-decr. $\boldsymbol{\sigma}_{\text{eff}} = f(\lvert\boldsymbol{\varepsilon}(\boldsymbol{u})\rvert)\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda(\boldsymbol{\nabla}\cdot\boldsymbol{u})\,\mathbf{I}$
Simple visco-elasto-plasticity (cf. [58])	$W(\boldsymbol{\varepsilon}(\boldsymbol{u})) = \int_0^{\lvert\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\rvert} sf(s)\,ds + K_{\text{dr}}(\boldsymbol{\nabla}\cdot\boldsymbol{u})^2,$ $f(s) = \begin{cases} 2\mu\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\,, & 2\mu\lvert\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\rvert < K \\ 2\mu + \frac{K}{\lvert\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\rvert}\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\,, & \text{else} \end{cases}$ $\boldsymbol{\sigma}_{\text{eff}} = \boldsymbol{\sigma}_{\text{eff}}^{\text{d}} + K_{\text{dr}}(\boldsymbol{\nabla}\cdot\boldsymbol{u})\,\mathbf{I}$ $\boldsymbol{\sigma}_{\text{eff}}^{\text{d}} = \begin{cases} 2\mu\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\,, & \lvert\boldsymbol{\sigma}_{\text{eff}}^{\text{d}}\rvert = 2\mu\lvert\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\rvert < K \\ \left(2\mu + \frac{K}{\lvert\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\rvert}\right)\boldsymbol{\varepsilon}^{\text{d}}(\boldsymbol{u})\,, & \text{else, i.e., } \lvert\boldsymbol{\sigma}_{\text{eff}}^{\text{d}}\rvert \geq K \end{cases}$

Table 1: Examples for the strain energy density W used for the definition of $\mathcal{E}_{\text{nl,eff}}$ and the corresponding effective stress $\boldsymbol{\sigma}_{\text{eff}}$.

Other than that, we employ the external work rate, the dissipation potential as for linear poro-elasticity, cf. Sec. 3.1. Inserting all components into the gradient flow framework from Sec. 2.1, yields the final model

$$(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_{\int}) = \underset{(\boldsymbol{v},\boldsymbol{z})\in\dot{\mathcal{V}}\times\dot{\mathcal{Z}}_{\int}}{\arg\min} \left\{\mathcal{D}_{\text{fluid}}(\boldsymbol{z}) + \left\langle \boldsymbol{\nabla}\mathcal{E}_{\text{nl}}(\boldsymbol{u}, \boldsymbol{q}_{\int}), (\boldsymbol{v}, \boldsymbol{z})\right\rangle - \mathcal{P}_{\text{ext}}(\boldsymbol{v}, \boldsymbol{z})\right\}. \tag{5.6}$$

Considering, e.g., constant shear modulus and non-linear compressibility, cf. Tab. 1, and the fluid pressure as defined by (5.4), the corresponding optimality conditions are consistent with the model considered by [30]

$$\begin{aligned} 2\mu\left\langle\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle + \left\langle l(\boldsymbol{\nabla}\cdot\boldsymbol{u}), \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle - \alpha\left\langle p, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle &= \mathcal{P}_{\text{ext,mech}}(\boldsymbol{v}) && \forall \boldsymbol{v} \in \mathcal{V}_0, \\ \left\langle \boldsymbol{\kappa}^{-1}\dot{\boldsymbol{q}}_{\int}, \boldsymbol{z}\right\rangle - \left\langle p, \boldsymbol{\nabla}\cdot\boldsymbol{z}\right\rangle &= \mathcal{P}_{\text{ext,fluid}}(\boldsymbol{z}), && \forall \boldsymbol{z} \in \mathcal{Z}_0, \\ b(p) + \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u} + \boldsymbol{\nabla}\cdot\boldsymbol{q}_{\int} &= \theta_0 + Q_\theta, && \text{in } L^2(\Omega). \end{aligned}$$

Well-posedness of non-linear poro-elasticity described by the generalized gradient flow (5.6) follows directly with same argumentation as in the case of linear poro-elasticity.

Lemma 5.1 (Well-posedness for non-linear poro-elasticity)**.** *Let $\mathcal{E}_{\mathrm{nl}}$, W and b as in* (5.1)*–*(5.3)*. Furthermore, let W be strongly convex in $\boldsymbol{\varepsilon}(\boldsymbol{u})$ with $W(\boldsymbol{\varepsilon}(\mathbf{0})) = 0$ and $\partial_{\boldsymbol{\varepsilon}} W(\boldsymbol{\varepsilon}(\mathbf{0})) = \mathbf{0}$, and let b^{-1} be uniformly increasing with $b^{-1}(0) = 0$. Consider homogeneous boundary conditions and conservation of mass, i.e., $\boldsymbol{u}_\Gamma = \mathbf{0}$, $q_{\Gamma,\mathrm{n}} = 0$, and $q_\theta = 0$. Other than that assume regularity as in Lemma 3.1. And assume finite energy $\mathcal{E}_{\mathrm{nl}}(\boldsymbol{u}_0, \mathbf{0}) < \infty$ and the compatibility condition*

$$\langle \partial_{\boldsymbol{u}} \mathcal{E}_{\mathrm{nl}}(\boldsymbol{u}_0, \mathbf{0}), \boldsymbol{v} \rangle = \mathcal{P}_{\mathrm{ext,mech}}(\boldsymbol{v}), \quad \forall \boldsymbol{v} \in \mathcal{V}_0.$$

Then there exists a unique solution $(\boldsymbol{u}, \boldsymbol{q}_\int)$ of (5.6)*. Its regularity is the same as in the case of linear poro-elasticity, cf.* (3.22)*–*(3.24)*.*

Proof. The proof goes along the lines of Lemma 3.1. We choose the partition $\{\boldsymbol{u}\}$ and $\{\boldsymbol{q}_\int\}$ and define

$$\Psi_{\mathrm{nl}}(\dot{\boldsymbol{q}}_\int) := \mathcal{D}_{\mathrm{fluid}}(\dot{\boldsymbol{q}}_\int) \qquad \mathcal{E}_{\mathrm{nl},1}(\boldsymbol{u}) := \mathcal{E}_{\mathrm{nl,eff}}(\boldsymbol{u}),$$

$$\mathcal{E}_{\mathrm{nl},2}(m) := \int_\Omega \int_0^m b^{-1}(s)\, ds\, dx, \qquad \Lambda_{\mathrm{nl}}(\boldsymbol{u}, \boldsymbol{q}_\int) := \theta_0 - \boldsymbol{\nabla} \cdot \boldsymbol{q}_\int - \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u}.$$

Employing this notation, (5.6) complies with the abstract structured discussed in Thm. 2.1. Furthermore, W strongly convex and b uniformly increasing guarantee growth conditions for the energy functionals, cf. (P4). The remaining properties (P1)–(P6) are simple to verify. Finally, the result is a consequence of Thm. 2.1 □

Remark 5.1. *All models, listed in Tab. 1, satisfy the assumptions from Lemma. 5.1, assuming $\mu > 0$.*

6 Extensions of Darcy flow in poro-elastic media as generalized gradient flow

In the presence of non-Newtonian, non-laminar or transitional flow between boundaries, the linear Darcy law is not appropriate anymore to relate the volumetric flux with the fluid pressure gradient. For that reason, extensions of Darcy's law have been established in the literature. In the following, we discuss extensions for non-Newtonian flow, Darcy-Forchheimer flow, and Darcy-Brinkman flow. We incorporate such in the gradient flow modelling framework by an adequate, alternative choice of a dissipation potential $\mathcal{D}_{\star,\mathrm{fluid}}$ corresponding to viscous flow. By keeping previous choices for energy functionals $\mathcal{E}$, external work rates $\mathcal{P}_{\mathrm{ext}}$ etc., and preserving the modelling ansatz

$$(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_\int) = \underset{(\boldsymbol{v},\boldsymbol{z}) \in \dot{\mathcal{V}} \times \dot{\mathcal{Z}}_\int}{\arg\min} \left\{ \mathcal{D}_{\star,\mathrm{fluid}}(\boldsymbol{z}) + \left\langle \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{u}, \boldsymbol{q}_\int), (\boldsymbol{v}, \boldsymbol{z}) \right\rangle - \mathcal{P}_{\mathrm{ext}}(\boldsymbol{v}, \boldsymbol{z}) \right\}, \tag{6.1}$$

the previously discussed poro-elasticity models get simply enhanced by the corresponding extension of Darcy's law.

Non-Newtonian flow: Explicitly distinguishing between the permeability $\boldsymbol{\kappa}$ and the fluid shear viscosity ν (not as in the previous sections), Darcy's law with potentially variable viscosity reads

$$\nu(|\boldsymbol{q}|)\boldsymbol{q} = -\boldsymbol{\kappa}(\boldsymbol{\nabla} p - \boldsymbol{g}_{\mathrm{ext}}). \tag{6.2}$$

Common, constitutive shear viscosity models employed in the literature, cf., e.g., [66], are given in Tab. 2. For non-constant viscosity we assume an isotropic, uniformly bounded permeability $\boldsymbol{\kappa} = \kappa \mathbf{I}$ – a commmon assumption in modelling non-Newtonian fluid flow in porous media, cf., e.g., [67]. The corresponding dissipation potential to be used in (6.1) is given by

$$\mathcal{D}_{\nu,\mathrm{fluid}}(\boldsymbol{q}) = \int_\Omega \kappa^{-1} \int_0^{|\boldsymbol{q}|} s\nu(s)\, ds\, dx. \tag{6.3}$$

Newtonian fluid	$\nu(s) = \nu_\infty$
Carreau model	$\nu(s) = \nu_\infty + \frac{\nu_0 - \nu_\infty}{(1+K_\mathrm{f}\lvert s\rvert^2)^{\frac{2-r}{2}}}$
Cross model	$\nu(s) = \nu_\infty + \frac{\nu_0 - \nu_\infty}{1+K_\mathrm{f}\lvert s\rvert^{2-r}}$
Power law	$\nu(s) = \frac{1}{K_\mathrm{f}\lvert s\rvert^{2-r}}$

Table 2: Constitutive models for the fluid shear viscosity ν [66]; let $0 < \nu_\infty < \nu_0$, $r \in (1,2)$ and $K_\mathrm{f} > 0$.

Darcy-Forchheimer flow: For flow in porous media with an elevated Reynolds number, Darcy's law is enhanced by the so-called *Forchheimer term*, accounting for inertial effects. The resulting non-linear, constitutive relation reads

$$\nu \boldsymbol{q} + \boldsymbol{\kappa} F |\boldsymbol{q}| \boldsymbol{q} = -\boldsymbol{\kappa}(\boldsymbol{\nabla} p - \boldsymbol{g}_\mathrm{ext}), \tag{6.4}$$

where $F \geq 0$ denotes the Forchheimer number [68]. The corresponding dissipation potential to be used in (6.1) is given by

$$\mathcal{D}_\mathrm{F,fluid}(\boldsymbol{q}) = \frac{\nu}{2} \int_\Omega \boldsymbol{\kappa}^{-1} \boldsymbol{q} \cdot \boldsymbol{q} \, dx + \frac{F}{2} \int_\Omega |\boldsymbol{q}|^3 \, dx. \tag{6.5}$$

Darcy-Brinkman flow: For transitional flow between boundaries, Darcy's law may be enhanced by the so-called *Brinkman term*. The resulting linear extension of Darcy's law reads

$$\nu \boldsymbol{q} - \boldsymbol{\kappa} \boldsymbol{\nabla} \cdot (\nu_\mathrm{eff} \boldsymbol{\nabla} \boldsymbol{q}) = -\boldsymbol{\kappa}(\boldsymbol{\nabla} p - \boldsymbol{g}_\mathrm{ext}), \tag{6.6}$$

where $\nu_\mathrm{eff} \geq 0$ denotes the effective viscosity related to the viscous drag effects [69]. The corresponding dissipation potential to be used in (6.1) is given by

$$\mathcal{D}_\mathrm{B,fluid}(\boldsymbol{q}) = \frac{\nu}{2} \int_\Omega \boldsymbol{\kappa}^{-1} \boldsymbol{q} \cdot \boldsymbol{q} \, dx + \frac{\nu_\mathrm{eff}}{2} \int_\Omega |\boldsymbol{\nabla} \boldsymbol{q}|^2 \, dx. \tag{6.7}$$

Independent of the specific choices of the energy functionals etc., well-posedness can be again discussed employing the abstract well-posedness result, cf. Thm. 2.1.

Lemma 6.1 (Well-posedness for extensions of linear Darcy flow in poro-elastic media)**.** *For $p \geq 1$, let $H^p_\mathrm{div}(\Omega) := \left\{ \boldsymbol{z} \in L^p(\Omega)^d \,\middle|\, \boldsymbol{\nabla} \cdot \boldsymbol{z} \in L^2(\Omega) \right\}$. Let the energy functional $\mathcal{E}$ be as in (5.1) with W and b as in Lemma 5.1. And assume finite energy $\mathcal{E}(\boldsymbol{u}_0, \boldsymbol{0}) < \infty$ and the compatibility condition*

$$\langle \partial_{\boldsymbol{u}} \mathcal{E}(\boldsymbol{u}_0, \boldsymbol{0}), \boldsymbol{v} \rangle = \mathcal{P}_\mathrm{ext,mech}(\boldsymbol{v}), \quad \forall \boldsymbol{v} \in \mathcal{V}_0.$$

Consider homogeneous boundary conditions and conservation of mass, i.e., $\boldsymbol{u}_\Gamma = \boldsymbol{0}$, $q_{\Gamma,\mathrm{n}} = 0$, and $q_\theta = 0$. Other than that assume regularity as in Lemma 3.1.

(1) Non-Newtonian fluid flow: Let $\nu = \nu(s)$ denote the fluid shear viscosity model. Let $s \mapsto s\nu(s)$ be non-decreasing, and assume there exists a $p \in (1, \infty)$ satisfying $\nu(s) \gtrsim s^{p-2}$, $s > 0$. Then there exists a solution $(\boldsymbol{u}, \boldsymbol{q}_\int)$ to (6.1) with the dissipation potential (6.3), satisfying

$$\begin{aligned} &\boldsymbol{u} \in L^\infty(0,T; \boldsymbol{H}^1(\Omega)), \\ &\boldsymbol{q}_\int \in H^1(0,T; L^p(\Omega)^d) \cap L^\infty\left(0,T; H^p_\mathrm{div}(\Omega)\right). \end{aligned}$$

In case the energy $\mathcal{E}$ is quadratic, the solution is unique.

(2) *Darcy-Forchheimer flow: There exists a solution* $(\boldsymbol{u}, \boldsymbol{q}_f)$ *to* (6.1) *with the dissipation potential* (6.4)*, satisfying*

$$\begin{aligned} \boldsymbol{u} &\in L^\infty(0,T;\boldsymbol{H}^1(\Omega)), \\ \boldsymbol{q}_f &\in H^1(0,T;L^3(\Omega)^d) \cap L^\infty\left(0,T;H^3_{\mathrm{div}}(\Omega)\right). \end{aligned}$$

In case the energy $\mathcal{E}$ is quadratic, the solution is unique.

(3) *Darcy-Brinkman flow: There exists a unique solution* $(\boldsymbol{u}, \boldsymbol{q}_f)$ *to* (6.1) *with the dissipation potential* (6.6)*, satisfying*

$$\begin{aligned} \boldsymbol{u} &\in L^\infty(0,T;\boldsymbol{H}^1(\Omega)), \\ \boldsymbol{q}_f &\in H^1(0,T;\boldsymbol{H}^1(\Omega)). \end{aligned}$$

Proof. The result is a direct consequence of Thm. 2.1, and the proof is analogous to the proofs of Lemma 3.1 and Lemma 5.1. It suffices to verify (P3) for the different dissipation potentials $\mathcal{D}_{\star,\mathrm{fluid}}$.

Non-Newtonian fluid flow: The dissipation potential $\mathcal{D}_{\nu,\mathrm{fluid}}$ is coercive wrt. $L^p(\Omega)^d$ since it holds

$$\mathcal{D}_{\nu,\mathrm{fluid}}(\boldsymbol{q}) \gtrsim \int_\Omega \int_0^{|\boldsymbol{q}|} s\nu(s)\,ds\,dx \gtrsim \int_\Omega \int_0^{|\boldsymbol{q}|} s^{p-1}\,ds\,dx \gtrsim \|\boldsymbol{q}\|^p_{L^p(\Omega)}.$$

Furthermore, $\mathcal{D}_{\nu,\mathrm{fluid}}$ is convex as composition of two convex maps; indeed, $\boldsymbol{q} \mapsto |\boldsymbol{q}|$ is convex, and $x \mapsto \int_0^x s\nu(s)\,ds$ is convex, since $s \mapsto s\nu(s)$ is increasing. All in all, (P3) is fulfilled.

Darcy-Forchheimer flow: The dissipation potential $\mathcal{D}_{\mathrm{F,fluid}}$ is by construction coercive wrt. $L^3(\Omega)^d$. As sum of convex functions, it is convex. Hence, (P3) is fulfilled.

Darcy-Brinkman flow: The dissipation potential $\mathcal{D}_{\mathrm{B,fluid}}$ defines a norm on $\boldsymbol{H}^1(\Omega)$. Hence, (P3) is fulfilled. Furthermore, $\mathcal{D}_{\mathrm{B,fluid}}$ is quadratic, and uniqueness of solutions to (6.1) is guaranteed.

□

Remark 6.1 (Well-posedness for different viscosity models from Tab. 2)**.** *All models mentioned in Tab. 2 satisfy the assumptions of Lemma 6.1. For fluid shear viscosities modelled by the Carreau model, the Cross model, as well as for Newtonian fluids, one can choose $p = 2$, since it holds $\nu(s) \geq \nu_\infty$, $s > 0$. For the power law, it holds $\nu(s) \gtrsim s^{r-2}$, $s > 0$. Hence, only reduced regularity is obtained with $p = r \in (1,2)$.*

7 Thermo-poro-elasticity as generalized gradient flow

Non-isothermal fluid flow in deformable porous media has in general a strongly non-linear, coupled character, compared to linear poro-elasticity. Even under the hypothesis of infinitesimal strains, three non-linearities may occur, cf., e.g. [22]: (i) thermal convection, coupled to the fluid problem; (ii) non-linear viscous dissipation, associated with Darcy's law, acting as a heat source; (iii) and a temperature weighted time derivative of the total entropy in the energy equation. In certain situations, those non-linearities can be neglected [22]: (i) for a small Péclet number, which quantifies the heat convectively transported by the fluid in comparison with the heat supplied by diffusion through the porous medium; (ii) for small Brinkman number, which quantifies the order of magnitude of the heat source due to viscous dissipation in comparison with heat supplied by conduction; and (iii) small variations of temperature. Under assumptions (ii) and (iii), the model for linear thermo-poro-elasticity with non-linear convection has been derived using homogenization [25, 26]. For a discussion of the general, fully non-linear model we refer to [22].

Assuming all three non-linear effects (i)–(iii) can be neglected, allows for linearizing the general thermo-poro-elasticity model. Using mechanical displacement $\boldsymbol{u}$, fluid pressure p and temperature T as primary variables, the linear, reduced thermo-poro-elasticity model including linearized fluid state equations reads

$$-\boldsymbol{\nabla}\cdot\left[\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u})-\alpha p\,\mathbf{I}-3\alpha_{\mathrm{T}}K_{\mathrm{dr}}T\,\mathbf{I}\right]=\boldsymbol{f}_{\mathrm{ext}}, \tag{7.1}$$

$$\tfrac{1}{M}\dot{p}+\alpha\boldsymbol{\nabla}\cdot\dot{\boldsymbol{u}}-3\alpha_{\phi}\dot{T}-\boldsymbol{\nabla}\cdot\left(\boldsymbol{\kappa}(\boldsymbol{\nabla}p-\boldsymbol{g}_{\mathrm{ext}})\right)=q_{\theta}, \tag{7.2}$$

$$C_{\mathrm{d}}\dot{T}+3\alpha_{\mathrm{T}}T_0K_{\mathrm{dr}}\boldsymbol{\nabla}\cdot\dot{\boldsymbol{u}}-3\alpha_{\phi}T_0\dot{p}-\boldsymbol{\nabla}\cdot\left(\boldsymbol{\kappa}_{\mathrm{F}}\boldsymbol{\nabla}T\right)=T_0q_S, \tag{7.3}$$

subject to suitable boundary and initial conditions, cf., e.g., [22]. Here, K_{dr} denotes the bulk modulus, α_{T} is the Biot coefficient associated with thermal effects, α_{ϕ} governs the pressure-temperature coupling of the fluid, C_{d} is the total volumetric heat capacity, T_0 is a constant reference temperature, and $\boldsymbol{\kappa}_{\mathrm{F}}$ denotes the thermal conductivity. Other than that, same notation is used as in the previous sections.

The linearized thermo-poro-elasticity model (7.1)–(7.3) has a similar structure as Biot's consolidation model. In the following, we provide a generalized gradient flow formulation of (7.1)–(7.3), which will be later exploited in the context of robust splitting schemes, cf. Sec. 12.

In the context of the abstract gradient flow modelling framework introduced in Sec. 2.1, we choose $(\boldsymbol{u},\theta,S)$ as state variables, i.e., the mechanical displacement, the fluid content and the total entropy. Motivated by (7.2)–(7.3), the latter two will be related to $(\boldsymbol{u},p,T)$ by

$$\theta=\tfrac{1}{M}p+\alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}-3\alpha_{\phi}T, \tag{7.4}$$

$$S=\tfrac{C_{\mathrm{d}}}{T_0}T+3\alpha_{\mathrm{T}}K_{\mathrm{dr}}\boldsymbol{\nabla}\cdot\boldsymbol{u}-3\alpha_{\phi}p. \tag{7.5}$$

Changes of states are associated to $(\dot{\boldsymbol{u}},\boldsymbol{q},\boldsymbol{j})$, i.e., the rate of mechanical displacement, the volumetric flux and the entropy flux, by conservation of volume and balance of entropy

$$\dot{\theta}+\boldsymbol{\nabla}\cdot\boldsymbol{q}=q_{\theta},$$
$$\dot{S}+\boldsymbol{\nabla}\cdot\boldsymbol{j}=q_S.$$

Under above, linearizing assumptions, the entropy flux can be identified with the heat flux scaled by T_0^{-1} such that due to Darcy's law and Fourier's law, we obtain

$$\boldsymbol{q}=-\boldsymbol{\kappa}(\boldsymbol{\nabla}p-\boldsymbol{g}_{\mathrm{ext}}), \tag{7.6}$$

$$\boldsymbol{j}=-\frac{\boldsymbol{\kappa}_{\mathrm{F}}}{T_0}\boldsymbol{\nabla}T. \tag{7.7}$$

Focussing on the inherent gradient flow structure of linearized thermo-poro-elasticity, we omit specifying the regularity of all variables; we refer to the formal function spaces including essential boundary conditions as defined in Sec. 2.1. Generalizing linear poro-elasticity, a natural choice for the dissipation potential is

$$\mathcal{D}_{\mathrm{th}}(\boldsymbol{q},\boldsymbol{j})=\frac{1}{2}\left\langle\boldsymbol{\kappa}^{-1}\boldsymbol{q},\boldsymbol{q}\right\rangle+\frac{1}{2}\left\langle T_0\boldsymbol{\kappa}_{\mathrm{F}}^{-1}\boldsymbol{j},\boldsymbol{j}\right\rangle.$$

The Helmholtz free energy associated to linearized thermo-poro-elasticity for given state is defined by

$$\mathcal{E}_{\mathrm{th}}(\boldsymbol{u},\theta,S):=\mathcal{E}_{\mathrm{eff}}(\boldsymbol{u})+\mathcal{E}_{\mathrm{th,fluid}}(\boldsymbol{u},\theta,S)$$

with $\mathcal{E}_\text{eff}$ defined as for linear poro-elasticity and the fluid contribution

$$\mathcal{E}_\text{th,fluid}(\boldsymbol{u},\theta,S) := \frac{1}{2}\left\langle \begin{bmatrix} [c]\frac{1}{M} & -3\alpha_\phi \\ -3\alpha_\phi & \frac{C_\text{d}}{T_0} \end{bmatrix}^{-1} \begin{bmatrix} [l]\theta - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u} \\ S - 3\alpha_\text{T}K_\text{dr}\boldsymbol{\nabla}\cdot\boldsymbol{u} \end{bmatrix}, \begin{bmatrix} [l]\theta - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u} \\ S - 3\alpha_\text{T}K_\text{dr}\boldsymbol{\nabla}\cdot\boldsymbol{u} \end{bmatrix} \right\rangle.$$

Using (7.4)–(7.5), $\mathcal{E}_\text{th,fluid}$ can be also written as function of the primary variables

$$\mathcal{E}_\text{th,fluid}(\boldsymbol{u},p,T) = \frac{1}{2}\left\langle \begin{bmatrix} \frac{1}{M} & -3\alpha_\phi \\ -3\alpha_\phi & \frac{C_\text{d}}{T_0} \end{bmatrix} \begin{bmatrix} p \\ T \end{bmatrix}, \begin{bmatrix} p \\ T \end{bmatrix} \right\rangle.$$

Finally, we formulate the linearized thermo-poro-elasticity model as generalized gradient flow: Given the current state $(\boldsymbol{u},\theta,S)$, its change is described by

$$\dot{\theta} = q_\theta - \boldsymbol{\nabla}\cdot\boldsymbol{q}, \tag{7.8}$$

$$\dot{S} = q_S - \boldsymbol{\nabla}\cdot\boldsymbol{j}, \tag{7.9}$$

$$\begin{aligned}(\dot{\boldsymbol{u}},\boldsymbol{q},\boldsymbol{j}) = \underset{(\boldsymbol{v},\boldsymbol{z},\boldsymbol{w})\in\dot{\mathcal{V}}\times\mathcal{Z}\times\mathcal{W}}{\arg\min} \Big\{ & \mathcal{D}_\text{th}(\boldsymbol{z},\boldsymbol{w}) + \langle \boldsymbol{\nabla}\mathcal{E}_\text{th}(\boldsymbol{u},\theta,S), [\boldsymbol{v}, -\boldsymbol{\nabla}\cdot\boldsymbol{z}, -\boldsymbol{\nabla}\cdot\boldsymbol{w}] \rangle \\ & - \mathcal{P}_\text{ext,th}(\boldsymbol{v},\boldsymbol{z},\boldsymbol{w}) \Big\}. \end{aligned} \tag{7.10}$$

Using simple calculations, one can verify that the corresponding optimality conditions are equivalent to the original problem formulation (7.1)–(7.3). Well-posedness can be established analogously to linear poro-elasticity, exploiting that linearized thermo-poro-elasticity is essentially a vectorized version of linear poro-elasticity.

Remark 7.1 (Including non-monotone perturbations)**.** *For larger Péclet and Brinkman numbers, the contributions*

$$W_{(i)}(\boldsymbol{q},\boldsymbol{j}) = -c\left\langle \boldsymbol{\kappa}_\text{F}^{-1}\boldsymbol{j},\boldsymbol{q}\right\rangle, \qquad (\textit{for some } c\in\mathbb{R}), \tag{7.11}$$

$$W_{(ii)}(\boldsymbol{q}) = \left(\tfrac{1}{T_0} - 3\alpha_\phi\right)\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q},\boldsymbol{q}\right\rangle, \tag{7.12}$$

corresponding to thermal convection and the heat production due to viscous dissipation, respectively, are non-negligible and have to be incorporated in the energy equation (7.3)*, which then becomes*

$$\tfrac{C_\text{d}}{T_0}\dot{T} + 3\alpha_\text{T}K_\text{dr}\boldsymbol{\nabla}\cdot\dot{\boldsymbol{u}} - 3\alpha_\phi\dot{p} + W_{(i)}(\boldsymbol{q},\boldsymbol{j}) + \boldsymbol{\nabla}\cdot\boldsymbol{j} = q_S + W_{(ii)}(\boldsymbol{q}).$$

Eq. (7.1)*–*(7.2) *remain unchanged. Based on the above discussion, the resulting equations can be interpreted as perturbed gradient flow (or doubly non-linear evolution equation with non-monotone perturbations). In the context of operator splitting schemes, we will discuss possibilities to still exploit the part containing a gradient flow structure for deriving robust splitting schemes for the general problem, cf. Sec. 12.3.*

Part II – Efficient discrete approximation schemes for thermo-poro-visco-elasticity

The gradient flow structure of thermo-poro-visco-elasticity revealed in Part I allows for a unified framework for deriving stable temporal and spatial discretizations, as well as the development and analysis of robust operator splitting schemes. The abstract workflow taken in this paper is visualized in Fig. 1 can be summarized as follows: Given a time-continuous gradient flow formulation, a time-discrete approximation is introduced by applying the minimizing movement scheme, resulting in a (convex and hence well-posed) minimization problem for each time step. A corresponding dual problem is derived by applying the Legendre-Fenchel duality theory [50].

Provided that the resulting minimization problems are block-separable, alternating minimization is utilized, decoupling physical subproblems – it comes with strong robustness under fairly weak assumptions and allows for an abstract convergence analysis [70, 71]. Furthermore, the underlying minimization structure of the coupled problems enables simple acceleration of iterative solvers via cheap line search strategies.

Although based on semi-discrete approximations, the results also immediately translate to fully-discrete approximations obtained by the Galerkin method, i.e., the well-posedness as convex minimization problems, and the efficient numerical solution by block-coordinate descent methods.

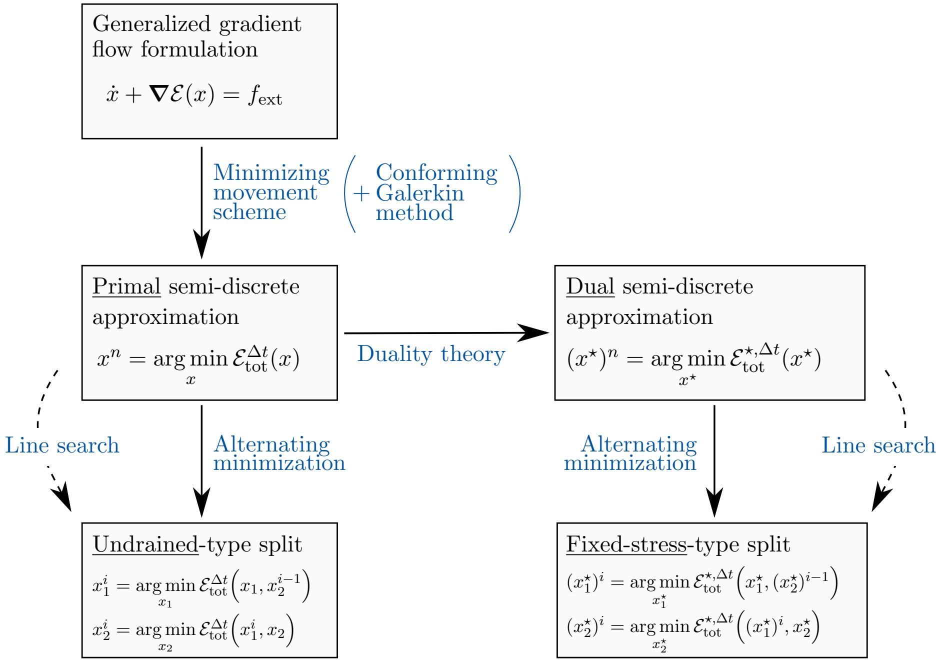

Generalized gradient flow formulation
$\dot{x} + \boldsymbol{\nabla}\mathcal{E}(x) = f_{\text{ext}}$
Minimizing movement scheme
$\left(+ \right.$ Conforming Galerkin method $\left. \right)$
Primal semi-discrete approximation
$x^n = \arg\min_x \mathcal{E}_{\text{tot}}^{\Delta t}(x)$
Duality theory
Dual semi-discrete approximation
$(x^\star)^n = \arg\min_{x^\star} \mathcal{E}_{\text{tot}}^{\star,\Delta t}(x^\star)$
Line search
Alternating minimization
Alternating minimization
Line search
Undrained-type split
$x_1^i = \arg\min_{x_1} \mathcal{E}_{\text{tot}}^{\Delta t}\left(x_1, x_2^{i-1}\right)$
$x_2^i = \arg\min_{x_2} \mathcal{E}_{\text{tot}}^{\Delta t}\left(x_1^i, x_2\right)$
Fixed-stress-type split
$(x_1^\star)^i = \arg\min_{x_1^\star} \mathcal{E}_{\text{tot}}^{\star,\Delta t}\left(x_1^\star, (x_2^\star)^{i-1}\right)$
$(x_2^\star)^i = \arg\min_{x_2^\star} \mathcal{E}_{\text{tot}}^{\star,\Delta t}\left((x_1^\star)^i, x_2^\star\right)$

Figure 1: Workflow for the derivation of splitting schemes for linear thermo-poro-visco-elasticity, illustrated for simplicity for classical gradient flows.

By applying the workflow in particular to linear poro-elasticity, we derive the well-known undrained and fixed-stress splits. Thereby, we provide a novel interpretation of the widely used splitting schemes as inexact minimization. Motivated by that, the abstract approach is further applied to distinct representatives of three generalizations of linear poro-elasticity: poro-visco-elasticity, non-linear poro-elasticity under infinitesimal strains, and thermo-poro-elasticity. Ultimately, novel splitting schemes are derived for poro-visco-elasticity and nonlinear poro-elasticity, structurally similar to the undrained and fixed-stress splits. In the context of thermo-poro-elasticity, the recently proposed undrained-adiabatic and extended fixed-stress splits [45] are derived and by that justified mathematically.

8 Energy-driven time discretization via minimizing movements

Gradient flows allow for stable time discretization. Throughout this part of the paper, we apply the so-called *minimizing movement* scheme [52], which is energy dissipating and closely related to the implicit Euler method, most often the first choice time discretization for poro-elasticity models. Utilizing a minimization formulation, the minimizing movement scheme retains the structure and possible convexity properties of the problem. However, we note, it does not preserve a discrete energy identity analogous to (A.3); for structure-preserving time-discretizations we refer, e.g., to [14] and the references within.

For simplicity, we consider an equidistant partition $0 = t_0 < t_1 < ... < t_N = T$ of the time interval $[0, T]$ with time step size Δt. Fields, functionals and function spaces evaluated at discrete time t_n are enhanced by an exponent n, e.g., $x^n := x(t_n, \cdot) \in \mathcal{X}^n := \mathcal{X}(t_n)$ and $\mathcal{P}^n_{\text{ext}}(\cdot) := \mathcal{P}_{\text{ext}}(t_n)$.

The minimizing movement scheme applied to the abstract, generalized gradient flow (1.2) is identical with a semi-implicit Euler method, where state-dependent functions are lagged in time. For time step n, it reads: Given $x^{n-1} \in \mathcal{X}^{n-1}$, find $x^n \in \mathcal{X}^n$ and $p^n \in \mathcal{P}^n$ satisfying

$$\frac{x^n - x^{n-1}}{\Delta t} + \mathcal{T}\left(x^{n-1}\right) p^n = 0 \tag{8.1}$$

$$p^n = \underset{p \in \mathcal{P}^n}{\arg\min} \left\{ \Delta t\, \mathcal{D}\left(x^{n-1}; p\right) + \mathcal{E}(x^n) - \Delta t \mathcal{P}^n_{\text{ext}}\left(x^{n-1}; p\right) \right\}. \tag{8.2}$$

As the structure of the original problem is retained, solvability of the time-discrete problem is automatically inherited from the continuous problem. In this work, all dissipation potentials, external work rates and process operators are state-independent. Thus, we omit the explicit dependence from now on.

The coupled problem (8.1)–(8.2) can be obviously decoupled by reducing (8.2) to a minimization problem for process vectors

$$\frac{x^n - x^{n-1}}{\Delta t} + \mathcal{T} p^n = 0 \tag{8.3}$$

$$p^n = \underset{p \in \mathcal{P}^n}{\arg\min} \left\{ \Delta t\, \mathcal{D}(p) + \mathcal{E}\left(x^{n-1} - \mathcal{T} p\right) - \Delta t\, \mathcal{P}^n_{\text{ext}}(p) \right\}. \tag{8.4}$$

Alternatively, provided that the change of state is directly associated to its rate, it is $\mathcal{T} = -Id$; in the context of poro-elasticity, this particular case occurs, e.g., for mechanical deformation. Consequently, (8.3)–(8.4) becomes a minimization problem for the state itself

$$x^n = \underset{x \in \mathcal{X}^n}{\arg\min} \left\{ \Delta t\, \mathcal{D}\left(\frac{x - x^{n-1}}{\Delta t}\right) + \mathcal{E}(x) - \Delta t\, \mathcal{P}^n_{\text{ext}}\left(\frac{x - x^{n-1}}{\Delta t}\right) \right\}. \tag{8.5}$$

The Euler-Lagrange equation is indeed equivalent to the *implicit Euler* scheme

$$\boldsymbol{\nabla}\mathcal{D}\left(\frac{x^n - x^{n-1}}{\Delta t}\right) + \boldsymbol{\nabla}\mathcal{E}(x^n) = \boldsymbol{\nabla}\mathcal{P}^n_{\text{ext}}\left(\frac{x^n - x^{n-1}}{\Delta t}\right), \quad \text{in } \mathcal{X}_0^\star,$$

where $\mathcal{X}_0$ is the linear tangent space to $\mathcal{X}^n$.

For the thermo-poro-visco-elasticity models discussed in this paper, we employ a combination of (8.3)–(8.4) and (8.5), depending on the nature of the particular variables and available process vectors. A fully-discrete approximation may then be obtained by the conforming Galerkin method; see Sec. 9.3 for an exemplary discussion in the context of linear poro-elasticity.

9 Minimization formulations of discrete linear poro-elasticity and robust splitting schemes via alternating minimization

In the literature, various formulations of linear poro-elasticity are employed, differing in the choice of primal variables. In this spirit, we present various minimization formulations of time-discrete, linear poro-elasticity, after applying the minimizing movement scheme (Sec. 8). In particular, we discuss the widely used two-, three-, and five-field formulations, as well as the primal and the dual formulations naturally arising from Part I. The specific minimization formulation is relevant when applying a line search strategy for the acceleration of iterative solvers as splitting schemes, cf. Sec. 13. Fully-discrete approximations with same properties are obtained by the conforming Galerkin method. We also note that minimization formulations can be derived in the context of the least-squares finite element method, cf., e.g., [34]; however, such usually do not stem naturally from a physical, gradient flow formulation but build directly on classical PDE-based models.

Following the workflow visualized in Fig. 1, we derive the widely used undrained split and fixed-stress split as alternating minimization applied to the primal and dual formulations, respectively. Splitting schemes for linear poro-elasticity have been well studied in recent literature, and as such, much of the material in this section represents a new perspective, and indeed also new proofs, of known results. However, even in this case the discussion in this section lays the foundation for the more advanced applications in the subsequent sections, wherein the gradient flow framework leads to new schemes not previously reported.

9.1 Minimization formulations of time-discrete linear poro-elasticity

In the following, we introduce various minimization formulations of time-discrete, linear poro-elasticity differing in the choice of the primary variables. We present the primal and dual formulations, as well as the more widely used two-, three-, and five-field formulations.

9.1.1 Primal formulation of time-discrete linear poro-elasticity

Time discretization of the continuous, primal formulation of linear poro-elasticity (3.12) via the minimizing movement scheme, yields the primal formulation of time-discrete, linear poro-elasticity. It is formulated as a series of minimization problems: At time step $n \geq 1$, let

$$\mathcal{V}^n := \left\{ \boldsymbol{v} \in H^1(\Omega) \,\middle|\, \boldsymbol{v} = \boldsymbol{u}_\Gamma^n \text{ on } \Gamma_{\boldsymbol{u}} \right\}, \tag{9.1}$$

$$\mathcal{Z}^n := \left\{ \boldsymbol{z} \in H(\mathrm{div};\Omega) \,\middle|\, \boldsymbol{z} \cdot \boldsymbol{n} = q_{\Gamma,\mathrm{n}}^n \text{ on } \Gamma_{\boldsymbol{q}} \right\}. \tag{9.2}$$

Then given θ^{n-1}, define $(\boldsymbol{u}^n, \boldsymbol{q}^n) \in \mathcal{V}^n \times \mathcal{Z}^n$ by

$$(\boldsymbol{u}^n, \boldsymbol{q}^n) := \underset{(\boldsymbol{u},\boldsymbol{q}) \in \mathcal{V}^n \times \mathcal{Z}^n}{\arg\min}\; \mathcal{E}_{\mathrm{tot}}^{\Delta t}(\theta^{n-1}; \boldsymbol{u}, \boldsymbol{q}), \quad \text{where} \tag{9.3}$$

$$\begin{aligned}\mathcal{E}_{\mathrm{tot}}^{\Delta t}(\theta^{n-1}; \boldsymbol{u}, \boldsymbol{q}) :=& \tfrac{1}{2} \left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u}) \right\rangle + \tfrac{\Delta t}{2} \left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{q} \right\rangle \\ &+ \tfrac{M}{2} \left\| \theta^{n-1} + \Delta t\, q_\theta^n - \Delta t \boldsymbol{\nabla} \cdot \boldsymbol{q} - \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u} \right\|^2 \\ &- \mathcal{P}_{\mathrm{ext,mech}}^n(\boldsymbol{u}) - \Delta t\, \mathcal{P}_{\mathrm{ext,fluid}}^n(\boldsymbol{q}),\end{aligned}$$

and set $\theta^n := \theta^{n-1} + \Delta t\, q_\theta^n - \Delta t \boldsymbol{\nabla} \cdot \boldsymbol{q}^n$. Since the energy $\mathcal{E}_{\mathrm{tot}}^{\Delta t}$ is strictly convex, existence and uniqueness of a solution to (9.3) follow by classical results from convex analysis, cf. Thm. A.2.

9.1.2 Dual formulation of time-discrete linear poro-elasticity

We introduce a dual formulation of (9.3). It can be derived using tools from convex analysis; or equivalently, by employing the minimizing movement scheme to the continuous, dual formulation (3.21): At time step $n \geq 1$, let

$$\mathcal{S}^n := \left\{ \boldsymbol{\tau} \in H(\mathrm{div};\Omega)^d \,\middle|\, \begin{array}{l} \boldsymbol{\tau}\boldsymbol{n} = \boldsymbol{\sigma}_{\Gamma,\mathrm{n}}^n \text{ on } \Gamma_{\boldsymbol{\sigma}}, \\ \boldsymbol{\nabla} \cdot \boldsymbol{\tau} + \boldsymbol{f}_{\mathrm{ext}}^n = \boldsymbol{0} \text{ in } L^2(\Omega), \\ \langle \boldsymbol{\tau}, \boldsymbol{\gamma} \rangle = 0 \;\forall \boldsymbol{\gamma} \in \boldsymbol{Q}_{\mathrm{AS}} \end{array} \right\},$$

$$\mathcal{Q}^n := \left\{ q \in H^1(\Omega) \,\middle|\, q = p_\Gamma^n \text{ on } \Gamma_p \right\},$$

where $\boldsymbol{Q}_{\mathrm{AS}}$ as defined in (3.17). Then given $(\boldsymbol{\sigma}^{n-1}, p^{n-1}) \in \mathcal{S}^{n-1} \times \mathcal{Q}^{n-1}$, set $\theta^{n-1} := \frac{1}{M} p^{n-1} + \alpha \operatorname{tr}\left(\mathbb{A}(\boldsymbol{\sigma}^{n-1} + \alpha p^{n-1}\mathbf{I})\right)$, and define $(\boldsymbol{\sigma}^n, p^n) \in \mathcal{S}^n \times \mathcal{Q}^n$ to be the solution of the dual minimization problem

$$(\boldsymbol{\sigma}^n, p^n) := \underset{(\boldsymbol{\sigma},p)\in\mathcal{S}^n\times\mathcal{Q}^n}{\arg\min}\ \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\theta^{n-1};\boldsymbol{\sigma},p), \quad \text{where} \tag{9.4}$$

$$\begin{aligned}\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\theta^{n-1};\boldsymbol{\sigma},p) := &\tfrac{1}{2}\left\langle \mathbb{A}(\boldsymbol{\sigma}+\alpha p\mathbf{I}), \boldsymbol{\sigma}+\alpha p\mathbf{I}\right\rangle \\ &+ \tfrac{1}{2M}\|p\|^2 + \tfrac{\Delta t}{2}\left\langle \boldsymbol{\kappa}(\boldsymbol{\nabla} p - \boldsymbol{g}_{\mathrm{ext}}^n), \boldsymbol{\nabla} p - \boldsymbol{g}_{\mathrm{ext}}^n\right\rangle \\ &- \left\langle \boldsymbol{u}_\Gamma^n, \boldsymbol{\sigma}\boldsymbol{n}\right\rangle_{\Gamma_{\boldsymbol{u}}} - \left\langle \theta^{n-1} + \Delta t\, q_\theta^n, p\right\rangle - \Delta t\left\langle q_{\Gamma,\mathrm{n}}^n, p\right\rangle_{\Gamma_q}.\end{aligned}$$

Since the energy $\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$ is strictly convex, and the feasible set $\mathcal{S}^n \times \mathcal{Q}^n$ is non-empty and convex, existence and uniqueness of a solution to (9.4) follow by classical results from convex analysis, cf. Thm. A.2.

9.1.3 Two-field saddle point formulation of time-discrete linear poro-elasticity

In the literature, linear poro-elasticity is often studied both numerically and analytically based on a two-field saddle point formulation of the linear Biot equations. It employs the mechanical displacement $\boldsymbol{u}$ and the fluid pressure p as primary variables, cf., e.g., [27, 37, 35, 48, 72]. Employing the implicit Euler method for time-discretization, time step $n \geq 1$ reads: Given $\theta^{n-1} := \frac{1}{M}p^{n-1} + \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}^{n-1}$, find $(\boldsymbol{u}^n, p^n) \in \mathcal{V}^n \times \mathcal{Q}^n$ satisfying for all $(\boldsymbol{v}, q) \in \mathcal{V}_0 \times \mathcal{Q}_0$

$$\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}^n), \boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle - \alpha\left\langle p^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle = \left\langle \boldsymbol{f}_{\mathrm{ext}}^n, \boldsymbol{v}\right\rangle + \left\langle \boldsymbol{\sigma}_{\Gamma,\mathrm{n}}^n, \boldsymbol{v}\right\rangle_{\Gamma_{\boldsymbol{\sigma}}},$$

$$\frac{1}{M}\left\langle p^n, q\right\rangle + \alpha\left\langle \boldsymbol{\nabla}\cdot\boldsymbol{u}^n, q\right\rangle + \Delta t\left\langle \boldsymbol{\kappa}(\boldsymbol{\nabla} p^n - \boldsymbol{g}_{\mathrm{ext}}^n), \boldsymbol{\nabla} q\right\rangle = \left\langle \theta^{n-1} + \Delta t\, q_\theta^n, q\right\rangle + \Delta t\left\langle q_{\Gamma,\mathrm{n}}^n, q\right\rangle_{\Gamma_{\boldsymbol{q}}},$$

where $\mathcal{Q}_0 := \{q \in H^1(\Omega)\,|\, q = 0 \text{ on } \Gamma_p\}$.

Saddle point formulations can be in general transformed to a constrained minimization problem [61]. In the context of linear poro-elasticity, the constraint has to explicitly impose one of the physical subproblems. In the following, we choose to impose the balance of linear momentum and define a suitable product space for each time step n

$$\tilde{\mathcal{H}}^n := \left\{ (\boldsymbol{u}, p) \in \mathcal{V}^n \times \mathcal{Q}^n \;\middle|\; \begin{aligned} &\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle - \alpha\left\langle p, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle \\ &\quad = \left\langle \boldsymbol{f}_{\mathrm{ext}}^n, \boldsymbol{v}\right\rangle + \left\langle \boldsymbol{\sigma}_{\Gamma,\mathrm{n}}^n, \boldsymbol{v}\right\rangle_{\Gamma_{\boldsymbol{\sigma}}} \qquad \forall \boldsymbol{v} \in \mathcal{V}_0 \end{aligned} \right\}.$$

The time-discrete, linear Biot equations at fixed time step n, formulated as constrained minimization problem, read: Given $(\boldsymbol{u}^{n-1}, p^{n-1}) \in \tilde{\mathcal{H}}^{n-1}$, set $\theta^{n-1} := \frac{1}{M}p^{n-1} + \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}^{n-1}$, and find $(\boldsymbol{u}^n, p^n) \in \tilde{\mathcal{H}}^n$, satisfying

$$(\boldsymbol{u}^n, p^n) := \underset{(\boldsymbol{u},p)\in\tilde{\mathcal{H}}^n}{\arg\min}\, \mathcal{E}_{\mathrm{tot}}^{\Delta t}(\theta^{n-1};\boldsymbol{u},p), \quad \text{where} \tag{9.5}$$

$$\begin{aligned}\mathcal{E}_{\mathrm{tot}}^{\Delta t}(\theta^{n-1};\boldsymbol{u},p) := &\tfrac{1}{2}\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u})\right\rangle + \tfrac{1}{2M}\|p\|^2 \\ &+ \tfrac{\Delta t}{2}\left\langle \boldsymbol{\kappa}(\boldsymbol{\nabla} p - \boldsymbol{g}_{\mathrm{ext}}^n), \boldsymbol{\nabla} p - \boldsymbol{g}_{\mathrm{ext}}^n\right\rangle \\ &- \left\langle \theta^{n-1} + \Delta t\, q_\theta^n, p\right\rangle - \Delta t\left\langle q_{\Gamma,\mathrm{n}}^n, p\right\rangle_{\Gamma_q}.\end{aligned}$$

Since the energy $\mathcal{E}_{\mathrm{tot}}^{\Delta t}$ is strictly convex and the feasible set $\tilde{\mathcal{H}}^n$ is non-empty and convex, existence and uniqueness of a solution to (9.4) follow by classical results from convex analysis, cf. Thm. A.2. We emphasize, that well-posedness also follows in the extreme case of an incompressible fluid and impermeable medium, as long as $\mathcal{V}_0 \times \mathcal{Q}_0$ is inf-sup stable such that $\tilde{\mathcal{H}}^n$ is essentially constrained by a one-dimensional relation. Finally, we note, compared to the primal and dual formulations, (9.5) is not block-separable.

Remark 9.1 (Mass conservation as constraint)**.** *Alternatively, mass conservation can imposed as constraint resulting in an alternative minimization formulation of the time-discrete, linear Biot equations. However, having splitting schemes accelerated by relaxation in mind, cf. Sec. 13, the particular choice matters. The constraint has to be satisfied after each splitting iteration; consequently, the above formulation* (9.5) *suits the fixed-stress split, whereas the use of mass conservation as constraint contrarily suits the undrained splitting scheme, cf. Sec. 9.4.*

9.1.4 Three-field formulation of time-discrete linear poro-elasticity

A conforming Galerkin finite element discretization of the classical two-field saddle point formulation is not locally mass conservative. Therefore, in the literature, often a mixed formulation of the fluid flow problem is employed, cf., e.g., [41, 73, 30, 39, 74, 42]; this results in a three-field formulation employing the mechanical displacement $\boldsymbol{u}$, the fluid pressure p and the volumetric flux $\boldsymbol{q}$ as primary variables. Based on the primal two-field minimization formulation of linear poro-elasticity (Sec. 9.1.1), we state an unconstrained minimization formulation corresponding to the three-field formulation of linear poro-elasticity. For this, we essentially modify slightly the energy used in (9.3) and define the pressure as a post-processed quantity. Consistent with (3.6), we define for given $\boldsymbol{u}$ and θ by

$$p := \Pi_{\tilde{\mathcal{Q}}}\big(M(\theta - \alpha \boldsymbol{\nabla}\cdot \boldsymbol{u})\big), \tag{9.6}$$

where $\Pi_{\tilde{\mathcal{Q}}}$ denotes the orthogonal projection onto $\tilde{\mathcal{Q}} := L^2(\Omega)$; the particular choice for the pressure space $\tilde{\mathcal{Q}}$ originates from the expected regularity, cf. Lemma 3.1.

Finally, we define the minimization formulation for time step n: Given $(\boldsymbol{u}^{n-1}, \boldsymbol{q}^{n-1}, p^{n-1}) \in \mathcal{V}^{n-1} \times \mathcal{Z}^{n-1} \times \tilde{\mathcal{Q}}$, set $\theta^{n-1} := \frac{1}{M}p^{n-1} + \alpha \boldsymbol{\nabla}\cdot \boldsymbol{u}^{n-1}$, and define $(\boldsymbol{u}^n, \boldsymbol{q}^n, p^n) \in \mathcal{V}^n \times \mathcal{Z}^n \times \tilde{\mathcal{Q}}$ as solution to

$$(\boldsymbol{u}^n, \boldsymbol{q}^n) := \underset{(\boldsymbol{u},\boldsymbol{q})\in\mathcal{V}^n\times\mathcal{Z}^n}{\arg\min}\ \mathcal{E}_{\text{tot}}^{\Delta t}(\theta^{n-1}; \boldsymbol{u}, \boldsymbol{q}), \tag{9.7}$$

$$p^n := \Pi_{\tilde{\mathcal{Q}}}\left(M(\theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\, \boldsymbol{\nabla}\cdot\boldsymbol{q}^n - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}^n)\right), \tag{9.8}$$

where

$$\begin{aligned}\mathcal{E}_{\text{tot}}^{\Delta t}(\theta^{n-1}; \boldsymbol{u}, \boldsymbol{q}) := &\tfrac{1}{2}\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u})\right\rangle + \tfrac{\Delta t}{2}\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{q}\right\rangle \\ &+ \tfrac{M}{2}\left\|\Pi_{\tilde{\mathcal{Q}}}(\theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\boldsymbol{\nabla}\cdot\boldsymbol{q} - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u})\right\|^2 \\ &- \mathcal{P}_{\text{ext,mech}}^n(\boldsymbol{u}) - \Delta t\, \mathcal{P}_{\text{ext,fluid}}^n(\boldsymbol{q}).\end{aligned}$$

The minimization problem is strictly convex and the projection is well-defined; existence and uniqueness of a solution to (11.1)–(11.2) follow by classical results from convex analysis, cf. Thm. A.2. Furthermore, it is simple to verify that the corresponding optimality conditions yield the classical three-field formulation of time-discrete, linear poro-elasticity, cf. Sec. 9.2.1.

Remark 9.2 (Constrained minimization formulation)**.** *Based on the inherent double saddle point structure of the three-field formulation, alternatively a non-block-separable minimization formulation, constrained by mass conservation, could be utilized, similar to Sec. 9.1.3. This would allow in particular for the discussion of the incompressible case* $M = \infty$*. However, unlike for* $M \in (0, \infty)$*, it becomes evident that inf-sup stability is required for the uniqueness of weak solutions.*

9.1.5 Five-field formulation of time-discrete linear poro-elasticity

So far, no minimization formulation presented above lays a foundation for a fully structure-preserving, conforming Galerkin finite element discretization, which is conserving locally both mass and linear momentum. In order to achieve this, a fully mixed five-field formulation can be used, i.e., mixed formulations for both the mechanical and the fluid flow subproblems, cf., e.g., [59, 33, 32]. Consequently, both independent subproblems incorporate themselves a saddle point structure; however, different to the two- and three-field formulations, the coupling of the two subproblems is symmetric.

After all, we combine ideas from previous sections and state a minimization formulation corresponding to the five-field formulation. In particular, starting from the dual formulation (9.1.2), we add the volumetric flux $\boldsymbol{q}$ as variable and impose Darcy's law as constraint. For fixed each time step n, we define a suitable product space for the fluid flow variables

$$\mathcal{F}^n := \left\{ (\boldsymbol{q}, p) \in \mathcal{Z}^n \times \tilde{\mathcal{Q}} \;\middle|\; \begin{array}{l} \left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{z} \right\rangle - \left\langle p, \boldsymbol{\nabla}\cdot\boldsymbol{z}\right\rangle \\ \quad = \left\langle \boldsymbol{g}_{\mathrm{ext}}^n, \boldsymbol{z}\right\rangle - \left\langle p_\Gamma^n, \boldsymbol{z}\cdot\boldsymbol{n}\right\rangle_{\Gamma_p} \quad \forall \boldsymbol{z}\in\mathcal{Z}_0 \end{array} \right\}.$$

Finally, the minimization formulation reads: Given $(\boldsymbol{\sigma}^{n-1}, \boldsymbol{q}^{n-1}, p^{n-1}) \in \mathcal{S}^{n-1}\times\mathcal{F}^{n-1}$, set $\theta^{n-1} := \frac{1}{M}p^{n-1} + \alpha\,\mathrm{tr}\left(\mathbb{A}(\boldsymbol{\sigma}^{n-1} + \alpha p^{n-1}\mathbf{I})\right)$, and define $(\boldsymbol{\sigma}^n, \boldsymbol{q}^n, p^n) \in \mathcal{S}^n\times\mathcal{F}^n$ to be the solution of the minimization problem

$$(\boldsymbol{\sigma}^n, \boldsymbol{q}^n, p^n) := \underset{(\boldsymbol{\sigma},\boldsymbol{q},p)\in\mathcal{S}^n\times\mathcal{F}^n}{\arg\min}\; \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\sigma}, \boldsymbol{q}, p), \quad \text{where} \tag{9.9}$$

$$\begin{aligned}\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\sigma}, \boldsymbol{q}, p) :=& \tfrac{1}{2}\left\langle \mathbb{A}(\boldsymbol{\sigma}+\alpha p\mathbf{I}), \boldsymbol{\sigma}+\alpha p\mathbf{I}\right\rangle + \tfrac{1}{2M}\|p\|^2 \\ &+ \tfrac{\Delta t}{2}\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{q}\right\rangle - \left\langle \theta^{n-1} + \Delta t\, q_\theta^n, p\right\rangle - \left\langle \boldsymbol{u}_\Gamma^n, \boldsymbol{\sigma}\boldsymbol{n}\right\rangle_{\Gamma_{\boldsymbol{u}}}.\end{aligned}$$

Since the energy $\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$ is strictly convex and the feasible set $\mathcal{S}^n\times\mathcal{F}^n$ is non-empty and convex, existence and uniqueness of a solution to (9.9) follow by classical results from convex analysis, cf. Thm. A.2. We refer to Sec. 9.2.2 for the derivation of the corresponding optimality conditions. We emphasize, that well-posedness also follows in the extreme case of an incompressible fluid, as long as $\mathcal{Z}_0\times\tilde{\mathcal{Q}}$ is inf-sup stable such that $\mathcal{F}^n$ is essentially constrained by a one-dimensional relation.

9.2 Optimality conditions

For each of the minimization formulations of the linear Biot equations presented in Sec. 9.1.1–9.1.5, the corresponding optimality conditions can be derived as the first variation. Constraints are handled via the method of Lagrange multipliers. For better illustration of the undrained and fixed-stress split in the following section, we derive the three-field and five-field formulation of the linear Biot equations below.

9.2.1 Three-field formulation of the linear Biot equations derived from minimization

We consider the minimization formulation of the linear Biot equations from Sec. 9.1.4. We recall that the mechanical displacement and volumetric flux $(\boldsymbol{u}^n, \boldsymbol{q}^n)$ are determined by minimization and the pressure p^n is defined using a post-processing. Hence, by definition of $p^n\in\tilde{\mathcal{Q}}$ and the orthogonal projection $\Pi_{\tilde{\mathcal{Q}}}$, it holds

$$\begin{aligned}\left\langle M\Pi_{\tilde{\mathcal{Q}}}(\theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\boldsymbol{\nabla}\cdot\boldsymbol{q}^n - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}^n), \Pi_{\tilde{\mathcal{Q}}}(\alpha\boldsymbol{\nabla}\cdot\boldsymbol{v})\right\rangle &= \alpha\left\langle p^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle, \\ \left\langle M\Pi_{\tilde{\mathcal{Q}}}(\theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\boldsymbol{\nabla}\cdot\boldsymbol{q}^n - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}^n), \Pi_{\tilde{\mathcal{Q}}}(\Delta t\boldsymbol{\nabla}\cdot\boldsymbol{z})\right\rangle &= \Delta t\left\langle p^n, \boldsymbol{\nabla}\cdot\boldsymbol{z}\right\rangle\end{aligned}$$

for all $(\boldsymbol{v},\boldsymbol{z})\in\mathcal{V}_0\times\mathcal{Z}_0$. From that, the optimality conditions for $(\boldsymbol{u}^n, \boldsymbol{q}^n)$ and the definition of p^n read for all $(\boldsymbol{v},\boldsymbol{z},q)\in\mathcal{V}_0\times\mathcal{Z}_0\times\tilde{\mathcal{Q}}$

$$\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}^n), \boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle - \alpha\left\langle p^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle = \mathcal{P}_{\mathrm{ext,mech}}^n(\boldsymbol{v}), \tag{9.10}$$

$$\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}^n, \boldsymbol{z}\right\rangle - \left\langle p^n, \boldsymbol{\nabla}\cdot\boldsymbol{z}\right\rangle = \mathcal{P}_{\mathrm{ext,fluid}}^n(\boldsymbol{z}), \tag{9.11}$$

$$\tfrac{1}{M}\left\langle p^n, q\right\rangle + \alpha\left\langle \boldsymbol{\nabla}\cdot\boldsymbol{u}^n, q\right\rangle + \Delta t\left\langle \boldsymbol{\nabla}\cdot\boldsymbol{q}^n, q\right\rangle = \left\langle \theta^{n-1} + \Delta t\, q_\theta^n, q\right\rangle, \tag{9.12}$$

which yields the classical three-field formulation of the linear Biot equations.

9.2.2 Five-field formulation of the linear Biot equations derived as optimality conditions

We consider the minimization formulation of the linear Biot equations from Sec. 9.1.5. First, we define function spaces for the stress variable

$$\begin{aligned} \tilde{\mathcal{S}}^n &= \left\{ \boldsymbol{\sigma} \in H(\mathrm{div};\Omega)^d \;\middle|\; \boldsymbol{\sigma}\boldsymbol{n} = \boldsymbol{\sigma}^n_{\Gamma,\mathrm{n}} \text{ on } \Gamma_{\boldsymbol{\sigma}} \right\}, \\ \tilde{\mathcal{S}}_0 &= \left\{ \boldsymbol{\sigma} \in H(\mathrm{div};\Omega)^d \;\middle|\; \boldsymbol{\sigma}\boldsymbol{n} = \boldsymbol{0} \text{ on } \Gamma_{\boldsymbol{\sigma}} \right\}. \end{aligned}$$

Additionally, we introduce the mechanical displacement $\boldsymbol{u} \in \boldsymbol{L}^2(\Omega)$, the rotation $\boldsymbol{\zeta} \in \boldsymbol{Q}_{\mathrm{AS}}$ and an artificial fluid flux $\tilde{\boldsymbol{q}} \in \mathcal{Z}^n$ as Lagrange multipliers. A suitable Lagrangian, incorporating the balance of linear momentum, weak symmetry of the stress tensor and Darcy's law, is given by

$$\begin{aligned} \mathcal{L}(\theta^{n-1}; \boldsymbol{\sigma}, \boldsymbol{q}, p, \boldsymbol{u}, \boldsymbol{\zeta}, \tilde{\boldsymbol{q}}) := {} & \mathcal{E}^{\star,\Delta t}_{\mathrm{tot}}(\theta^{n-1}; \boldsymbol{\sigma}, \boldsymbol{q}, p) + \langle \boldsymbol{\nabla} \cdot \boldsymbol{\sigma} + \boldsymbol{f}^n_{\mathrm{ext}}, \boldsymbol{u} \rangle + \langle \boldsymbol{\sigma}, \boldsymbol{\zeta} \rangle \\ & - \Delta t \left(\left\langle \boldsymbol{\kappa}^{-1} \boldsymbol{q}, \tilde{\boldsymbol{q}} \right\rangle - \langle p, \boldsymbol{\nabla} \cdot \tilde{\boldsymbol{q}} \rangle - \langle \boldsymbol{g}^n_{\mathrm{ext}}, \tilde{\boldsymbol{q}} \rangle + \langle p^n_\Gamma, \tilde{\boldsymbol{q}} \cdot \boldsymbol{n} \rangle_{\Gamma_p} \right). \end{aligned}$$

For the isotropic case, the corresponding saddle point $(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\zeta}, p, \tilde{\boldsymbol{q}}, \boldsymbol{q}) \in \tilde{\mathcal{S}}^n \times \times \boldsymbol{L}^2(\Omega) \times \boldsymbol{Q}_{\mathrm{AS}} \times \tilde{\mathcal{Q}} \times \mathcal{Z}^n \times \mathcal{Z}^n$ (omitting the index ^{n}) is characterized by

$$\begin{aligned} \langle \mathbb{A}\boldsymbol{\sigma}, \boldsymbol{\tau} \rangle + \langle \boldsymbol{u}, \boldsymbol{\nabla} \cdot \boldsymbol{\tau} \rangle + \langle \boldsymbol{\zeta}, \boldsymbol{\tau} \rangle + \tfrac{\alpha}{dK_{\mathrm{dr}}} \langle p, \mathrm{tr}\, \boldsymbol{\tau} \rangle &= \langle \boldsymbol{u}^n_\Gamma, \boldsymbol{\tau}\boldsymbol{n} \rangle_{\Gamma_{\boldsymbol{u}}}, \qquad (9.13) \\ \langle \boldsymbol{\nabla} \cdot \boldsymbol{\sigma}, \boldsymbol{v} \rangle &= - \langle \boldsymbol{f}^n_{\mathrm{ext}}, \boldsymbol{v} \rangle, \qquad (9.14) \\ \langle \boldsymbol{\sigma}, \boldsymbol{\gamma} \rangle &= 0, \qquad (9.15) \\ \left(\tfrac{1}{M} + \tfrac{\alpha^2}{K_{\mathrm{dr}}} \right) \langle p, q \rangle + \tfrac{\alpha}{dK_{\mathrm{dr}}} \langle \mathrm{tr}\, \boldsymbol{\sigma}, q \rangle + \Delta t \langle \boldsymbol{\nabla} \cdot \tilde{\boldsymbol{q}}, q \rangle &= \left\langle \theta^{n-1} + \Delta t\, q^n_\theta, q \right\rangle, \qquad (9.16) \\ \left\langle \boldsymbol{\kappa}^{-1} \boldsymbol{q}, \tilde{\boldsymbol{z}} \right\rangle - \left\langle \boldsymbol{\kappa}^{-1} \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{z}} \right\rangle &= 0, \qquad (9.17) \\ \left\langle \boldsymbol{\kappa}^{-1} \boldsymbol{q}, \boldsymbol{z} \right\rangle - \langle p, \boldsymbol{\nabla} \cdot \boldsymbol{z} \rangle &= \langle \boldsymbol{g}^n_{\mathrm{ext}}, \boldsymbol{z} \rangle - \langle p^n_\Gamma, \boldsymbol{z} \cdot \boldsymbol{n} \rangle_{\Gamma_p}. \qquad (9.18) \end{aligned}$$

for all variations $(\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{\gamma}, q, \tilde{\boldsymbol{z}}, \boldsymbol{z}) \in \tilde{\mathcal{S}}_0 \times \boldsymbol{L}^2(\Omega) \times \boldsymbol{Q}_{\mathrm{AS}} \times \tilde{\mathcal{Q}} \times \mathcal{Z}_0 \times \mathcal{Z}_0$. The artificial volumetric flux $\tilde{\boldsymbol{q}}$ can be identified with the actual volumetric flux $\boldsymbol{q}$, which finally yields the five-field formulation of the linear Biot equations.

9.3 Consequences for fully-discrete approximations

Exemplarily for all sections in Part II, we comment: The various minimization formulations of semi-discrete linear poro-(visco-thermo-)elasticity can be discretized in space by the finite element method using the conforming Galerkin method. Well-posedness of the different fully-discrete formulations follows then by the same arguments as for their semi-discrete versions. For variants described by constrained minimization, the need for inf-sup stable finite element pairs immediately translates from the continuous to the discrete setting in order to ensure non-empty feasible sets subject to one-dimensional constraints, and thereby strict convexity of the energies.

We note, the discussion on efficient splitting schemes for linear poro-(visco-thermo-)elasticity can be equally based on the semi- as well as the corresponding fully-discrete approximations.

9.4 Classical splitting schemes derived as alternating minimization

Summarizing the previous subsections: Time-discrete, linear poro-elasticity can be formulated as strictly convex, quadratic minimization problem, possibly subject to affine constraints depending on the choice of primary variables. By applying the conforming Galerkin method, the same translates to corresponding fully-discrete approximations. Various strategies can be applied to solve the resulting fully-discrete minimization problem numerically. We mention three:

(i) The corresponding optimality conditions are derived as in Sec. 9.2 and are solved in a monolithic fashion.

(ii) Popular in the poro-elasticity community, the coupled optimality conditions are solved using an iterative splitting scheme, decoupling the mechanics and fluid flow subproblems, cf., e.g., [35, 48, 37, 39, 41].

(iii) Based on the minimization formulation, some (inexact) minimization algorithm from the vast convex optimization literature, cf., e.g., [53, 58] is applied. In the poro-elasticity literature, such an approach has not yet been pursued.

Conforming simultaneously with options (ii) and (iii), we propose applying classical (exact) alternating minimization for solving linear poro-elasticity. Alternating minimization is equivalent to a two-block coordinate descent method as well as a successive two-subspace correction method for orthogonal space decompositions, alternating between minimizing the energy wrt. two different blocks of variables while constantly updating the complementary block. Partitioning the set of primal variables into a block of mechanical variables and a block of flow variables, yields a splitting scheme conforming with (ii).

As resulting schemes, we in fact obtain the previously introduced and now widely used *undrained split* and *fixed-stress split*, cf., e.g., [35, 48]. Originally, they have been physically motivated as predictor-corrector methods: For the undrained split, in the predictor step, the mechanical subproblem is solved under undrained conditions; in the corrector step, the unaltered fluid flow problem is solved with updated mechanical variables. Instead, for the fixed-stress split, in the predictor step, the fluid flow subproblem is solved under fixed volumetric stress; in the corrector step, the unaltered mechanics subproblem is solved. In order to explain their robustness and convergence properties, so far problem-specific analyses have been required, cf., e.g., [37, 41, 30, 72].

Originally, physically motivated, they are now endowed with a simple, mathematical intuition, providing an immediate understanding on their robust convergence properties. Alternating minimization exhibits guaranteed robustness under fairly weak assumptions; cf. [54, 75, 70, 71] from the perspective of block coordinate descent methods, or [76, 77, 78] from the perspective of successive subspace correction methods. Furthermore, under stronger convexity and continuity assumptions on the energy, theoretical convergence rates can be analyzed using abstract theory [79]. By improving the abstract result from the aforementioned work to constrained minimization problems in infinite dimensions, cf. Appendix B, we establish theoretical convergence rates for the undrained split and fixed-stress split consistent with problem-specific analyses in the literature, cf., e.g., [37, 41, 30, 72].

9.4.1 Derivation and analysis of the undrained split as alternating minimization

In this section, we identify the widely used undrained split [35] as alternating minimization applied to the primal two-field formulation of time-discrete, linear poro-elasticity, cf. Sec. 9.1.1. As the primal formulation is less frequently used in the literature, we illustrate the resulting scheme in the following with reference to the closely related three-field formulation, cf. Sec. 9.1.4. We note the undrained split can be in fact equivalently derived based on the three-field formulation, but the analysis requires unnecessarily more involved notation.

Alternating minimization is applied respecting the natural block structure of the problem, cf. Alg.1 for a single iteration. The first step is equivalent to solving a stabilized mechanics problem, cf. (9.10): For given $(\boldsymbol{u}^{n,i-1}, \boldsymbol{q}^{n,i-1}) \in \mathcal{V}^n \times \mathcal{Z}^n$, find $\boldsymbol{u}^{n,i} \in \mathcal{V}^n$ satisfying for all $\boldsymbol{v} \in \mathcal{V}_0$

$$\begin{aligned}\left\langle \mathbb{C}\boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i}\big), \boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle + \left\langle M\alpha^2 \operatorname{tr}\boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i} - \boldsymbol{u}^{n,i-1}\big), \operatorname{tr}\boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle& \\ -\alpha\left\langle p^{n,i-1}, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle &= \mathcal{P}^n_{\mathrm{ext,mech}}(\boldsymbol{v}),\end{aligned}$$

Algorithm 1: Single iteration of the undrained split

```
Input: $(\boldsymbol{u}^{n,i-1}, \boldsymbol{q}^{n,i-1}) \in \mathcal{V}^n \times \mathcal{Z}^n$
Determine $\boldsymbol{u}^{n,i} := \underset{\boldsymbol{u}\in\mathcal{V}^n}{\arg\min}\, \mathcal{E}_{\mathrm{tot}}^{\Delta t}(\theta^{n-1}; \boldsymbol{u}, \boldsymbol{q}^{n,i-1})$
Determine $\boldsymbol{q}^{n,i} := \underset{\boldsymbol{q}\in\mathcal{Z}^n}{\arg\min}\, \mathcal{E}_{\mathrm{tot}}^{\Delta t}(\theta^{n-1}; \boldsymbol{u}^{n,i}, \boldsymbol{q})$
```

where the pressure $p^{n,i-1}$ is formally defined, consistent with (4.6),

$$p^{n,i-1} := M\left(\theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\, \boldsymbol{\nabla}\cdot\boldsymbol{q}^{n,i-1} - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}^{n,i-1}\right).$$

The second step is equivalent to solving the fluid flow problem (9.11)–(9.12) with updated mechanical variables.

We highlight that the final iterative scheme is equivalent to the undrained split for linear poroelasticity. We establish convergence employing an abstract convergence result for alternating minimization leading to consistent previous problem-specific discussions [37, 30].

Lemma 9.1 (Linear convergence of the undrained split)**.** *The undrained split converges linearly, independent of the initial guess. Let* $\boldsymbol{U}^n := (\boldsymbol{u}^n, \boldsymbol{q}^n)$ *denote the solution of the coupled problem* (9.3) *and let* $\boldsymbol{U}^{n,i} := (\boldsymbol{u}^{n,i}, \boldsymbol{q}^{n,i})$ *denote the iterates defined by the undrained split, cf. Alg. 1. For* $i \in \mathbb{N}$, *define the errors* $\boldsymbol{e}_{\boldsymbol{u}}^{n,i} := \boldsymbol{u}^{n,i} - \boldsymbol{u}^n$, $\boldsymbol{e}_{\boldsymbol{q}}^{n,i} := \boldsymbol{q}^{n,i} - \boldsymbol{q}^n$. *Let* $|||\cdot|||$ *denote the norm induced by the quadratic part of* $\mathcal{E}_{\mathrm{tot}}^{\Delta t}$

$$|||(\boldsymbol{u},\boldsymbol{q})|||^2 := \tfrac{1}{2}\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u})\right\rangle + \tfrac{\Delta t}{2}\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{q}\right\rangle + \tfrac{M}{2}\left\|\Delta t\boldsymbol{\nabla}\cdot\boldsymbol{q} + \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}\right\|^2.$$

And let $K_{\mathrm{dr}}^\star \geq K_{\mathrm{dr}}$ *be largest constant such that*

$$K_{\mathrm{dr}}^\star \|\mathrm{tr}\,\boldsymbol{\varepsilon}(\boldsymbol{v})\|^2 \leq \left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle, \quad \forall\, \boldsymbol{v} \in \mathcal{V}_0. \tag{9.19}$$

It holds the a priori result

$$\left|\left|\left|(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \boldsymbol{e}_{\boldsymbol{q}}^{n,i})\right|\right|\right| \leq \left(\frac{\frac{\alpha^2}{K_{\mathrm{dr}}^\star}}{\frac{1}{M} + \frac{\alpha^2}{K_{\mathrm{dr}}^\star}}\right)^i \left(\mathcal{E}^{n,0} - \mathcal{E}^n\right)^{1/2}, \tag{9.20}$$

and the a posteriori result

$$\left|\left|\left|(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \boldsymbol{e}_{\boldsymbol{q}}^{n,i})\right|\right|\right| \leq \left(1 + \frac{\alpha^2}{K_{\mathrm{dr}}^\star} M\right) \left(\mathcal{E}^{n,i-1} - \mathcal{E}^{n,i}\right)^{1/2}, \tag{9.21}$$

where $\mathcal{E}^n := \mathcal{E}_{\mathrm{tot}}^{\Delta t}(\theta^{n-1}; \boldsymbol{U}^n)$, *and* $\mathcal{E}^{n,j} := \mathcal{E}_{\mathrm{tot}}^{\Delta t}(\theta^{n-1}; \boldsymbol{U}^{n,j})$, $j \in \mathbb{N}$.

Proof. We apply Lemma B.1. For this, we introduce two semi-norms

$$\begin{aligned}
\|\boldsymbol{U}\|_{1,\Delta t}^2 &:= \left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u})\right\rangle, \\
\|\boldsymbol{U}\|_{2,\Delta t}^2 &:= \Delta t\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{q}\right\rangle + \Delta t^2\left(\tfrac{1}{M} + \tfrac{\alpha^2}{K_{\mathrm{dr}}^\star}\right)^{-1}\|\boldsymbol{\nabla}\cdot\boldsymbol{q}\|^2.
\end{aligned}$$

We show that $\mathcal{E}_{\mathrm{tot}}^{\Delta t}$ is (i) strongly convex wrt. $\|\cdot\|_{1,\Delta t}$ and $\|\cdot\|_{2,\Delta t}$, (ii) $\boldsymbol{\nabla}_{\boldsymbol{u}}\mathcal{E}_{\mathrm{tot}}^{\Delta t}$ is Lipschitz continuous wrt. $\|\cdot\|_{1,\Delta t}$, and (iii) $\boldsymbol{\nabla}_{\boldsymbol{q}}\mathcal{E}_{\mathrm{tot}}^{\Delta t}$ is Lipschitz continuous $\|\cdot\|_{2,\Delta t}$. Throughout the proof, for lighter notation, we omit the explicit dependence of $\mathcal{E}_{\mathrm{tot}}^{\Delta t}$ on θ^{n-1}.

(i) Strong convexity of $\mathcal{E}_{\text{tot}}^{\Delta t}$. Let $\boldsymbol{U}_i = (\boldsymbol{v}_i, \boldsymbol{z}_i) \in \mathcal{V}^n \times \mathcal{Z}^n$, $i = 1, 2$. As $\mathcal{E}_{\text{tot}}^{\Delta t}$ is quadratic, it holds

$$\left\langle \boldsymbol{\nabla}\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}_1) - \boldsymbol{\nabla}\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}_2), \boldsymbol{U}_1 - \boldsymbol{U}_2 \right\rangle = 2\, |||\boldsymbol{U}_1 - \boldsymbol{U}_2|||^2 .$$

Clearly, $2\,|||\boldsymbol{U}|||^2 \geq \|\boldsymbol{U}\|_{1,\Delta t}^2$. By applying Young's inequality with optimally balanced weights and using (9.19), one can show for all $(\boldsymbol{u}, \boldsymbol{q}) \in \mathcal{V}_0 \times \mathcal{Z}_0$

$$\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u}) \rangle + M \left\| \Delta t \boldsymbol{\nabla} \cdot \boldsymbol{q} + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u} \right\|^2 \geq -\frac{\frac{\alpha^2}{K_{\text{dr}}^\star}}{\frac{1}{M} + \frac{\alpha^2}{K_{\text{dr}}^\star}} \left\| \Delta t \, \boldsymbol{\nabla} \cdot \boldsymbol{q} \right\|^2 . \tag{9.22}$$

Hence, $2\,|||\boldsymbol{U}_1 - \boldsymbol{U}_2|||^2 \geq \|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_{2,\Delta t}^2$. All in all, $\mathcal{E}_{\text{tot}}^{\Delta t}$ is strongly convex wrt. $\|\cdot\|_{i,\Delta t}$ with constant $\sigma_i = 1$, $i = 1, 2$.

(ii) Lipschitz continuity of $\boldsymbol{\nabla}_{\boldsymbol{u}}\mathcal{E}_{\text{tot}}^{\Delta t}$. Let $\boldsymbol{U} = (\boldsymbol{v}, \boldsymbol{z}) \in \mathcal{V}^n \times \mathcal{Z}^n$. It holds

$$\begin{aligned}
\sup_{\boldsymbol{h} \in \mathcal{V}_0} & \frac{\left\langle \boldsymbol{\nabla}_{\boldsymbol{u}}\mathcal{E}_{\text{tot}}^{\Delta t}\left(\boldsymbol{U} + (\boldsymbol{h}, \boldsymbol{0})\right) - \boldsymbol{\nabla}_{\boldsymbol{u}}\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}), (\boldsymbol{h}, \boldsymbol{0}) \right\rangle}{\|(\boldsymbol{h}, \boldsymbol{0})\|_{1,\Delta t}^2} \\
&= \sup_{\boldsymbol{h} \in \mathcal{V}_0} \frac{\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{h}), \boldsymbol{\varepsilon}(\boldsymbol{h}) \rangle + M\alpha^2 \|\boldsymbol{\nabla} \cdot \boldsymbol{h}\|^2}{\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{h}), \boldsymbol{\varepsilon}(\boldsymbol{h}) \rangle} \\
&\leq 1 + \frac{M\alpha^2}{K_{\text{dr}}^\star},
\end{aligned}$$

where we used (9.19). All in all, $\boldsymbol{\nabla}_{\boldsymbol{u}}\mathcal{E}_{\text{tot}}^{\Delta t}$ is Lipschitz continuous wrt. $\|\cdot\|_{1,\Delta t}$ with Lipschitz constant $L_1 = 1 + \frac{M\alpha^2}{K_{\text{dr}}^\star}$.

(iii) Lipschitz continuity of $\boldsymbol{\nabla}_{\boldsymbol{q}}\mathcal{E}_{\text{tot}}^{\Delta t}$. Let $\boldsymbol{U} = (\boldsymbol{v}, \boldsymbol{z}) \in \mathcal{V}^n \times \mathcal{Z}^n$. It holds

$$\begin{aligned}
\sup_{\boldsymbol{h} \in \mathcal{Z}_0} & \frac{\left\langle \boldsymbol{\nabla}_{\boldsymbol{q}}\mathcal{E}_{\text{tot}}^{\Delta t}\left(\boldsymbol{U} + (\boldsymbol{0}, \boldsymbol{h})\right) - \boldsymbol{\nabla}_{\boldsymbol{q}}\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}), (\boldsymbol{0}, \boldsymbol{h}) \right\rangle}{\|(\boldsymbol{0}, \boldsymbol{h})\|_{2,\Delta t}^2} \\
&= \sup_{\boldsymbol{h} \in \mathcal{Z}_0} \frac{\Delta t \left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{h}, \boldsymbol{h} \right\rangle + \Delta t^2 M \|\boldsymbol{\nabla} \cdot \boldsymbol{h}\|^2}{\Delta t \left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{h}, \boldsymbol{h} \right\rangle + \Delta t^2 \left(\frac{1}{M} + \frac{\alpha^2}{K_{\text{dr}}^\star} \right)^{-1} \|\boldsymbol{\nabla} \cdot \boldsymbol{h}\|^2} \\
&\leq 1 + \frac{M\alpha^2}{K_{\text{dr}}^\star}.
\end{aligned}$$

All in all, $\boldsymbol{\nabla}_{\boldsymbol{q}}\mathcal{E}_{\text{tot}}^{\Delta t}$ is Lipschitz continuous wrt. $\|\cdot\|_{2,\Delta t}$ with Lipschitz constant $L_2 = 1 + \frac{M\alpha^2}{K_{\text{dr}}^\star}$.

Consequences. By Lemma B.1, it follows

$$\begin{aligned}
\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^{n,i}) - \mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^n) &\leq \left(1 - \tfrac{1}{L_1}\right)\left(1 - \tfrac{1}{L_2}\right)\left(\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^{n,i-1}) - \mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^n)\right), \\
\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^{n,i}) - \mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^n) &\leq L_1 L_2 \left(\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^{n,i-1}) - \mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^{n,i})\right).
\end{aligned}$$

Moreover, as $\mathcal{E}_{\text{tot}}^{\Delta t}$ is quadratic and $\boldsymbol{\nabla}\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^n) = \boldsymbol{0}$, it holds

$$\mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^{n,i}) - \mathcal{E}_{\text{tot}}^{\Delta t}(\boldsymbol{U}^n) = \left|\left|\left|(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \boldsymbol{e}_{\boldsymbol{q}}^{n,i})\right|\right|\right|^2 .$$

The results (9.20) and (9.21) follow directly. □

Remark 9.3. *We emphasize that $K_{\text{dr}}^\star$ in (9.19) enters only the theoretical result, and exact knowledge is not required for practical use of the splitting scheme.*

9.4.2 Derivation and analysis of the fixed-stress split as alternating minimization

In this section, we identify the widely used fixed-stress split [35] as alternating minimization applied to the dual formulation of time-discrete, linear poro-elasticity, cf. Sec. 9.1.2. In the following, the resulting scheme is illustrated with reference to the closely related five-field formulation, cf. Sec. 9.1.5; in fact the five-field formulation (Sec. 9.1.5) can be equally used as basis leading to the same scheme.

Algorithm 2: Single iteration of the fixed-stress split

```
Input: $(\boldsymbol{\sigma}^{n,i-1}, p^{n,i-1}) \in \mathcal{S}^n \times \mathcal{Q}^n$
Determine $p^{n,i} := \underset{p\in\mathcal{Q}^n}{\arg\min}\, \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\sigma}^{n,i-1}, p)$
Determine $\boldsymbol{\sigma}^{n,i} := \underset{\boldsymbol{\sigma}\in\mathcal{S}^n}{\arg\min}\, \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\sigma}, p^{n,i})$
```

Alternating minimization is applied respecting the natural block structure of the dual problem, cf. Alg. 2 for a single iteration. It is important to note that the problem is block separable; constraints decouple into purely mechanical and fluid flow constraints. Hence, alternating minimization can be applied without violating constraints. The first step is equivalent to solving a stabilized flow problem, cf. (9.16) –(9.18): For given $(\boldsymbol{\sigma}^{n,i-1}, p^{n,i-1}) \in \mathcal{S}^n \times \mathcal{Q}^n$, find $p^{n,i} \in \mathcal{Q}^n$ satisfying for all $q \in \mathcal{Q}_0$

$$\tfrac{1}{M} \left\langle p^{n,i}, q \right\rangle + \left\langle \alpha^2 \left(\mathbf{I} : \mathbb{A} : \mathbf{I}\right) \left(p^{n,i} - p^{n,i-1}\right), q \right\rangle \tag{9.23}$$

$$+\, \alpha \left\langle \operatorname{tr} \boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i-1}, q \right\rangle + \Delta t \left\langle \boldsymbol{\nabla} \cdot \boldsymbol{q}^{n,i}, q \right\rangle = \left\langle \theta^{n-1} + \Delta t\, q_\theta^n, q \right\rangle, \tag{9.24}$$

where we have formally abbreviated the mechanical strain and the volumetric flux, consistent with (3.7) and Darcy's law

$$\boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i} := \mathbb{A}(\boldsymbol{\sigma}^{n,i} + \alpha p^{n,i}\,\mathbf{I}),$$
$$\boldsymbol{q}^{n,i} := -\boldsymbol{\kappa} \left(\boldsymbol{\nabla} p^{n,i} - \boldsymbol{g}_{\mathrm{ext}}^n\right).$$

For an isotropic, homogeneous material, the stabilization term equals

$$\tfrac{\alpha^2}{K_{\mathrm{dr}}} \left\langle p^{n,i} - p^{n,i-1}, q \right\rangle.$$

The second step is equivalent to solving the mechanics problem with updated fluid flow variables, cf. (9.13)–(9.15).

We highlight that the resulting scheme is equivalent to the fixed-stress split for linear poro-elasticity. As for the undrained split, we establish convergence based on an abstract convergence result for alternating minimization. After all, the following convergence result is consistent with results based on previous problem-specific *a priori* error analyses [37, 41, 72] and an *a posteriori* error analysis based on Ostrowski-type estimates for contraction mappings [80].

Lemma 9.2 (Linear convergence of the fixed-stress split)**.** *The fixed-stress split converges linearly, independent of the initial guess. Let $\boldsymbol{\Sigma}^n := (\boldsymbol{\sigma}^n, p^n)$ denote the solution of the coupled problem (9.4) and let $\boldsymbol{\Sigma}^{n,i} := (\boldsymbol{\sigma}^{n,i}, p^{n,i})$ denote the iterates defined by the fixed-stress split, cf. Alg. 2. For $i \in \mathbb{N}$, define the errors $\boldsymbol{e}_{\boldsymbol{\sigma}}^{n,i} := \boldsymbol{\sigma}^{n,i} - \boldsymbol{\sigma}^n$, $e_p^{n,i} := p^{n,i} - p^n$. Let $|||\cdot|||_\star$ denote the norm induced by the quadratic part of $\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$*

$$|||(\boldsymbol{\sigma}, p)|||_\star^2 := \tfrac{1}{2} \left\langle \mathbb{A}(\boldsymbol{\sigma} + \alpha p\mathbf{I}), \boldsymbol{\sigma} + \alpha p\mathbf{I} \right\rangle + \tfrac{1}{2M} \|p\|^2 + \tfrac{\Delta t}{2} \left\langle \boldsymbol{\kappa} \boldsymbol{\nabla} p, \boldsymbol{\nabla} p \right\rangle.$$

It holds the a priori result

$$\left|\left|\left|(\boldsymbol{e}_{\boldsymbol{\sigma}}^{n,i}, e_p^{n,i})\right|\right|\right|_\star \leq \left(\frac{\frac{\alpha^2}{K_{\mathrm{dr}}}}{\frac{1}{M} + \frac{\Delta t\, \kappa_{\mathrm{m}}}{C_\Omega^2} + \frac{\alpha^2}{K_{\mathrm{dr}}}} \right)^i \left(\mathcal{E}^{n,0} - \mathcal{E}^n\right)^{1/2}, \tag{9.25}$$

and the a posteriori result

$$\left|\left|\left|(\boldsymbol{e}_{\boldsymbol{\sigma}}^{n,i}, e_p^{n,i})\right|\right|\right|_\star \leq \left(1 + \frac{\alpha^2}{K_{\mathrm{dr}}} \left(\frac{1}{M} + \frac{\Delta t \kappa_m}{C_\Omega^2}\right)^{-1}\right) \left(\mathcal{E}^{n,i-1} - \mathcal{E}^{n,i}\right)^{1/2} \tag{9.26}$$

where $\mathcal{E}^n := \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\Sigma}^n)$ *and* $\mathcal{E}^{n,j} := \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\Sigma}^{n,j})$, $j \in \mathbb{N}$, *and* C_Ω *and* κ_m *denote the Poincaré constant and the smallest eigenvalue of* $\boldsymbol{\kappa}$*, respectively.*

Proof. We apply Lemma B.1. For this, we introduce two semi-norms

$$\begin{aligned}
\|\boldsymbol{\Sigma}\|_{1,\star,\Delta t}^2 &:= \tfrac{1}{M}\|p\|^2 + \Delta t \left\langle \boldsymbol{\kappa} \boldsymbol{\nabla} p, \boldsymbol{\nabla} p \right\rangle, \\
\|\boldsymbol{\Sigma}\|_{2,\star,\Delta t}^2 &:= \left\langle \mathbb{A}\boldsymbol{\sigma}, \boldsymbol{\sigma} \right\rangle - \frac{\frac{\alpha^2}{K_{\mathrm{dr}}}}{\frac{1}{M} + \frac{\Delta t \kappa_{\mathrm{m}}}{C_\Omega^2} + \frac{\alpha^2}{K_{\mathrm{dr}}}} \frac{1}{K_{\mathrm{dr}}} \|\sigma^{\mathrm{h}}\|^2,
\end{aligned}$$

where σ^h denotes the hydrostatic component of $\boldsymbol{\sigma}$. Positive semi-definiteness of $\|\cdot\|_{2,\star,\Delta \mathrm{t}}$ follows from (3.20). We show that $\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$ is (i) strongly convex wrt. $\|\cdot\|_{1,\star,\Delta t}$ and $\|\cdot\|_{2,\star,\Delta t}$, (ii) $\boldsymbol{\nabla}_p \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$ is Lipschitz continuous wrt. $\|\cdot\|_{1,\star,\Delta t}$, and (iii) $\boldsymbol{\nabla}_{\boldsymbol{\sigma}} \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$ is Lipschitz continuous $\|\cdot\|_{2,\star,\Delta t}$. Throughout the proof, for lighter notation, we omit the explicit dependence of $\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$ on θ^{n-1}.

(i) Strong convexity of $\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$. Let $\boldsymbol{\Sigma}_i = (\boldsymbol{\sigma}_i, p_i) \in \mathcal{S}^n \times \mathcal{Q}^n$, $i = 1, 2$. As $\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$ is quadratic, it holds

$$\left\langle \boldsymbol{\nabla} \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\boldsymbol{\Sigma}_1) - \boldsymbol{\nabla} \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 \right\rangle = 2 \left|\left|\left|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\right|\right|\right|_\star^2.$$

It follows directly, that $2 \left|\left|\left|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\right|\right|\right|_\star^2 \geq \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_{1,\star,\Delta t}^2$. Furthermore, utilizing the Poincaré inequality and decomposing the stress into its deviatoric and hydrostatic components, we obtain for all $(\boldsymbol{\sigma}, p) \in \tilde{\mathcal{S}}_0 \times \mathcal{Q}_0$

$$\begin{aligned}
2 \left|\left|\left|\boldsymbol{\Sigma}\right|\right|\right|_\star^2 &\geq \left\langle \mathbb{A}(\boldsymbol{\sigma} + \alpha p \mathbf{I}), \boldsymbol{\sigma} + \alpha p \mathbf{I} \right\rangle + \left(\tfrac{1}{M} + \tfrac{\Delta t \kappa_{\mathrm{m}}}{C_\Omega^2}\right) \|p\|^2 \\
&= \left\langle \mathbb{A}\boldsymbol{\sigma}, \boldsymbol{\sigma} \right\rangle + 2 \tfrac{\alpha}{K_{\mathrm{dr}}} \langle \sigma^{\mathrm{h}}, p \rangle + \left(\tfrac{1}{M} + \tfrac{\Delta t \kappa_{\mathrm{m}}}{C_\Omega^2} + \tfrac{\alpha^2}{K_{\mathrm{dr}}}\right) \|p\|^2 \\
&\geq \|\boldsymbol{\Sigma}\|_{2,\star,\Delta t}^2.
\end{aligned} \tag{9.27}$$

By applying Young's inequality adequately and rearranging terms, we obtain $2 \left|\left|\left|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\right|\right|\right|_\star^2 \geq \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_{2,\star,\Delta t}^2$. All in all, $\mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$ is strongly convex wrt. $\|\cdot\|_{i,\star,\Delta t}$ with constant $\sigma_i = 1$, $i = 1, 2$.

(ii) Lipschitz continuity of $\boldsymbol{\nabla}_p \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}$. Let $\boldsymbol{\Sigma} = (\boldsymbol{\sigma}, p) \in \mathcal{S}^n \times \mathcal{Q}^n$. It holds

$$\begin{aligned}
&\sup_{h \in \mathcal{Q}_0} \frac{\left\langle \boldsymbol{\nabla}_p \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}\left(\boldsymbol{\Sigma} + (\mathbf{0}, h)\right) - \boldsymbol{\nabla}_p \mathcal{E}_{\mathrm{tot}}^{\star,\Delta t}(\boldsymbol{\Sigma}), (\mathbf{0}, h) \right\rangle}{\|(\mathbf{0}, h)\|_{1,\star,\Delta}^2} \\
&\qquad = \sup_{h \in \mathcal{Q}_0} \frac{\frac{\alpha^2}{K_{\mathrm{dr}}} \|h\|^2 + \|(\mathbf{0}, h)\|_{1,\star,\Delta}^2}{\|(\mathbf{0}, h)\|_{1,\star,\Delta}^2}.
\end{aligned}$$

Decomposing and bounding $\|h\|$ optimally by $\|h\|$ and $\|\boldsymbol{\nabla} h\|$, using the Poincaré inequality, we obtain

$$\|h\|^2 \leq \left(\frac{1}{M} + \frac{\Delta t \kappa_\mathrm{m}}{C_\Omega^2} \right)^{-1} \|(\mathbf{0}, h)\|_{1,\star,\Delta t}^2.$$

All in all, $\boldsymbol{\nabla}_p \mathcal{E}_\mathrm{tot}^{\star,\Delta t}$ is Lipschitz continuous wrt. $\| \cdot \|_{1,\star,\Delta t}$ with Lipschitz constant $L_1 = 1 + \frac{\alpha^2}{K_\mathrm{dr}} \left(\frac{1}{M} + \frac{\Delta t \kappa_\mathrm{m}}{C_\Omega^2} \right)^{-1}$.

(iii) Lipschitz continuity of $\boldsymbol{\nabla}_{\boldsymbol{\sigma}} \mathcal{E}_\mathrm{tot}^{\star,\Delta t}$. Let $\boldsymbol{\Sigma} = (\boldsymbol{\sigma}, p) \in \mathcal{S}^n \times \mathcal{Q}^n$. It holds

$$\begin{aligned}
&\sup_{\boldsymbol{h} \in \tilde{\mathcal{S}}_0} \frac{\left\langle \boldsymbol{\nabla}_{\boldsymbol{\sigma}} \mathcal{E}_\mathrm{tot}^{\star,\Delta t} \left(\boldsymbol{\Sigma} + (\boldsymbol{h}, 0) \right) - \boldsymbol{\nabla}_{\boldsymbol{\sigma}} \mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}), (\boldsymbol{h}, 0) \right\rangle}{\|(\boldsymbol{h}, 0)\|_{2,\star,\Delta}^2} \\
&\quad = \sup_{\boldsymbol{h} \in \tilde{\mathcal{S}}_0} \frac{\frac{1}{2\mu}\|\boldsymbol{h}^\mathrm{d}\|^2 + \frac{1}{K_\mathrm{dr}}\|h^\mathrm{h}\|^2}{\frac{1}{2\mu}\|\boldsymbol{h}^\mathrm{d}\|^2 + \frac{1}{K_\mathrm{dr}} \left(1 - \frac{\frac{\alpha^2}{K_\mathrm{dr}}}{\frac{1}{M} + \frac{\Delta t \kappa_\mathrm{m}}{C_\Omega^2} + \frac{\alpha^2}{K_\mathrm{dr}}} \right) \|h^\mathrm{h}\|^2}.
\end{aligned}$$

We conclude, that $\boldsymbol{\nabla}_{\boldsymbol{\sigma}} \mathcal{E}_\mathrm{tot}^{\star,\Delta t}$ is Lipschitz continuous wrt. $\| \cdot \|_{2,\star,\Delta t}$ with Lipschitz constant $L_2 = 1 + \frac{\alpha^2}{K_\mathrm{dr}} \left(\frac{1}{M} + \frac{\Delta t \kappa_\mathrm{m}}{C_\Omega^2} \right)^{-1}$.

Consequences. By Lemma B.1, it follows

$$\begin{aligned}
\mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^{n,i}) - \mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^n) &\leq \left(1 - \tfrac{1}{L_1} \right) \left(1 - \tfrac{1}{L_2} \right) \left(\mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^{n,i-1}) - \mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^n) \right), \\
\mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^{n,i}) - \mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^n) &\leq L_1 L_2 \left(\mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^{n,i-1}) - \mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^{n,i}) \right).
\end{aligned}$$

Moreover, since $\mathcal{E}_\mathrm{tot}^{\star,\Delta t}$ is quadratic and $\left\langle \boldsymbol{\nabla} \mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^n), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^n \right\rangle \geq 0$ for all $\boldsymbol{\Sigma} \in \tilde{\mathcal{S}}^n \times \mathcal{Q}^n$, it holds

$$\mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^{n,i}) - \mathcal{E}_\mathrm{tot}^{\star,\Delta t}(\boldsymbol{\Sigma}^n) \geq \left|\left|\left| (\boldsymbol{e}_{\boldsymbol{\sigma}}^{n,i}, e_p^{n,i}) \right|\right|\right|_\star^2.$$

The results (9.25) and (9.26) follow directly. □

9.4.3 General remarks and implications

We close the section on splitting schemes for linear poro-elasticity with general remarks. Most remain true for the subsequent sections.

(i) *Order of minimization steps:* The order of the steps within the alternating minimization algorithm is not relevant for robust convergence; however, we have chosen the specific orders as above to demonstrate the closer connection to the undrained and the fixed-stress splits.

(ii) *Splitting schemes for particular formulation of the semi-discrete Biot equations:* The different, presented minimization formulations in Sec. 9.1.1–9.1.5 are all equivalent. Hence, for each specific formulation a splitting scheme can be constructed by equivalent reformulation of the splitting schemes presented above.

(iii) *Splitting schemes for fully-discrete linear Biot equations:* Fully-discrete Biot equations can be constructed by applying the conforming Galerkin method to the different minimization formulations from Sec. 9.1.1–9.1.5. In contrast to (ii), they are not equivalent. Hence, for a particular fully-discrete formulation, splitting schemes are derived from their corresponding semi-discrete versions, cf. (ii).

(iv) *Stable spatial discretization under splitting:* In practice it has been observed for the two-field saddle point formulation that inf-sup unstable pairs of finite elements are actually robust under the fixed-stress split [81]. Given that the fixed-stress split is equivalent to a two-block coordinate descent method, which converges already provided that each of the subproblems is uniquely solvable [70], this observation can now be theoretically explained. After all, it is nevertheless noteworthy that inf-sup stability can be beneficial for the performance of the fixed-stress split when applied to problems with a saddle point structure; e.g., for the two-field saddle point formulation, inf-sup stability adds artificial compressibility [72].

(v) *Different meshes for different subproblems:* The discussion in (v) also explains intuitively why splitting schemes allow the use of different meshes for different subproblems, without losing robustness [82]. In particular, the minimization structure allows for a natural development of specific two-mesh formulations retaining the symmetric character of the problem.

(vi) *Heterogeneous, anisotropic media:* The minimization structure of time-discrete poro-elasticity remains inherent for heterogeneous, anisotropic media. Consequently, alternating minimization can be again employed for constructing robust splitting schemes. In particular, convergence properties of the undrained split and fixed-stress split are retained, consistent with problem-specific analyses [41, 83].

(vii) *Inexact alternating minimization:* Clearly, instead of employing exact alternating minimization, each step may be also solved inexactly. As long as the energy is sufficiently decreased, convergence is still guaranteed. This allows for a more efficient implementation of splitting schemes employing, e.g., iterative solvers with coarse stopping criteria for each subproblem.

We will return to points (ii), (iii), and (vii) in the numerical examples in Sec. 14.

10 Robust splitting schemes for discrete linear poro-visco-elasticity

In the previous section, popular splitting schemes for linear poro-elasticity have been identified as alternating minimization applied to suitable minimization formulations of semi-discrete, linear poro-elasticity. In this section, we apply the same workflow, cf. Fig 1, and analogously derive novel extensions of the undrained and fixed-stress splits, now applicable to semi-discrete, linear poro-visco-elasticity. In this regard, we additionally establish for the first time guaranteed, linear convergence rates utilizing abstract optimization theory. After all, the key observation for the following efforts is the fact that semi-discrete, linear poro-visco-elasticity is a vectorized version of semi-discrete, linear poro-elasticity. Consequently, the subsequent discussion appears as a natural extension of Sec. 9. To highlight the analogy, we attempt to employ visually related notation.

10.1 Minimization formulations of time-discrete linear poro-visco-elasticity

We introduce two minimization formulations of time-discrete, linear poro-visco-elasticity. We obtain the primal formulation by applying the minimizing movement scheme to the primal formulation of time-continuous, linear poro-visco-elasticity (Sec. 4). A dual formulation is then proposed based on the close, structural connection between poro-visco-elasticity and poro-elasticity. Both formulations will serve as bases for the development of robust splitting schemes.

10.1.1 Primal formulation of time-discrete linear poro-visco-elasticity

The primal formulation of time-discrete, linear poro-visco-elasticity is directly obtained by applying the minimizing movement scheme to linear poro-visco-elasticity (4.1). By gathering terms, the resulting formulation can be interpreted as vectorized version of the primal formulation of time-discrete, linear poro-elasticity with a tensorial stiffness matrix (of sixth order) and Biot coefficient

$$\boldsymbol{\mathcal{C}}_{\mathrm{v}} := \begin{bmatrix} {[r]}1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \mathbb{C} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \left(\mathbb{C}_{\mathrm{v}} + \tfrac{1}{\Delta t}\mathbb{C}_{\mathrm{v}}'\right), \qquad \boldsymbol{\alpha}_{\mathrm{v}} := \begin{bmatrix} \alpha \\ \alpha_{\mathrm{v}} - \alpha \end{bmatrix},$$

such that for arbitrary $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2 \in \mathbb{R}^{d\times d}$, $\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$ is a third-order tensor and it holds

$$\begin{aligned} \boldsymbol{\mathcal{C}}_{\mathrm{v}} : \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} &= \begin{bmatrix} \mathbb{C}\left(\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\right) \\ -\mathbb{C}\left(\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\right) + \left(\mathbb{C}_{\mathrm{v}} + \tfrac{1}{\Delta t}\mathbb{C}_{\mathrm{v}}'\right)\boldsymbol{\varepsilon}_2 \end{bmatrix}, \\ \left(\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I}\right) : \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} &= \boldsymbol{\alpha}_{\mathrm{v}}^{\top} \begin{bmatrix} \operatorname{tr}\boldsymbol{\varepsilon}_1 \\ \operatorname{tr}\boldsymbol{\varepsilon}_2 \end{bmatrix}. \end{aligned}$$

Let the spaces $\mathcal{V}^n$ and $\mathcal{Z}^n$ be as defined in (9.1)–(9.2), and define additionally $\mathcal{T}^n := \mathcal{T}$. We obtain the time-discrete, primal formulation: For time step $n \geq 1$, given $\theta^{n-1}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}$, define $(\boldsymbol{u}^n, \boldsymbol{\varepsilon}_{\mathrm{v}}^n, \boldsymbol{q}^n) \in \mathcal{V}^n \times \mathcal{T}^n \times \mathcal{Z}^n$ to be the solution of the minimization problem

$$(\boldsymbol{u}^n, \boldsymbol{\varepsilon}_{\mathrm{v}}^n, \boldsymbol{q}^n) := \underset{(\boldsymbol{u}, \boldsymbol{\varepsilon}_{\mathrm{v}}, \boldsymbol{q}) \in \mathcal{V}^n \times \mathcal{T}^n \times \mathcal{Z}^n}{\arg\min} \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}(\theta^{n-1}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}; \boldsymbol{u}, \boldsymbol{\varepsilon}_{\mathrm{v}}, \boldsymbol{q}), \tag{10.1}$$

where

$$\begin{aligned} &\mathcal{E}_{\mathrm{v,tot}}^{\Delta t}(\theta^{n-1}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}; \boldsymbol{u}, \boldsymbol{\varepsilon}_{\mathrm{v}}, \boldsymbol{q}) \\ &\quad := \tfrac{1}{2} \left\langle \boldsymbol{\mathcal{C}}_{\mathrm{v}} : \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_{\mathrm{v}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_{\mathrm{v}} \end{bmatrix} \right\rangle + \tfrac{\Delta t}{2} \left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{q} \right\rangle \\ &\qquad + \tfrac{M}{2} \left\| \theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\, \boldsymbol{\nabla} \cdot \boldsymbol{q} - \left(\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I}\right) : \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_{\mathrm{v}} \end{bmatrix} \right\|^2 \\ &\qquad - \mathcal{P}_{\mathrm{ext,mech}}^n(\boldsymbol{u}) - \left\langle \tfrac{1}{\Delta t}\mathbb{C}_{\mathrm{v}}'\boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}, \boldsymbol{\varepsilon}_{\mathrm{v}} \right\rangle - \Delta t\, \mathcal{P}_{\mathrm{ext,fluid}}^n(\boldsymbol{q}), \end{aligned}$$

and set $\theta^n := \theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\, \boldsymbol{\nabla} \cdot \boldsymbol{q}^n$. Since the minimization problem is strictly convex and coercive, existence and uniqueness of a solution to (10.1) follow by classical results from convex analysis, cf. Thm. A.2.

Remark 10.1 (Explicit reduction to linear poro-elasticity)**.** *The first variation of* $\mathcal{E}_{\mathrm{v,tot}}^{\Delta t}(\theta^{n-1}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}; \boldsymbol{u}, \boldsymbol{\varepsilon}_{\mathrm{v}}, \boldsymbol{q})$ *wrt.* $\boldsymbol{\varepsilon}_{\mathrm{v}}$ *can be locally inverted for* $\boldsymbol{\varepsilon}_{\mathrm{v}}$*. Consequently, the coupled problem* (10.1) *can be easily reduced to a problem for* $(\boldsymbol{u}, \boldsymbol{q})$*. This allows in particular for reusing code written for linear poro-elasticity.*

10.1.2 Dual formulation of time-discrete linear poro-visco-elasticity

By adopting the analogy between the primal and dual formulations of time-discrete, linear poro-elasticity (Sec. 9) to its vectorized form, we introduce a natural dual minimization formulation of time-discrete, linear poro-visco-elasticity. Based on the corresponding primal formulation written as vectorized Biot equations (Sec. 10.1.1), we introduce natural dual variables $(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{\mathrm{v}}, p)$: a pair of stresses $(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{\mathrm{v}})$, consisting of the total stress $\boldsymbol{\sigma}$ and a stress-type field $\boldsymbol{\sigma}_{\mathrm{v}}$ enforcing the visco-elastic strain, and a fluid pressure p, formally related to the primal variables by

$$\begin{bmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\sigma}_{\mathrm{v}} \end{bmatrix} = \boldsymbol{\mathcal{C}}_{\mathrm{v}} : \left(\begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_{\mathrm{v}} \end{bmatrix} - \left(\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I}\right) p \right), \tag{10.2}$$

$$p = M \left(\theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\, \boldsymbol{\nabla} \cdot \boldsymbol{q} - \left(\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I}\right) : \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_{\mathrm{v}} \end{bmatrix} \right). \tag{10.3}$$

Analogous to linear poro-elasticity, constrained function spaces for the stress variables are used, with the constraints dictated by the primal formulation. Formally, it holds $\boldsymbol{\sigma}_{\mathrm{v}} = \frac{1}{\Delta t}\mathbb{C}'_{\mathrm{v}}\boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}$. Hence, given $\boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}$, we set

$$\mathcal{S}_{\mathrm{v}}^{n} := \left\{\tfrac{1}{\Delta t}\mathbb{C}'_{\mathrm{v}}\boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}\right\}. \tag{10.4}$$

Then $\mathcal{S}^n \times \mathcal{S}_{\mathrm{v}}^n \times \mathcal{Q}^n$ is a suitable function space for the dual variables $(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{\mathrm{v}}, p)$.

Let $\boldsymbol{\mathcal{A}}_{\mathrm{v}} := \boldsymbol{\mathcal{C}}_{\mathrm{v}}^{-1}$ denote the generalized compliance tensor. It satisfies for all $\boldsymbol{\sigma}, \boldsymbol{\sigma}_{\mathrm{v}}$ with deviatoric components $\boldsymbol{\sigma}^{\mathrm{d}}, \boldsymbol{\sigma}_{\mathrm{v}}^{\mathrm{d}}$ and hydrostatic components $\sigma^{\mathrm{h}}, \sigma_{\mathrm{v}}^{\mathrm{h}}$

$$\left\langle \boldsymbol{\mathcal{A}}_{\mathrm{v}} : \begin{bmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\sigma}_{\mathrm{v}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\sigma}_{\mathrm{v}} \end{bmatrix} \right\rangle = \left\langle (2\boldsymbol{\mathcal{M}})^{-1} : \begin{bmatrix} \boldsymbol{\sigma}^{\mathrm{d}} \\ \boldsymbol{\sigma}_{\mathrm{v}}^{\mathrm{d}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\sigma}^{\mathrm{d}} \\ \boldsymbol{\sigma}_{\mathrm{v}}^{\mathrm{d}} \end{bmatrix} \right\rangle + \left\langle \mathbf{K}^{-1} \begin{bmatrix} \sigma^{\mathrm{h}} \\ \sigma_{\mathrm{v}}^{\mathrm{h}} \end{bmatrix}, \begin{bmatrix} \sigma^{\mathrm{h}} \\ \sigma_{\mathrm{v}}^{\mathrm{h}} \end{bmatrix} \right\rangle \tag{10.5}$$

analogous to (3.20), where $\mathbb{1}$ is the fourth-order identity tensor, and

$$\begin{aligned} \boldsymbol{\mathcal{M}} &:= \mu \begin{bmatrix} [r]1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \mathbb{1} + \left(\mu_{\mathrm{v}} + \tfrac{1}{\Delta t}\mu'_{\mathrm{v}}\right) \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \mathbb{1}, \\ \mathbf{K} &:= K_{\mathrm{dr}} \begin{bmatrix} [r]1 & -1 \\ -1 & 1 \end{bmatrix} + \left(K_{\mathrm{dr,v}} + \tfrac{1}{\Delta t}K'_{\mathrm{dr,v}}\right) \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

Then the dual formulation of time-discrete, linear poro-visco-elasticity for time step $n \geq 1$ reads: Given $(\boldsymbol{\sigma}^{n-1}, \boldsymbol{\sigma}_{\mathrm{v}}^{n-1}, p^{n-1}) \in \mathcal{S}^{n-1} \times \mathcal{S}_{\mathrm{v}}^{n-1} \times \mathcal{Q}^{n-1}$, set

$$\begin{aligned} \theta^{n-1} &:= \tfrac{1}{M}p^{n-1} + (\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I}) : \boldsymbol{\mathcal{A}}_{\mathrm{v}} : \left(\begin{bmatrix} \boldsymbol{\sigma}^{n-1} \\ \boldsymbol{\sigma}_{\mathrm{v}}^{n-1} \end{bmatrix} + (\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I})\, p^{n-1} \right), \\ \boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1} &:= \begin{bmatrix} \mathbf{0}, & \mathbf{I} \end{bmatrix} \left(\boldsymbol{\mathcal{A}}_{\mathrm{v}} : \left(\begin{bmatrix} \boldsymbol{\sigma}^{n-1} \\ \boldsymbol{\sigma}_{\mathrm{v}}^{n-1} \end{bmatrix} + (\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I})\, p^{n-1} \right) \right), \end{aligned} \tag{10.6}$$

and define $(\boldsymbol{\sigma}^n, \boldsymbol{\sigma}_{\mathrm{v}}^n, p^n) \in \mathcal{S}^n \times \mathcal{S}_{\mathrm{v}}^n \times \mathcal{Q}^n$ to be the solution of the block-separable, constrained minimization problem

$$(\boldsymbol{\sigma}^n, \boldsymbol{\sigma}_{\mathrm{v}}^n, p^n) := \underset{(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{\mathrm{v}}, p) \in \mathcal{S}^n \times \mathcal{S}_{\mathrm{v}}^n \times \mathcal{Q}^n}{\arg\min} \mathcal{E}_{\mathrm{v,tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\sigma}, \boldsymbol{\sigma}_{\mathrm{v}}, p), \tag{10.7}$$

where

$$\begin{aligned} \mathcal{E}_{\mathrm{v,tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\sigma}, \boldsymbol{\sigma}_{\mathrm{v}}, p) :=& \tfrac{1}{2} \left\langle \boldsymbol{\mathcal{A}}_{\mathrm{v}} : \left(\begin{bmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\sigma}_{\mathrm{v}} \end{bmatrix} + (\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I})\, p \right), \begin{bmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\sigma}_{\mathrm{v}} \end{bmatrix} + (\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I})\, p \right\rangle \\ &+ \tfrac{1}{2M}\|p\|^2 + \tfrac{\Delta t}{2} \left\langle \boldsymbol{\kappa} \left(\boldsymbol{\nabla} p - \boldsymbol{g}_{\mathrm{ext}}^n \right), \boldsymbol{\nabla} p - \boldsymbol{g}_{\mathrm{ext}}^n \right\rangle \\ &- \left\langle \boldsymbol{u}_{\Gamma}^n, \boldsymbol{\sigma}\boldsymbol{n} \right\rangle_{\Gamma_{\boldsymbol{u}}} - \left\langle \theta^{n-1} + \Delta t\, q_{\theta}^n, p \right\rangle - \left\langle q_{\Gamma,\mathrm{n}}, p \right\rangle_{\Gamma_{\boldsymbol{q}}}. \end{aligned}$$

The minimization problem is strictly convex and the feasible set is non-empty and convex; existence and uniqueness of a solution to (10.7) follow by classical results from convex analysis, cf. Thm. A.2.

Remark 10.2 (Relations to previous formulations)**.** *Similar to the primal formulation, we highlight the vectorized character of* (10.7) *compared to the dual formulation of time-discrete, linear poro-elasticity* (9.4)*. Furthermore, it is evident, that including* $\boldsymbol{\sigma}_{\mathrm{v}}$ *as variable is redundant as it is determined beforehand. Hence, in practice, an equivalent formulation can be obtained by simple modification of the dual formulation of time-discrete poro-elasticity. Finally, along the lines of the five-field formulation of time-discrete, linear poro-elasticity (Sec. 9.1.5), a fully structure-preserving formulation can be also obtained for poro-visco-elasticity; for this essentially a flux variable has to be included as primary variable, and Darcy's law has to be enforced.*

10.2 Physical splitting schemes for time-discrete linear poro-visco-elasticity

Since time-discrete, linear poro-visco-elasticity is simply a vectorized generalization of time-discrete, linear poro-elasticity, the robust undrained split and fixed-stress split for poro-elasticity can be generalized to poro-visco-elasticity in a natural fashion. Again the detailed construction and analysis of the splitting schemes utilizes the natural interpretation as alternating minimization.

10.2.1 Undrained split for poro-visco-elasticity

We derive a robust splitting scheme by applying alternating minimization to the primal formulation of time-discrete poro-visco-elasticity (10.1). As before, we choose to minimize successively in the directions of the mechanical and fluid flow variables, cf. Alg. 3. The resulting scheme can be identified as undrained split for poro-visco-elasticity.

Algorithm 3: Single iteration of the undrained split for poro-visco-elasticity

```
1 Input: $(\boldsymbol{u}^{n,i-1}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i-1}, \boldsymbol{q}^{n,i-1}) \in \mathcal{V}^n \times \mathcal{T}^n \times \mathcal{Z}^n$
2 Determine $(\boldsymbol{u}^{n,i}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i}) := \underset{(\boldsymbol{u},\boldsymbol{\varepsilon}_{\mathrm{v}})\in\mathcal{V}^n\times\mathcal{T}^n}{\arg\min}\ \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}(\theta^{n-1}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}; \boldsymbol{u}, \boldsymbol{\varepsilon}_{\mathrm{v}}, \boldsymbol{q}^{n,i-1})$
3 Determine $\boldsymbol{q}^{n,i} := \underset{\boldsymbol{q}\in\mathcal{Z}^n}{\arg\min}\, \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}(\theta^{n-1}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}; \boldsymbol{u}^{n,i}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i}, \boldsymbol{q})$
```

As for linear poro-elasticity, the first step is equivalent to solving a stabilized mechanics problem: For given $(\boldsymbol{u}^{n,i-1}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i-1}, \boldsymbol{q}^{n,i-1}) \in \mathcal{V}^n \times \mathcal{T}^n \times \mathcal{Z}^n$, find $(\boldsymbol{u}^{n,i}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i}) \in \mathcal{V}^n \times \mathcal{T}^n$ satisfying for all $(\boldsymbol{v}, \boldsymbol{t}) \in \mathcal{V}_0 \times \mathcal{T}$

$$\left\langle \boldsymbol{\mathcal{C}}_{\mathrm{v}} : \begin{bmatrix} \boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i}\big) \\ \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{v}) \\ \boldsymbol{t} \end{bmatrix} \right\rangle + \left\langle M \boldsymbol{\alpha}_{\mathrm{v}} \boldsymbol{\alpha}_{\mathrm{v}}^{\top} \begin{bmatrix} \operatorname{tr} \boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i} - \boldsymbol{u}^{n,i-1}\big) \\ \operatorname{tr}\left(\boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i} - \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i-1}\right) \end{bmatrix}, \begin{bmatrix} \operatorname{tr} \boldsymbol{\varepsilon}(\boldsymbol{v}) \\ \operatorname{tr} \boldsymbol{t} \end{bmatrix} \right\rangle$$

$$- \left\langle (\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I})\, p^{n,i-1}, \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{v}) \\ \boldsymbol{t} \end{bmatrix} \right\rangle = \mathcal{P}_{\mathrm{ext,mech}}^{n}(\boldsymbol{v}) + \left\langle \tfrac{1}{\Delta t} \mathbb{C}_{\mathrm{v}}' \boldsymbol{\varepsilon}_{\mathrm{v}}^{n-1}, \boldsymbol{t} \right\rangle$$

where the pressure $p^{n,i-1}$ is formally defined, consistent with (4.6),

$$p^{n,i-1} := M \left(\theta^{n-1} + \Delta t\, q_{\theta}^{n} - \Delta t\, \boldsymbol{\nabla} \cdot \boldsymbol{q}^{n,i-1} - (\boldsymbol{\alpha}_{\mathrm{v}} \otimes \mathbf{I}) : \begin{bmatrix} \boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i-1}\big) \\ \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i-1} \end{bmatrix} \right).$$

We highlight a characteristic property: Tensorial stabilization is applied naturally. For instance, the stabilization term equals

$$\left\langle M \begin{bmatrix} \alpha^2 & \alpha(\alpha_{\mathrm{v}} - \alpha) \\ \alpha(\alpha_{\mathrm{v}} - \alpha) & (\alpha_{\mathrm{v}} - \alpha)^2 \end{bmatrix} \begin{bmatrix} \operatorname{tr} \boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i} - \boldsymbol{u}^{n,i-1}\big) \\ \operatorname{tr}\left(\boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i} - \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i-1}\right) \end{bmatrix}, \begin{bmatrix} \operatorname{tr} \boldsymbol{\varepsilon}(\boldsymbol{v}) \\ \operatorname{tr} \boldsymbol{t} \end{bmatrix} \right\rangle.$$

The second step is equivalent to solving the corresponding fluid flow problem with updated mechanical variables.

Global convergence follows immediately by abstract analysis on the two-block coordinate descent method. Furthermore, theoretical convergence rates can be derived as for linear poro-elasticity.

Lemma 10.1 (Linear convergence of the undrained split for poro-visco-elasticity)**.** *The undrained split converges linearly, independent of the initial guess. Let* $(\boldsymbol{u}^n, \boldsymbol{\varepsilon}_{\mathrm{v}}^n, \boldsymbol{q}^n)$ *denote the solution of the coupled problem* (10.1) *and let* $(\boldsymbol{u}^{n,i}, \boldsymbol{\varepsilon}_{\mathrm{v}}^{n,i}, \boldsymbol{q}^{n,i})$ *denote the iterates defined by the undrained* *split, cf. Alg. 3. For all $i \in \mathbb{N}$, define the errors $\boldsymbol{e}_{\boldsymbol{u}}^{n,i} := \boldsymbol{u}^{n,i} - \boldsymbol{u}^n$, $\boldsymbol{e}_{\boldsymbol{\varepsilon}_\mathrm{v}}^{n,i} := \boldsymbol{\varepsilon}_\mathrm{v}^{n,i} - \boldsymbol{\varepsilon}_\mathrm{v}^n$, $\boldsymbol{e}_{\boldsymbol{q}}^{n,i} := \boldsymbol{q}^{n,i} - \boldsymbol{q}^n$. Let $|||\cdot|||$ denote the norm induced by the quadratic part of $\mathcal{E}_{\mathrm{v,tot}}^{\Delta t}$*

$$|||(\boldsymbol{u}, \boldsymbol{\varepsilon}_\mathrm{v}, \boldsymbol{q})|||^2 := \tfrac{1}{2}\left\langle \boldsymbol{\mathcal{C}}_\mathrm{v} : \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_\mathrm{v} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_\mathrm{v} \end{bmatrix} \right\rangle + \tfrac{\Delta t}{2} \left\langle \boldsymbol{\kappa}^{-1} \boldsymbol{q}, \boldsymbol{q} \right\rangle$$
$$+ \tfrac{M}{2} \left\| \Delta t \boldsymbol{\nabla} \cdot \boldsymbol{q} + (\boldsymbol{\alpha}_\mathrm{v} \otimes \mathbf{I}) : \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_\mathrm{v} \end{bmatrix} \right\|^2 .$$

Let $K_\mathrm{dr}^\star$ as in (9.19), and $A_{\mathrm{K},\star}^2 := \frac{\alpha_\mathrm{v}^2}{K_\mathrm{dr,v} + \Delta t^{-1} K_\mathrm{dr,v}'} + \frac{\alpha^2}{K_\mathrm{dr}^\star}$. It holds the a priori result

$$|||(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \boldsymbol{e}_{\boldsymbol{\varepsilon}_\mathrm{v}}^{n,i}, \boldsymbol{e}_{\boldsymbol{q}}^{n,i})||| \leq \left(\frac{A_{\mathrm{K},\star}^2}{\frac{1}{M} + A_{\mathrm{K},\star}^2} \right)^i \left(\mathcal{E}^{n,0} - \mathcal{E}^n \right)^{1/2}, \tag{10.8}$$

and the a posteriori result

$$|||(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \boldsymbol{e}_{\boldsymbol{\varepsilon}_\mathrm{v}}^{n,i}, \boldsymbol{e}_{\boldsymbol{q}}^{n,i})||| \leq \left(1 + A_{\mathrm{K},\star}^2 M\right) \left(\mathcal{E}^{n,i-1} - \mathcal{E}^{n,i} \right)^{1/2}, \tag{10.9}$$

where

$$\mathcal{E}^n := \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}(\theta^{n-1}, \boldsymbol{\varepsilon}_\mathrm{v}^{n-1}; \boldsymbol{u}^n, \boldsymbol{\varepsilon}_\mathrm{v}^{n,j}, \boldsymbol{q}^n),$$
$$\mathcal{E}^{n,j} := \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}(\theta^{n-1}, \boldsymbol{\varepsilon}_\mathrm{v}^{n-1}; \boldsymbol{u}^{n,j}, \boldsymbol{\varepsilon}_\mathrm{v}^{n,j}, \boldsymbol{q}^{n,j}), \ j \in \mathbb{N}.$$

Proof. We follow the same strategy as in the proof of Lemma 9.1. Due to the similarities, we present only the main steps; we stress notation is attempted to look alike. We define two semi-norms

$$\|(\boldsymbol{u}, \boldsymbol{\varepsilon}_\mathrm{v}, \boldsymbol{q})\|_{v,1,\Delta t}^2 := \left\langle \boldsymbol{\mathcal{C}}_\mathrm{v} : \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_\mathrm{v} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_\mathrm{v} \end{bmatrix} \right\rangle,$$
$$\|(\boldsymbol{u}, \boldsymbol{\varepsilon}_\mathrm{v}, \boldsymbol{q})\|_{v,2,\Delta t}^2 := \Delta t \left\langle \boldsymbol{\kappa}^{-1} \boldsymbol{q}, \boldsymbol{q} \right\rangle + \Delta t^2 \left(\tfrac{1}{M} + A_{\mathrm{K},\star}^2 \right)^{-1} \|\boldsymbol{\nabla} \cdot \boldsymbol{q}\|^2 .$$

(i) Strong convexity of $\mathcal{E}_{\mathrm{v,tot}}^{\Delta t}$. The semi-norms $\|\cdot\|_{v,i,\Delta t}$ are chosen such that $\mathcal{E}_{\mathrm{v,tot}}^{\Delta t}$ is strongly convex with constant $\sigma_i = 1$, $i = 1, 2$. This is trivial for $\|\cdot\|_{v,1,\Delta t}$. For $\|\cdot\|_{v,2,\Delta t}$, one has to apply Young's inequality and balance weights optimally, similar to (9.22); for this, (9.19) has to be generalized: It holds

$$\left\| (\boldsymbol{\alpha}_\mathrm{v} \otimes \mathbf{I}) : \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_\mathrm{v} \end{bmatrix} \right\|^2 = \|\alpha_\mathrm{v} \operatorname{tr} \boldsymbol{\varepsilon}_\mathrm{v} + \alpha \operatorname{tr}(\boldsymbol{\varepsilon}(u) - \boldsymbol{\varepsilon}_\mathrm{v})\|^2 \tag{10.10}$$
$$\leq A_{\mathrm{K},\star}^2 \left(\left\langle \left(\tfrac{1}{\Delta t} \mathbb{C}_\mathrm{v}' + \mathbb{C}_\mathrm{v} \right) \boldsymbol{\varepsilon}_\mathrm{v}, \boldsymbol{\varepsilon}_\mathrm{v} \right\rangle + \left\langle \mathbb{C}(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{\varepsilon}_\mathrm{v}), \boldsymbol{\varepsilon}(\boldsymbol{u}) - \boldsymbol{\varepsilon}_\mathrm{v} \right\rangle \right)$$
$$= A_{\mathrm{K},\star}^2 \left\langle \boldsymbol{\mathcal{C}}_\mathrm{v} : \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_\mathrm{v} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{u}) \\ \boldsymbol{\varepsilon}_\mathrm{v} \end{bmatrix} \right\rangle, \qquad \forall (\boldsymbol{u}, \boldsymbol{\varepsilon}_\mathrm{v}) \in \mathcal{V}_0 \times \mathcal{T}.$$

(ii) Lipschitz continuity of $\boldsymbol{\nabla}_{(\boldsymbol{u},\boldsymbol{\varepsilon}_\mathrm{v})} \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}$ and $\boldsymbol{\nabla}_{\boldsymbol{q}} \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}$. By applying analogous steps as in the proof of Lemma 9.1, one can show that $\boldsymbol{\nabla}_{(\boldsymbol{u},\boldsymbol{\varepsilon}_\mathrm{v})} \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}$ and $\boldsymbol{\nabla}_{\boldsymbol{q}} \mathcal{E}_{\mathrm{v,tot}}^{\Delta t}$ are Lipschitz continuous wrt. $\|\cdot\|_{v,1,\Delta t}$ and $\|\cdot\|_{v,2,\Delta t}$, respectively, with Lipschitz constants $L_1 = L_2 = 1 + M A_{\mathrm{K},\star}^2$. For the first, utilize (10.10).

Consequences. The final thesis follows from the abstract convergence result Lemma B.1, and the fact that $\mathcal{E}_{\mathrm{v,tot}}^{\Delta t}$ is quadratic. □

Algorithm 4: Single iteration of the fixed-stress split for poro-visco-elasticity

```
Input: $(\boldsymbol{\sigma}^{n,i-1}, \boldsymbol{\sigma}_\mathrm{v}^{n,i-1}, p^{n,i-1}) \in \mathcal{S}^n \times \mathcal{S}_\mathrm{v}^n \times \mathcal{Q}^n$
Determine $p^{n,i} := \underset{p\in\mathcal{Q}^n}{\arg\min}\, \mathcal{E}_{\mathrm{v,tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\sigma}^{n,i-1}, \boldsymbol{\sigma}_\mathrm{v}^{n,i-1}, p)$
Determine $(\boldsymbol{\sigma}^{n,i}, \boldsymbol{\sigma}_\mathrm{v}^{n,i}) = \underset{(\boldsymbol{\sigma},\boldsymbol{\sigma}_\mathrm{v})\in\mathcal{S}^n\times\mathcal{S}_\mathrm{v}^n}{\arg\min}\, \mathcal{E}_{\mathrm{v,tot}}^{\star,\Delta t}(\theta^{n-1}; \boldsymbol{\sigma}, \boldsymbol{\sigma}_\mathrm{v}, p^{n,i})$
```

10.2.2 Fixed-stress split for poro-visco-elasticity

We derive a second, robust splitting scheme by applying alternating minimization to the dual formulation of time-discrete, linear poro-visco-elasticity (10.7). As before, we choose to minimize successively in the directions of mechanical and fluid flow variables, cf. Alg. 4. The resulting scheme can be interpreted as an fixed-stress split for poro-visco-elasticity.

The first step is equivalent to solving a stabilized flow problem: For given $(\boldsymbol{\sigma}^{n,i-1}, \boldsymbol{\sigma}_\mathrm{v}^{n,i-1}, p^{n,i-1}) \in \mathcal{S}^n \times \tilde{\mathcal{S}}^n \times \mathcal{Q}^n$, find $p^{n,i} \in \mathcal{Q}^n$ satisfying for all $q \in \mathcal{Q}_0$

$$\begin{aligned}\tfrac{1}{M}\left\langle p^{n,i}, q\right\rangle &+ \left\langle (\boldsymbol{\alpha}_\mathrm{v}\otimes\mathbf{I}) : \boldsymbol{\mathcal{A}}_\mathrm{v} : (\boldsymbol{\alpha}_\mathrm{v}\otimes\mathbf{I})\left(p^{n,i}-p^{n,i-1}\right), q\right\rangle \\ &+ \left\langle (\boldsymbol{\alpha}_\mathrm{v}\otimes\mathbf{I}) : \begin{bmatrix}\boldsymbol{\varepsilon}_\mathrm{v}^{n,i-1}\\ \boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i-1}\end{bmatrix}, q\right\rangle + \Delta t\left\langle \boldsymbol{\kappa}\left(\boldsymbol{\nabla}p^{n,i} - \boldsymbol{g}_\mathrm{ext}^n\right), \boldsymbol{\nabla}q\right\rangle = \left\langle \theta^{n-1} + \Delta t\, q_\theta^n, q\right\rangle,\end{aligned}$$

where we formally abbreviate the total and visco-elastic strains at the previous iteration

$$\begin{bmatrix}\boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i-1}\\ \boldsymbol{\varepsilon}_\mathrm{v}^{n,i-1}\end{bmatrix} := \boldsymbol{\mathcal{A}}_\mathrm{v} : \left(\begin{bmatrix}\boldsymbol{\sigma}^{n,i-1}\\ \boldsymbol{\sigma}_\mathrm{v}^{n,i-1}\end{bmatrix} + (\boldsymbol{\alpha}_\mathrm{v}\otimes\mathbf{I})\, p^{n,i-1}\right).$$

For homogeneous, isotropic materials, the stabilization term equals

$$\left\langle \boldsymbol{\alpha}_\mathrm{v}^\top \mathbf{K}^{-1}\boldsymbol{\alpha}_\mathrm{v}\left(p^{n,i}-p^{n,i-1}\right), q\right\rangle = \left(\tfrac{\alpha^2}{K_\mathrm{dr}} + \tfrac{\alpha_\mathrm{v}^2}{K_\mathrm{dr,v}+\Delta t^{-1}K'_\mathrm{dr,v}}\right)\left\langle p^{n,i}-p^{n,i-1}, q\right\rangle.$$

The second step is equivalent to solving the mechanics problem with updated fluid flow variables.

Linear convergence can be established based on an abstract convergence result for alternating minimization. Due to the structural similarities of semi-discrete, linear poro-visco-elasticity and poro-elasticity, the following lemma reads as corollary of Lemma 9.2.

Lemma 10.2 (Linear convergence of the fixed-stress split for poro-visco-elasticity)**.** *The fixed-stress split for poro-visco-elasticity converges linearly, independent of the initial guess. Let* $(\boldsymbol{\sigma}^n, \boldsymbol{\sigma}_\mathrm{v}^n, p^n)$ *denote the solution of the coupled problem* (10.7) *and let* $(\boldsymbol{\sigma}^{n,i}, \boldsymbol{\sigma}_\mathrm{v}^{n,i}, p^{n,i})$ *denote the iterates defined by the fixed-stress split, cf. Alg. 4. For* $i \in \mathbb{N}$*, define the errors* $\boldsymbol{e}_{\boldsymbol{\sigma}}^{n,i} := \boldsymbol{\sigma}^{n,i} - \boldsymbol{\sigma}^n$*,* $\boldsymbol{e}_{\boldsymbol{\sigma}_\mathrm{v}}^{n,i} := \boldsymbol{\sigma}_\mathrm{v}^{n,i} - \boldsymbol{\sigma}_\mathrm{v}^n$*,* $e_p^{n,i} := p^{n,i} - p^n$*. Let* $|||\cdot|||$ *denote the norm induced by the quadratic part of* $\mathcal{E}_{\mathrm{v,tot}}^{\star,\Delta t}$

$$\begin{aligned}|||(\boldsymbol{\sigma}, \boldsymbol{\sigma}_\mathrm{v}, p)|||_\star^2 := \tfrac{1}{2}\left\langle \boldsymbol{\mathcal{A}}_\mathrm{v} : \left(\begin{bmatrix}\boldsymbol{\sigma}\\ \boldsymbol{\sigma}_\mathrm{v}\end{bmatrix} + (\boldsymbol{\alpha}_\mathrm{v}\otimes\mathbf{I})\, p\right), \begin{bmatrix}\boldsymbol{\sigma}\\ \boldsymbol{\sigma}_\mathrm{v}\end{bmatrix} + (\boldsymbol{\alpha}_\mathrm{v}\otimes\mathbf{I})\, p\right\rangle \\ + \tfrac{1}{2M}\|p\|^2 + \tfrac{\Delta t}{2}\left\langle \boldsymbol{\kappa}\boldsymbol{\nabla}p, \boldsymbol{\nabla}p\right\rangle.\end{aligned}$$

Let $A_\mathrm{K}^2 := \frac{\alpha_\mathrm{v}^2}{K_\mathrm{dr,v}+\Delta t^{-1}K'_\mathrm{dr,v}} + \frac{\alpha^2}{K_\mathrm{dr}}$*. It holds the a priori result*

$$\left|\left|\left|(\boldsymbol{e}_{\boldsymbol{\sigma}}^{n,i}, \boldsymbol{e}_{\boldsymbol{\sigma}_\mathrm{v}}^{n,i}, e_p^{n,i})\right|\right|\right|_\star \leq \left(\frac{A_\mathrm{K}^2}{\frac{1}{M} + \frac{\Delta t\kappa_m}{C_\Omega^2} + A_\mathrm{K}^2}\right)^i \left(\mathcal{E}^{n,0} - \mathcal{E}^n\right)^{1/2},$$

and the a posteriori result

$$\left|\left|\left|(\boldsymbol{e}^{n,i}_{\boldsymbol{\sigma}}, \boldsymbol{e}^{n,i}_{\boldsymbol{\sigma}_\mathrm{v}}, e^{n,i}_{p})\right|\right|\right|_\star \leq \left(1 + A_\mathrm{K}^2 \left(\tfrac{1}{M} + \tfrac{\Delta t \kappa_m}{C_\Omega^2}\right)^{-1}\right) \left(\mathcal{E}^{n,i-1} - \mathcal{E}^{n,i}\right)^{1/2},$$

where

$$\begin{aligned}\mathcal{E}^{n} &:= \mathcal{E}^{\star,\Delta t}_{\mathrm{v,tot}}(\theta^{n-1}; \boldsymbol{\sigma}^n, \boldsymbol{\sigma}_\mathrm{v}^n, p^n),\\ \mathcal{E}^{n,j} &:= \mathcal{E}^{\star,\Delta t}_{\mathrm{v,tot}}(\theta^{n-1}; \boldsymbol{\sigma}^{n,j}, \boldsymbol{\sigma}_\mathrm{v}^{n,j}, p^{n,j}),\ j \in \mathbb{N},\end{aligned}$$

and C_Ω denotes a Poincaré-like constant and κ_m is the smallest eigenvalue of $\boldsymbol{\kappa}$.

Proof. We follow the same strategy as in the proof of Lemma 9.2. Due to the similarities, we present only the main steps; we stress notation is attempted to like alike. We define two semi-norms

$$\begin{aligned}\|(\boldsymbol{\sigma}, \boldsymbol{\sigma}_\mathrm{v}, p)\|^2_{1,\mathrm{v},\star,\Delta t} &:= \tfrac{1}{M}\|p\|^2 + \Delta t \left\langle \boldsymbol{\kappa}\boldsymbol{\nabla} p, \boldsymbol{\nabla} p \right\rangle,\\ \|(\boldsymbol{\sigma}, \boldsymbol{\sigma}_\mathrm{v}, p)\|^2_{2,\mathrm{v},\star,\Delta t} &:= \left\langle \boldsymbol{\mathcal{A}}_\mathrm{v} : \begin{bmatrix}\boldsymbol{\sigma}\\ \boldsymbol{\sigma}_\mathrm{v}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\sigma}\\ \boldsymbol{\sigma}_\mathrm{v}\end{bmatrix}\right\rangle - \frac{A_\mathrm{K}^2}{\frac{1}{M} + \frac{\Delta t \kappa_m}{C_\Omega^2} + A_\mathrm{K}^2} \left\langle \mathbf{K}^{-1} \begin{bmatrix}\sigma^\mathrm{d}\\ \sigma_\mathrm{v}^\mathrm{d}\end{bmatrix}, \begin{bmatrix}\sigma^\mathrm{d}\\ \sigma_\mathrm{v}^\mathrm{d}\end{bmatrix}\right\rangle.\end{aligned}$$

Positive semi-definiteness of $\|\cdot\|_{2,\mathrm{v},\star,\Delta\mathrm{t}}$ holds due to (10.5).

(i) Strong convexity of $\mathcal{E}^{\star,\Delta t}_{\mathrm{v,tot}}$. The semi-norms $\|\cdot\|_{i,\mathrm{v},\star,\Delta t}$ are chosen such that $\mathcal{E}^{\star,\Delta t}_{\mathrm{v,tot}}$ is strongly convex with constant $\sigma_i = 1$, $i = 1, 2$. This is trivial for $\|\cdot\|_{1,\mathrm{v},\star,\Delta t}$. For $\|\cdot\|_{2,\mathrm{v},\star,\Delta t}$, we employ an argument analogous to (9.27). Employing the Poincaré inequality, expanding the quadratic terms, and applying the Cauchy inequality and Young's inequality, yields for all $(\boldsymbol{\sigma}, stress_\mathrm{v}, p)$

$$\begin{aligned}&2\left|\left|\left|(\boldsymbol{\sigma}, \boldsymbol{\sigma}_\mathrm{v}, p)\right|\right|\right|^2_\star\\ &\quad\geq \left\langle \boldsymbol{\mathcal{A}}_\mathrm{v} : \begin{bmatrix}\boldsymbol{\sigma}\\ \boldsymbol{\sigma}_\mathrm{v}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\sigma}\\ \boldsymbol{\sigma}_\mathrm{v}\end{bmatrix}\right\rangle + 2\left\langle \boldsymbol{\alpha}_\mathrm{v}^\top (d\mathbf{K})^{-1} \begin{bmatrix}\sigma^\mathrm{h}\\ \sigma_\mathrm{v}^\mathrm{h}\end{bmatrix}, p\right\rangle + \left(\tfrac{1}{M} + \tfrac{\Delta t \kappa_m}{C_\Omega^2} + A_\mathrm{K}^2\right)\|p\|^2\\ &\quad\geq \left\langle \boldsymbol{\mathcal{A}}_\mathrm{v} : \begin{bmatrix}\boldsymbol{\sigma}\\ \boldsymbol{\sigma}_\mathrm{v}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\sigma}\\ \boldsymbol{\sigma}_\mathrm{v}\end{bmatrix}\right\rangle - \frac{\boldsymbol{\alpha}_\mathrm{v}^\top \mathbf{K}^{-1}\boldsymbol{\alpha}_\mathrm{v}}{\frac{1}{M} + \frac{\Delta t \kappa_m}{C_\Omega^2} + A_\mathrm{K}^2} \left\langle \mathbf{K}^{-1} \begin{bmatrix}\sigma^\mathrm{h}\\ \sigma_\mathrm{v}^\mathrm{h}\end{bmatrix}, \begin{bmatrix}\sigma^\mathrm{h}\\ \sigma_\mathrm{v}^\mathrm{h}\end{bmatrix}\right\rangle\\ &\quad= \|(\boldsymbol{\sigma}, \boldsymbol{\sigma}_\mathrm{v}, p)\|^2_{2,\mathrm{v},\star,\Delta\mathrm{t}}.\end{aligned}$$

(ii) Lipschitz continuity of $\boldsymbol{\nabla}_p \mathcal{E}^{\star,\Delta t}_{\mathrm{v,tot}}$ and $\boldsymbol{\nabla}_{(\boldsymbol{\sigma},\boldsymbol{\sigma}_\mathrm{v})} \mathcal{E}^{\star,\Delta t}_{\mathrm{tot}}$. Analogously to the proof of Lemma 9.2 it can be showed that $\boldsymbol{\nabla}_p \mathcal{E}^{\star,\Delta t}_{\mathrm{v,tot}}$ and $\boldsymbol{\nabla}_{(\boldsymbol{\sigma},\boldsymbol{\sigma}_\mathrm{v})} \mathcal{E}^{\star,\Delta t}_{\mathrm{tot}}$ are Lipschitz continuous wrt. $\|\cdot\|_{1,\mathrm{v},\star,\Delta t}$ and $\|\cdot\|_{2,\mathrm{v},\star,\Delta t}$, respectively, with Lipschitz constants $L_1 = L_2 = 1 + A_\mathrm{K}^2\left(\frac{1}{M} + \frac{\kappa_m \Delta t}{C_\Omega^2}\right)^{-1}$.

Consequences. The thesis follows by Lemma B.1 and the fact that $\mathcal{E}^{\star,\Delta t}_{\mathrm{v,tot}}$ is quadratic.

□

11 Robust splitting schemes for discrete non-linear poro-elasticity under infinitesimal strains

So far, part II dealt with quadratic minimization problems related to linear thermo-poro-visco-elasticity. In the following, we briefly demonstrate that the workflow illustrated in Fig. 1 can be likewise utilized for discussing non-linear poro-elasticity originating from convex minimization. As an example for non-quadratic, convex minimization problems, we consider non-linear poro-elasticity under infinitesimal strain (Sec. 5), and provide the first mathematically justified derivation of a fixed-stress split. Contrary to the previous sections, the coupled problem is decoupled into non-linear subproblems. Therefore, considering inexact solutions of those, different linearization techniques effectively lead to different splitting schemes. In the course of this work, we mention Newton's method and the so-called L-scheme, employing constant approximations of derivatives.

We just remark, a corresponding undrained split can be easily derived within the general framework, based on the primal formulation in Sec. 5. Employing inexact solution of the resulting non-linear subproblems by single L-scheme iterations, yields essentially a specific splitting scheme recently derived and analyzed by [30].

11.1 Minimization formulation for the three-field formulation

We recall the primal formulation (5.6) of non-linear poro-elasticity under infinitesimal strains, allowing for non-linear mechanics and fluid compressibility

$$(\dot{\boldsymbol{u}}, \dot{\boldsymbol{q}}_{\int}) = \underset{(\boldsymbol{v},\boldsymbol{z})\in\dot{\mathcal{V}}\times\dot{\mathcal{Z}}_{\int}}{\arg\min} \Big\{\mathcal{D}_{\text{fluid}}(\boldsymbol{z}) + \Big\langle \boldsymbol{\nabla}\mathcal{E}_{\text{nl}}(\boldsymbol{u}, \boldsymbol{q}_{\int}), (\boldsymbol{v}, \boldsymbol{z})\Big\rangle - \mathcal{P}_{\text{ext}}(\boldsymbol{v}, \boldsymbol{z})\Big\}.$$

A semi-discrete approximation is directly obtained by applying the minimizing movement schemes (Sec. 8). By explicit introduction of the fluid pressure consistent with (2.8), we consider the more common three-field saddle point formulation, incorporating the structural displacement, volumetric flux and fluid pressure as primary variables. All in all, we obtain a generalization of the three-field formulation of linear poro-elasticity (Sec. 9.1.4).

Reusing notation, we define the minimization formulation for time step n: Given $(\boldsymbol{u}^{n-1}, \boldsymbol{q}^{n-1}, p^{n-1}) \in \mathcal{V}^{n-1} \times \mathcal{Z}^{n-1} \times \tilde{\mathcal{Q}}$, set $\theta^{n-1} := b(p^{n-1}) + \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}^{n-1}$, and define $(\boldsymbol{u}^n, \boldsymbol{q}^n, p^n) \in \mathcal{V}^n \times \mathcal{Z}^n \times \tilde{\mathcal{Q}}$ as solution to

$$\begin{aligned} (\boldsymbol{u}^n, \boldsymbol{q}^n) &:= \underset{(\boldsymbol{u},\boldsymbol{q})\in\mathcal{V}^n\times\mathcal{Z}^n}{\arg\min}\ \mathcal{E}^{\Delta t}_{\text{nl,tot}}(\theta^{n-1}; \boldsymbol{u}, \boldsymbol{q}), && (11.1)\\ p^n &:= \Pi_{\tilde{\mathcal{Q}}}\left(b^{-1}\left(\Pi_{\tilde{\mathcal{Q}}}(\theta^{n-1} + \Delta t\, q_\theta^n - \Delta t\, \boldsymbol{\nabla}\cdot\boldsymbol{q}^n - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}^n)\right)\right), && (11.2) \end{aligned}$$

where

$$\begin{aligned} \mathcal{E}^{\Delta t}_{\text{tot}}(\theta^{n-1}; \boldsymbol{u}, \boldsymbol{q}) := &\int_\Omega W(\boldsymbol{\varepsilon}(\boldsymbol{u}))\, dx + \tfrac{\Delta t}{2}\left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}, \boldsymbol{q}\right\rangle \\ &+ \int_\Omega \int_0^{\Pi_{\tilde{\mathcal{Q}}}(\theta^{n-1}+\Delta t\, q_\theta^n - \Delta t\boldsymbol{\nabla}\cdot\boldsymbol{q} - \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u})} b^{-1}(s)\, ds\, dx \\ &- \mathcal{P}^n_{\text{ext,mech}}(\boldsymbol{u}) - \Delta t\, \mathcal{P}^n_{\text{ext,fluid}}(\boldsymbol{q}). \end{aligned}$$

Provided that W is strictly convex and b is Lipschitz continuous, the minimization problem is strictly convex. Since also the projection is well-defined; existence and uniqueness of a solution to (11.1)–(11.2) follow by classical results from convex analysis, cf. Thm. A.2. Following Sec. 9.2.1, the corresponding optimality conditions are given by

$$\begin{aligned} \left\langle \boldsymbol{\nabla}W(\boldsymbol{\varepsilon}(\boldsymbol{u}^n)), \boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle - \alpha\left\langle p^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle &= \mathcal{P}^n_{\text{ext,mech}}(\boldsymbol{v}) && \forall \boldsymbol{v}\in\mathcal{V}_0, && (11.3)\\ \left\langle \boldsymbol{\kappa}^{-1}\boldsymbol{q}^n, \boldsymbol{z}\right\rangle - \left\langle p^n, \boldsymbol{\nabla}\cdot\boldsymbol{z}\right\rangle &= \mathcal{P}^n_{\text{ext,fluid}}(\boldsymbol{z}), && \forall \boldsymbol{z}\in\mathcal{Z}_0, && (11.4)\\ \left\langle b(p^n) + \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u} + \Delta t\,\boldsymbol{\nabla}\cdot\boldsymbol{q}^n, q\right\rangle &= \left\langle \theta^{n-1} + \Delta t q_\theta^n, q\right\rangle, && \forall q\in\tilde{\mathcal{Q}}. && (11.5) \end{aligned}$$

11.2 Foundation for an exact fixed-stress split for the dual formulation

For the derivation of a fixed-stress split for the three-field formulation (11.3)–(11.5), we utilize a natural dual formulation, generalizing the dual formulation for linear poro-elasticity (Sec. 9.1.2).

For this, we first note that $\boldsymbol{\nabla} W$ is invertible for strictly convex W, and there exists a dual scalar potential $U : \mathbb{R}^{d\times d} \to \mathbb{R}$, which by the inverse function theorem satisfies for all $\boldsymbol{\sigma} \in \mathbb{R}^{d\times d}$

$$\begin{aligned}\boldsymbol{\nabla} U(\boldsymbol{\sigma}) &= (\boldsymbol{\nabla} W)^{-1}(\boldsymbol{\sigma}),\\ \boldsymbol{\nabla}^2 U(\boldsymbol{\sigma}) &= \boldsymbol{\nabla}^2 W\left((\boldsymbol{\nabla} W)^{-1}(\boldsymbol{\sigma})\right)^{-1}.\end{aligned}$$

Similarly, let $B : \mathbb{R} \to \mathbb{R}$ a primitive of b, satisfying $B' = b$.

Then the dual minimization formulation reads: Given $(\boldsymbol{\sigma}^{n-1}, p^{n-1}) \in \mathcal{S}^{n-1} \times \mathcal{Q}^{n-1}$, set $\theta^{n-1} := b(p^{n-1}) + \alpha \,\mathrm{tr}\, \boldsymbol{\nabla} U\left(\boldsymbol{\sigma}^{n-1} + \alpha p^{n-1}\mathbf{I}\right)$, and define $(\boldsymbol{\sigma}^n, p^n) \in \mathcal{S}^n \times \mathcal{Q}^n$ to be the solution of the dual minimization problem

$$(\boldsymbol{\sigma}^n, p^n) := \underset{(\boldsymbol{\sigma},p)\in\mathcal{S}^n\times\mathcal{Q}^n}{\arg\min}\, \mathcal{E}^{\star,\Delta t}_{\mathrm{nl,tot}}(\theta^{n-1}; \boldsymbol{\sigma}, p), \quad \text{where} \tag{11.6}$$

$$\begin{aligned}\mathcal{E}^{\star,\Delta t}_{\mathrm{tot}}(\theta^{n-1}; \boldsymbol{\sigma}, p) := &\int_\Omega U(\boldsymbol{\sigma} + \alpha p\mathbf{I})\, dx\\ &+ \int_\Omega B(p)\, dx + \tfrac{\Delta t}{2} \left\langle \boldsymbol{\kappa}(\boldsymbol{\nabla} p - \boldsymbol{g}^n_{\mathrm{ext}}), \boldsymbol{\nabla} p - \boldsymbol{g}^n_{\mathrm{ext}}\right\rangle\\ &- \langle \boldsymbol{u}^n_\Gamma, \boldsymbol{\sigma}\boldsymbol{n}\rangle_{\Gamma_{\boldsymbol{u}}} - \left\langle \theta^{n-1} + \Delta t\, q^n_\theta, p\right\rangle - \Delta t \left\langle q^n_{\Gamma,\mathrm{n}}, p\right\rangle_{\Gamma_q}.\end{aligned}$$

The *exact fixed-stress split* is then defined as (exact) alternating minimization applied to (11.6), cf. Alg. 5. Convergence follows directly, and theoretical convergence rates can be studied as previously. When employing inexact minimization in one of the steps, we refer to an *inexact fixed-stress split.*

Algorithm 5: Single iteration of the exact fixed-stress split for non-linear poroelasticity under infinitesimal strain

```
Input: $(\boldsymbol{\sigma}^{n,i-1}, p^{n,i-1}) \in \mathcal{S}^n \times \mathcal{Q}^n$
Determine $p^{n,i} := \arg\min_{p\in\mathcal{Q}^n} \mathcal{E}^{\star,\Delta t}_{\mathrm{nl,tot}}(\theta^{n-1}; \boldsymbol{\sigma}^{n,i-1}, p)$
Determine $\boldsymbol{\sigma}^{n,i} := \arg\min_{\boldsymbol{\sigma}\in\mathcal{S}^n} \mathcal{E}^{\star,\Delta t}_{\mathrm{nl,tot}}(\theta^{n-1}; \boldsymbol{\sigma}, p^{n,i})$
```

11.3 Fixed-stress splits for the three-field formulation of non-linear poroelasticity under infinitesimal strains

Pursuing the previous philosophy, the fixed stress split for the three-three field formulation (11.3)–(11.5) is equivalent with solving first a pressure-stabilized version of the flow problem (11.4)–(11.5), and second the mechanics problem (11.3) with updated fluid flow variables. The stabilization term can be concluded from the discussion in Sec. 11.2. In particular, the first step of the exact fixed-stress split for the dual problem, cf. Alg. 5, reads: Find $p^{n,i} \in \mathcal{Q}^n$, satisfying

$$\begin{aligned}&\left\langle b(p^{n,i}), q\right\rangle + \alpha \left\langle \mathbf{I} : \boldsymbol{\nabla} U(\boldsymbol{\sigma}^{n,i-1} + \alpha p^{n,i}\mathbf{I}), q\right\rangle\\ &\quad + \Delta t \left\langle \boldsymbol{\kappa}\boldsymbol{\nabla}(p^{n,i} - \boldsymbol{g}_{\mathrm{ext}}), \boldsymbol{\nabla} q\right\rangle = \left\langle \theta^{n-1} + \Delta t\, q^n_\theta, q\right\rangle \quad \forall q \in \mathcal{Q}_0.\end{aligned} \tag{11.7}$$

Utilizing the natural linearization of the non-linear coupling term

$$
\begin{aligned}
&\alpha \left\langle \mathbf{I} : \boldsymbol{\nabla} U(\boldsymbol{\sigma}^{n,i-1} + \alpha p^{n,i}\mathbf{I}), q \right\rangle \\
&\quad \approx \alpha \left\langle \underbrace{\mathbf{I} : \boldsymbol{\nabla} U(\boldsymbol{\sigma}^{n,i-1} + \alpha p^{n,i-1}\mathbf{I})}_{\hat{=}\operatorname{tr}\boldsymbol{\varepsilon}(\boldsymbol{u}^{n,i-1})}, q \right\rangle \\
&\qquad + \alpha^2 \left\langle \underbrace{\left(\mathbf{I} : \boldsymbol{\nabla}^2 U(\boldsymbol{\sigma}^{n,i-1} + \alpha p^{n,i-1}\mathbf{I}) : \mathbf{I}\right)}_{\hat{=}\mathbf{I}:\boldsymbol{\nabla}^2 W(\boldsymbol{\varepsilon}(\boldsymbol{u}^{n,i-1}))^{-1}:\mathbf{I}=:K_{\mathrm{dr}}(\boldsymbol{\varepsilon}(\boldsymbol{u}^{n,i-1}))^{-1}} (p^{n,i} - p^{n,i-1}), q \right\rangle
\end{aligned}
\tag{11.8}
$$

combined with different linearization techniques, we propose, two versions of the exact fixed-stress split for the three-field formulation (11.3)–(11.5). For direct comparison, we define natural residuals; for $(\boldsymbol{v}, q) \in \mathcal{V}_0 \times \tilde{\mathcal{Q}}$, let

$$
\begin{aligned}
R_{\boldsymbol{u}}^n(\boldsymbol{u}, \boldsymbol{q}, p; \boldsymbol{v}) &:= \langle \boldsymbol{\nabla} W(\boldsymbol{\varepsilon}(\boldsymbol{u})), \boldsymbol{\varepsilon}(\boldsymbol{v}) \rangle - \alpha \langle p, \boldsymbol{\nabla} \cdot \boldsymbol{v} \rangle - \mathcal{P}_{\text{ext,mech}}^n(\boldsymbol{v}), \\
R_p^n(\boldsymbol{u}, \boldsymbol{q}, p; q) &:= \left\langle b(p) + \alpha \boldsymbol{\nabla} \cdot \boldsymbol{u} + \Delta t\, \boldsymbol{\nabla} \cdot \boldsymbol{q} - \theta^{n-1} - \Delta t q_\theta^n, q \right\rangle.
\end{aligned}
$$

Newton-based fixed-stress split. In the first step, set $(\boldsymbol{q}^{n,i,0}, p^{n,i,0}) = (\boldsymbol{q}^{n,i-1}, p^{n,i-1})$, and iterate over $j \geq 1$ until convergence: Given $\boldsymbol{u}^{n,i-1} \in \mathcal{V}^n$ and $(\boldsymbol{q}^{n,i,j-1}, p^{n,i,j-1}) \in \mathcal{Z}^n \times \tilde{\mathcal{Q}}$, find $(\boldsymbol{q}^{n,i,j}, p^{n,i,j}) \in \mathcal{Z}^n \times \tilde{\mathcal{Q}}$, satisfying for all $(\boldsymbol{z}, q) \in \mathcal{Z}_0 \times \tilde{\mathcal{Q}}$

$$
\begin{aligned}
&\left\langle \boldsymbol{\kappa}^{-1} \boldsymbol{q}^{n,i,j}, \boldsymbol{z} \right\rangle - \left\langle p^{n,i,j}, \boldsymbol{\nabla} \cdot \boldsymbol{z} \right\rangle = \mathcal{P}_{\text{ext,fluid}}^n(\boldsymbol{z}), \\
&\left\langle \left[b'(p^{n,i,j-1}) + \frac{\alpha^2}{K_{\mathrm{dr}}(\boldsymbol{\varepsilon}(\boldsymbol{u}^{n,i-1}))} \right] (p^{n,i,j} - p^{n,i,j-1}), q \right\rangle \\
&\qquad + \Delta t \left\langle \boldsymbol{\nabla} \cdot (\boldsymbol{q}^{n,i,j} - \boldsymbol{q}^{n,i,j-1}), q \right\rangle = R_p^n(\boldsymbol{u}^{n,i-1}, \boldsymbol{q}^{n,i,j-1}, p^{n,i,j-1}; q).
\end{aligned}
$$

In the second step, set $\boldsymbol{u}^{n,i,0} = \boldsymbol{u}^{n,i-1}$, and iterate over $k \geq 1$ until convergence: Given $\boldsymbol{u}^{n,i,k-1} \in \mathcal{V}^n$ and $(\boldsymbol{q}^{n,i}, p^{n,i}) \in \mathcal{Z}^n \times \tilde{\mathcal{Q}}$, find $\boldsymbol{u}^{n,i,k} \in \mathcal{V}^n$, satisfying for all $\boldsymbol{v} \in \mathcal{V}_0$

$$
\left\langle \boldsymbol{\nabla}^2 W(\boldsymbol{\varepsilon}\left(\boldsymbol{u}^{n,i,k-1}\right)) \boldsymbol{\varepsilon}\left(\boldsymbol{u}^{n,i,k} - \boldsymbol{u}^{n,i,k-1}\right), \boldsymbol{\varepsilon}(\boldsymbol{v}) \right\rangle = R_u^n(\boldsymbol{u}^{n,i,k-1}, \boldsymbol{q}^{n,i}, p^{n,i}; \boldsymbol{v}).
$$

L-scheme-based fixed-stress split. Having in mind the inexact solution of non-linear subproblems, and motivated by the fact, that any fixed-stress split at most is linearly convergent, we disregard Newton's method and choose a very simple linearization instead – the so-called L-scheme, which employs a constant Jacobian. Let L_b, $L_{\mathrm{FS}} \geq 0$ and $\mathbb{L} \in \mathbb{R}^{d \times d \times d \times d}$ symmetric positive definite (in the same sense as $\mathbb{C}$). In the first step, set $(\boldsymbol{q}^{n,i,0}, p^{n,i,0}) = (\boldsymbol{q}^{n,i-1}, p^{n,i-1})$, and iterate over $j \geq 1$ until convergence: Given $\boldsymbol{u}^{n,i-1} \in \mathcal{V}^n$ and $(\boldsymbol{q}^{n,i,j-1}, p^{n,i,j-1}) \in \mathcal{Z}^n \times \tilde{\mathcal{Q}}$, find $(\boldsymbol{q}^{n,i,j}, p^{n,i,j}) \in \mathcal{Z}^n \times \tilde{\mathcal{Q}}$, satisfying for all $(\boldsymbol{z}, q) \in \mathcal{Z}_0 \times \tilde{\mathcal{Q}}$

$$
\begin{aligned}
&\left\langle \boldsymbol{\kappa}^{-1} \boldsymbol{q}^{n,i,j}, \boldsymbol{z} \right\rangle - \left\langle p^{n,i,j}, \boldsymbol{\nabla} \cdot \boldsymbol{z} \right\rangle = \mathcal{P}_{\text{ext,fluid}}^n(\boldsymbol{z}), \\
&\left\langle (L_b + L_{\mathrm{FS}})(p^{n,i,j} - p^{n,i,j-1}), q \right\rangle + \Delta t \left\langle \boldsymbol{\nabla} \cdot (\boldsymbol{q}^{n,i,j} - \boldsymbol{q}^{n,i,j-1}), q \right\rangle \\
&\qquad = R_p^n(\boldsymbol{u}^{n,i-1}, \boldsymbol{q}^{n,i,j-1}, p^{n,i,j-1}; q).
\end{aligned}
$$

In the second step, set $\boldsymbol{u}^{n,i,0} = \boldsymbol{u}^{n,i-1}$, and iterate over $k \geq 1$ until convergence: Given $\boldsymbol{u}^{n,i,k-1} \in \mathcal{V}^n$ and $(\boldsymbol{q}^{n,i}, p^{n,i}) \in \mathcal{Z}^n \times \tilde{\mathcal{Q}}$, find $\boldsymbol{u}^{n,i,k} \in \mathcal{V}^n$, satisfying for all $\boldsymbol{v} \in \mathcal{V}_0$

$$
\left\langle \mathbb{L} \boldsymbol{\varepsilon}\left(\boldsymbol{u}^{n,i,k} - \boldsymbol{u}^{n,i,k-1}\right), \boldsymbol{\varepsilon}(\boldsymbol{v}) \right\rangle = R_u^n(\boldsymbol{u}^{n,i,k-1}, \boldsymbol{q}^{n,i}, p^{n,i}; \boldsymbol{v}).
$$

Following previous studies on the L-scheme, cf., e.g. [30, 84], choosing $L_b, L_{\mathrm{FS}}, \mathbb{L}$ sufficiently large may be expected to yield robust convergence. For instance, for Lipschitz continuous non-linearities, the Lipschitz constants are suitable candidates; or solution-dependent choices as $L_b = \max\limits_{x,t} |b'(p(x,t))|$ and $L_{\mathrm{FS}} = \frac{\alpha^2}{\min\limits_{x,t} K_{\mathrm{dr}}(\boldsymbol{\varepsilon}(\boldsymbol{u}(x,t)))}$.

Remark 11.1 (Inexact fixed-stress splits). *By choosing coarse tolerances or applying only a fixed amount of non-linear iterations in each of the two steps yields an inexact version of the fixed-stress split. In particular, for strongly coupled problems, one can expect that inexact fixed-stress splits are potentially more efficient than the exact fixed-stress split.*

12 Robust splitting schemes for discrete thermo-poro-elasticity

As a result of the gradient flow structure of linear thermo-poro-elasticity, cf. Sec. 7, robust iterative splitting schemes can be derived for the implicit Euler time-discrete approximation employing the workflow visualized in Fig. 1. We in particular observe that semi-discrete, linear thermo-poro-elasticity can be formulated as vectorized, semi-discrete, linear poro-elasticity, similar to linear poro-visco-elasticity (Sec. 10). Thus, technicalities can be immediately adopted from linear poro-elasticity including the construction of a dual problem, the derivation of two-stage splitting schemes and their analyses. After all, we identify the recently proposed undrained-adiabatic and extended fixed-stress splits proposed for non-linear thermo-poro-elasticity [45] as alternating minimization. This new perspective endows the originally physically motivated schemes with mathematical justification. Motivated by the three-way coupling of thermo-poro-elasticity, we also derive a novel, robust three-stage splitting scheme by applying a cyclic three-block coordinate descent method. Finally, we close the section, commenting on possible applications of the splitting schemes to non-linear thermo-poro-elasticity including for instance thermal convection.

12.1 Minimization formulations for time-discrete linear thermo-poro-elasticity

Following the abstract workflow visualized in Fig. 1, we introduce a primal and a dual formulation for time-discrete linear thermo-poro-elasticity. In the second part of this section, both formulations will serve as bases for the derivation of practical operator splitting schemes.

12.1.1 Primal formulation of time-discrete linear thermo-poro-elasticity

The primal formulation of time-discrete, linear thermo-poro-elasticity is obtained by applying the minimizing movement scheme to the time-continuous model (7.8)–(7.10). Similar to the case of semi-discrete poro-visco-elasticity, the resulting formulation can be interpreted as vectorized version of the primal formulation of time-discrete, linear poro-elasticity, but now with a vectorized flow problem – a key characteristic which will be utilized throughout the entire section. For this, we introduce a tensorial diffusion, compressibility and Biot coefficient, respectively, by

$$\mathbf{K}_\mathrm{T} := \begin{bmatrix} \boldsymbol{\kappa} & \mathbf{0} \\ \mathbf{0} & \frac{\boldsymbol{\kappa}_\mathrm{F}}{T_0} \end{bmatrix}, \qquad \mathbf{M}_\mathrm{T}^{-1} := \begin{bmatrix} [c]\frac{1}{M} & -3\alpha_\phi \\ -3\alpha_\phi & \frac{C_\mathrm{d}}{T_0} \end{bmatrix}, \qquad \boldsymbol{\alpha}_\mathrm{T} := \begin{bmatrix} \alpha \\ 3\alpha_\mathrm{T} K_\mathrm{dr} \end{bmatrix}.$$

Let the spaces $\mathcal{V}^n$ and $\mathcal{Z}^n$ be as defined in (9.1)–(9.2), and define additionally

$$\mathcal{W}^n := \{ \boldsymbol{w} \in H(\mathrm{div};\Omega) \mid \boldsymbol{w} \cdot \boldsymbol{n} = j_\Gamma^n \text{ on } \Gamma_{\boldsymbol{j}} \}.$$

Finally, we state the time-discrete, primal formulation for time step $n \geq 1$: Given θ^{n-1} and S^{n-1}, define $(\boldsymbol{u}^n, \boldsymbol{q}^n, \boldsymbol{j}^n) \in \mathcal{V}^n \times \mathcal{Z}^n \times \mathcal{W}^n$ to be the solution of the minimization problem

$$(\boldsymbol{u}^n, \boldsymbol{q}^n, \boldsymbol{j}^n) := \underset{(\boldsymbol{u},\boldsymbol{q},\boldsymbol{j}) \in \mathcal{V}^n \times \mathcal{Z}^n \times \mathcal{W}^n}{\arg\min} \; \mathcal{E}_{\mathrm{th,tot}}^{\Delta t}(\theta^{n-1}, S^{n-1}; \boldsymbol{u}, \boldsymbol{q}, \boldsymbol{j}), \tag{12.1}$$

where

$$\begin{aligned}
&\mathcal{E}^{\Delta t}_{\mathrm{th,tot}}(\theta^{n-1}, S^{n-1}; \boldsymbol{u}, \boldsymbol{q}, \boldsymbol{j}) \\
&\quad := \frac{1}{2}\left\langle \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{u}) \right\rangle + \frac{\Delta t}{2}\left\langle \mathbf{K}_{\mathrm{T}}^{-1}\begin{bmatrix}\boldsymbol{q}\\ \boldsymbol{j}\end{bmatrix}, \begin{bmatrix}\boldsymbol{q}\\ \boldsymbol{j}\end{bmatrix}\right\rangle \\
&\quad + \frac{1}{2}\left\langle \mathbf{M}_{\mathrm{T}}\left(\begin{bmatrix}{[l]}\theta^{n-1}{+}\Delta t\, q_\theta^n\\ S^{n-1}\ {+}\Delta t\, q_S^n\end{bmatrix} - \Delta t\begin{bmatrix}\boldsymbol{\nabla}\cdot\boldsymbol{q}\\ \boldsymbol{\nabla}\cdot\boldsymbol{j}\end{bmatrix} - (\boldsymbol{\alpha}_{\mathrm{T}}\otimes\mathbf{I}) : \boldsymbol{\varepsilon}(\boldsymbol{u})\right),\right. \\
&\qquad\qquad\qquad \left.\begin{bmatrix}{[l]}\theta^{n-1}{+}\Delta t\, q_\theta^n\\ S^{n-1}\ {+}\Delta t\, q_S^n\end{bmatrix} - \Delta t\begin{bmatrix}\boldsymbol{\nabla}\cdot\boldsymbol{q}\\ \boldsymbol{\nabla}\cdot\boldsymbol{j}\end{bmatrix} - (\boldsymbol{\alpha}_{\mathrm{T}}\otimes\mathbf{I}) : \boldsymbol{\varepsilon}(\boldsymbol{u})\right\rangle \\
&\quad - \mathcal{P}^n_{\mathrm{ext,mech}}(\boldsymbol{u}) - \Delta t\, \mathcal{P}^n_{\mathrm{ext,fluid}}(\boldsymbol{q}) - \Delta t\, \mathcal{P}^n_{\mathrm{ext,temp}}(\boldsymbol{j}),
\end{aligned}$$

and set

$$\begin{bmatrix}\theta^n\\ S^n\end{bmatrix} := \begin{bmatrix}\theta^{n-1}{+}\Delta t\, q_\theta^n\\ S^{n-1}{+}\Delta t\, q_S^n\end{bmatrix} - \Delta t\begin{bmatrix}\boldsymbol{\nabla}\cdot\boldsymbol{q}^n\\ \boldsymbol{\nabla}\cdot\boldsymbol{j}^n\end{bmatrix}.$$

For $\mathbf{K}_{\mathrm{T}}$ and $\mathbf{M}_{\mathrm{T}}$ positive definite, the resulting minimization problem is strictly convex and coercive; existence and uniqueness of a solution to (12.1) follow by classical results from convex analysis, cf. Thm. A.2.

12.1.2 Dual formulation of time-discrete linear thermo-poro-elasticity

Given the primal formulation of time-discrete, linear thermo-poro-elasticity in vectorized form, we utilize the insights gained from linear poro-elasticity and poro-visco-elasticity and impose a corresponding dual formulation. First, we introduce natural dual variables: the total stress $\boldsymbol{\sigma}$, the fluid pressure p and the temperature of the bulk T, formally related to the primal variables by

$$\begin{aligned}
\boldsymbol{\sigma} &= \mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - (\mathbf{I}\otimes\boldsymbol{\alpha}_{\mathrm{T}}) : \begin{bmatrix}p\\ T\end{bmatrix}\right), \\
\begin{bmatrix}p\\ T\end{bmatrix} &= \mathbf{M}_{\mathrm{T}}\left(\begin{bmatrix}\theta^{n-1}{+}\Delta t\, q_\theta^n\\ S^{n-1}{+}\Delta t\, q_S^n\end{bmatrix} - \Delta t\begin{bmatrix}\boldsymbol{\nabla}\cdot\boldsymbol{q}\\ \boldsymbol{\nabla}\cdot\boldsymbol{j}\end{bmatrix} - (\boldsymbol{\alpha}_{\mathrm{T}}\otimes\mathbf{I}) : \boldsymbol{\varepsilon}(\boldsymbol{u})\right).
\end{aligned}$$

For fixed time step n, we introduce suitable trial and test function spaces

$$\begin{aligned}
\mathcal{R}^n &:= \left\{ r \in H^1(\Omega) \,|\, r = T^n_\Gamma \text{ on } \Gamma_T \right\}, \\
\mathcal{R}_0 &:= \left\{ r \in H^1(\Omega) \,|\, r = 0 \text{ on } \Gamma_T \right\},
\end{aligned}$$

corresponding to the temperature variable. Then $\mathcal{S}^n \times \mathcal{Q}^n \times \mathcal{R}^n$ yields a suitable function space for the dual variables $(\boldsymbol{\sigma}, p, T)$.

Finally, the dual formulation of time-discrete, linear thermo-poro-elasticity for time step $n \geq 1$ reads: Given $(\boldsymbol{\sigma}^{n-1}, p^{n-1}, T^{n-1}) \in \mathcal{S}^{n-1} \times \mathcal{Q}^{n-1} \times \mathcal{R}^{n-1}$, set

$$\begin{aligned}
\boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n-1} &:= \mathbb{A}\left(\boldsymbol{\sigma}^{n-1} + (\mathbf{I}\otimes\boldsymbol{\alpha}_{\mathrm{T}}) : \begin{bmatrix}p^{n-1}\\ T^{n-1}\end{bmatrix}\right), \\
\begin{bmatrix}\theta^{n-1}\\ S^{n-1}\end{bmatrix} &:= \mathbf{M}_{\mathrm{T}}^{-1}\begin{bmatrix}p^{n-1}\\ T^{n-1}\end{bmatrix} + (\boldsymbol{\alpha}_{\mathrm{T}}\otimes\mathbf{I}) : \boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n-1}\big),
\end{aligned}$$

and define $(\boldsymbol{\sigma}^n, p^n, T^n) \in \mathcal{S}^n \times \mathcal{Q}^n \times \mathcal{R}^n$ to be the solution of the block-separable, constrained minimization problem

$$(\boldsymbol{\sigma}^n, p^n, T^n) := \underset{(\boldsymbol{\sigma},p,T)\in\mathcal{S}^n\times\mathcal{Q}^n\times\mathcal{R}^n}{\arg\min} \mathcal{E}^{\star,\Delta t}_{\mathrm{th,tot}}(\theta^{n-1}, S^{n-1}; \boldsymbol{\sigma}, p, T), \quad \text{where} \tag{12.2}$$

$$\begin{aligned}
&\mathcal{E}^{\star,\Delta t}_{\mathrm{th,tot}}(\theta^{n-1}, S^{n-1}; \boldsymbol{\sigma}, p, T) \\
&\quad := \frac{1}{2}\left\langle \mathbb{A}\left(\boldsymbol{\sigma} + (\mathbf{I}\otimes\boldsymbol{\alpha}_{\mathrm{T}}) : \begin{bmatrix} p \\ T \end{bmatrix}\right), \boldsymbol{\sigma} + (\mathbf{I}\otimes\boldsymbol{\alpha}_{\mathrm{T}}) : \begin{bmatrix} p \\ T \end{bmatrix}\right\rangle \\
&\quad + \frac{1}{2}\left\langle \mathbf{M}_{\mathrm{T}}^{-1}\begin{bmatrix} p \\ T \end{bmatrix}, \begin{bmatrix} p \\ T \end{bmatrix}\right\rangle + \frac{\Delta t}{2}\left\langle \mathbf{K}_{\mathrm{T}}\begin{bmatrix} [l]\boldsymbol{\nabla} p - \boldsymbol{g}^n_{\mathrm{ext}} \\ \boldsymbol{\nabla} T \end{bmatrix}, \begin{bmatrix} [l]\boldsymbol{\nabla} p - \boldsymbol{g}^n_{\mathrm{ext}} \\ \boldsymbol{\nabla} T \end{bmatrix}\right\rangle \\
&\quad - \langle \boldsymbol{u}^n_\Gamma, \boldsymbol{\sigma}\boldsymbol{n}\rangle_{\Gamma_{\boldsymbol{u}}} - \left\langle \begin{bmatrix} \theta^{n-1} + \Delta t\, q^n_\theta \\ S^{n-1} + \Delta t\, q^n_S \end{bmatrix}, \begin{bmatrix} p \\ T \end{bmatrix}\right\rangle - \Delta t \left\langle q^n_{\Gamma,\mathrm{n}}, p\right\rangle_{\Gamma_{\boldsymbol{q}}} - \Delta t \left\langle j^n_{\mathrm{F},\Gamma}, T\right\rangle_{\Gamma_{\boldsymbol{j}}}.
\end{aligned}$$

The minimization problem is strictly convex and the feasible set is non-empty and convex; existence and uniqueness of a solution to (12.2) follow by classical results from convex analysis, cf. Thm. A.2.

12.2 Splitting schemes for linear thermo-poro-elasticity derived as alternating minimization

Due to the convexity properties, any cyclic block coordinate descent method applied to either the primal or the dual formulation, which respects the block structure of the problem, is globally convergent [70, 71]; in particular two- and three-block coordinate descent methods, decoupling the fully-coupled problem into its physical subproblems. Based on that fact, we derive the undrained-adiabatic and extended fixed-stress splits [45] as two-block coordinate descent methods, following the abstract workflow, cf. Fig. 1, and additionally propose a robust three-block coordinate descent method for linear thermo-poro-elasticity. Theoretical convergence can be showed by adjusting the proofs for the corresponding results in the context of linear poro-elasticity.

12.2.1 Undrained-adiabatic split based on primal thermo-poro-elasticity

Applying alternating minimization to the primal formulation of semi-discrete thermo-poro-elasticity yields a generalized undrained split, decoupling the mechanics problem from the rest. For this, the primal variables corresponding to the fluid flow and thermal subproblems are considered a single block, cf. Alg. 6 for a single iteration of the resulting scheme.

Algorithm 6: Single iteration of undrained-adiabatic split

```
Input: $(\boldsymbol{u}^{n,i-1}, \boldsymbol{q}^{n,i-1}, \boldsymbol{j}^{n,i-1}) \in \mathcal{V}^n \times \mathcal{Z}^n \times \mathcal{W}^n$
Determine $\boldsymbol{u}^{n,i} := \underset{\boldsymbol{u}\in\mathcal{V}^n}{\arg\min}\, \mathcal{E}^{\Delta t}_{\mathrm{th,tot}}(\theta^{n-1}, S^{n-1}; \boldsymbol{u}, \boldsymbol{q}^{n,i-1}, \boldsymbol{j}^{n,i-1})$
Determine $(\boldsymbol{q}^{n,i}, \boldsymbol{j}^{n,i}) := \underset{(\boldsymbol{q},\boldsymbol{j})\in\mathcal{Z}^n\times\mathcal{W}^n}{\arg\min} \mathcal{E}^{\Delta t}_{\mathrm{th,tot}}(\theta^{n-1}, S^{n-1}; \boldsymbol{u}^{n,i}, \boldsymbol{q}, \boldsymbol{j})$
```

By construction, the resulting splitting scheme is equivalent to a predictor-corrector method, solving the mechanics problem under undrained and adiabatic conditions in the predictor step. This is equivalent to the stabilized mechanics problem: Find $\boldsymbol{u}^{n,i} \in \mathcal{V}^n$ satisfying for all $\boldsymbol{v} \in \mathcal{V}_0$

$$\begin{aligned}
\left\langle \mathbb{C}\boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i}\big), \boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle + \left\langle \boldsymbol{\alpha}_{\mathrm{T}}^\top \mathbf{M}_{\mathrm{T}} \boldsymbol{\alpha}_{\mathrm{T}}\, \mathrm{tr}\,\boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i} - \boldsymbol{u}^{n,i-1}\big), \mathrm{tr}\,\boldsymbol{\varepsilon}(\boldsymbol{v})\right\rangle& \\
- \left\langle \boldsymbol{\alpha}_{\mathrm{T}}^\top \begin{bmatrix} p^{n,i-1} \\ T^{n,i-1} \end{bmatrix}, \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle &= \mathcal{P}^n_{\mathrm{ext,mech}}(\boldsymbol{v}),
\end{aligned}$$

where we formally abbreviated the fluid pressure and temperature by

$$\begin{bmatrix} p^{n,i-1} \\ T^{n,i-1} \end{bmatrix} := \mathbf{M}_\mathrm{T} \left(\begin{bmatrix} \theta^{n-1} {+} \Delta t\, q_\theta^n \\ S^{n-1} {+} \Delta t\, q_S^n \end{bmatrix} - \Delta t \begin{bmatrix} \boldsymbol{\nabla} \cdot \boldsymbol{q}^{n,i-1} \\ \boldsymbol{\nabla} \cdot \boldsymbol{j}^{n,i-1} \end{bmatrix} - (\boldsymbol{\alpha}_\mathrm{T} \otimes \mathbf{I}) : \boldsymbol{\varepsilon}\big(\boldsymbol{u}^{n,i-1}\big) \right).$$

For homogeneous, isotropic materials, the stabilization equals

$$\left(M\alpha^2 + 9 \frac{\big(\alpha_\mathrm{T} K_\mathrm{dr} + M \alpha \alpha_\phi\big)^2}{\frac{C_\mathrm{d}}{T_0} - 9 M \alpha_\phi^2} \right) \left\langle \boldsymbol{\nabla} \cdot (\boldsymbol{u}^{n,i} - \boldsymbol{u}^{n,i-1}), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle.$$

The second step of Alg. 6 (the corrector step) is equivalent to solving the unmodified, coupled fluid flow and thermal subproblems with updated displacement. After all, the resulting stabilization term is identical with that employed within the undrained-adiabatic split for thermo-poro-elasticity including thermal convection [45].

By adopting the ideas of the proof for the undrained split for poro-elasticity, cf. Lemma 9.1, to vectorized poro-elasticity, analogous convergence results can be deduced for the undrained-adiabatic split for linear thermo-poro-elasticity.

Corollary 12.1 (Linear convergence of the undrained-adiabatic split)**.** *The undrained-adiabatic split for linear thermo-poro-elasticity converges linearly, independent of the initial guess . Let* $\boldsymbol{e}_{\boldsymbol{j}}^{n,i} := \boldsymbol{j}^{n,i} - \boldsymbol{j}^n$*,* $n, i \in \mathbb{N}$*, and let* $|||\cdot|||$ *denote the norm induced by the quadratic part of* $\mathcal{E}_{\mathrm{th,tot}}^{\Delta t}$*. Let* $K_\mathrm{dr}^\star$ *as in* (9.19)*. It holds the a priori result*

$$\left|\left|\left| (\boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \boldsymbol{e}_{\boldsymbol{q}}^{n,i}, \boldsymbol{e}_{\boldsymbol{j}}^{n,i}) \right|\right|\right| \leq \left(\frac{\frac{|\boldsymbol{\alpha}_\mathrm{T}|^2}{K_\mathrm{dr}^\star}}{\frac{|\boldsymbol{\alpha}_\mathrm{T}|^2}{\boldsymbol{\alpha}_\mathrm{T}^\top \mathbf{M}_\mathrm{T} \boldsymbol{\alpha}_\mathrm{T}} + \frac{|\boldsymbol{\alpha}_\mathrm{T}|^2}{K_\mathrm{dr}^\star}} \right)^i \left(\mathcal{E}^{n,0} - \mathcal{E}^n \right)^{1/2},$$

where $\mathcal{E}^{n,0}$ *and* $\mathcal{E}^n$ *are the energies of the initial iterate and the solution, resp.*

12.2.2 Extended fixed-stress split based on dual thermo-poro-elasticity

A generalized fixed-stress split for thermo-poro-elasticity is derived by applying alternating minimization to the dual formulation of time-discrete thermo-poro-elasticity (12.2). For this, the energy is successively minimized for fixed total stress, and simultaneously fixed fluid pressure and temperature variables; cf. Alg. 7 for a single iteration of the resulting scheme.

Algorithm 7: Single iteration of the extended fixed-stress split

```
1 Input: $(\boldsymbol{\sigma}^{n,i-1}, p^{n,i-1}, T^{n,i-1}) \in \mathcal{S}^n \times \mathcal{Q}^n \times \mathcal{R}^n$
2 Determine $(p^{n,i}, T^{n,i}) := \underset{(p,T) \in \mathcal{Q}^n \times \mathcal{R}^n}{\arg\min}\ \mathcal{E}_{\mathrm{th,tot}}^{\star,\Delta t}(\theta^{n-1}, S^{n-1}; \boldsymbol{\sigma}^{n,i-1}, p, T)$
3 Determine $\boldsymbol{\sigma}^{n,i} := \underset{\boldsymbol{\sigma} \in \mathcal{S}^n}{\arg\min}\, \mathcal{E}_{\mathrm{th,tot}}^{\star,\Delta t}(\theta^{n-1}, S^{n-1}; \boldsymbol{\sigma}, p^{n,i}, T^{n,i})$
```

By construction the generalized fixed-stress split is equivalent to a predictor-corrector method, simultaneously solving the coupled fluid flow and thermal subproblems under fixed stress conditions in the predictor step. This is equivalent to the stabilized problem: Find $(p^{n,i}, T^{n,i}) \in \mathcal{Q}^n \times \mathcal{R}^n$ satisfying for all $(q, r) \in \mathcal{Q}_0 \times \mathcal{R}_0$

$$\begin{aligned} &\left\langle \mathbf{M}_\mathrm{T}^{-1} \begin{bmatrix} p^{n,i} \\ T^{n,i} \end{bmatrix}, \begin{bmatrix} q \\ r \end{bmatrix} \right\rangle + \left\langle \boldsymbol{\alpha}_\mathrm{T} \boldsymbol{\alpha}_\mathrm{T}^\top \ (\mathbf{I} : \mathbb{A} : \mathbf{I}) \begin{bmatrix} p^{n,i} - p^{n,i-1} \\ T^{n,i} - T^{n,i-1} \end{bmatrix}, \begin{bmatrix} q \\ r \end{bmatrix} \right\rangle \\ &\quad + \left\langle (\boldsymbol{\alpha}_\mathrm{T} \otimes \mathbf{I}) : \boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i-1}, \begin{bmatrix} q \\ r \end{bmatrix} \right\rangle + \Delta t \left\langle \mathbf{K}_\mathrm{T} \begin{bmatrix} [l] \boldsymbol{\nabla} p^{n,i} - \boldsymbol{g}_\mathrm{ext}^n \\ \boldsymbol{\nabla} T^{n,i} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\nabla} q \\ \boldsymbol{\nabla} r \end{bmatrix} \right\rangle \\ &\quad = \left\langle \begin{bmatrix} \theta^{n-1} {+} \Delta t\, q_\theta^n \\ S^{n-1} {+} \Delta t\, q_S^n \end{bmatrix}, \begin{bmatrix} q \\ r \end{bmatrix} \right\rangle + \Delta t \left\langle q_{\Gamma,\mathrm{n}}^n, q \right\rangle_{\Gamma_{\boldsymbol{q}}} + \Delta t \left\langle j_{\Gamma,\mathrm{n}}^n, r \right\rangle_{\Gamma_{\boldsymbol{j}}}, \end{aligned}$$

where we used the formal abbreviation of the mechanical strain

$$\boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i-1} := \mathbb{A}\left(\boldsymbol{\sigma}^{n,i-1} + (\mathbf{I}\otimes\boldsymbol{\alpha}_{\mathrm{T}}) : \begin{bmatrix} p^{n,i-1} \\ T^{n,i-1}\end{bmatrix}\right).$$

A characteristic property: Tensorial stabilization is applied. For instance, for a homogeneous, isotropic material, the stabilization term equals

$$\left\langle \begin{bmatrix} \frac{\alpha^2}{K_{\mathrm{dr}}} & 3\alpha\alpha_{\mathrm{T}} \\ 3\alpha\alpha_{\mathrm{T}} & 9\alpha_{\mathrm{T}}^2 K_{\mathrm{dr}} \end{bmatrix} \begin{bmatrix} p^{n,i} - p^{n,i-1} \\ T^{n,i} - T^{n,i-1} \end{bmatrix}, \begin{bmatrix} q \\ r \end{bmatrix} \right\rangle.$$

The second step of Alg. 7 (the corrector step) is equivalent to solving the unmodified, mechanical problem with updated pressure and temperature. The resulting stabilization terms are identical with those employed within the extended fixed-stress split for thermo-poro-elasticity with thermal convection [45].

By adopting the ideas of the proof for the undrained split for poro-elasticity, cf. Lemma 9.1, to vectorized poro-elasticity, analogous convergence results can be deduced for the undrained-adiabatic split for linear thermo-poro-elasticity.

As for the undrained-adiabatic split, the structural similarities to poro-elasticity allows for adopting the convergence results for the standard fixed-stress split, cf. Lemma 9.2, and deduce analogous results for the extended fixed-stress split. For instance, without presenting the analogous proof, we state the generalized *a priori* convergence result.

Corollary 12.2 (Linear convergence of the extended fixed-stress split)**.** *The extended fixed-stress split for linear thermo-poro-elasticity converges linearly, independent of the initial guess. Let $e_T^{n,i} := T^{n,i} - T^n$, $n, i \in \mathbb{N}$, and let $|||\cdot|||_\star$ denote the norm induced by the quadratic part of $\mathcal{E}_{\mathrm{th,tot}}^{\star,\Delta t}$. Assume for brevity, $\boldsymbol{\kappa} = \kappa\mathbf{I}$ and $\boldsymbol{\kappa}_{\mathrm{F}} = \kappa_{\mathrm{F}}\mathbf{I}$ constant in space. It holds the a priori result*

$$\left|\left|\left|(\boldsymbol{e}_{\boldsymbol{\sigma}}^{n,i}, e_p^{n,i}, e_T^{n,i})\right|\right|\right|_\star \leq \left(\frac{\frac{|\boldsymbol{\alpha}_{\mathrm{T}}|^2}{K_{\mathrm{dr}}}}{\frac{\boldsymbol{\alpha}_{\mathrm{T}}^\top \left(\mathbf{M}_{\mathrm{T}}^{-1} + \Delta t\, C_\Omega^{-2} \begin{bmatrix} \kappa & 0 \\ 0 & \frac{\kappa_{\mathrm{F}}}{T_0} \end{bmatrix}\right) \boldsymbol{\alpha}_{\mathrm{T}}}{|\boldsymbol{\alpha}_{\mathrm{T}}|^2} + \frac{|\boldsymbol{\alpha}_{\mathrm{T}}|^2}{K_{\mathrm{dr}}}} \right)^i \left(\mathcal{E}^{n,0} - \mathcal{E}^n\right)^{1/2},$$

where $\mathcal{E}^{n,0}$ and $\mathcal{E}^n$ are the energies of the initial iterate and the solution, resp.

By the Cauchy-Schwarz inequality, the convergence rate of the extended fixed-stress split is lower than for the undrained-adiabatic split – even for $\mathbf{K}_{\mathrm{T}} = \mathbf{0}$.

12.2.3 Three-block coordinate descent for thermo-poro-elasticity

By definition thermo-hydro-mechanical models couple three processes. Thus, in the context of splitting schemes, it is a natural ambition to decouple all three subproblems from each other – with a potential benefit increase of the same kind as two-stage decoupling methods. Three-stage decoupling methods for thermo-poro-elasticity with thermal convection have been recently proposed by [85], including a rigorous convergence analysis. In the following, we briefly demonstrate that similar methods can be derived by applying three-block coordinate descent methods, a natural generalization of alternating minimization.

Since both the primal and the dual formulations of linear thermo-poro-elasticity are block-separable and convex, any cyclic three-block coordinate descent is globally convergent which respects the block structure of the coupled problem, cf. [70, 71]. We exemplarily state one candidate of six possible combinations based on the dual problem – we solve successively for pressure, temperature and stress, cf. Alg. 8 for a single iteration. Similarly, the primal problem

Algorithm 8: Single iteration of the three-block coordinate descent for dual thermo-poro-elasticity

```
Input: $(\boldsymbol{\sigma}^{n,i-1}, p^{n,i-1}, T^{n,i-1}) \in \mathcal{S}^n \times \mathcal{Q}^n \times \mathcal{R}^n$
Determine $p^{n,i} := \underset{p \in \mathcal{Q}^n}{\arg\min}\, \mathcal{E}^{\star,\Delta t}_{\mathrm{th,tot}}(\theta^{n-1}, S^{n-1}; \boldsymbol{\sigma}^{n,i-1}, p, T^{n,i-1})$
Determine $T^{n,i} := \underset{T \in \mathcal{R}^n}{\arg\min}\, \mathcal{E}^{\star,\Delta t}_{\mathrm{th,tot}}(\theta^{n-1}, S^{n-1}; \boldsymbol{\sigma}^{n,i-1}, p^{n,i}, T)$
Determine $\boldsymbol{\sigma}^{n,i} := \underset{\boldsymbol{\sigma} \in \mathcal{S}^n}{\arg\min}\, \mathcal{E}^{\star,\Delta t}_{\mathrm{th,tot}}(\theta^{n-1}, S^{n-1}; \boldsymbol{\sigma}, p^{n,i}, T^{n,i})$
```

can serve as basis; we choose an algorithm closer to the extended fixed-stress split expecting better performance.

The first step of Alg. 8 is equivalent to solving a fluid flow problem with fixed-stress type pressure stabilization: Find $p^{n,i} \in \mathcal{Q}^n$ satisfying for all $q \in \mathcal{Q}_0$

$$\begin{aligned} &\tfrac{1}{M} \left\langle p^{n,i}, q \right\rangle - 3\alpha_\phi \left\langle T^{n,i-1}, q \right\rangle + \tfrac{\alpha^2}{K_{\mathrm{dr}}} \left\langle p^{n,i} - p^{n,i-1}, q \right\rangle + \alpha \left\langle \operatorname{tr} \boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i-1}, q \right\rangle \\ &\quad + \Delta t \left\langle \boldsymbol{\kappa}(\boldsymbol{\nabla} p^{n,i} - \boldsymbol{g}^n_{\mathrm{ext}}), \boldsymbol{\nabla} q \right\rangle = \left\langle \theta^{n-1} + \Delta t\, q^n_\theta, q \right\rangle + \Delta t \left\langle q^n_{\Gamma,\mathrm{n}}, q \right\rangle_{\Gamma_{\boldsymbol{q}}}. \end{aligned}$$

The second step of Alg. 8 is equivalent to solving a thermal problem with fixed-stress type temperature stabilization: Find $T^{n,i} \in \mathcal{R}^n$ satisfying

$$\begin{aligned} &\tfrac{C_{\mathrm{d}}}{T_0} \left\langle T^{n,i}, r \right\rangle - 3\alpha_\phi \left\langle p^{n,i}, r \right\rangle + 9\alpha_{\mathrm{T}}^2 K_{\mathrm{dr}} \left\langle T^{n,i} - T^{n,i-1}, r \right\rangle + 3\alpha_{\mathrm{T}} K_{\mathrm{dr}} \left\langle \operatorname{tr} \boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i-1/2}, r \right\rangle \\ &\quad + \Delta t \left\langle \tfrac{\boldsymbol{\kappa}_{\mathrm{F}}}{T_0} \boldsymbol{\nabla} T^{n,i}, \boldsymbol{\nabla} r \right\rangle = \left\langle S^{n-1} + \Delta t\, q^n_S, r \right\rangle + \Delta t \left\langle j^n_{\mathrm{F},\Gamma}, r \right\rangle_{\Gamma_{\boldsymbol{j}}}, \end{aligned}$$

for all $r \in \mathcal{R}_0$, where we formally abbreviated the updated mechanical strain

$$\boldsymbol{\varepsilon}_{\boldsymbol{u}}^{n,i-1/2} := \mathbb{A} \left(\boldsymbol{\sigma}^{n,i-1} + (\mathbf{I} \otimes \boldsymbol{\alpha}_{\mathrm{T}}) : \begin{bmatrix} [l]p^{n,i} \\ T^{n,i-1} \end{bmatrix} \right).$$

The final step of Alg. 8 is identical with solving the pure mechanical problem for updated pressure and temperature. All in all, the main difference of the resulting scheme to the extended fixed-stress split is the diagonal instead of tensorial stabilization due to further decoupling.

12.3 Comments on splitting schemes for non-linear thermo-poro-elasticity

The splitting schemes derived in this section are in first place only guaranteed to be robust for semi-discrete thermo-poro-elasticity models with an underlying convex minimization structure. As discussed in the modelling section, general thermo-poro-elasticity models do only satisfy a perturbed gradient flow structure, cf. Remark 7.1. Therefore the minimizing movement scheme does not apply immediately, and implicit semi-discrete thermo-poro-elasticity models do generally not stem from convex minimization. Evidently, by explicitly lagging the perturbations in time, the symmetric character of linear thermo-poro-elasticity can be retained, and the above splitting schemes are robust.

Nonetheless, the splitting schemes derived for the simplified, linear case above may as well assist in the construction of splitting schemes for the fully non-linear problem. We mention two possible strategies:

(i) After decomposing the time-continuous, coupled problem into a sum of a linear and parabolic, and a convective problem, an operator splitting [86], e.g., Strang splitting, is utilized. Then the parabolic problem, essentially identical to linear thermo-poro-elasticity, may be solved efficiently using the above splitting schemes, and the convective problem may be solved by a separate, tailored scheme.

(ii) Consider the semi-discrete problem obtained after applying the implicit Euler method. Provided that the perturbations and the time step size are sufficiently small, the semi-discrete problem exhibits a non-symmetric but elliptic character. Under that assumptions iterative two- and three-stage splitting schemes with sufficient diagonal stabilization have been rigorously showed to be linearly convergent [85]. Consequently, robust convergence may be also expected for stabilization terms replaced by those resulting from the above discussions, i.e., effectively by applying the undrained-adiabatic and extended fixed-stress split as recently proposed by [45]. Numerically, this has been demonstrated by the aforementioned work.

13 Acceleration of splitting schemes by optimal relaxation

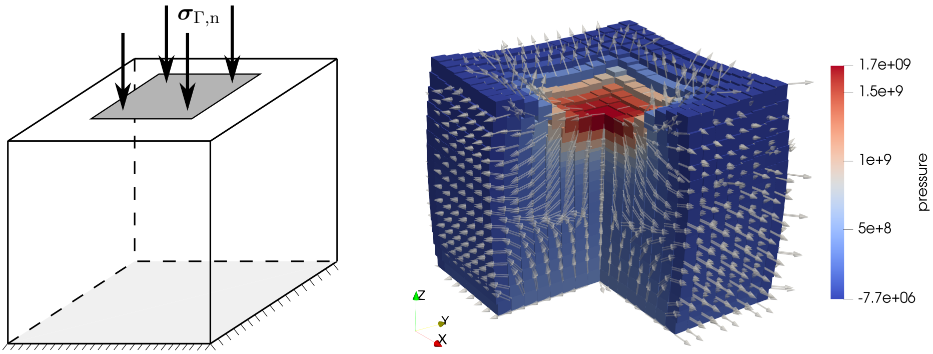
Due to the minimization character of the fully coupled, semi-discrete thermo-poro-visco-elasticity equations, the convergence of splitting schemes for such can be effectively improved by relaxation. Alg. 9 formulates relaxation by exact line search for a general, inexact minimization algorithm for solving semi-discrete generalized gradient flows discretized by the minimizing movement scheme (Sec. 8). For quadratic minimization problems with affine constraints (i.e., e.g., linear thermo-poro-visco-elasticity), optimal relaxation in the sense of a classical, exact line search strategy is feasible; minimizing the quadratic interpolation of three energy values is sufficient for computing the optimal weight. However, also for nonlinear thermo-poro-visco-elasticity stemming from non-quadratic, but convex minimization under affine constraints, we propose the same simple (now inexact) line search strategy.

Algorithm 9: Relaxation of inexact minimization $\mathcal{IM}$ by exact line search for solving time-discrete generalized gradient flows (8.5)

```
Given $\mathcal{X}^n$ affine, $x^{n-1} \in \mathcal{X}^{n-1}$, define $\mathcal{E}^\Delta(x) := \Delta t\, \mathcal{D}\left(\frac{x-x^{n-1}}{\Delta t}\right) + \mathcal{E}(x) - \mathcal{P}^n_{\text{ext}}(x)$
Let $\mathcal{IM} : \mathcal{X}^n \to \mathcal{X}^n$ such that $\mathcal{E}^\Delta(\mathcal{IM}(x)) < \mathcal{E}^\Delta(x)$, where wlog. $x$ is not the minimizer
$x^{n,0} \leftarrow x^{n-1}$, $i \leftarrow 1$
while 'stopping criterion not satisfied' do
    Compute $x^{n,i-1/2} \leftarrow \mathcal{IM}(x^{n,i-1}) \in \mathcal{X}^n$
    Obtain descent direction $\Delta x^{n,i} \leftarrow x^{n,i-1/2} - x^{n,i-1}$
    Solve $\alpha^{n,i} \leftarrow \arg\min_\alpha \mathcal{E}^\Delta\left(x^{n,i-1/2} + \alpha \Delta x^{n,i}\right)$
    Update $x^{n,i} \leftarrow x^{n,i-1/2} + \alpha^{n,i} \Delta x^{n,i} \in \mathcal{X}^n$
    $i \leftarrow i + 1$
end while
```

14 Numerical examples – Performance of the relaxed fixed-stress split for a 3D footing problem

Splitting schemes for solving thermo-hydro-mechanical processes have been numerically studied from various angles in the literature. In the following, we focus on three of the main new contributions obtained from the gradient flow analysis, not previously reported in literature, and study: (i) the impact of relaxation of splitting schemes by exact line search also put in context to the optimization of splitting schemes, (ii) the performance of splitting schemes for



(a) Geometry

(b) Simulation result for test case I

Figure 2: Initial configuration incl. boundary conditions for the 3D footing problem; deformed, poro-elastic configuration after 5 time steps for $E = 10^{11}$ [Pa], incl. pressure profile and outflow; the deformation is scaled by factor 15.

poro-visco-elasticity, and (iii) the performance of splitting schemes for nonlinear poro-elasticity. Due to its larger popularity, we restrict the study to fixed-stress-type splits.

All in all, we consider four test cases based on the same geometry but with slightly differing mechanical material behavior – a unit cube, cf. Fig. 2a, fixed at the bottom and subject to a ramped up, normal force at the top, i.e.,

$$\boldsymbol{u}_\Gamma = \mathbf{0} \text{ on } [0,1]^2 \times \{0\}, \quad \boldsymbol{\sigma}_{\Gamma,\mathrm{n}}(t) = 10^9 t\,[\mathrm{N/m^2s}]\,\boldsymbol{e}_3, \text{ on } [0.25, 0.75]^2 \times \{1\}.$$

No-flow is imposed on the same parts of the boundary. No-stress and zero-pressure boundary conditions are applied on the remaining boundary. Body forces are absent. Similar setups have been considered by [87, 72, 88].

If not mentioned otherwise, the geometry is discretized by a structured $16 \times 16 \times 16$ hexahedral mesh, and 5 time steps of constant time step size $\Delta t = 0.1$ [s] are simulated. For the numerical solution the plain and the relaxed fixed-stress splits are applied. The performance of those is measured in terms of the average number of iterations per time step required for convergence and run times, where as stopping criterion a relative $L^2(\Omega)$ error with tolerance $\epsilon_r = 10^{-6}$ is employed. For the implementation of the numerical examples, we use the DUNE project [89], with extensive use of the DUNE-functions module [90, 91].

14.1 Poro-elastic test case I – Line search under varying coupling strength

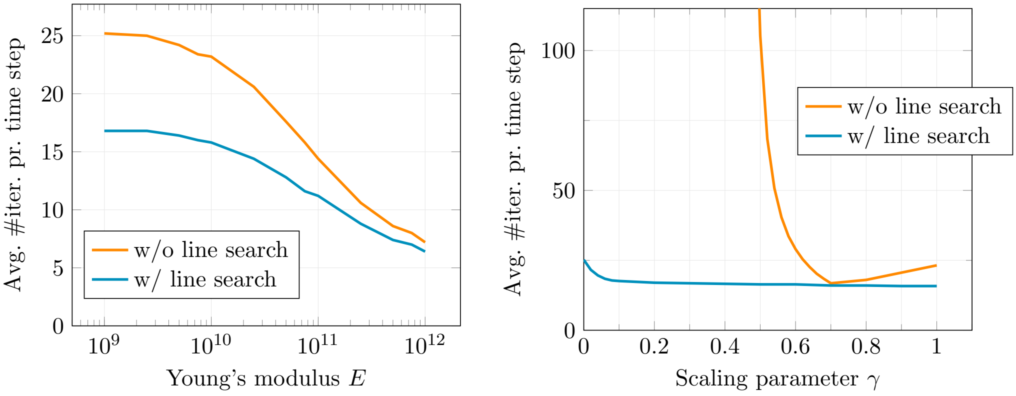
The material is assumed to be poro-elastic, homogeneous and isotropic with material parameters as in Tab. 3. In this first part, we study the impact of the relaxation by line search of the fixed-stress split under varying coupling strength. For this, we vary the Young's modulus E, which is inversely proportional to the coupling strength. A simulation result for $E = 10^{11}$ [Pa] is visualized in Fig. 2b.

By applying the Galerkin method to the five-field formulation of the semi-discrete, linear Biot equations, cf. Sec. 9.1.5, a fully structure-preserving spatial discretization is employed. As conforming finite element spaces for the mechanical problem, we utilize lowest order Brezzi-Douglas-Marini elements, cf., e.g., [61], for the (unsymmetric) stress tensor, piecewise constants for the mechanical displacements and piecewise constant, skew-symmetric tensors for the rotation. For the fluid flow problem, we employ lowest order Brezzi-Douglas-Marini elements for the volumetric flux and piecewise constant elements for the fluid pressure. However, we note, the subsequent results are not strongly depending on the particular formulation or spatial discretization.

Name	Symbol	Value (Test case I–III)	Value (Test case IV)	Unit
Young's modulus	E	$10^9..10^{12}$	10^{10}	Pa
Poisson's ratio	ν	0.2	$0.2, 0.495$	–
Biot-Willis constant	α	1	1	–
Compressibility coefficient	M	10^{11}	10^{11}	Pa
Permeability	κ	10^{-13}	10^{-11}	m^2

Table 3: Poro-elasticity-specific material parameters for the 3D footing problem, used in test cases I–IV.

The performance of the plain and the relaxed fixed-stress splits is displayed in Fig. 3a. We observe that the relaxation by line search allows for reducing the number of iterations up to a factor of 30%. A greater impact can be observed for more strongly coupled problems. On the other hand, only small improvement is observed for weakly coupled problems. This is related to the result of the following test case.



(a) Test case I

(b) Test case II

Figure 3: Number of (poro-elasticity) fixed-stress split iterations for varying coupling strength (Test case I) and varying stabilization parameter (Test case II).

14.2 Poro-elastic test case II – Line search vs. stabilization tuning

It has been previously emphasized [48, 37] that the fixed-stress split can be tuned by appropriate weighting of the stabilization parameter $\frac{\alpha^2}{K_{\mathrm{dr}}}$ in the fluid flow problem, cf. (9.23). *A priori* knowledge on optimal tuning however is lacking due to a strong dependence on the specific geometry, material parameters and applied boundary conditions [92]. It has been numerically demonstrated that optimal weighting may differ substantially from test case to test case [73]. Hence, in general, it is difficult to tune the parameter in practice; in [72] the authors discuss a brute-force optimization strategy utilizing a coarse mesh.

In the following, we demonstrate that the application of exact line search yields a flexible, black box-type alternative to tuning the stabilization parameter. For this, we replace the stabilization parameter by $\gamma\frac{\alpha^2}{K_{\mathrm{dr}}}$ with $\gamma \in [0, 1]$ and apply again both plain and relaxed fixed-stress splits in order to solve the 3D footing problem. Here, we choose the same parameters as in test case I, but with fixed $E = 10^{10}$ [Pa]. The number of iterations required for convergence for varying γ are displayed in Fig. 3b. We make two observations:

- For the plain fixed-stress split we observe practical convergence only for $\gamma \in [0.5, 1]$. This is consistent with theoretical considerations, cf., e.g., [37, 41, 72]. The line-search enhanced fixed-stress split however shows very robust behavior wrt. γ; despite the strong coupling, convergence is even observed for lacking stabilization ($\gamma = 0$).
- For optimally chosen weighting ($\gamma \approx 0.7$) there is no difference in the number of iterations between the plain and the relaxed fixed-stress splits.

Altogether, line search acts here as black-box tuning of the stabilization parameter. However, we note, there is no theoretical guarantee for the optimality of relaxed splitting schemes compared to optimized splitting schemes.

14.3 Poro-visco-elastic test case III – Line search under varying coupling strength

In the following test case, we demonstrate the convergence of the fixed-stress split for poro-visco-elasticity. For this, we re-consider test case I now for a poro-visco-elastic material, and enhance the poro-elastic material parameters (Tab. 3) by visco-elasticity-specific parameters displayed in Tab. 4. A simulation result for $E = 10^{11}$ [Pa] is visualized in Fig. 4a.

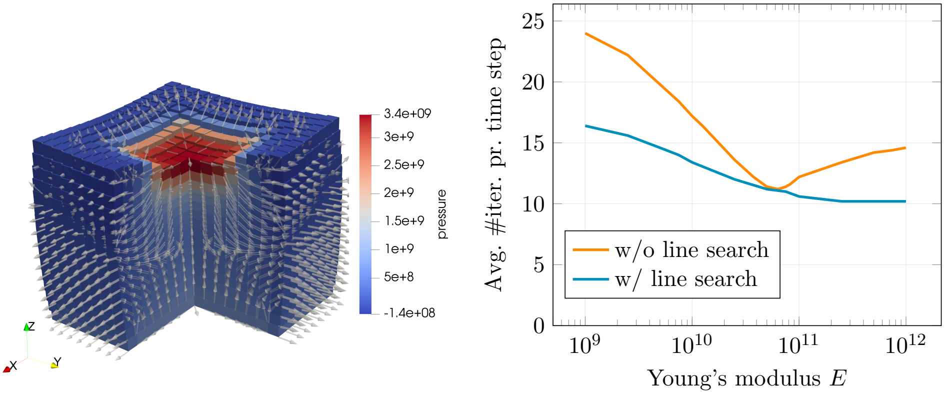
Name	Symbol	Value	Unit
Young's modulus	E_v	10^{10}	Pa
Poisson's ratio	ν_v	0.3	–
Shear modulus	μ'_v	0	Pa
Lamé constant	λ'_v	10^9	Pa
Biot-Willis constant	α_v	0.8	–

Table 4: Poro-visco-elasticity-specific material parameters for the 3D footing problem.

For the spatial discretization, we again utilize a fully-structure preserving formulation based on the dual formulation, cf. Remark 10.2. In particular, the visco-elastic stress $\boldsymbol{\sigma}_\text{v}$ is explicitly introduced, cf. (10.4), with the visco-elastic strain computed from (10.6) by projection onto piecewise constant, symmetric tensors. Hence, the resulting, spatial discretization has the same complexity as in the case of poro-elasticity.

The number of iterations for the plain and relaxed fixed-stress splits required for convergence is displayed in Fig. 4b. At first glance, the performances of both splitting schemes look qualitatively differently. The relaxed fixed-stress split exhibits a monotone relation between its performance and the coupling strength, consistent with the theoretical convergence result, cf. Lemma 10.2. In contrast, the plain fixed-stress split reveals a worsening of the performance for weaker coupling. This can be explained by the findings from test case II. For varying Young's modulus, the overall, structural behavior of the material alters due to $\nu \neq \nu_\text{v}$. As a consequence, considering the optimized fixed-stress split, the optimal tuning parameter changes with E. For smaller and larger E, it is further off the natural stabilization parameter employed within the plain fixed-stress split; for intermediate Young's modulus ($E \approx 5 \cdot 10^{11}$ [Pa]), both parameters are relatively close. This can be justified by the fact that for that configuration line search relaxation does not yield any improvement of the convergence.

After all, if the optimal tuning parameter had been employed for each Young's modulus, the plain fixed-stress split would exhibit the same monotone behavior as under relaxation. Again,



(a) Simulation result

(b) Performance result

Figure 4: Test case III: The deformed, poro-visco-elastic configuration after 5 time steps for $E = 10^{11}$ [Pa], incl. pressure profile and outflow. And the number of (poro-visco-elasticity) fixed-stress split iterations for varying coupling strength.

line search relaxation proves successful as black box tuning without *a priori* knowledge of the physical behavior of the medium.

14.4 Non-linear poro-elastic test case IV – Acceleration and robustness increase of splitting schemes by line search relaxation

In the final numerical test case, we demonstrate the convergence of the fixed-stress split for nonlinear poro-elasticity under infinitesimal strains (Sec. 5). In particular, we study the impact of line search relaxation for various, inexact fixed-stress splits (Sec. 11.3) in comparison to the exact fixed-stress split.

For this, we re-consider test case I now for a nonlinearly elastic material. Differently from before, for the spatial discretization, we consider a structured 32 × 32 × 32 hexahedral mesh, inducing a greater challange to the nonlinear and liner solvers. Furthermore, a three-field formulation, consistent with Sec. 11, is considered. We employ linear elements for the structural displacement, lowest order Raviart-Thomas elements for the volumetric flux and piecewise constant elements for the fluid pressure.

In order to pinpoint the impact of the non-linear character of the equations, we introduce only a single non-linearity compared to test case I – a non-linear (effective) stress-strain relationship corresponding to the non-quadratic p-Laplacian-type energy [63, 65]

$$\mathcal{E}_{\mathrm{nl,eff}}(\boldsymbol{u}) = \int_\Omega W(\boldsymbol{\varepsilon}(\boldsymbol{u}))\,dx = \int_\Omega \left(\mu|\boldsymbol{\varepsilon}(\boldsymbol{u})|^2 + \frac{\lambda}{3}|\boldsymbol{\nabla}\cdot\boldsymbol{u}|^3 \right) dx.$$

Apart from that we choose the same model as in test case I, with material parameters from Tab. 3. We consider two setups with two different Poisson ratios: Setup A with $\nu = 0.2$, and Setup B with $\nu = 0.495$, inducing a comparatively stronger coupling strength and stronger nonlinearity, respectively. The simulation result for $\nu = 0.2$ is illustrated in Fig. 5; the qualitative difference in the flow field compared to test case I, cf. Fig 2b, originates from significantly different permeability fields.

The agenda is similar as before. We apply the fixed-stress split in order to solve the coupled problem, and study the impact of line search relaxation. The non-linear character of the problem allows for choosing various exact or inexact non-linear solvers for solving the mechanics subproblems. In addition, we point out, that the exact fixed-stress split (Sec. 11.3) introduces a displacement-dependent pressure stabilization of the flow equation via the solution-dependent bulk modulus $K_{\mathrm{dr}}(\boldsymbol{\varepsilon}(u)) = \frac{2\mu}{d} + 2\lambda|\boldsymbol{\nabla}\cdot\boldsymbol{u}|$, cf. (11.8). Hence, despite the linear character of the flow equation, the exact Jacobian of the stabilized pressure equation alters with each fixed-stress iteration.

In the following, we apply Newton- and L-scheme-based fixed-stress splits, as introduced in Sec. 11.3, with the latter chosen due to the low computational cost per iteration; however, the L-scheme requires choosing L_b, L_{FS} and $\mathbb{L}$. Given user-defined parameters $0 \leq |\boldsymbol{\nabla}\cdot\boldsymbol{u}|_{\min} \leq |\boldsymbol{\nabla}\cdot\boldsymbol{u}|_{\max} < \infty$, we set

$$L_b = \frac{1}{M}, \quad L_{\mathrm{FS}} = \frac{\alpha^2}{\frac{2\mu}{d} + 2\lambda|\boldsymbol{\nabla}\cdot\boldsymbol{u}|_{\min}}, \quad \mathbb{L} = 2\mu\mathbb{I} + 2\lambda|\boldsymbol{\nabla}\cdot\boldsymbol{u}|_{\max}\,\mathbf{I}\otimes\mathbf{I}. \tag{14.1}$$

Detailed descriptions of the non-linear solvers used in this section are given in Tab. 5. In addition, we apply three relaxation techniques, cf. Table 6. In particular, we also consider applying line search after each non-linear iteration.

Abbreviation	Description
$\mathrm{N}_{\max}$	Newton's method until convergence, i.e., $\|r^i\|/\|r^0\| < 10^{-5}$, where r^i is the residual of the subproblem in the i-th Newton iteration; the Jacobian of the flow equation is reassembled.
N_1	As $\mathrm{N}_{\max}$ but employing only a single Newton iteration.
$\mathrm{L}_m^{\mathrm{ex}}$	m L-scheme iterations if convergence is not met before (see $\mathrm{N}_{\max}$), with L_b, L_{FS} and $\mathbb{L}$ as in (14.1) with $|\boldsymbol{\nabla}\cdot\boldsymbol{u}|_{\min} = \min\limits_{x,t}|\boldsymbol{\nabla}\cdot\boldsymbol{u}|$ and $|\boldsymbol{\nabla}\cdot\boldsymbol{u}|_{\max} = \max\limits_{x,t}|\boldsymbol{\nabla}\cdot\boldsymbol{u}|$
$\mathrm{L}_m^{\mathrm{opt}}$	As $\mathrm{L}_m^{\mathrm{ex}}$ but with $|\boldsymbol{\nabla}\cdot\boldsymbol{u}|_{\min} = |\boldsymbol{\nabla}\cdot\boldsymbol{u}|_{\max} = \frac{1}{10}\max\limits_{x,t}|\boldsymbol{\nabla}\cdot\boldsymbol{u}|$

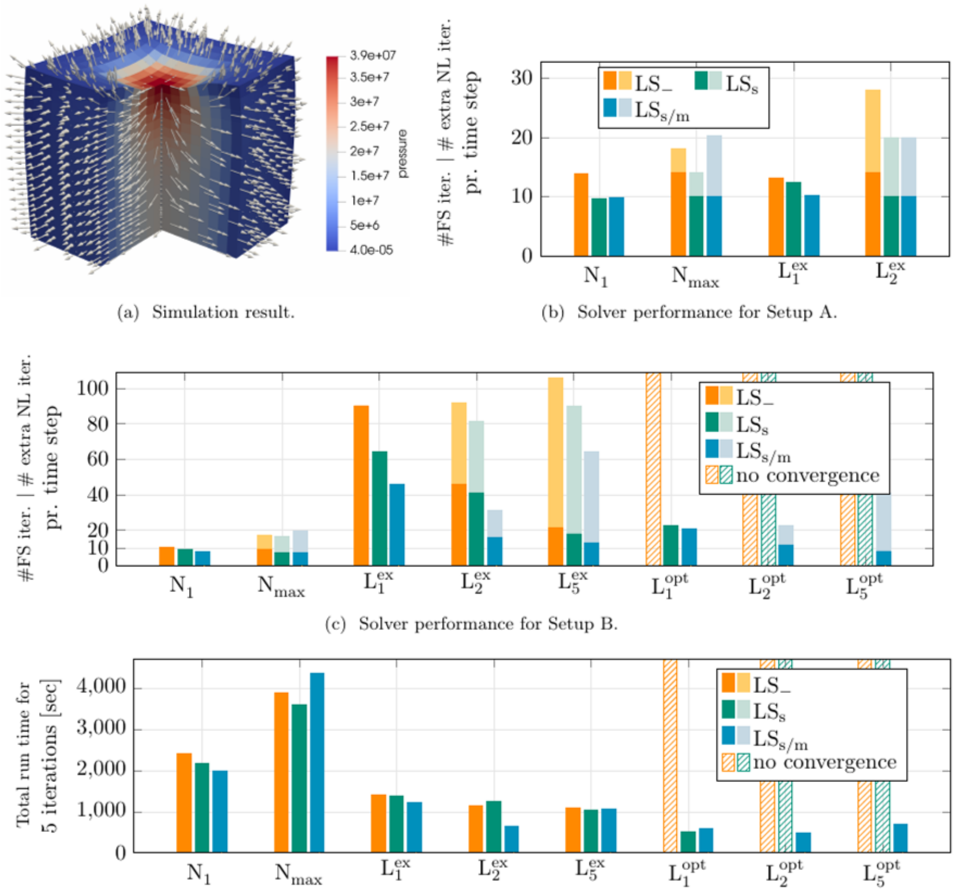
Table 5: Non-linear solvers employed in test case IV.

Abbr.	Description of the relaxation strategy
LS_{-}	Plain splitting scheme and non-linear solver without any relaxation.
LS_{s}	Line search based on quadratic interpolation applied for the splitting solver.
$\mathrm{LS}_{\mathrm{s/m}}$	Same as LS_{s}, but with the same strategy also applied on the inner non-linear solver for the mechanics subproblem.

Table 6: Relaxation strategies employed in test case IV.

The solver performances of various relevant combinations of non-linear solvers and relaxation strategies for Setup A and Setup B are displayed in Fig. 5. Those include the plain number of outer fixed-stress iterations and potential inner extra non-linear iterations if more than one iteration has been applied; in addition, total run times are displayed for Setup B, including run times for assembling matrices and right hand sides, as well as the application of linear solvers. We stress, we use serial, direct solvers. Hence, the Jacobian employed for L-scheme-based splittings is factorized only once, but not for Newton-based splits. Moreover, we mention observations not indicated in the figures:

- For Setup A, the number of fixed-stress iterations per time step is approximately the same for all schemes, indicating a dominant coupling strength.



(a) Simulation result.

(b) Solver performance for Setup A.

(c) Solver performance for Setup B.

(d) Run times for Setup B.

Figure 5: Test case IV: (a) Deformed, poro-elastic configuration after 5 time steps for $\nu = 0.2$, incl. pressure profile and outflow; the deformation is scaled by factor 5. (b) and (c): Performance of different non-linear solvers (Tab. 5) combined with different relaxation strategies (Tab. 6), measured in average number of fixed-stress (FS) iterations and extra non-linear (NL) iterations per time step used for solving the mechanics problem, if more than one non-linear iterations per fixed-stress iteration is utilized; they are displayed on top of each other, illustrating the total amount of non-linear iterations required. (d) Total run times (incl. assembly and solver) for five time steps corresponding to (c).

- For Setup B, the number of fixed-stress iterations per time step decreases for the Newton- and L^{ex}-type methods; it increases for the more optimistic choice L^{opt}. Under relaxation on both levels, the iteration counts are practically constant for all methods.

We conclude, most importantly, inexact alternating minimization can outperform exact alternating minimization. The number of outer fixed-stress iterations might decrease the more accurately the non-linear problems are solved, but on the other hand, the total amount of required inner non-linear iterations increases much more. This makes relaxation by (inexact) line search attractive, which allows for improved solution of the non-linear subproblems and the overall performance of the splitting scheme without requiring to solve a linear system. We observe, relaxation does not only accelerate convergence but it also increases the robustness; similar effects have been previously observed for relaxation by Anderson acceleration for the fixed-stress split [93].

Finally, if applicable, simple linearizations as the L-scheme might outperform Newton-based linearization techniques. In particular, when combining them with relaxation. The main drawback of the L-scheme is that it includes tuning parameters. Optimal choices may lead to good performance, whereas bad choices might even lead to no convergence. Suitable choices being solution-dependent, makes the final choice rather difficult; however, line search may allow convergence for a wide range of parameters, potentially even faster than more conservative choices of the tuning parameters, for which the plain scheme converges. In the present example, by choosing an L-scheme-based fixed-stress split with optimistic tuning parameters and full line search relaxation, run times 1/8 of those for the non-relaxed, exact Newton-based fixed-stress split have been achieved. The finer the mesh the more drastic the difference as direct solvers are employed in this study.

15 Concluding remarks and discussion

The aim of the present work was to examine the inherent gradient flow structures of thermo-hydro-mechanical processes in porous media with focus on consequences for the well-posedness analysis and construction of numerical approximations and solvers. A major finding was that various, existing PDE models from the literature can be formulated as generalized gradient flows utilizing thermodynamic interpretation of energies and dissipation potentials – for instance, linear poro-elasticity, linear poro-visco-elasticity, non-linear poro-elasticity in the infinitesimal strain regime, non-Newtonian Darcy and non-Darcy flows in poro-elastic media, and thermo-poro-elasticity without thermal convection. Moreover, well-posedness has been established for those models utilizing a unified framework introduced for doubly non-linear evolution equations.

One further significant finding to emerge from this work is that robust, physically based operator splitting schemes for time-discrete approximations are a consequence of a suitable *choice* of primary variables (in fact dictated by the gradient flow structure) and a simple application of plain alternating minimization. Robustness is then an immediate consequence of the naturally underlying minimization structure of the semi-discrete problem arising from suitable time-discretization of gradient flows; in that light, e.g., the undrained and the fixed-stress splits appear to be the natural splitting schemes for linear poro-elasticity. Moreover, abstract convergence theory allows to quantify the energy decrease for each iteration of the splitting schemes only utilizing convexity and Lipschitz continuity properties of the problem – a fairly simple machinery compared to previous analyses in the literature and also immediately applicable to heterogeneous, anisotropic materials. We derive novel splitting schemes and establish *a priori* and *a posteriori* convergence results in the context of linear poro-elasticity, linear poro-visco-elasticity and linear thermo-poro-elasticity.

The results of this work support the idea that splitting schemes for models with a vector structure ought to utilize tensorial stabilization instead of diagonal stabilization; such has been previously proposed either based on physical intuition or rather more ad hoc, cf., e.g., [45] in the context of thermo-poro-elasticity or [94] in the context of multiple-network poro-elastic theory. The latter has not been covered in this work, but it is essentially a generalization of linear thermo-poro-elasticity; our results can be immediately extrapolated.

Additionally, we highlight the known and simple fact that a minimization formulation enables relaxation of iterative solvers by line search strategies. Such have not been utilized before in the field of poro-elasticity. Our numerical experiments suggest that line search acts as black box optimization of the stabilization and possibly linearization in the context of optimized splitting schemes [72], which is especially practical for problems with changing geometries or boundary conditions.

Throughout the entire work, we utilize linear poro-elasticity as proof of concept and verify that the provided framework yields consistent results with the literature, but from a new perspective. After all, it seems promising for handling further models as also demonstrated for various extensions of linear poro-elasticity.

The most important limitation lies in the fact that, evidently, not all thermo-hydro-mechanical processes are suitably modelled by gradient flows, e.g., convective-dominated processes, or materials with limit behavior as incompressible fluids or solids. However, at least in the context of the numerical solution, non-monotone perturbations of gradient flows may be discussed using operator splitting techniques as Strang's splitting or semi-implicit time-discretization, and limit cases may be handled employing duality theory. After all, the provided theory may still assist in various situations – to what extent is topic of future research. Moreover, in this sense, interesting areas of applications and model extensions include finite strain poro-elasticity, poro-elasticity for fractured media, poroplasticity, and compositional and multi-phase flow in poro-elastic media. In terms of numerical solvers, for strongly non-linear and possibly non-convex problems, a further study could assess the need for more advanced optimization algorithms as primal-dual methods, alternating direction method of multipliers, or proximal alternating minimization for deriving robust linearization or non-linear preconditioners. This would be a fruitful area for further work.

A Abstract well-posedness results

The theoretical results in this work are mostly deduced by application of abstract results from literature; we recall two results for doubly non-linear evolution equations and convex optimization.

The following well-posedness result for doubly non-linear evolution equation can be understood as a corollary or refined discussion of previous classical results, e.g., [11]. The main improvement to previous results is a weaker regularity assumption on the external loading. This is compensated with stronger, structural assumptions on the functions spaces, as well as the dissipation potential and energy functional. Here, we consider an energy functional which does not explicitly depend on time. In order to incorporate time-dependent energy functionals, assumptions and proof techniques as, e.g., by [9], can be additionally applied.

Theorem A.1 (Well-posedness for doubly evolution equations with weakly regular load)**.** *Consider the doubly non-linear evolution equation*

$$\boldsymbol{\nabla}\Psi(\dot{x}(t)) + \boldsymbol{\nabla}\mathcal{E}(t, x(t)) = f(t) \text{ in } \mathcal{V}^\star \text{ a.e. in } (0,T); \quad x(0) = x_0. \tag{A.1}$$

where

- $p_\psi, p_\mathcal{E} \in (1,\infty)$*;* $p := \min\{p_\psi, p_\mathcal{E}\}$*;* $p^\star \in (1,\infty)$ *such that* $\frac{1}{p} + \frac{1}{p^\star} = 1$.
- $\mathcal{B}$ *is a separable, reflexive Banach space with norm* $\|\cdot\|_\mathcal{B}$.

- *$\mathcal{V}$ is a separable, reflexive Banach space with a semi-norm $|\cdot|_{\mathcal{V}}$, such that*

$$\|x\|_{\mathcal{V}} := \left(\|x\|_{\mathcal{B}}^p + |x|_{\mathcal{V}}^p\right)^{1/p}, \quad x \in \mathcal{V}. \tag{A.2}$$

  defines a norm on $\mathcal{V}$. Furthermore, $\mathcal{V}$ is dense and compactly embedded in $\mathcal{B}$.

- *$\Psi : \mathcal{B} \to [0,\infty)$ is convex and continuously differentiable. There exists a constant $C > 0$ such that*

$$\Psi(x) \geq C\|x\|_{\mathcal{B}}^{p_\psi}, \quad x \in \mathcal{B}.$$

- *$\mathcal{E} : [0,T] \times \mathcal{V} \to [0,\infty)$, such that there exist constants $C_1 > 0$, $C_2 \geq 0$, satisfying*

$$\mathcal{E}(t,x) \geq C_1|x|_{\mathcal{V}}^{p_\mathcal{E}} - C_2 \quad \textit{for all } (t,x) \in [0,T] \times \mathcal{V}.$$

  Furthermore, $\mathcal{E}(t,\cdot) : \mathcal{V} \to (-\infty,\infty)$ is convex, lower-semicontinuous, and continuously differentiable for all $t \in [0,T]$; and $\mathcal{E}(\cdot,x) : [0,T] \to (-\infty,\infty)$ is differentiable a.e. for all $x \in \mathcal{V}$ such that there exists a constant $C > 0$, satisfying for a.e. $t \in (0,T)$

$$|\partial_t\mathcal{E}(t,x)| \leq C(1 + \mathcal{E}(t,x)) \quad \textit{for all } x \in \mathcal{V}.$$

- *$f \in C(0,T;\mathcal{V}^\star) \cap W^{1,p^\star}(0,T;\mathcal{V}^\star)$.*

- *$x_0 \in \mathcal{V}$ such that $\mathcal{E}(0,x_0) < \infty$.*

Then there exists a solution $x \in W^{1,p}(0,T;\mathcal{B}) \cap L^\infty(0,T;\mathcal{V})$ of (A.1)*, satisfying $\mathcal{E}(x) \in L^\infty(0,T)$ and the energy identity*

$$\begin{aligned}&\int_0^T \Psi(\dot{x}(t))\,dt + \mathcal{E}(x(T)) - \langle f(T), x(T)\rangle \\ &\quad = \mathcal{E}(x_0) - \langle f(0), x(0)\rangle + \int_0^T \partial_t\mathcal{E}(t,x(t))\,dt - \int_0^T \left\langle \dot{f}(t), x(t)\right\rangle\,dt.\end{aligned} \tag{A.3}$$

If $\boldsymbol{\nabla}\Psi$ or $\mathcal{E}$ are linear and self-adjoint, it is unique.

Proof. The proof is analogous to the proof of Thm. 1 by [11], enhanced by discussions of the time-dependence of the energy functional by [95]: First, the doubly non-linear evolution equation (A.1) is discretized in time by consecutive convex minimization problems, and second, stability bounds are derived, and finally, compactness arguments are employed in order to pass to the limit, obtaining a solution to the time-continuous problem. Due to the weaker regularity assumptions on the load term, the second step of [11] is not applicable here. In the following, we derive stability for the time-discrete approximation under the above assumptions.

As in [11, 95], we use the minimizing movement scheme to discretize (A.1) in time. Let $0 = t_0 < t_1 < \dots < t_N = T$ of $[0,T]$ denote a partition of $[0,T]$ with constant time step size Δt. Set $x^0 = x(0)$ and define consecutively

$$x^n := \operatorname*{arg\,min}_{x \in \mathcal{B}} \left\{\Delta t\,\Psi\left(\frac{x - x^{n-1}}{\Delta t}\right) + \tilde{\mathcal{E}}^n(x)\right\}$$

where $f^n := \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} f(t)\,dt$, and $\tilde{\mathcal{E}}^n : \mathcal{B} \to (-\infty,\infty]$ defined by

$$\tilde{\mathcal{E}}^n(x) = \begin{cases} \mathcal{E}(t_n,x) - \langle f^n, x\rangle & x \in \mathcal{V}, \\ \infty, & \text{otherwise,}\end{cases}$$

is a proper, convex, lower-semicontinuous function. By Thm. A.2, x^n is well-defined. Furthermore, for all n, it holds

$$\Delta t\,\Psi\left(\frac{x^n - x^{n-1}}{\Delta t}\right) + \tilde{\mathcal{E}}^n(x^n) \leq \tilde{\mathcal{E}}^n\left(x^{n-1}\right)$$

and hence, by induction $x^n \in \mathcal{V}$ for all n, since $x^0 \in \mathcal{V}$. Summing over all time steps, employing the definition of $\tilde{\mathcal{E}}$ and manipulating the sum over the load terms, yields

$$\begin{aligned}
&\sum_n \Delta t\,\Psi\left(\frac{x^n - x^{n-1}}{\Delta t}\right) + \mathcal{E}(t_N, x^N) \\
&\quad\leq \mathcal{E}(0, x(0)) + \sum_n \int_{t_{n-1}}^{t_n} \partial_t \mathcal{E}(t, x^{n-1})\,dt \\
&\qquad + \left\langle f^0, x(0)\right\rangle - \left\langle f^N, x^N\right\rangle - \sum_n \Delta t\left\langle \frac{f^n - f^{n-1}}{\Delta t}, x^{n-1}\right\rangle .
\end{aligned} \tag{A.4}$$

As in [95], we employ the bound on $\partial_t\mathcal{E}$ together with a Grönwall inequality and obtain

$$\sum_n \int_{t_{n-1}}^{t_n} \partial_t \mathcal{E}(t, x^{n-1})\,dt \leq C\left(1 + \sum_n \Delta t\,\mathcal{E}(t_{n-1}, x^{n-1})\right),$$

where $C > 0$ depends on T and the stability bound on $\partial_t\mathcal{E}$. Inserting into (A.4), and, furthermore, utilizing the assumptions on Ψ, $\mathcal{E}$ and f yields for arbitrary $\delta > 0$

$$\begin{aligned}
&\sum_n \Delta t\left\|\frac{x^n - x^{n-1}}{\Delta t}\right\|_{\mathcal{B}}^{p_\psi} + \left|x^N\right|_{\mathcal{V}}^{p_\mathcal{E}} + \mathcal{E}(t_N, x^N) \\
&\quad\leq C\left(1 + \sum_n \Delta t\,\mathcal{E}(t_{n-1}, x^{n-1})\right) + \delta\left(\left\|x^N\right\|_{\mathcal{V}}^{p} + \sum_n \Delta t\left\|x^{n-1}\right\|_{\mathcal{V}}^{p}\right).
\end{aligned}$$

where $C > 0$ depends on T, δ, the initial data, and regularity of the loading. Using Young's inequality and the definition of $\|\cdot\|_{\mathcal{V}}$, it holds

$$\begin{aligned}
&\sum_n \Delta t\left\|\frac{x^n - x^{n-1}}{\Delta t}\right\|_{\mathcal{B}}^{p_\psi} + \left|x^N\right|_{\mathcal{V}}^{p_\mathcal{E}} + \mathcal{E}(t_N, x^N) \\
&\leq C\left(1 + \sum_n \Delta t\,\mathcal{E}(t_{n-1}, x^{n-1})\right) \\
&\qquad + \delta\left(\left\|x^N\right\|_{\mathcal{B}}^{p_\psi} + \sum_n \Delta t\left\|x^{n-1}\right\|_{\mathcal{B}}^{p_\psi} + \left|x^N\right|_{\mathcal{V}}^{p_\mathcal{E}} + \sum_n \Delta t\left|x^{n-1}\right|_{\mathcal{V}}^{p_\mathcal{E}}\right).
\end{aligned}$$

By constructing a telescope sum, exploiting the convexity of $x \mapsto x^{p_\psi}$ and applying Hölder inequalities, we obtain

$$\left\|x^N\right\|_{\mathcal{B}}^{p_\psi} + \sum_n \Delta t\left\|x^{n-1}\right\|_{\mathcal{B}}^{p_\psi} \leq C\left(1 + \sum_n \Delta t\left\|\frac{x^n - x^{n-1}}{\Delta t}\right\|_{\mathcal{B}}^{p_\psi}\right)$$

for $C > 0$ depending on p_ψ, x_0 and T. Hence, for δ sufficiently small it holds

$$\begin{aligned}
&\sum_n \Delta t\left\|\frac{x^n - x^{n-1}}{\Delta t}\right\|_{\mathcal{B}}^{p_\psi} + \left|x^N\right|_{\mathcal{V}}^{p_\mathcal{E}} + \mathcal{E}(t_N, x^N) \\
&\quad\leq C\left(1 + \sum_n \Delta t\,\mathcal{E}(t_{n-1}, x^{n-1}) + \sum_n \Delta t\left|x^{n-1}\right|_{\mathcal{V}}^{p_\mathcal{E}}\right).
\end{aligned}$$

Finally, by employing a Grönwall inequality, we obtain uniform stability for the left hand side. Based on that, the proof can be continued along the lines of [11, 95], utilizing compactness arguments in order to pass to the limit $\Delta t \to 0$ and obtain a solution to the time-continuous doubly non-linear evolution equation, that in particular satisfies the energy identity (A.3). □

Theorem A.2 (Well-posedness for convex minimization [50])**.** *Consider the problem*

$$\begin{aligned}&\textit{minimize } f(x)\\&\textit{subject to } x \in \mathcal{C},\end{aligned} \tag{A.5}$$

where $f : \mathcal{X} \to \mathbb{R}$ is a proper, convex, lower semi-continuous function, and $\mathcal{C} \subset \mathcal{X}$ is non-empty, closed, convex subset of $\mathcal{X}$, a reflexive Banach space. If $\mathcal{C}$ is bounded or f is coercive over $\mathcal{C}$, i.e., $f(x) \to \infty$ for $x \in \mathcal{C}$ with $\|x\| \to \infty$, then (A.5) *has a solution. It is unique if f is strictly convex.*

B Alternating minimization for block-separable constrained convex minimization in infinitely dimensional Hilbert spaces

In [79], the authors establish an abstract convergence result for alternating minimization, applied to a constrained, strongly convex minimization problem in finite dimensions. Furthermore, convexity and Lipschitz continuity are solely considered wrt. Euclidean norms. We generalize the abstract result, allowing for a constrained minimization problem in infinitely dimensional Hilbert spaces. Convexity and Lipschitz continuity are considered wrt. the semi-norms.

Hilbert space structure. Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ be a product of Hilbert spaces, equipped with an inner product $\langle \cdot, \cdot \rangle$. Assume it is induced by separate inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ on $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, such that

$$\langle (x_1, x_2), (y_1, y_2) \rangle = \langle x_1, y_1 \rangle_1 + \langle x_2, y_2 \rangle_2 , \qquad (x_1, x_2), (y_1, y_2) \in \mathcal{X}_1 \times \mathcal{X}_2.$$

The inner product $\langle \cdot, \cdot \rangle$ acts naturally also as duality pairing on $\mathcal{X}^\star \times \mathcal{X}$. Additionally, let $|\cdot|_\star$ on $\mathcal{X}$ denote some semi-norm on $\mathcal{X}$.

Function properties. Let $f : \mathcal{X} \to \mathbb{R}$ be differentiable. We introduce two properties:

(i) We call f *strongly convex wrt.* $|\cdot|_\star$ if there exists a constant $\sigma > 0$ such that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{\nabla} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\sigma}{2} |\boldsymbol{y} - \boldsymbol{x}|_\star^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \tag{B.1}$$

which is equivalent to (see, e.g., [96])

$$\langle \boldsymbol{\nabla} f(\boldsymbol{y}) - \boldsymbol{\nabla} f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \geq \sigma |\boldsymbol{y} - \boldsymbol{x}|_\star^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}.$$

(ii) We call the k-th block $\boldsymbol{\nabla}_k f$ of the gradient of f *Lipschitz continuous wrt.* $|\cdot|_\star$ if there exists a constant $L_k < \infty$ such that for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{h}_k \in \mathcal{X}$, it holds

$$\langle \boldsymbol{\nabla}_k f(\boldsymbol{x} + \boldsymbol{h}_k) - \boldsymbol{\nabla}_k f(\boldsymbol{x}), \boldsymbol{h}_k \rangle \leq L_k |\boldsymbol{h}_k|_{\star,1}^2, \tag{B.2}$$

where $\boldsymbol{h}_1 = (\tilde{h}_1, 0)$ and $\boldsymbol{h}_2 = (0, \tilde{h}_2)$ for some $\tilde{h}_k \in \mathcal{X}_k$, $k = 1, 2$. The condition (B.2) is equivalent to (see, e.g., [96])

$$f(\boldsymbol{x} + \boldsymbol{h}_k) \leq f(\boldsymbol{x}) + \langle \boldsymbol{\nabla}_k f(\boldsymbol{x}), \boldsymbol{h}_k \rangle + \frac{L_k}{2} |\boldsymbol{h}_k|_{\star,1}^2. \tag{B.3}$$

Alternating minimization. Let $\mathcal{X}$ as above and $f : \mathcal{X} \to \mathbb{R}$. Furthermore, let $\tilde{\mathcal{X}}_1 \subset \mathcal{X}_1$ and $\tilde{\mathcal{X}}_2 \subset \mathcal{X}_2$ be non-empty, convex subsets. We consider the constrained minimization problem

$$\inf_{(x_1,x_2)\in\tilde{\mathcal{X}}_1\times\tilde{\mathcal{X}}_2} f(x_1, x_2). \tag{B.4}$$

Under certain assumptions on f and $|\cdot|_\star$, we can show global, linear convergence for alternating minimization, cf. Alg. 10.

Algorithm 10: Single iteration of alternating minimization

```
Input: $\boldsymbol{x}^{i-1} = (x_1^{i-1}, x_2^{i-1}) \in \tilde{\mathcal{X}}_1 \times \tilde{\mathcal{X}}_2$
Determine $x_1^i := \arg\min_{x_1\in\tilde{\mathcal{X}}_1} f(x_1, x_2^{i-1})$
Determine $x_2^i := \arg\min_{x_2\in\tilde{\mathcal{X}}_2} f(x_1^i, x_2)$
```

Lemma B.1 (Linear convergence for alternating minimization)**.** *Let $\tilde{\mathcal{X}} := \tilde{\mathcal{X}}_1 \times \tilde{\mathcal{X}}_2 \subset \mathcal{X}$ as above, and let $f : \mathcal{X} \to \mathbb{R}$ be a differentiable function. Furthermore, let $|\cdot|_{\star,1}$ denote semi-norm on $\mathcal{X}$ satisfying:*

- $|(x_1, x_2)|_{\star,1} \geq |(x_1, 0)|_{\star,1}$ *for all* $(x_1, x_2) \in \mathcal{X}$,
- f *is strongly convex wrt.* $|\cdot|_{\star,1}$ *with constant* σ_1,
- $\boldsymbol{\nabla}_1 f$ *is Lipschitz continuous wrt.* $|\cdot|_{\star,1}$ *with Lipschitz constant* L_1.

Then the alternating minimization Alg. 10 is globally, linearly convergent. In particular, let $(\boldsymbol{x}^i)_i \subset \tilde{\mathcal{X}}_1 \times \tilde{\mathcal{X}}_2$ be the sequence generated by the alternating minimization Alg. 10 for given initial value $\boldsymbol{x}^0 \in \mathcal{X}$. And let $\boldsymbol{x}^\star \in \mathcal{X}$ denotes the unique solution of (B.4)*. It holds for all $i \in \mathbb{N}$*

$$\begin{aligned} f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) &\leq \left(1 - \frac{\sigma_1}{L_1}\right) \left(f(\boldsymbol{x}^{i-1}) - f(\boldsymbol{x}^\star)\right) \leq \left(1 - \frac{\sigma_1}{L_1}\right)^i \left(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^\star)\right), \\ f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) &\leq \frac{L_1}{\sigma_1} \left(f(\boldsymbol{x}^{i-1}) - f(\boldsymbol{x}^i)\right). \end{aligned}$$

Assume there additionally exists a second semi-norm $|\cdot|_{\star,2}$ on $\mathcal{X}$ satisfying:

- $|(x_1, x_2)|_{\star,2} \geq |(0, x_2)|_{\star,2}$ *for all* $(x_1, x_2) \in \mathcal{X}$,
- f *is strongly convex wrt.* $|\cdot|_{\star,2}$ *with constant* σ_2,
- $\boldsymbol{\nabla}_2 f$ *is Lipschitz continuous wrt.* $|\cdot|_{\star,2}$ *with Lipschitz constant* L_2.

Then it holds for all $i \in \mathbb{N}$

$$\begin{aligned} f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) &\leq \prod_{j=1}^{2} \left(1 - \frac{\sigma_j}{L_j}\right) \left(f(\boldsymbol{x}^{i-1}) - f(\boldsymbol{x}^\star)\right) \leq \prod_{j=1}^{2} \left(1 - \frac{\sigma_j}{L_j}\right)^i \left(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^\star)\right), \\ f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) &\leq \prod_{j=1}^{2} \frac{L_j}{\sigma_j} \left(f(\boldsymbol{x}^{i-1}) - f(\boldsymbol{x}^i)\right). \end{aligned}$$

Proof. The proof follows the same line of argumentation as the proof of Theorem 5.2 [79], but carefully tailored to the more general setting used above.

Consequence from strong convexity. Consider (B.1) for $\boldsymbol{x} = \boldsymbol{x}^i$. Minimizing both sides wrt. $\boldsymbol{y} \in \tilde{\mathcal{X}}$ yields

$$f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) \leq - \inf_{\boldsymbol{y}\in\tilde{\mathcal{X}}} \left(\left\langle \boldsymbol{\nabla} f(\boldsymbol{x}^i), \boldsymbol{y} - \boldsymbol{x}^i \right\rangle + \frac{\sigma_1}{2} |\boldsymbol{y} - \boldsymbol{x}^i|_{\star,1}^2 \right). \tag{B.5}$$

By definition of alternating minimization it holds

$$\langle \boldsymbol{\nabla}_2 f(\boldsymbol{x}^i), y_2 - x_2^i \rangle_2 \geq 0 \quad \forall y_2 \in \tilde{\mathcal{X}}_2.$$

Hence, together with (A1), (B.5) becomes

$$f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) \leq - \inf_{y_1 \in \tilde{\mathcal{X}}_1} \left(\left\langle \boldsymbol{\nabla}_1 f(\boldsymbol{x}^i), y_1 - x_1^i \right\rangle_1 + \frac{\sigma_1}{2} \left| (y_1 - x_1^i, 0) \right|_{\star,1}^2 \right). \tag{B.6}$$

Since $\tilde{\mathcal{X}}_1$ is convex and $0 < \sigma_1 \leq L_1$ by definition, it holds for all $x_1 \in \tilde{\mathcal{X}}_1$

$$\mathcal{B}_{L_1/\sigma_1}(x_1) := \left\{ h \in \mathcal{X}_1 \,\middle|\, x_1 + \frac{L_1}{\sigma_1} h \in \tilde{\mathcal{X}}_1 \right\} \subset \left\{ h \in \mathcal{X}_1 \,\middle|\, x_1 + h \in \tilde{\mathcal{X}}_1 \right\} =: \mathcal{B}_1(x_1)$$

Hence, we obtain

$$\begin{aligned}
&\inf_{y_1 \in \tilde{\mathcal{X}}_1} \left(\left\langle \boldsymbol{\nabla}_1 f(\boldsymbol{x}^i), y_1 - x_1^i \right\rangle_1 + \frac{\sigma_1}{2} \left| (y_1 - x_1^i, 0) \right|_{\star,1}^2 \right) \\
&\quad = \inf_{h \in \mathcal{B}_{L_1/\sigma_1}(x_1^i)} \left(\left\langle \boldsymbol{\nabla}_1 f(\boldsymbol{x}^i), \frac{L_1}{\sigma_1} h \right\rangle_1 + \frac{\sigma_1}{2} \left| \left(\frac{L_1}{\sigma_1} h, 0 \right) \right|_{\star,1}^2 \right) \\
&\quad \geq \frac{L_1}{\sigma_1} \inf_{h \in \mathcal{B}_1(x_1^i)} \left(\left\langle \boldsymbol{\nabla}_1 f(\boldsymbol{x}^i), h \right\rangle_1 + \frac{\sigma_1}{2} \left| (h, 0) \right|_{\star,1}^2 \right).
\end{aligned}$$

Altogether, it holds

$$f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) \leq - \frac{L_1}{\sigma_1} \inf_{x_1^i + h \in \tilde{\mathcal{X}}_1} \left(\left\langle \boldsymbol{\nabla}_1 f(\boldsymbol{x}^i), h \right\rangle_1 + \frac{L_1}{2} \left| (h, 0) \right|_{\star,1}^2 \right). \tag{B.7}$$

Consequence from Lipschitz continuity. Consider (B.3) for $\boldsymbol{x} = \boldsymbol{x}^i$. Minimizing both sides wrt. $\boldsymbol{h}_1 = (h_1, 0)$ such that $x_1^i + \boldsymbol{h}_1 \in \tilde{\mathcal{X}}_1$ yields

$$f(\boldsymbol{x}^i) - f(x_1^{i+1}, x_2^i) \geq - \inf_{x_1^i + h_1 \in \tilde{\mathcal{X}}_1} \left(\left\langle \boldsymbol{\nabla}_1 f(\boldsymbol{x}^i), h_1 \right\rangle_1 + \frac{L_1}{2} \left| (h_1, 0) \right|_{\star,1}^2 \right). \tag{B.8}$$

Consequences for alternating minimization. By putting together (B.7) and (B.8), and exploiting the definition of alternating minimization, we obtain the *a posteriori* estimate

$$\begin{aligned}
&f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) \\
&\quad \leq \frac{L_1}{\sigma_1} \left(f(\boldsymbol{x}^i) - f(x_1^{i+1}, x_2^i) \right) \\
&\quad \leq \frac{L_1}{\sigma_1} \left(f(\boldsymbol{x}^i) - f(\boldsymbol{x}^{i+1}) \right).
\end{aligned}$$

Adding and subtracting $f(\boldsymbol{x}^\star)$ on the right hand side and reordering terms, yields

$$f(\boldsymbol{x}^{i+1}) - f(\boldsymbol{x}^\star) \leq \left(1 - \frac{\sigma_1}{L_1} \right) \left(f(\boldsymbol{x}^i) - f(\boldsymbol{x}^\star) \right).$$

The *a priori* result follows immediately. The second part of the thesis is proved analogously with focus on the second step of the alternating minimization. □

C Nomenclature

Space and time

x	Spatial coordinate
t	Time
d	Space dimension
Ω	Domain
Γ	Boundary of Ω
C_Ω	Poincaré constant
T	Final time
Δt	Time increment
t_n	n-th time step

Generalized gradient flows

$\mathcal{X}$	State space
$\mathcal{P}_{\dot{\mathcal{X}}}$	Process space
$\mathcal{E}$	Free energy
$\mathcal{D}$	Dissipation potential
$\mathcal{P}_{\mathrm{ext}}$	External work rate

Physical fields

$\boldsymbol{u}$	Structural displacement
$\boldsymbol{\varepsilon}(\boldsymbol{u})$	Linear strain / symmetric gradient of $\boldsymbol{u}$
$\boldsymbol{\sigma}$	Total stress
$\boldsymbol{\sigma}^{\mathrm{d}}$	Deviatoric stress
$\boldsymbol{\sigma}^{\mathrm{h}}$	Hydrostatic stress
$\boldsymbol{\sigma}_{\mathrm{eff}}$	Effective stress
$\boldsymbol{\zeta}$	Rotation
θ	Fluid content
p	Fluid pressure
$\boldsymbol{q}$	Volumetric flux
S	Entropy
$\boldsymbol{j}$	Entropy flux
$\boldsymbol{\varepsilon}_{\mathrm{v}}$	Visco-elastic strain
$\boldsymbol{q}_{\int},\ \boldsymbol{j}_{\int}$	Accumulated volumetric and entropy fluxes

External sources

q_θ	Source for mass conservation
q_S	Entropy source
$Q_\theta,\ Q_S$	Accumulated mass and entropy sources
$\boldsymbol{f}_{\mathrm{ext}}$	External body force acting on the bulk
$\boldsymbol{g}_{\mathrm{ext}}$	External body force acting on the fluid
$\boldsymbol{u}_\Gamma$	Prescribed displacement
$\boldsymbol{\sigma}_{\Gamma,\mathrm{n}}$	Prescribed surface force onto boundary
p_Γ	Prescribed pressure
$q_{\Gamma,\mathrm{n}}$	Prescribed normal volumetric flux
T_Γ	prescribed temperature
$j_{\Gamma,\mathrm{n}}$	Prescribed normal entropy flux

Function spaces

Ω	Porous medium
d	Spatial dimension
$\mathcal{X}^\star$	Dual space of some function space $\mathcal{X}$
$\mathcal{X}^n$	Space $\mathcal{X}$ evaluated at time t_n
$\mathcal{X}_0$	Tangent space of some function space $\mathcal{X}$
$\mathcal{V}$	Space for structural displacement
$\mathcal{S}$	Space for total stress including the balance of momentum
$\tilde{\mathcal{S}}$	Space for total stress without the balance of momentum
$\boldsymbol{Q}_{\mathrm{AS}}$	Space of skew-symmetric tensors in $\mathbb{R}^{d\times d}$
$\mathcal{Q}, \tilde{\mathcal{Q}}$	Space for fluid pressure
$\Pi_{\tilde{\mathcal{Q}}}$	Orthogonal projection onto $\tilde{\mathcal{Q}}$
$\mathcal{Z}$	Space for volumetric flux
$\mathcal{Z}_\int$	Space for accumulated flux
$\mathcal{W}$	Space for entropy flux
$\mathcal{W}_\int$	Space for accumulated entropy flux
$\mathcal{T}$	Space for visco-elastic strains

Material parameters

$\mathbb{C},\ \mathbb{C}_{\mathrm{v}},\ \mathbb{C}_{\mathrm{v}}'$	Stiffness tensors
$\mathbb{A}$	Compliance tensor
$\mathcal{C}_{\mathrm{v}}$	Generalized stiffness tensor
$\mathcal{A}_{\mathrm{v}}$	Generalized compliance tensor
$\mu,\ \lambda$	Lamé parameters
$\mu_{\mathrm{v}},\ \lambda_{\mathrm{v}},\ \mu_{\mathrm{v}}',\ \lambda_{\mathrm{v}}'$	Visco-elasticity-specific Lamé parameters
E	Young's modulus
ν	Poisson's ratio
K_{dr}	Drained bulk modulus
$W(\boldsymbol{\varepsilon}(\boldsymbol{u}))$	Strain energy density
$\frac{1}{M}$	Storage coefficient
$b(p)$	Nonlinear compressibility
$\alpha,\ \alpha_{\mathrm{v}},\ \alpha_{\mathrm{T}}$	Biot coefficients
α_φ	Thermo-hydro coupling coefficient
$\boldsymbol{\kappa}$	Hydraulic conductivity / permeability
$\boldsymbol{\kappa}_{\mathrm{F}}$	Thermal conductivity
$\nu(\lvert\boldsymbol{q}\rvert)$	Fluid viscosity
C_{d}	Total volumetric heat capacity

Acknowledgements

This work is supported in part by the Research Council of Norway Project 250223. The authors also acknowledge the support from the University of Bergen.

References

[1] Y. Komura, "Nonlinear semi-groups in Hilbert space," *Journal of the Mathematical Society of Japan*, vol. 19, no. 4, pp. 493–507, 1967.

[2] M. G. Crandall and A. Pazy, "Semi-groups of nonlinear contractions and dissipative sets," *Journal of Functional Analysis*, vol. 3, no. 3, pp. 376 – 418, 1969.

[3] H. Brezis, "Monotonicity Methods in Hilbert Spaces and Some Applications to Nonlinear Partial Differential Equations," in *Contributions to Nonlinear Functional Analysis* (E. H. Zarantonello, ed.), pp. 101 – 156, Academic Press, 1971.

[4] H. Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert.* North-Holland Mathematics Studies, Elsevier Science, 1973.

[5] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures (LECTURES IN MATHEMATICS ETH ZURICH).* Birkhäuser Basel, 2005.

[6] F. Otto, "The geometry of dissipative evolution equations: The porous medium equation," *Communications in Partial Differential Equations*, vol. 26, no. 1-2, pp. 101–174, 2001.

[7] M. A. Peletier, "Variational Modelling: Energies, gradient flows, and large deviations," *arXiv e-prints*, p. arXiv:1402.1990, Feb 2014.

[8] C. Cancès, T. O. Gallouët, and L. Monsaingeon, "The gradient flow structure for incompressible immiscible two-phase flows in porous media," *Comptes Rendus Mathematique*, vol. 353, no. 11, pp. 985–989, 2015.

[9] A. Mielke, R. Rossi, and G. Savaré, "Nonsmooth analysis of doubly nonlinear evolution equations," *Calculus of Variations and Partial Differential Equations*, vol. 46, pp. 253–310, Jan 2013.

[10] A. Mielke, R. Rossi, and G. Savaré, "Global existence results for viscoplasticity at finite strain," *Archive for Rational Mechanics and Analysis*, vol. 227, no. 1, pp. 423–475, 2018.

[11] P. Colli, "On some doubly nonlinear evolution equations in Banach spaces," *Japan Journal of Industrial and Applied Mathematics*, vol. 9, p. 181, Jun 1992.

[12] R. H. Nochetto, G. Savaré, and C. Verdi, "A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations," *Communications on Pure and Applied Mathematics*, vol. 53, no. 5, pp. 525–589, 2000.

[13] S. Bartels, R. Nochetto, and A. Salgado, "Discrete Total Variation Flows without Regularization," *SIAM Journal on Numerical Analysis*, vol. 52, no. 1, pp. 363–385, 2014.

[14] A. Jüngel, U. Stefanelli, and L. Trussardi, "Two time discretizations for gradient flows exactly replicating energy dissipation," *arXiv e-prints*, p. arXiv:1811.06033, Nov. 2018.

[15] P. Segall and S. D. Fitzgerald, "A note on induced stress changes in hydrocarbon and geothermal reservoirs," *Tectonophysics*, vol. 289, no. 1, pp. 117 – 128, 1998.

[16] T. I. Bjørnarå, J. M. Nordbotten, and J. Park, "Vertically integrated models for coupled two-phase flow and geomechanics in porous media," *Water Resources Research*, vol. 52, no. 2, pp. 1398–1417, 2016.

[17] W. Hong, X. Zhao, J. Zhou, and Z. Suo, "A theory of coupled diffusion and large deformation in polymeric gels," *Journal of the Mechanics and Physics of Solids*, vol. 56, no. 5, pp. 1779 – 1793, 2008.

[18] S. C. Cowin, "Bone poroelasticity," *Journal of Biomechanics*, vol. 32, no. 3, pp. 217 – 238, 1999.

[19] K. Terzaghi and O. Fröhlich, *Theorie der Setzung von Tonschichten: Eine Einführung in die analytische Tonmechanik.* Franz Deuticke, 1936.

[20] M. Biot, "General theory of three-dimensional consolidation," *Journal of applied physics*, vol. 12, no. 2, pp. 155–164, 1941.

[21] R. Lewis and B. Schrefler, *The finite element method in the static and dynamic deformation and consolidation of porous media.* Numerical methods in engineering, John Wiley, 1998.

[22] O. Coussy, *Poromechanics.* Wiley, 2004.

[23] R. Burridge and J. B. Keller, "Poroelasticity equations derived from microstructure," *The Journal of the Acoustical Society of America*, vol. 70, no. 4, pp. 1140–1146, 1981.

[24] A. Mikelić and M. F. Wheeler, "Theory of the dynamic Biot-Allard equations and their link to the quasi-static Biot system," *Journal of Mathematical Physics*, vol. 53, no. 12, p. 123702, 2012.

[25] M. K. Brun, I. Berre, J. M. Nordbotten, and F. A. Radu, "Upscaling of the Coupling of Hydromechanical and Thermal Processes in a Quasi-static Poroelastic Medium," *Transport in Porous Media*, vol. 124, pp. 137–158, Aug 2018.

[26] C. J. Van Duijn, A. Mikelic, M. Wheeler, and T. Wick, "Thermoporoelasticity via homogenization I. Modeling and formal two-scale expansions ." working paper or preprint, Nov. 2017.

[27] R. Showalter, "Diffusion in Poro-Elastic Media," *Journal of Mathematical Analysis and Applications*, vol. 251, no. 1, pp. 310 – 340, 2000.

[28] M. K. Brun, E. Ahmed, J. M. Nordbotten, and F. A. Radu, "Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport," *Journal of Mathematical Analysis and Applications*, vol. 471, no. 1, pp. 239 – 266, 2019.

[29] L. Bociu, G. Guidoboni, R. Sacco, and J. T. Webster, "Analysis of Nonlinear Poro-Elastic and Poro-Visco-Elastic Models," *Archive for Rational Mechanics and Analysis*, vol. 222, Dec 2016.

[30] M. Borregales, F. A. Radu, K. Kumar, and J. M. Nordbotten, "Robust iterative schemes for non-linear poromechanics," *Computational Geosciences*, vol. 22, pp. 1021–1038, Aug 2018.

[31] J. B. Haga, H. Osnes, and H. P. Langtangen, "On the causes of pressure oscillations in low-permeable and low-compressible porous media," *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 36, no. 12, pp. 1507–1522, 2012.

[32] E. Keilegavlen and J. M. Nordbotten, "Finite volume methods for elasticity with weak symmetry," *International Journal for Numerical Methods in Engineering*, vol. 112, no. 8, pp. 939–962.

[33] I. Ambartsumyan, E. Khattatov, J. M. Nordbotten, and I. Yotov, "A multipoint stress mixed finite element method for elasticity I: Simplicial grids," *arXiv e-prints*, p. arXiv:1805.09920, May 2018.

[34] J. Korsawe and G. Starke, "A Least-Squares Mixed Finite Element Method for Biot's Consolidation Problem in Porous Media," *SIAM Journal on Numerical Analysis*, vol. 43, no. 1, pp. 318–339, 2005.

[35] J. Kim, H. Tchelepi, and R. Juanes, "Stability and convergence of sequential methods for coupled flow and geomechanics: Drained and undrained splits," *Computer Methods in Applied Mechanics and Engineering*, vol. 200, no. 23, pp. 2094 – 2116, 2011.

[36] J. B. Haga, H. Osnes, and H. P. Langtangen, "A parallel block preconditioner for large-scale poroelasticity with highly heterogeneous material parameters," *Computational Geosciences*, vol. 16, pp. 723–734, Jun 2012.

[37] A. Mikelić and M. F. Wheeler, "Convergence of iterative coupling for coupled flow and geomechanics," *Computational Geosciences*, vol. 17, pp. 455–461, Jun 2013.

[38] T. Almani, K. Kumar, A. Dogru, G. Singh, and M. Wheeler, "Convergence analysis of multirate fixed-stress split iterative schemes for coupling flow with geomechanics," *Computer Methods in Applied Mechanics and Engineering*, vol. 311, pp. 180 – 207, 2016.

[39] N. Castelletto, J. A. White, and M. Ferronato, "Scalable algorithms for three-field mixed finite element coupled poromechanics," *Journal of Computational Physics*, vol. 327, pp. 894 – 918, 2016.

[40] J. A. White, N. Castelletto, and H. A. Tchelepi, "Block-partitioned solvers for coupled poromechanics: A unified framework," *Computer Methods in Applied Mechanics and Engineering*, vol. 303, pp. 55 – 74, 2016.

[41] J. Both, M. Borregales, J. Nordbotten, K. Kumar, and F. Radu, "Robust fixed stress splitting for Biot's equations in heterogeneous media," *Applied Mathematics Letters*, vol. 68, pp. 101 – 108, 2017.

[42] M. Bause, F. Radu, and U. Köcher, "Space–time finite element approximation of the Biot poroelasticity system with iterative coupling," *Computer Methods in Applied Mechanics and Engineering*, vol. 320, pp. 745 – 768, 2017.

[43] J. H. Adler, F. J. Gaspar, X. Hu, C. Rodrigo, and L. T. Zikatanov, "Robust block preconditioners for Biot's model," in *International Conference on Domain Decomposition Methods*, pp. 3–16, Springer, 2017.

[44] N. Castelletto, S. Klevtsov, H. Hajibeygi, and H. A. Tchelepi, "Multiscale two-stage solver for biot's poroelasticity equations in subsurface media," *Computational Geosciences*, vol. 23, pp. 207–224, Apr 2019.

[45] J. Kim, "Unconditionally stable sequential schemes for all-way coupled thermoporomechanics: Undrained-adiabatic and extended fixed-stress splits," *Computer Methods in Applied Mechanics and Engineering*, vol. 341, pp. 93 – 112, 2018.

[46] J. Kim, "A new numerically stable sequential algorithm for coupled finite-strain elastoplastic geomechanics and flow," *Computer Methods in Applied Mechanics and Engineering*, vol. 335, pp. 538 – 562, 2018.

[47] C. Miehe, S. Mauthe, and S. Teichtmeister, "Minimization principles for the coupled problem of Darcy–Biot-type fluid transport in porous media linked to phase field modeling of fracture," *Journal of the Mechanics and Physics of Solids*, vol. 82, pp. 186 – 217, 2015.

[48] J. Kim, H. Tchelepi, and R. Juanes, "Stability and convergence of sequential methods for coupled flow and geomechanics: Fixed-stress and fixed-strain splits," *Computer Methods in Applied Mechanics and Engineering*, vol. 200, no. 13, pp. 1591 – 1606, 2011.

[49] R. Rossi and G. Savaré, "Gradient flows of non convex functionals in hilbert spaces and applications," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 12, no. 3, p. 564–614, 2006.

[50] I. Ekeland and R. Temam, *Convex analysis and variational problems*, vol. 28. Siam, 1999.

[51] P. Colli and A. Visintin, "On A Class Of Doubly Nonlinear Evolution Equations," *Communications in Partial Differential Equations*, vol. 15, no. 5, pp. 737–756, 1990.

[52] L. Ambrosio, "Minimizing movements," *Rend. Accad. Naz. Sci. XL Mem. Mat. Appl.(5)*, vol. 19, pp. 191–246, 1995.

[53] D. Bertsekas, *Nonlinear Programming.* 1999.

[54] Z. Q. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *Journal of Optimization Theory and Applications*, vol. 72, pp. 7–35, Jan 1992.

[55] R. Adams and J. Fournier, *Sobolev Spaces.* Pure and Applied Mathematics, Elsevier Science, 2003.

[56] R. T. Rockafellar, *Convex Analysis.* Princeton University Press, 1970.

[57] P. G. Ciarlet, G. Geymonat, and F. Krasucki, "Legendre–Fenchel duality in elasticity," *Comptes Rendus Mathematique*, vol. 349, no. 9, pp. 597 – 602, 2011.

[58] R. Temam and A. Miranville, *Mathematical Modeling in Continuum Mechanics.* Cambridge University Press, 2000.

[59] E. Ahmed, F. A. Radu, and J. M. Nordbotten, "Adaptive poromechanics computations based on a posteriori error estimates for fully mixed formulations of Biot's consolidation model," *Computer Methods in Applied Mechanics and Engineering*, vol. 347, pp. 264 – 294, 2019.

[60] T. Bærland, J. J. Lee, K.-A. Mardal, and R. Winther, "Weakly Imposed Symmetry and Robust Preconditioners for Biot's Consolidation Model," *Computational Methods in Applied Mathematics*, vol. 17, no. 3, pp. 377–396, 2017.

[61] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications.* Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 2013.

[62] W. Van der Knaap *et al.*, "Nonlinear behavior of elastic porous media," 1959.

[63] M. A. Biot, "Nonlinear and semilinear rheology of porous solids," *Journal of Geophysical Research*, vol. 78, no. 23, pp. 4924–4937, 1973.

[64] B. O. Hardin and V. P. Drnevich, "Shear modulus and damping in soils: design equations and curves," *Journal of Soil Mechanics & Foundations Div*, vol. 98, no. sm7, 1972.

[65] H. Barucq, M. Madaune-Tort, and P. Saint-Macary, "On nonlinear Biot's consolidation models," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 63, no. 5-7, pp. e985–e995, 2005.

[66] R. G. Owens and T. N. Phillips, *Computational Rheology.* Imperial College Press, 2002.

[67] I. Ambartsumyan, V. J. Ervin, T. Nguyen, and I. Yotov, "A nonlinear Stokes-Biot model for the interaction of a non-Newtonian fluid with poroelastic media I: well-posedness of the model," *arXiv e-prints*, p. arXiv1803.00947, Mar. 2018.

[68] P. Forchheimer, "Wasserbewegung durch boden," *Z. Ver. Deutsch, Ing.*, vol. 45, pp. 1782–1788, 1901.

[69] H. C. Brinkman, "A calculation of the viscous force exerted by a flowing fluid on a dense swarm of particles," *Flow, Turbulence and Combustion*, vol. 1, p. 27, Dec 1949.

[70] L. Grippof and M. Sciandrone, "Globally convergent block-coordinate techniques for unconstrained optimization," *Optimization Methods and Software*, vol. 10, no. 4, pp. 587–637, 1999.

[71] L. Grippo and M. Sciandrone, "On the Convergence of the Block Nonlinear Gauss-Seidel Method Under Convex Constraints," *Oper. Res. Lett.*, vol. 26, pp. 127–136, Apr. 2000.

[72] E. Storvik, J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu, "On the optimization of the fixed-stress splitting for Biot's equations," *International Journal for Numerical Methods in Engineering*, vol. 0, no. ja.

[73] J. W. Both and U. Köcher, "Numerical investigation on the fixed-stress splitting scheme for Biot's equations: Optimality of the tuning parameter," *ArXiv e-prints*, Jan. 2018.

[74] M. Ferronato, N. Castelletto, and G. Gambolati, "A fully coupled 3-D mixed finite element model of Biot consolidation," *Journal of Computational Physics*, vol. 229, no. 12, pp. 4813 – 4830, 2010.

[75] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: a general approach," *Annals of Operations Research*, vol. 46, pp. 157–178, Mar 1993.

[76] X.-C. Tai and M. Espedal, "Rate Of Convergence Of Some Space Decomposition Methods For Linear And Nonlinear Problems," *SIAM J. Numer. Anal*, vol. 35, pp. 1558–1570, 1998.

[77] X.-C. Tai and J. Xu, "Global and Uniform Convergence of Subspace Correction Methods for Some Convex Optimization Problems," *Math. Comp*, vol. 71, pp. 105–124, 2001.

[78] J. Xu, "The method of subspace corrections," *Journal of Computational and Applied Mathematics*, vol. 128, no. 1, pp. 335 – 362, 2001. Numerical Analysis 2000. Vol. VII: Partial Differential Equations.

[79] A. Beck and L. Tetruashvili, "On the Convergence of Block Coordinate Descent Type Methods," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2037–2060, 2013.

[80] K. Kumar, S. Matculevich, J. Nordbotten, and S. Repin, "Guaranteed and computable bounds of approximation errors for the semi-discrete Biot problem," *arXiv e-prints*, p. arXiv:1808.08036, Aug 2018.

[81] H. C. Yoon and J. Kim, "Spatial stability for the monolithic and sequential methods with various space discretizations in poroelasticity," *International Journal for Numerical Methods in Engineering*, vol. 114, no. 7, pp. 694–718.

[82] S. Dana and M. F. Wheeler, "Convergence analysis of two-grid fixed stress split iterative scheme for coupled flow and deformation in heterogeneous poroelastic media," *Computer Methods in Applied Mechanics and Engineering*, vol. 341, pp. 788 – 806, 2018.

[83] S. Dana and M. Wheeler, "Convergence analysis of fixed stress split iterative scheme for anisotropic poroelasticity with tensor Biot parameter," *Computational Geosciences*, 04 2018.

[84] F. List and F. A. Radu, "A study on iterative methods for solving Richards' equation," *Computational Geosciences*, vol. 20, pp. 341–353, Apr 2016.

[85] M. K. Brun, E. Ahmed, I. Berre, J. M. Nordbotten, and F. A. Radu, "Monolithic and splitting based solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport," *arXiv e-prints*, p. arXiv:1902.05783, Feb 2019.

[86] H. Holden, K. Karlsen, and K. Lie, *Splitting Methods for Partial Differential Equations with Rough Solutions: Analysis and MATLAB Programs.* EMS series of lectures in mathematics, European Mathematical Society, 2010.

[87] F. J. Gaspar, J. L. Gracia, F. J. Lisbona, and C. W. Oosterlee, "Distributive smoothers in multigrid for problems with dominating grad–div operators," *Numerical Linear Algebra with Applications*, vol. 15, no. 8, pp. 661–683, 2008.

[88] J. H. Adler, F. J. Gaspar, X. Hu, P. Ohm, C. Rodrigo, and L. T. Zikatanov, "Robust preconditioners for a new stabilized discretization of the poroelastic equations," *arXiv e-prints*, p. arXiv:1905.10353, May 2019.

[89] M. Blatt, A. Burchardt, A. Dedner, C. Engwer, J. Fahlke, B. Flemisch, C. Gersbacher, C. Gräser, F. Gruber, C. Grüninger, *et al.*, "The distributed and unified numerics environment, version 2.4," *Archive of Numerical Software*, vol. 4, no. 100, pp. 13–29.

[90] C. Engwer, C. Gräser, S. Müthing, and O. Sander, "The interface for functions in the dune-functions module," *arXiv preprint arXiv:1512.06136*, 2015.

[91] C. Engwer, C. Gräser, S. Müthing, and O. Sander, "Function space bases in the dune-functions module," *arXiv e-prints*, p. arXiv:1806.09545, Jun 2018.

[92] N. Castelletto, J. A. White, and H. A. Tchelepi, "Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics," *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 39, no. 14, pp. 1593–1618, 2015.

[93] J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu, "Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media," *Computers & Mathematics with Applications*, 2018.

[94] Q. Hong, J. Kraus, M. Lymbery, and M. Fanett Wheeler, "Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems," *arXiv e-prints*, p. arXiv:1812.11809, Dec 2018.

[95] G. Francfort and A. Mielke, "Existence results for a class of rate-independent material models with nonconvex elastic energies," *Journal fur die Reine und Angewandte Mathematik*, no. 595, pp. 55–91, 2006.

[96] Y. Nesterov, *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

Paper B

Robust fixed stress splitting for Biot's equations in heterogeneous media

Both, J.W, Borregales, M., Kumar, K., Nordbotten, J.M., and Radu, F.A.

Applied Mathematics Letters **68** (2016), pg. 101–108.

doi:10.1016/j.aml.2016.12.019

Applied Mathematics Letters 68 (2017) 101–108

Contents lists available at ScienceDirect

Applied Mathematics Letters

www.elsevier.com/locate/aml

Robust fixed stress splitting for Biot's equations in heterogeneous media

Jakub Wiktor Both[a,*], Manuel Borregales[a], Jan Martin Nordbotten[a,b], Kundan Kumar[a], Florin Adrian Radu[a]

[a] *Department of Mathematics, University of Bergen, Bergen, Norway*
[b] *Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA*

ARTICLE INFO

Article history:
Received 31 October 2016
Received in revised form 30 December 2016
Accepted 30 December 2016
Available online 12 January 2017

Keywords:
Linear poroelasticity
Biot's equations
Iterative coupling
Heterogeneous porous media

ABSTRACT

We study the iterative solution of coupled flow and geomechanics in heterogeneous porous media, modeled by a three-field formulation of the linearized Biot's equations. We propose and analyze a variant of the widely used Fixed Stress Splitting method applied to heterogeneous media. As spatial discretization, we employ linear Galerkin finite elements for mechanics and mixed finite elements (lowest order Raviart–Thomas elements) for flow. Additionally, we use implicit Euler time discretization. The proposed scheme is shown to be globally convergent with optimal theoretical convergence rates. The convergence is rigorously shown in energy norms employing a new technique. Furthermore, numerical results demonstrate robust iteration counts with respect to the full range of Lamé parameters for homogeneous and heterogeneous media. Being in accordance with the theoretical results, the iteration count is hardly influenced by the degree of heterogeneities.

1. Introduction

The coupling of mechanics and flow in porous media is relevant for many applications ranging from environmental engineering to biomedical engineering. The simplest model of real applied importance is the quasi-static linearized Biot system, applicable for infinitesimally deforming, fully saturated porous media. Existence, uniqueness and regularity for Biot's equations have been investigated first by Showalter [1].

There are two approaches currently employed for solving Biot's equations. They are referred to as fully-implicit and iterative coupling [2]. The fully-implicit approach involves solving the fully coupled system of governing equations simultaneously, providing the benefit of unconditional stability. It requires advanced and efficient preconditioners. For this purpose, (Schur complement based) block preconditioners appear to be a sound choice [3–6]. The iterative coupling approach involves the sequential-implicit solution of flow and

[*] Corresponding author.
E-mail addresses: jakub.both@uib.no (J.W. Both), manuel.borregales@uib.no (M. Borregales), jan.nordbotten@uib.no (J.M. Nordbotten), kundan.kumar@uib.no (K. Kumar), florin.radu@uib.no (F.A. Radu).

http://dx.doi.org/10.1016/j.aml.2016.12.019
0893-9659/© 2017 Elsevier Ltd. All rights reserved.

mechanics using the latest solution information, iterating the procedure at each time step until convergence. The sequential-implicit approach offers greater flexibility in code design than the fully-implicit approach. On the other hand, being equivalent to a preconditioned Richardson method [7], sequential-implicit approaches also provide a basis to design efficient block preconditioners for the fully-implicit approach [8,9]. Among iterative coupling schemes, the widely used Fixed Stress Splitting method has been rigorously shown to be unconditionally stable in the sense of a Von Neumann analysis [10] and globally convergent [11], when considering slightly compressible flow in a homogeneous porous medium.

The new contributions of this work are:

- We prove global, linear convergence in energy norms of the Fixed Stress Splitting method applied to the fully discretized three-field formulation of Biot's equations for heterogeneous media, where linear finite elements are employed for mechanics, mixed finite elements (lowest order Raviart–Thomas elements) are employed for flow, and backward Euler time discretization is applied.
- We propose a new, optimized tuning parameter for heterogeneous media.

In the case of homogeneous media, the results are in consistency with previous numerical studies, cf., e.g., [12]. To the best of our knowledge, this is the first time the convergence of the Fixed Stress Splitting method is rigorously shown for energy norms and considering heterogeneous media.

2. Mathematical model — Biot's equations

We consider the quasi-static Biot's equations [13,14], modeling a linearly elastic porous medium $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, saturated with a slightly compressible fluid. On the space–time domain $\Omega \times (0,T)$, the governing equations read

$$-\boldsymbol{\nabla}\cdot[2\mu\boldsymbol{\varepsilon}(\boldsymbol{u})+\lambda\boldsymbol{\nabla}\cdot\boldsymbol{u}]+\alpha\boldsymbol{\nabla}p=\boldsymbol{f}, \qquad \partial_t\left(\frac{p}{M}+\alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}\right)+\boldsymbol{\nabla}\cdot\boldsymbol{w}=S_f, \qquad \boldsymbol{K}^{-1}\boldsymbol{w}+\boldsymbol{\nabla}p=\rho_f\boldsymbol{g}. \quad (1)$$

Here, $\boldsymbol{u}$ is the displacement, p is the fluid pressure, $\boldsymbol{w}$ is the Darcy flux, $\boldsymbol{\varepsilon}(\boldsymbol{u})=0.5(\boldsymbol{\nabla}\boldsymbol{u}+\boldsymbol{\nabla}\boldsymbol{u}^\top)$ is the linearized strain tensor, μ, λ are the Lamé parameters, α is the Biot coefficient, M is the Biot modulus, ρ_f is the fluid density, $\boldsymbol{K}$ is the permeability tensor divided by fluid viscosity, $\boldsymbol{g}$ is the gravity vector, and S_f is a volume source term. For simplicity, we assume homogeneous boundary $\boldsymbol{u}=\boldsymbol{0}$, $p=0$ on $\partial\Omega\times[0,T]$ and initial conditions $\boldsymbol{u}=\boldsymbol{u}_0$, $p=p_0$ in $\Omega\times\{0\}$. We make the following assumptions on the effective coefficients:

(A1) Let $\rho_f \in \mathbb{R}$, $\boldsymbol{g}\in\mathbb{R}^d$ be constant.

(A2) Let $M,\alpha,\mu,\lambda \in L^\infty(\Omega)$ be positive, uniformly bounded, with the lower bound strictly positive.

(A3) Let $\boldsymbol{K}\in L^\infty(\Omega)^{d\times d}$ be a symmetric matrix, which is constant in time and has uniformly bounded eigenvalues, i.e., there exist constants $k_m, k_M \in \mathbb{R}$, satisfying for all $\boldsymbol{x}\in\Omega$ and for all $\boldsymbol{z}\in\mathbb{R}^d\setminus\{\boldsymbol{0}\}$

$$0 < k_m\boldsymbol{z}^\top\boldsymbol{z} \le \boldsymbol{z}^\top\boldsymbol{K}(x)\boldsymbol{z} \le k_M\boldsymbol{z}^\top\boldsymbol{z} < \infty.$$

Below, we consider a numerical approximation of the weak solution of Biot's equations as described above.

3. Fixed stress splitting for the fully discretized system

Let $\mathcal{T}_h$ be a regular decomposition of mesh size h of the domain Ω. Furthermore, let $0=t_0<t_1<\cdots<t_N=T$, $N\in\mathbb{N}$, define a partition of the time interval $(0,T)$ with constant time step size $\tau=t_{k+1}-t_k$, $k\ge 0$. In order to discretize Biot's equations in space, we use linear, constant and lowest order Raviart–Thomas

elements to approximate the displacement, pressure and flux, respectively. The corresponding discrete spaces are given by

$$\begin{aligned}
&\boldsymbol{V}_h = \left\{ \boldsymbol{v}_h \in [H_0^1(\Omega)]^d \,\middle|\, \forall T \in \mathcal{T}_h, \boldsymbol{v}_h|_T \in [\mathbb{P}_1]^d \right\}, \qquad Q_h = \left\{ q_h \in L^2(\Omega) \,\middle|\, \forall T \in \mathcal{T}_h, q_h|_T \in \mathbb{P}_0 \right\},\\
&\boldsymbol{Z}_h = \left\{ \boldsymbol{z}_h \in H(\mathrm{div};\Omega) \,\middle|\, \forall T \in \mathcal{T}_h, \boldsymbol{z}_h|_T(\boldsymbol{x}) = \boldsymbol{a} + b\boldsymbol{x}, \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R} \right\},
\end{aligned}$$

where $\mathbb{P}_0$ and $\mathbb{P}_1$ denote the spaces of scalar piecewise constant and piecewise linear functions, respectively. Additionally, we use backward Euler time discretization in order to discretize Biot's equations in time.

Let $\langle \cdot, \cdot \rangle$ denote the standard $L^2(\Omega)$ scalar product. Then for given initial values $(\boldsymbol{u}_h^0, \boldsymbol{v}_h^0, \boldsymbol{w}_h^0) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$, the fully-implicit discretization reads: For all $n \in \mathbb{N}$, $n \geq 1$, given $(\boldsymbol{u}_h^{n-1}, p_h^{n-1}, \boldsymbol{w}_h^{n-1}) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$, find the current displacement, pressure and flux fields $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{w}_h^n) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$, satisfying for all $(\boldsymbol{v}_h, q_h, \boldsymbol{z}_h) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$

$$\langle 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\rangle + \langle \lambda \boldsymbol{\nabla}\cdot\boldsymbol{u}_h^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle - \langle \alpha p_h^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle = \langle \boldsymbol{f}, \boldsymbol{v}_h\rangle, \tag{2}$$

$$\left\langle \frac{1}{M} p_h^n, q_h \right\rangle + \langle \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^n, q_h\rangle + \tau\langle \boldsymbol{\nabla}\cdot\boldsymbol{w}_h^n, q_h\rangle = \tau\langle S_f, q_h\rangle + \left\langle \frac{1}{M} p_h^{n-1}, q_h\right\rangle + \langle \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n-1}, q_h\rangle, \tag{3}$$

$$\langle \boldsymbol{K}^{-1}\boldsymbol{w}_h^n, \boldsymbol{z}_h\rangle - \langle p_h^n, \boldsymbol{\nabla}\cdot\boldsymbol{z}_h\rangle = \langle \rho_f \boldsymbol{g}, \boldsymbol{z}_h\rangle. \tag{4}$$

Instead of solving system (2)–(4) in a fully coupled manner, a popular alternative is to use iterative methods, which decouple mechanics and flow problems and allow for an efficient solution of the separate subproblems. Here, we limit our considerations to the widely used Fixed Stress Splitting method and adapt the idea by Mikelić and Wheeler [11], which considers keeping an artificial volumetric stress constant. Nevertheless, the same ideas can be also used to prove the convergence of the optimized Undrained Splitting scheme.

The iterative scheme defines a sequence $(\boldsymbol{u}_h^{n,i}, p_h^{n,i}, \boldsymbol{w}_h^{n,i})$, $i \geq 0$. After initialization $\boldsymbol{u}_h^{n,0} = \boldsymbol{u}_h^{n-1}$, $p_h^{n,0} = p_h^{n-1}$, and $\boldsymbol{w}_h^{n,0} = \boldsymbol{w}_h^{n-1}$, each iterate is defined in two steps. First, the flow problem is solved independently, keeping the artificial volumetric stress $\sigma_\beta = \sigma_0 + K_{dr}\boldsymbol{\nabla}\cdot\boldsymbol{u} - \alpha p$ constant, which introduces a tuning parameter $K_{dr} \in L^\infty(\Omega)$ (classically the drained bulk modulus). Equivalently, we consider the tuning parameter $\beta_{FS} = \alpha^2/K_{dr}$. Second, the mechanics problem is solved using updated pressure and flux. For fixed $n, i \in \mathbb{N}$, the detailed splitting scheme reads as follows:

Step 1: Given $(\boldsymbol{u}_h^{n,i-1}, p_h^{n,i-1}, \boldsymbol{w}_h^{n,i-1}) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$, find $(p_h^{n,i}, \boldsymbol{w}_h^{n,i}) \in Q_h \times \boldsymbol{Z}_h$ s.t. for all $(q_h, \boldsymbol{z}_h) \in Q_h \times \boldsymbol{Z}_h$ it holds

$$\begin{aligned}
\left\langle \left(\frac{1}{M} + \beta_{FS}\right) p_h^{n,i}, q_h \right\rangle + \tau\langle \boldsymbol{\nabla}\cdot\boldsymbol{w}_h^{n,i}, q_h\rangle &= \tau\langle S_f, q_h\rangle + \left\langle \frac{1}{M} p_h^{n-1}, q_h\right\rangle + \langle \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n-1}, q_h\rangle\\
&\quad + \langle \beta_{FS} p_h^{n,i-1}, q_h\rangle - \langle \alpha\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n,i-1}, q_h\rangle,
\end{aligned} \tag{5}$$

$$\langle \boldsymbol{K}^{-1}\boldsymbol{w}_h^{n,i}, \boldsymbol{z}_h\rangle - \langle p_h^{n,i}, \boldsymbol{\nabla}\cdot\boldsymbol{z}_h\rangle = \langle \rho_f \boldsymbol{g}, \boldsymbol{z}_h\rangle. \tag{6}$$

Step 2: Given $p_h^{n,i} \in Q_h$, find $\boldsymbol{u}_h^{n,i} \in \boldsymbol{V}_h$ such that for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$ it holds

$$\langle 2\mu\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\rangle + \langle \lambda\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n,i}, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle = \langle \boldsymbol{f}, \boldsymbol{v}_h\rangle + \langle \alpha p_h^{n,i}, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle. \tag{7}$$

In the following, we consider three tuning parameters — the classical, physically motivated choice β_{FS}^{cl}, cf., e.g., [10], and the parameters β_{FS}^{λ}, β_{FS}^{opt}, revealed by the analysis of Mikelić and Wheeler [11,12], which is valid for homogeneous Lamé parameters. The latter parameter is also revealed by the present convergence analysis, valid for heterogeneous Lamé parameters. More precisely, the parameters are given by

$$\beta_{FS}^{cl} = \frac{\alpha^2}{\frac{2\mu}{d} + \lambda}, \qquad \beta_{FS}^{\lambda} = \frac{\alpha^2}{2\lambda}, \qquad \beta_{FS}^{opt} = \frac{\alpha^2}{2\left(\frac{2\mu}{d} + \lambda\right)}. \tag{8}$$

4. Convergence analysis

We prove linear convergence of the Fixed Stress Splitting method, when applied to Biot's equations in heterogeneous media. For this purpose, we show a contraction with respect to energy norms, making use of the following lemma and remark. We refer to the Supplementary material (see Appendix A) for further standard lemmas used in the proof. Furthermore, in the Supplementary material, the proof is repeated for homogeneous media in a simpler, but a more detailed form.

Lemma 1 (*Thomas' Lemma, [15]*). *There exists a constant $C_{\Omega,d} > 0$ not depending on the mesh size h, such that given an arbitrary $q_h \in Q_h$ there exists $\boldsymbol{z}_h \in \boldsymbol{Z}_h$, satisfying $\boldsymbol{\nabla} \cdot \boldsymbol{z}_h = q_h$ and $\|\boldsymbol{z}_h\| \leq C_{\Omega,d}\|q_h\|$.*

Remark 1 (*Weighted $L^2(\Omega)^d$ Norms*). Let $k \in \{1, d\}$. Let further $\boldsymbol{A} \in [L^\infty(\Omega)]^{k\times k}$ be a symmetric, uniformly positive definite matrix and let its eigenvalues be uniformly bounded, i.e., there exist constants $a_m, a_M \in \mathbb{R}$ such that for all eigenvalues $\lambda(\boldsymbol{x})$ of matrix $\boldsymbol{A}(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, it holds $0 < a_m \leq \lambda(x) \leq a_M \leq \infty$. Then, we define a weighted scalar product $\langle \cdot, \cdot \rangle_{\boldsymbol{A}}$ on $L^2(\Omega)^d$ by $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\boldsymbol{A}} = \langle \boldsymbol{A}\boldsymbol{f}, \boldsymbol{g} \rangle$, $\boldsymbol{f}, \boldsymbol{g} \in L^2(\Omega)^d$. Let $\|\cdot\|_{\boldsymbol{A}}$ denote the corresponding norm. Then it holds $\forall \boldsymbol{f}, \boldsymbol{g} \in L^2(\Omega)^d$

$$a_m\|\boldsymbol{f}\|^2 \leq \|\boldsymbol{f}\|_{\boldsymbol{A}}^2 \leq a_M\|\boldsymbol{f}\|^2, \qquad \langle \boldsymbol{f}, \boldsymbol{g} \rangle \leq \|\boldsymbol{f}\|_{\boldsymbol{A}}\|\boldsymbol{g}\|_{\boldsymbol{A}^{-1}}.$$

Theorem 2 (*Linear Convergence for Fixed Stress Splitting*). *Assume* (A1)–(A3). *Let $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{w}_h^n)$ and $(\boldsymbol{u}_h^{n,i}, p_h^{n,i}, \boldsymbol{w}_h^{n,i})$ be the solutions of Eqs.* (2)–(4) *and Eqs.* (5)–(7)*, respectively. Let $e_{\boldsymbol{u}}^i = \boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^n$, $e_p^i = p_h^{n,i} - p_h^n$ and $e_{\boldsymbol{w}}^i = \boldsymbol{w}_h^{n,i} - \boldsymbol{w}_h^n$ denote the errors at current iteration. Then for all $\beta_{FS} \in L^\infty(\Omega)$, satisfying*

$$\beta_{FS} \geq \frac{\alpha^2}{2(\frac{2\mu}{d} + \lambda)} \quad \text{on } \Omega, \tag{9}$$

for all $i \geq 1$, it holds

$$\|e_p^i\|_{\beta_{FS}}^2 \leq \left\| \frac{\frac{\beta_{FS}}{2}}{\frac{1}{M} + \frac{\beta_{FS}}{2} + \frac{\tau k_m}{C_{\Omega,d}^2}} \right\|_\infty \|e_p^{i-1}\|_{\beta_{FS}}^2, \tag{10}$$

$$\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^i)\|_{2\mu}^2 + \|\boldsymbol{\nabla} \cdot e_{\boldsymbol{u}}^i\|_\lambda^2 \leq \|e_p^i\|_{\frac{\alpha^2}{\frac{2\mu}{d}+\lambda}}^2. \tag{11}$$

Optimal convergence rates are obtained in case of equality in Eq. (9).

Proof. Due to Assumptions (A1)–(A3), all effective coefficients fulfill the requirements for defining weighted $L^2(\Omega)$-norms, cf. Remark 1. Throughout the proof we make use of weighted norms without further comment.

Step 1: Flow and mechanics. By taking the differences of corresponding Eqs. (5)–(7) and Eqs. (2)–(4), testing with $\boldsymbol{v}_h = e_{\boldsymbol{u}}^{i-1} \in \boldsymbol{V}_h$, $q_h = e_p^i \in Q_h$ and $\boldsymbol{z}_h = \tau e_{\boldsymbol{w}}^i \in \boldsymbol{Z}_h$ and adding all together, we obtain

$$\langle \boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^i), \boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^{i-1}) \rangle_{2\mu} + \langle \boldsymbol{\nabla} \cdot e_{\boldsymbol{u}}^i, \boldsymbol{\nabla} \cdot e_{\boldsymbol{u}}^{i-1} \rangle_\lambda + \|e_p^i\|_{\frac{1}{M}}^2 + \tau\|e_{\boldsymbol{w}}^i\|_{\boldsymbol{K}^{-1}}^2 + \langle e_p^i - e_p^{i-1}, e_p^i \rangle_{\beta_{FS}} = 0.$$

Using a polarization and binomial identity yields

$$\begin{aligned} &\frac{1}{4}\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^i + e_{\boldsymbol{u}}^{i-1})\|_{2\mu}^2 + \frac{1}{4}\|\boldsymbol{\nabla} \cdot (e_{\boldsymbol{u}}^i + e_{\boldsymbol{u}}^{i-1})\|_\lambda^2 - \frac{1}{4}\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_{2\mu}^2 - \frac{1}{4}\|\boldsymbol{\nabla} \cdot (e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_\lambda^2 \\ &\quad + \|e_p^i\|_{\frac{1}{M}}^2 + \tau\|e_{\boldsymbol{w}}^i\|_{\boldsymbol{K}^{-1}}^2 + \|e_p^i\|_{\frac{\beta_{FS}}{2}}^2 + \|e_p^i - e_p^{i-1}\|_{\frac{\beta_{FS}}{2}}^2 - \|e_p^{i-1}\|_{\frac{\beta_{FS}}{2}}^2 = 0. \end{aligned} \tag{12}$$

Step 2*: Mechanics*. Evaluating Eq. (7) at iteration i and $i-1$, taking the difference and testing with $\boldsymbol{v}_h = e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}$ yields

$$\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|_{2\mu}^2 + \|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_\lambda^2 = \langle e_p^i - e_p^{i-1}, \boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\rangle_\alpha.$$

Let $\gamma \in L^\infty(\Omega)$ with $\gamma(\Omega) \subset [0,1]$ and $f_\mu, f_\lambda \in L^\infty(\Omega)$, satisfying the assumptions of Remark 1. Then by applying weighted Cauchy–Schwarz inequalities, cf. Remark 1, and an arithmetic mean-root mean square inequality (AM-QM inequality), we obtain

$$\begin{aligned}
&\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|_{2\mu}^2 + \|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_\lambda^2 \\
&\quad= \langle e_p^i - e_p^{i-1}, \boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\rangle_{\gamma\alpha} + \langle e_p^i - e_p^{i-1}, \boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\rangle_{(1-\gamma)\alpha} \\
&\quad\le \|e_p^i - e_p^{i-1}\|_{\gamma\alpha^2 f_\mu^{-1}} \|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_{\gamma f_\mu} + \|e_p^i - e_p^{i-1}\|_{(1-\gamma)\alpha^2 f_\lambda^{-1}} \|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_{(1-\gamma)f_\lambda} \\
&\quad\le \|e_p^i - e_p^{i-1}\|_{\gamma\alpha^2 f_\mu^{-1}} \|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|_{\gamma d f_\mu} + \|e_p^i - e_p^{i-1}\|_{(1-\gamma)\alpha^2 f_\lambda^{-1}} \|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_{(1-\gamma)f_\lambda}.
\end{aligned}$$

By applying Young's inequality, rearranging terms and scaling, it holds for $c \in (0,\infty)$ and γ, f_μ, f_λ as above

$$\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|_{c(2\mu - \frac{1}{2}\gamma d f_\mu)}^2 + \|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_{c(\lambda - \frac{1}{2}(1-\gamma)f_\lambda)}^2 \le \|e_p^i - e_p^{i-1}\|_{\frac{\alpha^2}{2}c(\gamma f_\mu^{-1} + (1-\gamma)f_\lambda^{-1})}^2.$$

By choosing $c, \gamma, f_\mu, f_\lambda$ optimally, we finally obtain

$$\frac{1}{4}\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|_{2\mu}^2 + \frac{1}{4}\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|_\lambda^2 \le \|e_p^i - e_p^{i-1}\|_{\frac{\alpha^2}{4\left(\frac{2\mu}{d}+\lambda\right)}}^2. \tag{13}$$

Step 3*: Darcy*. Taking the difference of Eqs. (6) and (4) yields for any $\boldsymbol{z}_h \in \boldsymbol{Z}_h$

$$\langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, \boldsymbol{z}_h\rangle - \langle e_p^i, \boldsymbol{\nabla}\cdot\boldsymbol{z}_h\rangle = 0. \tag{14}$$

Using Thomas' Lemma, there exists a constant $C_{\Omega,d} > 0$ and a function $\tilde{\boldsymbol{z}}_h \in \boldsymbol{Z}_h$ satisfying $\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{z}}_h = e_p^i$ and $\|\tilde{\boldsymbol{z}}_h\| \le C_{\Omega,d}\|e_p^i\|$. Then, together with Eq. (14) and Assumption (A3), after some rearranging, we obtain

$$\frac{k_m}{C_{\Omega,d}^2}\|e_p^i\|^2 \le \langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, e_{\boldsymbol{w}}^i\rangle. \tag{15}$$

Step 4*: Combining Step* 1–3. Discarding the first two terms in Eq. (12), using Assumption (9), Eq. (13) and inserting Eq. (15) yields

$$\|e_p^i\|_{\frac{1}{M} + \frac{\beta_{FS}}{2} + \frac{\tau k_m}{C_{\Omega,d}^2}}^2 \le \|e_p^{i-1}\|_{\frac{\beta_{FS}}{2}}^2.$$

By employing Remark 1, we obtain Eq. (10).

Step 5*: Mechanics revisited*. Taking the difference of Eqs. (7) and (2), tested with $\boldsymbol{v}_h = e_{\boldsymbol{u}}^i$ yields

$$\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i\big)\|_{2\mu}^2 + \|\boldsymbol{\nabla}\cdot e_{\boldsymbol{u}}^i\|_\lambda^2 = \langle e_p^i, \boldsymbol{\nabla}\cdot e_{\boldsymbol{u}}^i\rangle_\alpha.$$

We repeat all steps from Step 2. Due to linearity, we obtain Eq. (11) analogously. □

Remark 2 (*Discussion*)**.** The above analysis covers global convergence in energy norms for all considered tuning parameters β_{FS}^{cl}, β_{FS}^{λ} and β_{FS}^{opt}, where the first two only yield sub-optimal convergence rates in the energy norms and the latter yields optimal rates, as shown by our proof. The parameter β_{FS}^{λ} recovers optimality in the limit of $\frac{\mu}{\lambda} \ll 1$. For soft materials, i.e., in the limit of $\frac{\mu}{\lambda} \gg 1$, we expect deteriorating convergence rates due to lack of dependence on μ.

Table 1
Problem parameters, chosen identically to [12].

Symb.	Quantity	Values [Unit]
E	Bulk modulus	0.594 [GPa]
α	Biot's coefficient	1
M	Biot's modulus	1.65e10 [Pa]
$\boldsymbol{K}$	Permeability tensor divided by fluid viscosity	$100\boldsymbol{I}$ [mD/cP]
$\boldsymbol{g}$	Gravity vector	$\mathbf{0}$ [m/s^2]
Δx, Δy	Grid spacing in x and y	0.025 [m]
τ	Time step size	1 [s]
δ_a	Absolute error tolerance	1e−6
δ_r	Relative error tolerance	1e−6

5. Numerical results

We analyze the robustness of the Fixed Stress Splitting scheme with respect to different Lamé parameters and compare the convergence behavior for the tuning parameters β_{FS}^{cl}, β_{FS}^{λ} and β_{FS}^{opt}. For further test cases with $d \in \{2,3\}$, we refer to the Supplementary material (see Appendix A). Note, that convergence has been already demonstrated by Mikelić et al. [12]. Focusing on the performance of the splitting scheme, we employ direct solvers for all occurring subproblems. Furthermore, let $(\mathbf{u}^i, \mathbf{p}^i, \mathbf{w}^i)$ denote the solution coefficient vector in step i. Then given tolerances $\delta_a, \delta_r > 0$, we employ the stopping criterion $\|(\mathbf{u}^i, \mathbf{p}^i, \mathbf{w}^i) - (\mathbf{u}^{i-1}, \mathbf{p}^{i-1}, \mathbf{w}^{i-1})\| \leq \delta_a + \delta_r \|(\mathbf{u}^i, \mathbf{p}^i, \mathbf{w}^i)\|$. For the implementation we used the Dune libraries [16].

5.1. Two-dimensional homogeneous medium — Constant Poisson's ratio

Let $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. For given $\xi \in \mathbb{R}$, we prescribe displacement, pressure and flux fields

$$\boldsymbol{u}(x,y,t) = tx(1-x)y(1-y)\begin{bmatrix}1 & 1\end{bmatrix}^\top, \qquad p(x,y,t) = \xi \cdot tx(1-x)y(1-y), \qquad \boldsymbol{w} = -\boldsymbol{K}\boldsymbol{\nabla}p \tag{16}$$

and choose source terms, initial and Dirichlet boundary conditions such that Eq. (16) is the solution of problem (1). We choose the same set of physical parameters as [12] apart from varying mechanical parameters (see Table 1). Instead of considering the full range of Lamé parameters, it is equivalent to consider the range $\nu \in (0, 0.5)$ for the Poisson's ratio as $\nu = (2(1+\mu/\lambda))^{-1}$. For the rather realistic parameters, we choose $\xi = 1e8$ to achieve convergence of the discretization.

The iteration count for different Poisson's ratios and different tuning parameters is illustrated in Fig. 1. Both β_{FS}^{cl} and β_{FS}^{opt} are robust with respect to the full range of Poisson's ratios, whereas the parameter β_{FS}^{λ} shows deteriorating convergence rates for soft materials, demonstrating the general necessity of the dependence of the tuning parameter on both Lamé parameters. As expected, in the limit, i.e., for $\nu \to 0.5$, both parameters β_{FS}^{λ} and β_{FS}^{opt} yield identical iteration counts.

5.2. Three-dimensional heterogeneous medium — Jumping Poisson's ratio

We compare Fixed Stress Splitting iteration counts for three-dimensional, heterogeneous media with constant and non-constant Poisson's ratios. We consider a cube $\Omega = (0,1) \times (0,1) \times (0,1) \subset \mathbb{R}^3$ discretized by $20 \times 20 \times 20$ hexahedra. For given $\xi \in \mathbb{R}$, we prescribe displacement and pressure fields

$$\boldsymbol{u}(x,y,z,t) = tx(1-x)y(1-y)z(1-z)\begin{bmatrix}1 & 1 & 1\end{bmatrix}^\top, \quad p(x,y,z,t) = \xi \cdot tx(1-x)y(1-y)z(1-z) \tag{17}$$

and a corresponding flux field $\boldsymbol{w} = -\boldsymbol{K}\boldsymbol{\nabla}p$. Further, we proceed analogously to Section 5.1, also considering the same physical parameters besides a locally varying Poisson's ratio. For chosen $\Delta\nu \in \{0.0, 0.05, 0.1, 0.2\}$

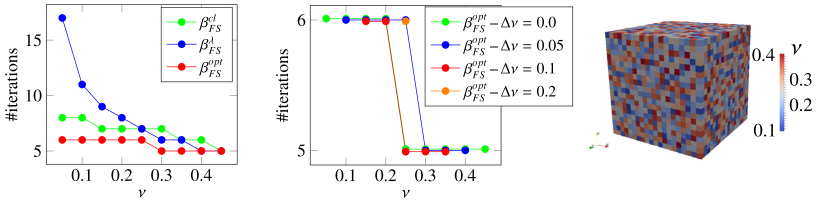


Fig. 1. *Left and Center:* Number of Fixed Stress iterations against Poisson's ratio for the first time step for (*left*) the homogeneous (Section 5.1) and (*center*) the heterogeneous (Section 5.2) test case. *Right:* Example for Poisson's ratio distribution in the interval $[0.1, 0.4]$ (Section 5.2).

and $\nu \in (\Delta\nu, 0.5 - \Delta\nu)$, we consider uniformly distributed Poisson's ratios in the interval $[\nu - \Delta\nu, \nu + \Delta\nu]$. An example distribution is shown in Fig. 1. We note, that for $\Delta\nu = 0$ the medium is homogeneous.

The iteration counts for different values for ν and $\Delta\nu$ are visualized in Fig. 1. We make two observations. For homogeneous media, the iteration count is robust with respect to different Poisson's ratios as it remains almost constant, as already seen for the two-dimensional test case in Section 5.1. Furthermore, we note that for heterogeneous media, the iteration count is bounded by the maximum of numbers of iterations obtained for homogeneous media over all Poisson's ratio values taken in the heterogeneous medium. This is in accordance with the theoretical convergence result, as the theoretical convergence rate includes a infinity norm, evaluating the worst case.

6. Conclusion

We have proposed an optimized Fixed Stress Splitting method for heterogeneous media. Its global convergence has been shown in weighted energy norms. The optimized tuning parameter depends on all mechanical parameters and shows stable iteration counts on the full range of Poisson's ratios. Numerical test cases show no significant increase of iterations when switching from a homogeneous to a heterogeneous medium or from two to three dimensions, demonstrating the robustness of the splitting scheme with respect to heterogeneities.

Acknowledgments

The research was partially supported by the NFR-DAADppp grant 255715 and the NFR-Toppforsk project 250223.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.aml.2016.12.019.

References

[1] R. Showalter, Diffusion in poro-elastic media, J. Math. Anal. Appl. 251 (1) (2000) 310–340. http://dx.doi.org/10.1006/jmaa.2000.7048.

[2] A. Settari, F.M. Mourits, A coupled reservoir and geomechanical simulation system, SPE J. 3 (3) (1998) 219–226. http://dx.doi.org/10.2118/50939-PA.

[3] K.K. Phoon, K.C. Toh, S.H. Chan, F.H. Lee, An efficient diagonal preconditioner for finite element solution of Biot's consolidation equations, Int. J. Numer. Methods Eng. 55 (4) (2002) 377–400. http://dx.doi.org/10.1002/nme.500.

[4] J.A. White, R.I. Borja, Block-preconditioned Newton–Krylov solvers for fully coupled flow and geomechanics, Comput. Geosci. 15 (4) (2011) 647. http://dx.doi.org/10.1007/s10596-011-9233-7.

[5] J.B. Haga, H. Osnes, H.P. Langtangen, Efficient block preconditioners for the coupled equations of pressure and deformation in highly discontinuous media, Int. J. Numer. Anal. Methods Geomech. 35 (13) (2011) 1466–1482. http://dx.doi.org/10.1002/nag.973.

[6] J.J. Lee, K.-A. Mardal, R. Winther, Parameter-robust discretization and preconditioning of Biot's consolidation model, 2015. arXiv:1507.03199 [math.NA].

[7] N. Castelletto, J.A. White, H.A. Tchelepi, Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics, Int. J. Numer. Anal. Methods Geomech. 39 (14) (2015) 1593–1618. http://dx.doi.org/10.1002/nag.2400.

[8] N. Castelletto, J.A. White, M. Ferronato, Scalable algorithms for three-field mixed finite element coupled poromechanics, J. Comput. Phys. 327 (2016) 894–918. http://dx.doi.org/10.1016/j.jcp.2016.09.063.

[9] J.A. White, N. Castelletto, H.A. Tchelepi, Block-partitioned solvers for coupled poromechanics: A unified framework, Comput. Methods Appl. Mech. Engrg. 303 (2016) 55–74. http://dx.doi.org/10.1016/j.cma.2016.01.008.

[10] J. Kim, H.A. Tchelepi, R. Juanes, Stability, Accuracy, and Efficiency of Sequential Methods for Coupled Flow and Geomechanics. Society of Petroleum Engineers, 2011. http://dx.doi.org/10.2118/119084-PA.

[11] A. Mikelić, M.F. Wheeler, Convergence of iterative coupling for coupled flow and geomechanics, Comput. Geosci. 17 (3) (2013) 455–461. http://dx.doi.org/10.1007/s10596-012-9318-y.

[12] A. Mikelić, B. Wang, M.F. Wheeler, Numerical convergence study of iterative coupling for coupled flow and geomechanics, Comput. Geosci. 18 (3) (2014) 325–341. http://dx.doi.org/10.1007/s10596-013-9393-8.

[13] M. Biot, General theory of three-dimensional consolidation, J. Appl. Phys. 12 (2) (1941) 155–164.

[14] O. Coussy, Mechanics of Porous Continua, Wiley, 1995.

[15] J. Thomas, Sur l'analyse numérique des méthodes d'éléments finis hybrides et mixtes, Thése d'Etat, Université Pierre & Marie Curie, Paris, 1977.

[16] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, O. Sander, A generic grid interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE, Computing 82 (2–3) (2008) 121–138. http://dx.doi.org/10.1007/s00607-008-0004-9.

Robust fixed stress splitting for Biot's equations in heterogeneous media: Supplementary Material

Jakub Both, jakub.both@uib.no,
Manuel Borregales, manuel.borregales@uib.no,
Jan M. Nordbotten, jan.nordbotten@uib.no,
Kundan Kumar, kundan.kumar@uib.no,
Florin A. Radu, florin.radu@uib.no
University of Bergen, Norway

The supplementary material contains

- a review of the fully discretized problem and the Fixed Stress Splitting method for reference in Section 3,
- a review of lemmas used in the convergence analysis,
- convergence analysis of the Fixed Stress Splitting method for homogeneous media,
- additional numerical test cases.

1 Review – Fixed Stress Splitting method for the fully discretized problem

Let $\langle\cdot,\cdot\rangle$ denote the standard $L^2(\Omega)$ scalar product. For homogeneous media, the fully implicit discretization then reads: Given $(\boldsymbol{u}_h^{n-1}, p_h^{n-1}, \boldsymbol{w}_h^{n-1}) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$, find $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{w}_h^n) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$, satisfying for all $(\boldsymbol{v}_h, q_h, \boldsymbol{z}_h) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$

$$2\mu\langle\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\rangle + \lambda\langle\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle - \alpha\langle p_h^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle = \langle\boldsymbol{f}, \boldsymbol{v}_h\rangle, \tag{1}$$

$$\frac{1}{M}\langle p_h^n, q_h\rangle + \alpha\langle\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^n, q_h\rangle + \tau\langle\boldsymbol{\nabla}\cdot\boldsymbol{w}_h^n, q_h\rangle = \tau\langle S_f, q_h\rangle + \frac{1}{M}\langle p_h^{n-1}, q_h\rangle + \alpha\langle\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n-1}, q_h\rangle, \tag{2}$$

$$\langle\boldsymbol{K}^{-1}\boldsymbol{w}_h^n, \boldsymbol{z}_h\rangle - \langle p_h^n, \boldsymbol{\nabla}\cdot\boldsymbol{z}_h\rangle = \langle\rho_f\boldsymbol{g}, \boldsymbol{z}_h\rangle. \tag{3}$$

The Fixed Stress Splitting method reads:
Step 1: Given $(\boldsymbol{u}_h^{n,i-1}, p_h^{n,i-1}, q_h^{n,i-1}) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{Z}_h$, find $(p_h^{n,i}, \boldsymbol{w}_h^{n,i}) \in Q_h \times \boldsymbol{Z}_h$ such that for all $(q_h, \boldsymbol{z}_h) \in Q_h \times \boldsymbol{Z}_h$ it holds

$$\left(\frac{1}{M} + \beta_{FS}\right)\langle p_h^{n,i}, q_h\rangle + \tau\langle\boldsymbol{\nabla}\cdot\boldsymbol{w}_h^{n,i}, q_h\rangle = \tau\langle S_f, q_h\rangle + \frac{1}{M}\langle p_h^{n-1}, q_h\rangle + \alpha\langle\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n-1}, q_h\rangle \tag{4}$$
$$+ \beta_{FS}\langle p_h^{n,i-1}, q_h\rangle - \alpha\langle\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n,i-1}, q_h\rangle,$$

$$\langle\boldsymbol{K}^{-1}\boldsymbol{w}_h^{n,i}, \boldsymbol{z}_h\rangle - \langle p_h^{n,i}, \boldsymbol{\nabla}\cdot\boldsymbol{z}_h\rangle = \langle\rho_f\boldsymbol{g}, \boldsymbol{z}_h\rangle. \tag{5}$$

Step 2: Given $p_h^{n,i} \in Q_h$, find $\boldsymbol{u}_h^{n,i} \in \boldsymbol{V}_h$ such that for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$ it holds

$$2\mu\langle\boldsymbol{\varepsilon}\Big(\boldsymbol{u}_h^{n,i}\Big), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\rangle + \lambda\langle\boldsymbol{\nabla}\cdot\boldsymbol{u}_h^{n,i}, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle = \langle\boldsymbol{f}, \boldsymbol{v}_h\rangle + \alpha\langle p_h^{n,i}, \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle. \tag{6}$$

2 Preliminaries

For the convergence analysis in Section 3, we make use of the following lemmas.

Lemma 1 (Polarization identity)**.** *Let $(X, \langle\cdot,\cdot\rangle_X)$ be a Hilbert space and $x, y \in X$. Then it holds*

$$\langle x, y\rangle_X \;=\; \frac{1}{4}\|x+y\|_X^2 - \frac{1}{4}\|x-y\|_X^2.$$

Lemma 2 (Binomial identity)**.** *Let $(X, \langle\cdot,\cdot\rangle_X)$ be a Hilbert space and $x, y \in X$. Then it holds*

$$\langle x-y, x\rangle_X \;=\; \frac{1}{2}\|x\|_X^2 + \frac{1}{2}\|x-y\|_X^2 - \frac{1}{2}\|y\|_X^2.$$

Lemma 3 (Cauchy-Schwarz inequality)**.** *Let $(X, \langle\cdot,\cdot\rangle_X)$ be a Hilbert space and $x, y \in X$. Then it holds*

$$|\langle x, y\rangle_X| \;\le\; \|x\|_X\|y\|_X.$$

Lemma 4 (Young's inequality)**.** *Let $a, b, \delta \in \mathbb{R}$, $\delta > 0$. Then it holds*

$$|ab| \;\le\; \frac{1}{2\delta}a^2 + \frac{\delta}{2}b^2.$$

Lemma 5 (Arithmetic mean-root mean square inequality, AM-QM inequality)**.** *Let $n \in \mathbb{N}$ and $\{x_j\}_{j=1}^n \subset \mathbb{R}$. Then it holds*

$$\frac{1}{n}\sum_{j=1}^{n} x_j \;\le\; \sqrt{\frac{1}{n}\sum_{j=1}^{n} x_j^2}.$$

3 Convergence analysis for constant Lamé parameters

We prove global convergence of the Fixed Stress Splitting method applied to Biot's equations in a homogeneous porous medium. The proof has the same character as for heterogeneous media, but simpler notation is used. Furthermore, slightly more details are presented.

Theorem 6 (Linear convergence for Fixed Stress Splitting)**.** *Assume spatially constant effective coefficients, satisfying Assumptions (A1)–(A3). Let $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{w}_h^n)$ and $(\boldsymbol{u}_h^{n,i}, p_h^{n,i}, \boldsymbol{w}_h^{n,i})$ be the solutions of Eq.* (1)–(3) *and Eq.* (4)–(6)*, respectively. Let $e_{\boldsymbol{u}}^i = \boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^n$, $e_p^i = p_h^{n,i} - p_h^n$ and $e_{\boldsymbol{w}}^i = \boldsymbol{w}_h^{n,i} - \boldsymbol{w}_h^n$ denote the errors at current iteration. Then for all*

$$\beta_{FS} \;\ge\; \frac{\alpha^2}{2\left(\frac{2}{d}\mu + \lambda\right)} \tag{7}$$

it holds

$$\|e_p^i\|^2 \le \frac{\frac{\beta_{FS}}{2}}{\frac{1}{M} + \frac{\beta_{FS}}{2} + \frac{\tau k_m}{C_{\Omega,d}}}\|e_p^{i-1}\|^2$$

and

$$2\mu\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^i)\|^2 + \lambda\|\boldsymbol{\nabla}\cdot e_{\boldsymbol{u}}^i\|^2 \;\le\; \frac{\alpha^2}{\left(\frac{2}{d}\mu + \lambda\right)}\|e_p^i\|^2.$$

Optimal convergence rates are obtained when equality holds in Eq. (7)*.*

Proof. **Step 1: Flow and mechanics**

By taking the differences of corresponding Eq. (4)–(6) and Eq. (1)–(3), testing with $\boldsymbol{v}_h = e_{\boldsymbol{u}}^{i-1} \in \boldsymbol{V}_h$, $q_h = e_p^i \in Q_h$ and $\boldsymbol{z}_h = \tau e_{\boldsymbol{w}}^i \in \boldsymbol{Z}_h$ and adding all together, we obtain

$$2\mu\langle \boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i\big), \boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^{i-1}\big)\rangle + \lambda\langle \boldsymbol{\nabla}\cdot e_{\boldsymbol{u}}^i, \boldsymbol{\nabla}\cdot e_{\boldsymbol{u}}^{i-1}\rangle + \frac{1}{M}\|e_p^i\|^2 + \tau\langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, e_{\boldsymbol{w}}^i\rangle + \beta_{FS}\langle e_p^i - e_p^{i-1}, e_p^i\rangle = 0.$$

Using the polarization and binomial identities (cf. Lemma 1 and 2) yields

$$\begin{aligned}\frac{\mu}{2}\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i + e_{\boldsymbol{u}}^{i-1}\big)\|^2 + \frac{\lambda}{4}\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i + e_{\boldsymbol{u}}^{i-1})\|^2 - \frac{\mu}{2}\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|^2 - \frac{\lambda}{4}\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|^2 \\ + \frac{1}{M}\|e_p^i\|^2 + \tau\langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, e_{\boldsymbol{w}}^i\rangle + \frac{\beta_{FS}}{2}\left(\|e_p^i\|^2 + \|e_p^i - e_p^{i-1}\|^2 - \|e_p^{i-1}\|^2\right) \;=\; 0.\end{aligned} \tag{8}$$

Step 2: Mechanics

Taking the difference of Eq. (6) evaluated at iteration i and $i-1$, tested with $\boldsymbol{v}_h = e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}$ yields

$$2\mu\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|^2 + \lambda\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|^2 \;=\; \alpha\langle e_p^i - e_p^{i-1}, \boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\rangle. \tag{9}$$

Applying the Cauchy-Schwarz inequality and the AM-QM inequality (cf. Lemma 3 and 5) to the right hand side, yields for any $\gamma \in (0,1)$

$$\begin{aligned}&2\mu\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|^2 + \lambda\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|^2 \\ \le\, &\alpha\|e_p^i - e_p^{i-1}\|\,\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\| \\ \le\, &\alpha\|e_p^i - e_p^{i-1}\|\left(\gamma\sqrt{d}\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\| + (1-\gamma)\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|\right).\end{aligned}$$

By applying Young's inequality (cf. Lemma 4), rearranging terms and scaling, for any δ_1, δ_2, $c > 0$ and $\gamma \in (0,1)$, it holds

$$\begin{aligned}&c(2-\delta_1\gamma)\mu\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1}\big)\|^2 + c(1-\delta_2(1-\gamma))\lambda\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\|^2 \\ \le\, &\frac{\alpha^2 c}{4}\left(\frac{\gamma d}{\delta_1\mu} + \frac{1-\gamma}{\delta_2\lambda}\right)\|e_p^i - e_p^{i-1}\|^2.\end{aligned} \tag{10}$$

Now we choose the parameters optimally. We consider the minimization problem

$$\begin{aligned}\min_{\gamma,c,\delta_1,\delta_2}\; &\frac{\alpha^2 c}{4}\left(\frac{\gamma d}{\delta_1\mu} + \frac{1-\gamma}{\delta_2\lambda}\right) \\ s.t.\quad &c, \delta_1, \delta_2 > 0 \\ &\gamma \in (0,1) \\ &c(2-\delta_1\gamma) = \tfrac{1}{2} \\ &c(1-\delta_2(1-\gamma)) = \tfrac{1}{4}\end{aligned}$$

By substituting both equality constraints and rearranging terms, we obtain a reduced problem

$$\begin{aligned}&\min_{\gamma,c}\frac{\alpha^2}{4}\frac{c}{1-\frac{1}{4c}}\left(\frac{d}{2\mu}\gamma^2 + \frac{1}{\lambda}(1-\gamma)^2\right) \\ &s.t.\; c > \frac{1}{4}, \gamma \in (0,1)\end{aligned}$$

which can be separated into two independent problems yielding optimal

$$c^\star = \frac{1}{2}, \qquad \gamma^\star = \frac{2\mu}{2\mu + d\lambda},$$

and hence

$$\delta_1^\star = \frac{2\mu + d\lambda}{2\mu}, \qquad \delta_2^\star = \frac{2\mu + d\lambda}{2d\lambda}.$$

Inserting the optimal values into Eq. (10) yields

$$\frac{\mu}{2}\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^{i+1} - e_{\boldsymbol{u}}^{i}\big)\|^2 + \frac{\lambda}{4}\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^{i+1} - e_{\boldsymbol{u}}^{i})\|^2 \;\leq\; \frac{\alpha^2}{4\left(\frac{2}{d}\mu + \lambda\right)}\|e_p^i - e_p^{i-1}\|^2. \tag{11}$$

Step 3: Darcy

Taking the difference of Eq. (5) and Eq. (3) yields

$$\langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, \boldsymbol{z}_h\rangle - \langle e_p^i, \boldsymbol{\nabla}\cdot \boldsymbol{z}_h\rangle \;=\; 0 \tag{12}$$

for any $\boldsymbol{z}_h \in \boldsymbol{Z}_h$. Using Thomas' Lemma, there exists a constant $C_{\Omega,d} > 0$ and a function $\tilde{\boldsymbol{z}}_h \in \boldsymbol{Z}_h$ satisfying

$$\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{z}}_h = e_p^i, \quad \|\tilde{\boldsymbol{z}}_h\| \leq C_{\Omega,d}\|e_p^i\|.$$

Hence, together with Eq. (12), Assumption (A3) and by applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\|e_p^i\|^2 \;&=\; \langle e_p^i, \boldsymbol{\nabla}\cdot\tilde{\boldsymbol{z}}_h\rangle \\
&=\; \langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, \tilde{\boldsymbol{z}}_h\rangle \\
&\leq\; \langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, e_{\boldsymbol{w}}^i\rangle^{\frac{1}{2}} \,\langle \boldsymbol{K}^{-1}\tilde{\boldsymbol{z}}_h, \tilde{\boldsymbol{z}}_h\rangle^{\frac{1}{2}} \\
&\leq\; \frac{1}{\sqrt{k_m}}\langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, e_{\boldsymbol{w}}^i\rangle^{\frac{1}{2}}\,\|\tilde{\boldsymbol{z}}_h\| \\
&\leq\; \frac{C_{\Omega,d}}{\sqrt{k_m}}\langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, e_{\boldsymbol{w}}^i\rangle^{\frac{1}{2}}\,\|e_p^i\|.
\end{aligned}$$

After cancellation and rearranging, we obtain

$$\frac{k_m}{C_{\Omega,d}^2}\|e_p^i\|^2 \;\leq\; \langle \boldsymbol{K}^{-1} e_{\boldsymbol{w}}^i, e_{\boldsymbol{w}}^i\rangle. \tag{13}$$

Step 4: Combining Step 1–3

Discarding the first two positive terms in Eq. (8), inserting Eq. (13) and Eq. (11) yields

$$\begin{aligned}
&\left(\frac{1}{M} + \frac{\beta_{FS}}{2} + \frac{\tau k_m}{C_{\Omega,d}^2}\right)\|e_p^i\|^2 + \frac{\beta_{FS}}{2}\|e_p^i - e_p^{i-1}\|^2 \\
&\leq \frac{\beta_{FS}}{2}\|e_p^{i-1}\|^2 + \frac{\alpha^2}{4\left(\frac{2}{d}\mu + \lambda\right)}\|e_p^i - e_p^{i-1}\|^2.
\end{aligned} \tag{14}$$

By setting $\beta_{FS} \geq \frac{\alpha^2}{2\left(\frac{2}{d}\mu + \lambda\right)}$, we obtain

$$\left(\frac{1}{M} + \frac{\beta_{FS}}{2} + \frac{\tau k_m}{C_{\Omega,d}^2}\right)\|e_p^i\|^2 \;\leq\; \frac{\beta_{FS}}{2}\|e_p^{i-1}\|^2.$$

Step 5: Mechanics revisited

Taking the difference of Eq. (6) and Eq. (1), tested with $\boldsymbol{v}_h = e_{\boldsymbol{u}}^i$ yields

$$2\mu\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i\big)\|^2 + \lambda\|\boldsymbol{\nabla}\cdot e_{\boldsymbol{u}}^i\|^2 \;=\; \alpha\langle e_p^i, \boldsymbol{\nabla}\cdot e_{\boldsymbol{u}}^i\rangle.$$

We repeat all steps from Step 2. Due to linearity, we obtain analogously

$$\frac{\mu}{2}\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i\big)\|^2 + \frac{\lambda}{4}\|\boldsymbol{\nabla}\cdot e_{\boldsymbol{u}}^i\|^2 \;\le\; \frac{\alpha^2}{4\left(\frac{2}{d}\mu+\lambda\right)}\|e_p^i\|^2.$$

□

Remark 1 (Alternative proof)**.** *Using previous calculations, we give an alternative proof to obtain Eq.* (14) *in Step 4 of the previous proof. Applying the AM-QM inequality to the left hand side, and the Cauchy-Schwarz inequality and Young's inequality to the right hand side of Eq.* (9)*, it follows after some algebraic manipulation*

$$\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\| \le \frac{\alpha}{\left(\frac{2\mu}{d}+\lambda\right)}\|e_p^i - e_p^{i-1}\|. \tag{15}$$

Inserting Eq. (9) *and Eq.* (15) *into Eq.* (8)*, yields*

$$\begin{aligned}
&\frac{\mu}{2}\|\boldsymbol{\varepsilon}\big(e_{\boldsymbol{u}}^i + e_{\boldsymbol{u}}^{i-1}\big)\|^2 + \frac{\lambda}{4}\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i + e_{\boldsymbol{u}}^{i-1})\|^2 + \left(\frac{1}{M}+\frac{\beta_{FS}}{2}\right)\|e_p^i\|^2 + \tau\langle\boldsymbol{K}^{-1}e_{\boldsymbol{w}}^i, e_{\boldsymbol{w}}^i\rangle + \frac{\beta_{FS}}{2}\|e_p^i - e_p^{i-1}\|^2 \\
&\quad= \frac{\beta_{FS}}{2}\|e_p^{i-1}\|^2 + \frac{\alpha}{4}\langle e_p^i - e_p^{i-1}, \boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\rangle \\
&\quad\le \frac{\beta_{FS}}{2}\|e_p^{i-1}\|^2 + \frac{\alpha}{4}\|e_p^i - e_p^{i-1}\|\,\|\boldsymbol{\nabla}\cdot(e_{\boldsymbol{u}}^i - e_{\boldsymbol{u}}^{i-1})\| \\
&\quad\le \frac{\beta_{FS}}{2}\|e_p^{i-1}\|^2 + \frac{\alpha^2}{4\left(\frac{2\mu}{d}+\lambda\right)}\|e_p^i - e_p^{i-1}\|^2
\end{aligned}$$

By dropping some positive terms and inserting Eq. (13) *on the left hand side, we finally obtain Eq.* (14)*.*

4 Numerical results

In the following section, we present additional test cases demonstrating the performance of the Fixed Stress Splitting method. We consider the classical Mandel's problem, a modification, a three-dimensional version of the test case shown in the original work and a two- and three-dimensional medium with highly jumping mechanical parameters.

4.1 Mandel's problem

Mandel's problem is a classical benchmark for linearized Biot's equations, for which an analytical solution is known. See, e.g., Coussy [1] for details. Briefly, Mandel's problem considers a two-dimensional, rectangular, fully saturated poroelastic material, loaded by a constant compressive force, which is applied instantaneously. Gravity is neglected. Due to symmetry, only the top right quarter is considered, discretized by a regular quadrilateral grid (cf. Fig. 1). The top, left and bottom boundary is treated as impermeable, while zero pressure is implied at the right boundary. For the simplicity of numerical implementation, a displacement in y-direction is prescribed on the top. The type of the remaining boundary conditions for the mechanical problem is illustrated in Fig. 1.

We consider the domain $\Omega = (0,100)\times(0,10)\subset\mathbb{R}^2$ and choose the same set of parameters as chosen by Mikelić et al. [2] apart from varying Poisson's ratio, cf. Table 1.

Symb.	Quantity	Values [Unit]
E	bulk modulus	0.594 [GPa]
ν	Poisson's ratio	[0.05, ..., 0.45]
α	Biot's coefficient	1
M	Biot's modulus	1.65e10 [Pa]
$\boldsymbol{K}$	permeability tensor divided by fluid viscosity	100$\boldsymbol{I}$ [mD/cP]
$\boldsymbol{g}$	gravity vector	$\mathbf{0}$ [m/s^2]
$\Delta x, \Delta y$	grid spacing in x and y	2.5 [m]
$\Delta x, \Delta y$	grid spacing in x and y	0.25 [m]
τ	time step size	10 [s]
δ_a	absolute error tolerance	1e-6
δ_r	relative error tolerance	1e-6

Table 1: Problem parameters, chosen identically to [2].

The number of iterations for the first time step is shown in Figure 2. The tuning parameter β_{FS}^{λ} shows large iteration counts for small Poisson's ratios, i.e., large μ/λ, which is due to the lack of dependence of the tuning parameter on μ. On the other side, if we choose the tuning parameter β_{FS}^{opt}, we obtain stable iteration counts with respect to different Poisson's ratios. However, in the range of $\nu \in [0.3, 0.45]$ the parameter β_{FS}^{λ} yields slightly fewer iterations than β_{FS}^{opt}.

Due to the oedometric character of Mandel's problem, there exists an optimal tuning parameter yielding nonlinear convergence on the full range of Lamé parameters. Employing the vertical uniaxial bulk modulus yields the tuning parameter $\beta_{FS}^{v} = \frac{\alpha^2}{2\mu+\lambda}$, for which the Fixed Stress Splitting method converges within three iterations independent of the mechanical parameters. Due to its specific character, Mandel's problem is unsuitable for analyzing the general performance of different tuning parameters for the Fixed Stress Splitting method applied to two-dimensional problems.

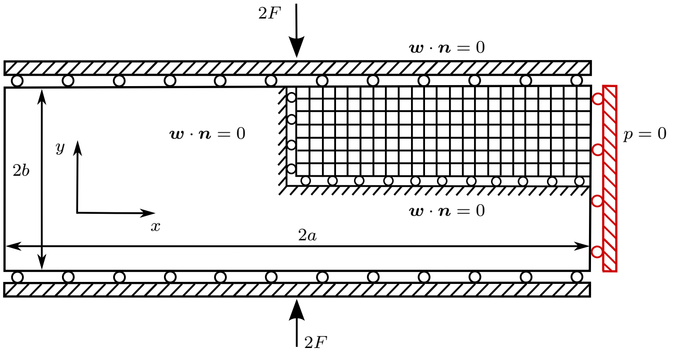

2F
$\boldsymbol{w} \cdot \boldsymbol{n} = 0$
$\boldsymbol{w} \cdot \boldsymbol{n} = 0$
$p = 0$
y
2b
$\boldsymbol{w} \cdot \boldsymbol{n} = 0$
x
2a
2F

Figure 1: Reduced domain for Mandel's problem. Red boundary condition only valid for test case in Section 4.2.

4.2 Modified Mandel's problem

The performance of the Fixed Stress Splitting scheme for Mandel's problem is biased by the existence of an optimal tuning parameter β_{FS}^{v} (cf. Section 4.1). Therefore, we repeat the simulation of Mandel's problem with a modified boundary condition on the right boundary of the domain. We consider a zero Dirichlet boundary condition in x-direction for the displacement, illustrated by the red element in Fig. 1. Apart from

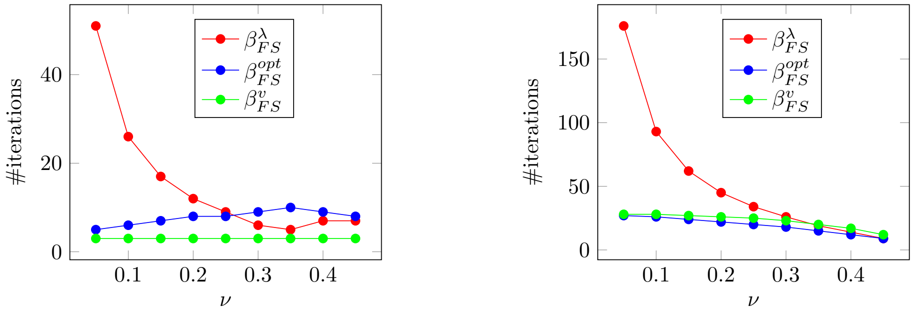


Figure 2: *Left:* Number of Fixed Stress iterations against Poisson's ratio for the first time step of Mandel's problem (*left*) and modified Mandel's problem (*right*).

that, we use the same setting as in Section 4.1. The number of iterations for the first time step is shown in Figure 2.

In this test case, we observe a better performance for the new tuning parameter β^{opt}_{FS} compared to the value β^{λ}_{FS} on the full range for the Poisson's ratio. However, the number of iterations depends weakly on the Poisson's ratio, as it decreases by a factor of three over the interval $[0.05, 0.45]$. Furthermore, the previously empirically chosen tuning parameter β^{v}_{FS} does not yield better performance than the tuning parameter β^{opt}_{FS}, demonstrating the extraordinary character of the classical Mandel's problem.

4.3 Revisit three-dimensional heterogeneous medium – Jumping Poisson's ratio

We revisit the test case from Sec. 5.2 of our original work, which considers a three-dimensional medium with a spatially-varying, randomly distributed Poisson's ratio, cf. Fig. 3. We consider the same settings and compare number of iterations needed for convergence for the tuning parameters β^{cl}_{FS}, β^{λ}_{FS}, β^{opt}_{FS} and for different Poisson's ratio distributions. We use same notation as in Sec. 5.2 of our original work.

The iteration counts for different values for ν and $\Delta\nu$ are visualized in Fig. 3. We make two observations. For homogeneous media, i.e., $\Delta\nu = 0.0$, we observe the same convergence behavior as for the two-dimensional test case in Sec. 5.1 of our original work. Both β^{cl}_{FS} and β^{opt}_{FS} show stable iteration counts with respect to the full range of Poisson's ratios, whereas β^{λ}_{FS} shows deteriorating convergence rates for soft materials. When switching to heterogeneous media, $\Delta\nu = 0.1$, β^{opt}_{FS} shows iteration counts almost not deviating from the iteration counts for the homogeneous analogon. However, both non-optimal parameters β^{cl}_{FS} and β^{λ}_{FS} show slightly worse convergence behavior for heterogeneous media than for homogeneous media. Again, all in all, the choice β^{opt}_{FS} outperforms β^{cl}_{FS} and β^{λ}_{FS} for homogeneous and heterogeneous media.

4.4 Two- and three-dimensional test cases with jumping mechanical parameters

In the following, we enhance two mechanics test cases from [3], a two-dimensional test case (2D) and a three-dimensional test case (3D).

(2D) We consider a heterogeneous medium $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$ discretized by 40×40 quadrilaterals. Further, we prescribe a discontinuous solution which yields a differentiable stress field and hence a computable mechanical driving force. Let

$$\chi_{2D}(x,y) = \begin{cases} 1, & \min(x,y) > \frac{1}{2} \\ 0, & \text{otherwise.} \end{cases}$$

be a characteristic function and define discontinuous parameters

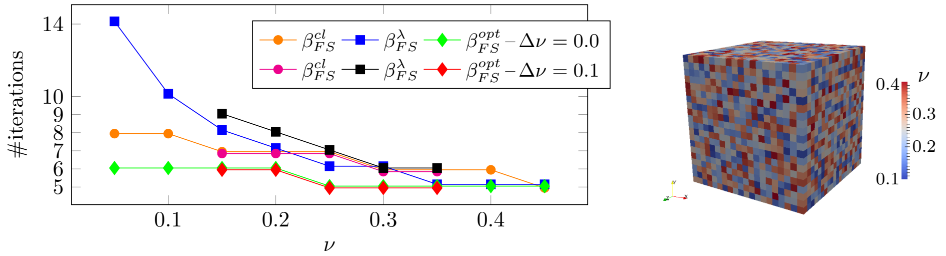


Figure 3: *Left:* Number of Fixed Stress iterations against Poisson's ratio for the first time step for the heterogeneous test case (Sec. 4.3). *Right:* Example for Poisson's ratio distribution in the interval $[0.1, 0.4]$ (Sec. 4.3).

- $\nu = (1 - \chi_{2D})\nu_1 + \chi_{2D}\nu_2$, for given ν_1, ν_2, see, e.g., Figure 4,
- $E = (1 - \chi_{2D})E_1 + \chi_{2D}E_2$, for given E_1, E_2,
- $M = (1 - \chi_{2D})M_1 + \chi_{2D}M_2$, where for given coupling strength τ we choose M_i s.t. $\tau = \frac{\alpha^2 M_i}{K_{dr,i}}$, where $K_{dr,i} = \mu_i + \lambda_i$ is the bulk modulus, $i \in \{1, 2\}$.

Apart from that, we choose the parameters as before. Then, with $\xi = 1e9$, we prescribe a solution:

$$\begin{aligned}
\boldsymbol{u}(x, y, t) &= \xi \cdot \frac{t}{\mu(x, y)} \begin{bmatrix} (x - 0.5)^2(y - 0.5)^2 \\ -\frac{2}{3}(x - 0.5)(y - 0.5)^3 \end{bmatrix} \\
p(x, y, t) &= \xi \cdot tx(1 - x)y(1 - y) \\
\boldsymbol{w}(x, y, t) &= -\boldsymbol{K}\boldsymbol{\nabla} p(x, y, t)
\end{aligned}$$

(3D) We consider a heterogeneous medium $\Omega = (0, 1) \times (0, 1) \times (0, 1) \subset \mathbb{R}^3$ discretized by $10 \times 10 \times 10$ hexahedra. Further, we prescribe a discontinuous solution which yields a differentiable stress field and hence a computable mechanical driving force. Let

$$\chi_{3D}(x, y, z) = \begin{cases} 1, & \min(x, y, z) > \frac{1}{2} \\ 0, & \text{otherwise.} \end{cases}$$

be a characteristic function and define discontinuous parameters

- $\nu = (1 - \chi_{3D})\nu_1 + \chi_{3D}\nu_2$, for given ν_1, ν_2, see, e.g., Figure 4,
- $E = (1 - \chi_{3D})E_1 + \chi_{3D}E_2$, for given E_1, E_2,
- $M = (1 - \chi_{3D})M_1 + \chi_{3D}M_2$, where for given coupling strength τ we choose M_i s.t. $\tau = \frac{\alpha^2 M_i}{K_{dr,i}}$, where $K_{dr,i} = \frac{2\mu_i}{3} + \lambda_i$ is the bulk modulus, $i \in \{1, 2\}$.

Apart from that, we choose the parameters as before. Then, with $\xi = 1e9$, we prescribe a solution:

$$\begin{aligned}
\boldsymbol{u}(x, y, z, t) &= \xi \cdot \frac{t}{\mu(x, y, z)} \begin{bmatrix} (x - 0.5)^2(y - 0.5)^2(z - 0.5)^2 \\ (x - 0.5)^2(y - 0.5)^2(z - 0.5)^2 \\ -\frac{2}{3}\left((x - 0.5)(y - 0.5)^2 + (x - 0.5)^2(y - 0.5)\right)(z - 0.5)^3 \end{bmatrix} \\
p(x, y, z, t) &= \xi \cdot tx(1 - x)y(1 - y)z(1 - z) \\
\boldsymbol{w}(x, y, z, t) &= -\boldsymbol{K}\boldsymbol{\nabla} p(x, y, z, t)
\end{aligned}$$

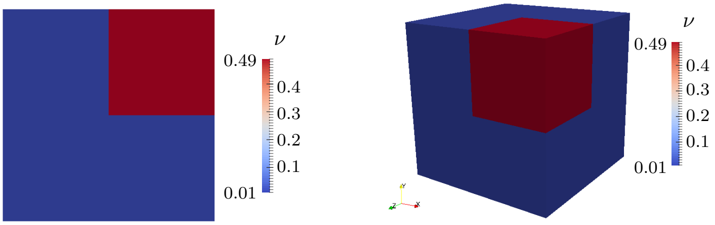


Figure 4: Heterogeneous medium: Poisson's ratio distribution. *Left:* (2D). *Right:* (3D).

ν_1	ν_2	E_1	E_2	τ	#It. (2D)	#It. (3D)
0.01	0.01	5.94e9	5.94e9	5	6	6
0.25	0.25	5.94e9	5.94e9	5	6	5
0.49	0.49	5.94e9	5.94e9	5	4	3
0.01	0.01	5.94e10	5.94e10	5	4	4
0.25	0.25	5.94e10	5.94e10	5	4	4
0.49	0.49	5.94e10	5.94e10	5	3	3
0.01	0.01	5.94e9	5.94e10	5	6	6
0.25	0.25	5.94e9	5.94e10	5	5	5
0.49	0.49	5.94e9	5.94e10	5	4	3
0.01	0.49	5.94e9	5.94e9	5	6	6
0.01	0.49	5.94e10	5.94e10	5	4	4
0.01	0.49	5.94e9	5.94e10	5	6	6

Table 2: Number of Fixed Stress iterations (#It.) using β_{FS}^{opt} for test cases (2D) and (3D) with jumping mechanical parameters.

For both test cases, we only consider the tuning parameter β_{FS}^{opt}, analyzing the performance for different heterogeneities. The resulting iteration counts for various combinations of parameters is presented in Table 2. It can be observed, that the number of iterations counts is robust with respect to jumping mechanical parameters. Almost equal results are obtained for two dimensions (2D) and three dimensions (3D). We note, that poor approximations are given for E_1, E_2 of smaller order. Hence, no big jumps for the Young's modulus have been tested.

References

[1] O. Coussy, *Mechanics of Porous Continua*. Wiley, 1995.

[2] A. Mikelić, B. Wang, and M. F. Wheeler, "Numerical convergence study of iterative coupling for coupled flow and geomechanics," *Computational Geosciences*, vol. 18, no. 3, pp. 325–341, 2014.

[3] E. Keilegavlen and J. M. Nordbotten, "Finite volume methods for elasticity with weak symmetry," *arXiv:1512.01042v1*, Dec. 2015.

Paper D

Global existence of a weak solution to unsaturated poroelasticity

Both, J.W., Pop, I.S., and Yotov, I.

arXiv:1909.06679 [math.NA]

Global existence of a weak solution to unsaturated poroelasticity

Jakub W. Both* Iuliu Sorin Pop† Ivan Yotov‡

Abstract

In this paper, we consider unsaturated poroelasticity, i.e., coupled hydro-mechanical processes in unsaturated porous media, modeled by a non-linear extension of Biot's quasi-static consolidation model. The coupled, elliptic-parabolic system of partial differential equations is a simplified version of the general model for multi-phase flow in deformable porous media obtained under similar assumptions as usually considered for Richards' equation. In this work, the existence of a weak solution is established using regularization techniques, the Galerkin method, and compactness arguments. The final result holds under non-degeneracy conditions and natural continuity properties for the non-linearities. The assumptions are demonstrated to be reasonable in view of geotechnical applications.

1 Introduction

Strongly coupled hydro-mechanical processes in porous media are occurring in various applications of societal relevance within, e.g., geotechnical, structural, and biomechanical engineering. Examples for instance are soil subsidence due to groundwater withdrawal, geothermal energy storage in fractured rocks, swelling and drying shrinkage of concrete, and deformation of soft, biological tissue components.

In the field of porous media, such microscopically complex processes are typically modeled by a continuum mechanics approach [1]. The multi-phasic solid-fluid mixture is considered a homogenized continuum, and both geometry, skeleton, and fluid properties are averaged over representative elementary volumes, consisting of a mixture of solid and fluid particles. Ultimately, the microscopic interaction of the different constituents is described by macroscopic, effective equations. The simplest, macroscopic model accounting for the coupling of single-phase flow and elastic deformation in a porous medium is *Biot's linear, quasi-static consolidation model.* Its phenomenological derivation dates back to the seminal works by Terzaghi [2] and Biot [3]. In the course of the last century, many more advanced models have been developed, accounting, e.g., for the presence of different interacting fluids, thermal effects, or chemical reactions. We refer to the textbooks [4, 5] for an introduction and their derivation.

In this paper, we consider a non-linear, coupled system of partial differential equations, modelling the quasi-static consolidation of variably saturated porous media, also called *unsaturated poroelasticity* – in particular relevant in soil mechanics. The model can be obtained by simplifying the more general model for two-phase flow in deformable porous media, founded on macroscopic momentum and mass balances combined with constitutive relations [4]. It is assumed that one fluid phase can be simply neglected. This is a common practice for fluids with high viscosity ratios if the negligible fluid phase is continuous and connected to the atmosphere, i.e., the same hypotheses as for Richards' equation [6, 7]. Finally, the resulting model generalizes Biot's quasi-static, linear consolidation model, combining Richards' equation and linear elasticity equations

*Department of Mathematics, University of Bergen, Bergen, Norway; {jakub.both@uib.no}

†Faculty of Science, Hasselt University, Hasselt, Belgium; and
Department of Mathematics, University of Bergen, Bergen, Norway; {sorin.pop@uhasselt.be}

‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, USA; {yotov@math.pitt.edu}

with non-linear coupling. It is highly non-linear, potentially strongly coupled, and potentially degenerate, which makes its analysis complicated.

Regarding the mathematical theory of poroelasticity, in particular Biot's quasi-static, linear consolidation model has been well-studied. Well-posedness including the existence, uniqueness, and regularity of solutions, has been established [8–10]; recent advances in the numerical analysis include, e.g., stable finite discretizations [11–18], efficient numerical iterative solvers [19–24], and a posteriori error estimates [25–27]. Lately, linear and non-linear extensions have become of increased interest. Well-posedness and the efficient numerical solution have been analyzed for the dynamic Biot-Allard system [28], Biot-Stokes systems [29–31], the Biot model with deformation dependent permeability [32, 33], poroelasticity in fractured media [34–37], poroelasticity with non-linear solid and fluid compressibility [38, 39], general non-linear single-phase poroelasticity [40], poro-visco-elasticity [33, 38], thermoporoelasticity [38, 41–44], poroelasticity from a gradient flow perspective [38], and multiple-permeability poroelasticity systems [45–47], among others. In all problems, the coupling is linear.

Despite the large interest, rather few theoretical results have been established for unsaturated or multi-phase poroelasticity. We highlight [48], in which the first ever mathematical analysis of the consolidation of a variably saturated, porous medium has been presented. In the aforementioned work, the existence of a weak solution is established under two strict model assumptions: (i) the coupling term in the fluid flow equation is linear; and (ii) after introducing a new pressure variable by applying the Kirchhoff transformation the coupling and the diffusion terms in the mass balance simultaneously become linear. The second assumption implies a specific, artificial form of the so-called pore pressure, a non-linearity arising in the linear momentum balance. Ultimately, the result does not apply to the general model for unsaturated poroelasticity. On the other hand, the analysis accounts for non-linearly variable densities and porosities, and allows for degenerate situations. In addition, we mention efforts on studying the efficient numerical solution for unsaturated poroelasticity [49] and multi-phase poroelasticity [50–52].

In this paper, the existence of weak solutions for the general model of unsaturated poroelasticity is established. In order to deal with the non-linear character, the problem is first transformed utilizing the Kirchhoff transformation, a technique commonly used for the analysis of non-linear diffusion problems [53]. By this, the diffusion component of the mass balance becomes linear – a fully non-linear coupling and a non-linear storage coefficient are still present. The analysis then employs regularization techniques and compactness arguments in six steps and goes as follows. First, a physically motivated double regularization is introduced, adding a non-degenerate parabolic character to both balance equations. Regularization is required in order to allow the discussion of the non-linear coupling terms. Ultimately, the regularized model accounts for primary and secondary consolidation of variably saturated, porous media with compressible grains. Second, the problem is discretized combining an implicit time stepping, the finite element method (FEM) for the mechanics equation, and the finite volume method involving a two-point flux approximation (TPFA) for the flow equation. The motivation for the chosen discretization is two-fold: (i) it is a common discretization in the field of poroelasticity [13, 54], also closely related to mixed finite element discretizations [11]; moreover, finite volume methods [55–58] and mixed finite element methods [59, 60] are widely used for discretizing Richards' equation. Even more importantly, (ii) the specific choice of the discretization becomes crucial for the subsequent step of the proof, allowing for straightforward cancelling of the coupling terms. In the third step of the proof, stability of the discrete solution is showed, and compactness arguments are utilized for deriving a weak solution of the doubly regularized problem. For this, on the one hand the Legendre transformation is exploited as in [53] and specific finite volume techniques are employed for discussing the limit of the spatial discretization parameters, inspired by [61, 62]. Fourth, improved regularity is showed for the weak solution of the doubly regularized problem. Fifth and finally sixth, the limit of vanishing regularization in the momentum and mass balances are discussed, respectively.

Difficulties arise in the last steps of the proof due to a possible degenerate character of the problem for vanishing saturation. Our analysis requires an overall parabolic character of the coupled problem and natural continuity properties for the non-linearities. Those are ensured under specific material assumptions and a non-vanishing, minimal amount of fluid saturation. In the appendix, the assumptions are demonstrated to be satisfied for constitutive relations typically utilized in real-life applications. Furthermore, for simplicity, the porous material is assumed to be isotropic, gravity has been neglected and homogeneous, essential boundary conditions have been considered. The focus of this work is on the involved, non-linear, coupled character of the governing equations.

The rest of the paper is organized as follows. In Section 2, the model is introduced as derived in the engineering literature, and the model is transformed using the Kirchhoff transformation. In Section 3, the notion of a weak solution to the transformed problem is introduced, and the main result is stated: existence of a weak solution to the transformed problem under certain model assumptions and non-degeneracy conditions. The idea of the proof, consisting of six steps, is presented. The details of those six steps are the subject of the remaining Sections 4–9. In the appendix, the feasibility of the required assumptions for the main result are discussed for widely used constitutive models from the literature. In addition, technical results from the literature used in the proof of the main result are recalled for a comprehensive presentation.

2 Mathematical model for unsaturated poroelasticity

We consider a continuum mechanics model for unsaturated poroelasticity, a particular simplification of general multi-phase poroelasticity [4, 5]. It is based on the fundamental principles of momentum and mass balance combined with constitutive relations. The model is valid under the assumptions of infinitesimal strains and the presence of two fluid phases, an active and a passive phase; the displacement of the passive phase does not impede the advance of the active phase and can be therefore neglected. Finally, the model couples non-linearly the Richards equation and the linear elasticity equations utilizing an effective stress approach.

In the following, we recall the mathematical model employing the mechanical displacement and fluid pressure as primary variables. Additionally, the problem is transformed by the Kirchhoff transformation, a standard tool for the analysis of non-linear diffusion problems, cf., e.g., [53]. The latter will be subject of the subsequent analysis.

2.1 The original formulation

We consider a poroelastic medium occupying the open, connected, and bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. Let $T > 0$ denote the final time and $(0, T)$ denote the time interval of interest. Let $Q_T := \Omega \times (0, T)$ denote the space-time domain.

The balance equations as derived in [4] (note, we use an arbitrary pore pressure, whereas the specific *average pore pressure* has been used in the aforementioned work) reads on Q_T:

$$-\boldsymbol{\nabla} \cdot \left[2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\boldsymbol{\nabla} \cdot \boldsymbol{u}\mathbf{I} - \alpha p_{\mathrm{pore}}(p_{\mathrm{w}})\mathbf{I}\right] = \boldsymbol{f}, \tag{2.1}$$

$$\phi\partial_t s_{\mathrm{w}}(p_{\mathrm{w}}) + \phi c_{\mathrm{w}}\partial_t p_{\mathrm{w}} + \frac{1}{N} s_{\mathrm{w}}(p_{\mathrm{w}})\partial_t p_{\mathrm{pore}}(p_{\mathrm{w}}) + \alpha s_{\mathrm{w}}(p_{\mathrm{w}})\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u} + \boldsymbol{\nabla} \cdot \boldsymbol{q} = h, \tag{2.2}$$

where $\boldsymbol{u}$ is the mechanical displacement and p_{w} is the fluid pressure (of the active phase). Furthermore, $\boldsymbol{q}$ is the volumetric flux described by the generalized Darcy law

$$\boldsymbol{q} = -\kappa_{\mathrm{abs}}\, \kappa_{\mathrm{rel}}(s_{\mathrm{w}}(p_{\mathrm{w}}))\, (\boldsymbol{\nabla} p_{\mathrm{w}} - \rho_{\mathrm{w}}\boldsymbol{g})\,. \tag{2.3}$$

Constitutive laws are given for the pore pressure p_{pore}, the fluid saturation s_{w} and the relative permeability κ_{rel}; the latter two are assumed to be homogeneous, i.e., they do not vary explicitly in space. Furthermore, $\boldsymbol{f}$ and h are external load and source terms; μ, λ are the Lamé parameters;

$\alpha \in [0,1]$ is the Biot constant; $c_{\mathrm{w}} \in [0,\infty)$ is the storage coefficient associated to fluid compressibility; $N \in (0,\infty]$ is the Biot modulus associated to the compressibility of solid grains; κ_{abs} is the absolute permeability; ρ_{w} is a reference fluid density and $\boldsymbol{g}$ is the gravitational acceleration. Finally, ϕ is the porosity. Under the hypothesis of small perturbations of the porosity [5], often applied along with the assumptions of linear elasticity, we can assume that the porosity ϕ acting as weight is constant in time, equal to some reference porosity field ϕ_0.

From now on, we consider a compact form of (2.1)–(2.3). Specifically, we seek $(\boldsymbol{u}, p_{\mathrm{w}})$ such that on Q_T

$$-\boldsymbol{\nabla}\cdot\left[2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\boldsymbol{\nabla}\cdot\boldsymbol{u}\mathbf{I} - \alpha p_{\mathrm{pore}}(p_{\mathrm{w}})\mathbf{I}\right] = \boldsymbol{f}, \tag{2.4}$$

$$\partial_t b(p_{\mathrm{w}}) + \alpha s_{\mathrm{w}}(p_{\mathrm{w}})\partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u} - \boldsymbol{\nabla}\cdot\left(\kappa_{\mathrm{abs}}\kappa_{\mathrm{rel}}(s_{\mathrm{w}}(p_{\mathrm{w}}))\left(\boldsymbol{\nabla}p_{\mathrm{w}} - \rho_{\mathrm{w}}\boldsymbol{g}\right)\right) = h, \tag{2.5}$$

where the function b is defined as

$$b(p_{\mathrm{w}}) = \phi_0 s_{\mathrm{w}}(p_{\mathrm{w}}) + c_{\mathrm{w}}\phi_0\int_0^{p_{\mathrm{w}}} s_{\mathrm{w}}(p)\,dp + \frac{1}{N}\int_0^{p_{\mathrm{w}}} s_{\mathrm{w}}(p)p'_{\mathrm{pore}}(p)\,dp. \tag{2.6}$$

We note that the subsequent analysis is not dependent on specific choices for b, s_{w}, p_{pore} and κ_{rel}.

In order to close the system (2.4)–(2.5), we impose: boundary conditions

$$\boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}} \qquad \text{on } \Gamma_{\mathrm{D}}^{\mathrm{m}}\times(0,T), \tag{2.7}$$

$$\left(2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\boldsymbol{\nabla}\cdot\boldsymbol{u}\mathbf{I} - \alpha p_{\mathrm{pore}}(p_{\mathrm{w}})\mathbf{I}\right)\boldsymbol{n} = \boldsymbol{\sigma}_{\mathrm{N}} \qquad \text{on } \Gamma_{\mathrm{N}}^{\mathrm{m}}\times(0,T), \tag{2.8}$$

$$p_{\mathrm{w}} = p_{\mathrm{w,D}} \qquad \text{on } \Gamma_{\mathrm{D}}^{\mathrm{f}}\times(0,T), \tag{2.9}$$

$$-\kappa_{\mathrm{abs}}\,\kappa_{\mathrm{rel}}(s_{\mathrm{w}}\,(p_{\mathrm{w}}))\left(\boldsymbol{\nabla}p_{\mathrm{w}} - \rho_{\mathrm{w}}\boldsymbol{g}\right)\cdot\boldsymbol{n} = q_N \qquad \text{on } \Gamma_{\mathrm{N}}^{\mathrm{f}}\times(0,T), \tag{2.10}$$

for the partitions $\{\Gamma_{\mathrm{D}}^{\mathrm{m}}, \Gamma_{\mathrm{N}}^{\mathrm{m}}\}$ and $\{\Gamma_{\mathrm{D}}^{\mathrm{f}}, \Gamma_{\mathrm{N}}^{\mathrm{f}}\}$ of the boundary $\partial\Omega$, where $\Gamma_{\mathrm{D}}^{\mathrm{m}}$ and $\Gamma_{\mathrm{D}}^{\mathrm{f}}$ have positive measure; as well as initial conditions

$$\boldsymbol{u} = \boldsymbol{u}_0 \qquad \text{in } \Omega\times\{0\}, \tag{2.11}$$

$$p_{\mathrm{w}} = p_{\mathrm{w},0}, \qquad \text{in } \Omega\times\{0\}. \tag{2.12}$$

Putting the focus on the non-linear and coupled character of the balance equations, in the subsequent, mathematical analysis, we consider a simplified setting. We neglect gravity and non-homogeneous, essential boundary conditions, which in particular simplifies notation.

2.2 The mathematical model under the Kirchhoff transformation

The Kirchhoff transformation defines a new pressure-like variable

$$\chi(p_{\mathrm{w}}) = \int_0^{p_{\mathrm{w}}} \kappa_{\mathrm{rel}}(s_{\mathrm{w}}(\tilde{p}))\,d\tilde{p}. \tag{2.13}$$

Assuming the constitutive laws satisfy $\kappa_{\mathrm{rel}}(s_{\mathrm{w}}(p)) > 0$, for all $p \in \mathbb{R}$, (2.13) can be inverted. We redefine all functions in p_{w} as functions in χ

$$\hat{p}_{\mathrm{w}} := \chi^{-1}, \quad \hat{b} := b\circ\chi^{-1}, \quad \hat{s}_{\mathrm{w}} := s_{\mathrm{w}}\circ\chi^{-1}, \quad \hat{p}_{\mathrm{pore}} := p_{\mathrm{pore}}\circ\chi^{-1}, \quad \hat{\kappa}_{\mathrm{rel}} := \kappa_{\mathrm{rel}}\circ\hat{s}_{\mathrm{w}}. \tag{2.14}$$

Then under the assumption of a homogeneous relative permeability and saturation, the non-linear Biot equations (2.4)–(2.5) reduces to finding $(\boldsymbol{u}, \chi)$, satisfying

$$-\boldsymbol{\nabla}\cdot\left(2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\boldsymbol{\nabla}\cdot\boldsymbol{u}\mathbf{I} - \alpha\hat{p}_{\mathrm{pore}}(\chi)\mathbf{I}\right) = \boldsymbol{f}, \tag{2.15}$$

$$\partial_t\hat{b}(\chi) + \alpha\hat{s}_{\mathrm{w}}(\chi)\partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u} - \boldsymbol{\nabla}\cdot\left(\kappa_{\mathrm{abs}}\boldsymbol{\nabla}\chi\right) = h, \tag{2.16}$$

on Q_T, and subject to the adapted boundary conditions

$$\begin{aligned} \boldsymbol{u} &= \mathbf{0} && \text{on } \Gamma_{\mathrm{D}}^{\mathrm{m}} \times (0,T), && (2.17)\\ (2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\boldsymbol{\nabla}\cdot\boldsymbol{u}\mathbf{I} - \alpha\hat{p}_{\mathrm{pore}}(\chi)\mathbf{I})\,\boldsymbol{n} &= \boldsymbol{\sigma}_{\mathrm{N}} && \text{on } \Gamma_{\mathrm{N}}^{\mathrm{m}} \times (0,T), && (2.18)\\ \chi &= 0 && \text{on } \Gamma_{\mathrm{D}}^{\mathrm{f}} \times (0,T), && (2.19)\\ -\kappa_{\mathrm{abs}}\boldsymbol{\nabla}\chi\cdot\boldsymbol{n} &= w_N && \text{on } \Gamma_{\mathrm{N}}^{\mathrm{f}} \times (0,T), && (2.20) \end{aligned}$$

and the initial conditions

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{u}_0 && \text{in } \Omega\times\{0\}, && (2.21)\\ \chi &= \chi_0, && \text{in } \Omega\times\{0\}. && (2.22) \end{aligned}$$

3 Main result – existence of a weak solution for the unsaturated poroelasticity model

The main result of this work is the existence result of a weak solution for the unsaturated poroelasticity model under the Kirchhoff transformation, cf. Section 2.2. In this section, we state the main result. This includes the notion of a weak solution, required assumptions and the idea of the proof. The details of the proof are the subject of the remainder of this paper.

3.1 Definition of a weak solution

Let $Q_T := \Omega \times (0,T)$ denote the space-time domain. We use the standard notation for L^p, Sobolev and Bochner spaces, together with their inherent norms and scalar products. Let $\langle\cdot,\cdot\rangle$ denote the standard $L^2(\Omega)$ scalar product for scalars, vectors and tensors. For shorter notation, we use $\|\cdot\| := \|\cdot\|_{L^2(\Omega)}$. Let

$$\begin{aligned} \boldsymbol{V} &= \left\{\boldsymbol{v}\in H^1(\Omega)^d \,\middle|\, \boldsymbol{v}_{|\Gamma_{\mathrm{D}}^{\mathrm{m}}} = \mathbf{0}\right\},\\ Q &= \left\{q\in H^1(\Omega) \,\middle|\, q_{|\Gamma_{\mathrm{D}}^{\mathrm{f}}} = 0\right\}, \end{aligned}$$

denote the function spaces corresponding to mechanical displacement and fluid pressure, respectively, incorporating essential boundary conditions. We abbreviate the bilinear form associated to linear elasticity

$$a(\boldsymbol{u},\boldsymbol{v}) = 2\mu\int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{u})\,dx + \lambda\int_\Omega \boldsymbol{\nabla}\cdot\boldsymbol{u}\,\boldsymbol{\nabla}\cdot\boldsymbol{v}\,dx, \quad \boldsymbol{u},\boldsymbol{v}\in\boldsymbol{V},$$

and define $\|\cdot\|_{\boldsymbol{V}} := a(\cdot,\cdot)^{1/2}$, which induces a norm on $\boldsymbol{V}$ due to Korn's inequality. Moreover, we combine the external body and surface sources as elements in $\boldsymbol{V}^\star$ and $Q^\star$, the duals of $\boldsymbol{V}$ and Q, respectively. Let $\boldsymbol{f}_{\mathrm{ext}} = (\boldsymbol{f},\boldsymbol{\sigma}_{\mathrm{N}})$ and $h_{\mathrm{ext}} = (h, w_{\mathrm{N}})$ be defined by

$$\begin{aligned} \langle\boldsymbol{f}_{\mathrm{ext}},\boldsymbol{v}\rangle &= \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx + \int_{\Gamma_{\mathrm{N}}^{\mathrm{m}}} \boldsymbol{\sigma}_{\mathrm{N}}\cdot\boldsymbol{v}\,ds, \quad && \boldsymbol{v}\in\boldsymbol{V},\\ \langle h_{\mathrm{ext}},q\rangle &= \int_\Omega h\,q\,dx + \int_{\Gamma_{\mathrm{N}}^{\mathrm{f}}} w_{\mathrm{N}}\,q\,ds, && q\in Q. \end{aligned}$$

Definition 3.1 (Weak solution of the unsaturated poroelasticity model)**.** *A weak solution to* (2.15)–(2.22) *is a pair* $(\boldsymbol{u},\chi)\in L^2(0,T;\boldsymbol{V})\times L^2(0,T;Q)$ *satisfying the following:*

(W1) $\hat{p}_{\mathrm{pore}}(\chi)\in L^2(Q_T)$, $\hat{s}_{\mathrm{w}}(\chi)\in L^\infty(Q_T)$.

(W2) *$\hat{b}(\chi) \in L^\infty(0,T;L^1(\Omega))$ and $\partial_t \hat{b}(\chi) \in L^2(0,T;Q^\star)$ such that*

$$\int_0^T \left\langle \partial_t \hat{b}(\chi), q \right\rangle \, dt + \int_0^T \left\langle \hat{b}(\chi) - \hat{b}(\chi_0), \partial_t q \right\rangle \, dt = 0,$$

for all $q \in L^2(0,T;Q)$ with $\partial_t q \in L^1(0,T;L^\infty(\Omega))$ and $q(T) = 0$.

(W3) *$\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u} \in L^2(Q_T)$ such that*

$$\int_0^T \left\langle \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}, q \right\rangle \, dt + \int_0^T \left\langle \boldsymbol{\nabla} \cdot \boldsymbol{u} - \boldsymbol{\nabla} \cdot \boldsymbol{u}_0, \partial_t q \right\rangle \, dt = 0,$$

for all $q \in H^1(0,T;L^2(\Omega))$ with $q(T) = 0$.

(W4) *$(\boldsymbol{u}, \chi)$ satisfies the variational equations*

$$\int_0^T \left[a(\boldsymbol{u}, \boldsymbol{v}) - \alpha \left\langle \hat{p}_{\mathrm{pore}}(\chi), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle \right] dt = \int_0^T \left\langle \boldsymbol{f}_{\mathrm{ext}}, \boldsymbol{v} \right\rangle \, dt, \tag{3.1}$$

$$\int_0^T \left[\left\langle \partial_t \hat{b}(\chi), q \right\rangle + \alpha \left\langle \hat{s}_{\mathrm{w}}(\chi) \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}, q \right\rangle + \left\langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi, \boldsymbol{\nabla} q \right\rangle \right] dt = \int_0^T \left\langle h_{\mathrm{ext}}, q \right\rangle \, dt, \tag{3.2}$$

for all $(\boldsymbol{v}, q) \in L^2(0,T;\boldsymbol{V}) \times L^2(0,T;Q)$.

We note that the weak formulation of the initial conditions (W3) of the mechanical displacement immediately allow for a stronger formulation. See Lemma 9.6 for more information.

3.2 Assumptions on model and data

For proving the existence of a weak solution, we require several assumptions on the model, including the constitutive laws, model parameters, source terms and initial conditions:

(A0) $s_{\mathrm{w}} : \mathbb{R} \to [0,1]$ and $\kappa_{\mathrm{rel}} : [0,1] \to [0,1]$ such that $\kappa_{\mathrm{rel}}(s_{\mathrm{w}}(p)) > 0$, for all $p \in \mathbb{R}$ allowing for defining $\hat{p}_{\mathrm{w}}$, $\hat{b}$, $\hat{s}_{\mathrm{w}}$, $\hat{p}_{\mathrm{pore}}$, and $\hat{\kappa}_{\mathrm{rel}}$ as in (2.14).

(A1) $\hat{b} : \mathbb{R} \to \mathbb{R}$ is continuous and non-decreasing, and it holds that $\hat{b}(0) = 0$.

(A2) $\hat{s}_{\mathrm{w}} : \mathbb{R} \to (0,1]$ continuous and differentiable a.e., and $\hat{s}_{\mathrm{w}}(\chi) = 1$ for $\chi \geq 0$.

(A3) $\hat{p}_{\mathrm{pore}} : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, non-decreasing, and it holds that $\hat{p}_{\mathrm{pore}}(0) = 0$.

(A4) $\frac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}} : \mathbb{R} \to \mathbb{R}$ is invertible and uniformly increasing, i.e., there exists a constant $c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}} > 0$ satisfying $\left(\frac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}} \right)'(x) \geq c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}}$ for all $x \in \mathbb{R}$.

Assumptions (A0)–(A4) are valid for standard constitutive laws, cf. Appendix A. The assumptions on the model parameters read:

(A5) $\mu > 0$, $\lambda \geq 0$, $\alpha \geq 0$ are constant, and define the bulk modulus $K_{\mathrm{dr}} := \frac{2\mu}{d} + \lambda$.

(A6) κ_{abs} is uniformly bounded from below and above, such that there exist constants $0 < \kappa_{\mathrm{m,abs}} \leq \kappa_{\mathrm{M,abs}} < \infty$ with $\kappa_{\mathrm{abs}} \in [\kappa_{\mathrm{m,abs}}, \kappa_{\mathrm{M,abs}}]$ on Ω.

We note, (A5) is stated only for simplicity. The assumptions on the external load and source terms read:

(A7) $\boldsymbol{f}_{\mathrm{ext}} \in H^1(0,T;\boldsymbol{V}^\star) \cap C(0,T;\boldsymbol{V}^\star)$ and $h_{\mathrm{ext}} \in H^1(0,T;Q^\star) \cap C(0,T;Q^\star)$, where

$$\|\boldsymbol{f}_{\mathrm{ext}}\|^2_{L^p(0,T;\boldsymbol{V}^\star)} := \|\boldsymbol{f}\|^2_{L^p(0,T;\boldsymbol{V}^\star)} + \|\boldsymbol{\sigma}_{\mathrm{N}}\|^2_{L^p(0,T;\boldsymbol{V}^\star)}, \qquad p \in \{2,\infty\},$$
$$\|h_{\mathrm{ext}}\|^2_{L^p(0,T;Q^\star)} := \|h\|^2_{L^p(0,T;L^2(\Omega))} + \|w_{\mathrm{N}}\|^2_{L^p(0,T;L^2(\Gamma^{\mathrm{f}}_{\mathrm{N}}))}, \; p \in \{2,\infty\},$$

and analogously $\|\boldsymbol{f}_{\mathrm{ext}}\|_{\boldsymbol{V}^\star}$, $\|\partial_t \boldsymbol{f}_{\mathrm{ext}}\|_{L^2(0,T;\boldsymbol{V}^\star)}$, $\|\boldsymbol{f}_{\mathrm{ext}}\|_{H^1(0,T;\boldsymbol{V}^\star)}$, and $\|h_{\mathrm{ext}}\|_{Q^\star}$, $\|\partial_t h_{\mathrm{ext}}\|_{L^2(0,T;Q^\star)}$, $\|h_{\mathrm{ext}}\|_{H^1(0,T;Q^\star)}$.

The assumptions on the initial data read:

(A8) The initial data $(\boldsymbol{u}_0, \chi_0) \in \boldsymbol{V} \times Q$ is sufficient regular such that there exists a constant C_0 satisfying

$$\|\boldsymbol{u}_0\|^2_{\boldsymbol{V}} + \|\boldsymbol{\nabla}\chi_0\|^2 + \left\|\hat{b}\left(\chi_0\right)\right\|_{L^1(\Omega)} + \left\|\hat{B}\left(\chi_0\right)\right\|_{L^1(\Omega)}$$
$$+ \left\|\bar{B}\left(\frac{\hat{p}_{\mathrm{pore}}(\chi_0)}{\hat{s}_{\mathrm{w}}(\chi_0)}\right)\right\|_{L^1(\Omega)} + \|\hat{p}_{\mathrm{pore}}(\chi_0)\|^2 \leq C_0,$$

where $\hat{B}$ and $\bar{B}$ are the Legendre transformations of $\hat{b}$ and $\bar{b} := \hat{b} \circ \left(\frac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}}\right)^{-1}$, respectively:

$$\hat{B}(z) := \int_0^z (\hat{b}(z) - \hat{b}(s))\, ds \geq 0, \tag{3.3}$$
$$\bar{B}(z) := \int_0^z (\bar{b}(z) - \bar{b}(s))\, ds \geq 0. \tag{3.4}$$

(A9) The initial data $(\boldsymbol{u}_0, \chi_0)$ satisfies the compatibility condition: $\hat{p}_{\mathrm{pore}}(\chi_0) \in Q$ and

$$a(\boldsymbol{u}_0, \boldsymbol{v}) - \alpha \left\langle \hat{p}_{\mathrm{pore}}(\chi_0), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle = \left\langle \boldsymbol{f}_{\mathrm{ext}}(0), \boldsymbol{v} \right\rangle, \quad \text{for all } \boldsymbol{v} \in \boldsymbol{V},$$

i.e., the mechanics equation at initial time.

Additionally, the following non-degeneracy conditions are required:

(ND1) There exists a constant $C_{\mathrm{ND},1} > 0$ such that

$$\left|\frac{\hat{p}_{\mathrm{pore}}(\chi)}{\hat{s}_{\mathrm{w}}(\chi)\chi}\right| \leq C_{\mathrm{ND},1}, \quad \text{for all } \chi \in \mathbb{R}.$$

(ND2) There exists a constant $C_{\mathrm{ND},2} > 0$ such that

$$C^{-1}_{\mathrm{ND},2} \leq \hat{p}'_{\mathrm{pore}}(\chi) \leq C_{\mathrm{ND},2}, \quad \text{for all } \chi \in \mathbb{R}.$$

(ND3) There exists a constant $C_{\mathrm{ND},3} \in (0,1)$ such that

$$K_{\mathrm{dr}} - \frac{\alpha^2}{4}\left(\frac{\hat{s}_{\mathrm{w}}(\chi)}{\hat{p}'_{\mathrm{pore}}(\chi)} - 1\right)^2 \frac{\left(\hat{p}'_{\mathrm{pore}}(\chi)\right)^2}{\hat{b}'(\chi)} \geq C_{\mathrm{ND},3} K_{\mathrm{dr}}, \quad \text{for all } \chi \in \mathbb{R}.$$

In Appendix A, it is demonstrated that for the van Genuchten model for s_{w} and κ_{rel} [63], and the equivalent pore pressure model for p_{pore} [5], (ND1) and (ND2) follow if the saturation takes values above a residual saturation. Thus, (ND1) and (ND2) may be implicitly satisfied assuming (ND3) holds true. Furthermore, the calculations in Appendix A illustrate that for materials typically present in geotechnical application, the condition (ND3) is satisfied in saturation regimes above 1 to 10 percent (depending on the material parameters). Thereby, the practical saturation regime is covered for a wide range of applications. After all, (ND3) is the most restrictive assumption of all assumptions. It essentially requires the mechanical system to be sufficiently stiff in relation to the saturation profile. The lower the minimal saturation value, the stiffer the system has to be.

3.3 Existence of solutions for the unsaturated poroelasticity model

This section is presenting the main result together with the main steps of the proof.

Theorem 3.2 (Existence of a weak solution to the unsaturated poroelasticity model)**.** *Under the model assumptions* (A0)–(A9) *and the non-degeneracy conditions* (ND1)–(ND3)*, there exists a weak solution of* (2.15)–(2.22) *in the sense of Definition 3.1.*

The main idea of the proof of Theorem 3.2 is to use the Galerkin method in combination with compactness arguments. The main difficulty here is the control over the non-linear coupling terms. For this a regularization approach is used. After all, the proof consists of six steps. In the following, we present the idea of each step. Details are subject of the remainder of the article and will be presented in the six, subsequent sections.

Step 1: Double physical regularization. Applying the Galerkin method along with compactness arguments for the original problem (3.1)–(3.2) is challenging due to the coupling terms. A simple way to control the term $\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}$ is to add a suitable regularization term in the mechanics equation (3.1). As the coupling terms also involve non-linearities in the Kirchhoff pressure, strong compactness is required. Therefore, we add a coercive term in the flow equation, which allows for controlling the term $\partial_t \chi$. In this way, one can control the coupling terms, and eventually leading to convergence.

From a physical point of view, the regularized model accounts for secondary consolidation and compressible solid grains. In mathematical terms, it reads as follows. For given regularization parameters $\zeta, \eta > 0$, find $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ to be the solution to the variational equations

$$\int_0^T \Big[\zeta a(\partial_t \boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) + a(\boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) - \alpha \left\langle \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta}), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle \Big]\, dt = \int_0^T \left\langle \boldsymbol{f}_{\mathrm{ext}}, \boldsymbol{v} \right\rangle \, dt, \tag{3.5}$$

$$\int_0^T \Big[\Big\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), q \Big\rangle + \alpha \left\langle \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta}) \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta}, q \right\rangle + \left\langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}, \boldsymbol{\nabla} q \right\rangle \Big]\, dt = \int_0^T \left\langle h_{\mathrm{ext}}, q \right\rangle \, dt, \tag{3.6}$$

for all $(\boldsymbol{v}, q) \in L^2(0,T; \boldsymbol{V}) \times L^2(0,T; Q)$, where $\hat{b}_\eta$ is a strictly increasing regularization of $\hat{b}$ (see (A1*) for further properties). The next two steps prove that the regularized problem has a weak solution in an analogous sense to Definition 3.1.

Step 2: Discretization in space and time. We employ the implicit Euler scheme and a Galerkin method based on an inf-sup stable finite element/finite volume method to obtain a fully discrete counterpart of (3.5)–(3.6). In particular, the pressure variables are discretized by piecewise constant elements, and for the diffusion term a discrete gradient $\boldsymbol{\nabla}_h$ is employed corresponding to a two-point flux approximation of the volumetric fluxes [62, 64].

Given an admissible mesh $\mathcal{T}$, cf. Definition 5.1, the conforming and non-conforming, discrete spaces $\boldsymbol{V}_h \subset \boldsymbol{V}$ and $Q_h \not\subset Q$, respectively, and a partition $\{t_n\}_n$ of the interval $(0,T)$, the discretization for time steps n reads: given the solution at the previous time step $(\boldsymbol{u}_h^{n-1}, \chi_h^{n-1}) \in \boldsymbol{V}_h \times Q_h$, find $(\boldsymbol{u}_h^n, \chi_h^n) \in \boldsymbol{V}_h \times Q_h$ satisfying for all $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h \times Q_h$

$$\zeta \tau^{-1} a(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) + a(\boldsymbol{u}_h^n, \boldsymbol{v}_h) - \alpha \langle \hat{p}_{\mathrm{pore}}(\chi_h^n), \boldsymbol{\nabla} \cdot \boldsymbol{v}_h \rangle \; = \; \langle \boldsymbol{f}_{\mathrm{ext}}^n, \boldsymbol{v}_h \rangle, \tag{3.7}$$

$$\langle \hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^{n-1}), q_h \rangle + \alpha \langle \hat{s}_{\mathrm{w}}(\chi_h^n) \boldsymbol{\nabla} \cdot (\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}), q_h \rangle + \tau \langle \boldsymbol{\nabla}_h \chi_h^n, \boldsymbol{\nabla}_h q_h \rangle_{\kappa_{\mathrm{abs}}} \; = \; \tau \langle h_{\mathrm{ext}}^n, q_h \rangle. \tag{3.8}$$

The reason for this particular choice of a discretization is two-fold: (i) the piecewise constant approximation of the pressure allows for the simple handling of non-linearities; (ii) the discrete gradients $\boldsymbol{\nabla}_h$ retain the local character of the differential operator. This together allows for simultaneously cancelling the coupling terms and utilizing the coercivity of the diffusion term. This is required, e.g., for proving the existence of a discrete solution employing a corollary of Brouwer's fixed point theorem, or in Step 3.

Step 3: Existence of a weak solution to the regularized model. Based on the discrete values $\{(\boldsymbol{u}_h^n, \chi_h^n)\}_n$, we define suitable interpolations in time, $(\boldsymbol{u}_{h\tau}, \chi_{h\tau})$, yielding approximations of $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$. We remark that various interpolations are in fact introduced in the course of step 3 and 4. To avoid an excess in notations and for the ease of the presentation, we use the same notation, $(\boldsymbol{u}_{h\tau}, \chi_{h\tau})$, for all interpolations.

The goal is to show convergence (in a certain sense) of $\{(\boldsymbol{u}_{h\tau}, \chi_{h\tau})\}_{h,\tau}$ along a monotonically decreasing sequence of pairs $(h,\tau) \to (0,0)$ (from now on denoted $h,\tau \to 0$) towards a solution of (3.5)–(3.6). This is achieved using compactness arguments; however, given the coupled and non-linear nature of (3.5)–(3.6), several terms require careful discussion:

- Non-linearities as $\hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta})$ or products of independent variables as $\hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta}$ require partially strong convergence.
- Since $\hat{b}_\eta$ is not necessarily Lipschitz continuous, it is not sufficient to show uniform stability for $\{\partial_t \chi_{h\tau}\}$ to conclude weak convergence of $\{\partial_t \hat{b}_\eta(\chi_{h\tau})\}_{h,\tau}$ towards $\partial_t \hat{b}_\eta(\chi_{\varepsilon\eta})$. Instead, we apply techniques by [53] utilizing the Legendre transformation, $\hat{B}_\eta$, of $\hat{b}_\eta$, analogously defined as in (3.3).
- The diffusion term is discretized using discrete gradients. Thus, weak convergence $\boldsymbol{\nabla}_h \chi_{h\tau} \to \boldsymbol{\nabla} \chi_{\varepsilon\eta}$ is not an obvious consequence of uniform stability for $\{\boldsymbol{\nabla}_h \chi_{h,\tau}\}_{h,\tau}$. For this, we apply techniques from finite volume literature [61, 62].

Motivated by that, we first derive stability estimates that are uniform wrt. the discretization parameters

$$\begin{aligned}
&\|\boldsymbol{u}_{h\tau}\|_{H^1(0,T;\boldsymbol{V})} + \operatorname*{ess\,sup}_{t\in(0,T)} \|\chi_{h\tau}(t)\|_{1,\mathcal{T}} + \|\hat{p}_{\mathrm{pore}}(\chi_{h\tau})\|_{L^2(Q_T)} \\
&\quad + \left\|\hat{B}_\eta(\chi_{h\tau})\right\|_{L^\infty(0,T;L^1(\Omega))} + \left\|\partial_t \hat{b}_\eta(\chi_{h\tau})\right\|_{L^2(0,T;H^{-1}(\Omega))} + \|\partial_t \chi_{h\tau}\|_{L^2(Q_T)} \leq C_{\zeta\eta}
\end{aligned}$$

for some constant $C_{\zeta\eta} > 0$ independent of h, τ. Therefore, one obtains weak convergence for subsequences (denoted the same as before) for $h, \tau \to 0$

$$\begin{aligned}
\boldsymbol{u}_{h\tau} &\rightharpoonup \boldsymbol{u}_{\varepsilon\eta} && \text{in } L^2(0,T;\boldsymbol{V}), \\
\partial_t \boldsymbol{u}_{h\tau} &\rightharpoonup \partial_t \boldsymbol{u}_{\varepsilon\eta} && \text{in } L^2(0,T;\boldsymbol{V}), \\
\hat{p}_{\mathrm{pore}}(\chi_{h\tau}) &\rightharpoonup \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta}) && \text{in } L^2(Q_T), \\
\partial_t \hat{b}_\eta(\chi_{h\tau}) &\rightharpoonup \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}) && \text{in } L^2(0,T;Q^\star), \\
\hat{s}_{\mathrm{w}}(\chi_{h\tau})\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{h\tau} &\rightharpoonup \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta} && \text{in } L^2(Q_T), \\
\boldsymbol{\nabla}_h \chi_{h\tau} &\rightharpoonup \boldsymbol{\nabla} \chi_{\varepsilon\eta} && \text{in } L^2(Q_T).
\end{aligned}$$

Moreover, by employing finite volume techniques the following convergence of the discrete diffusion term can be showed

$$\int_0^T \langle \boldsymbol{\nabla}_h \chi_{h\tau}, \boldsymbol{\nabla}_h q_h \rangle_{\kappa_{\mathrm{abs}}} \, dt \to \int_0^T \langle \boldsymbol{\nabla} \chi_{\varepsilon\eta}, \boldsymbol{\nabla} q \rangle_{\kappa_{\mathrm{abs}}} \, dt,$$

for arbitrary discrete test functions q_h, which strongly converge towards continuous functions q. Finally, the limit, $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$, can be identified as weak solution of the regularized problem (3.5)–(3.6).

Step 4: Increased regularity for the weak solution of the regularized model. When discussing the limit $\zeta \to 0$ in step 5, it will be beneficial to have access to the derivative in time of the mechanics equation (3.5) . Under the additional non-degeneracy condition (ND2), i.e., that $\hat{p}_{\text{pore}}$ is Lipschitz continuous, an increased regularity can be showed for the weak solution of the regularized model, $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$. For instance, for all $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V})$ it holds that

$$\int_0^T \left[\zeta a(\partial_{tt}\boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) + a(\partial_t \boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) - \alpha \left\langle \partial_t \hat{p}_{\text{pore}}(\chi_{\varepsilon\eta}), \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle\right] dt = \int_0^T \left\langle \partial_t \boldsymbol{f}_{\text{ext}}, \boldsymbol{v}\right\rangle \, dt. \tag{3.9}$$

The proof follows the same line of argumentation as step 3. First a fully discrete counterpart of (3.9) is constructed by considering differences of (3.7) between subsequent time steps

$$\begin{aligned}&\zeta\tau^{-1} a(\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}, \boldsymbol{v}_h) + a(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h)\\ &\quad - \alpha\left\langle \hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1}), \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\right\rangle = \left\langle \boldsymbol{f}_{\text{ext}}^n - \boldsymbol{f}_{\text{ext}}^{n-1}, \boldsymbol{v}_h\right\rangle \quad \text{for all } \boldsymbol{v}_h \in \boldsymbol{V}_h.\end{aligned}$$

In addition, suitable interpolations $\hat{\boldsymbol{u}}_{t,h\tau}$ and $\hat{p}_{\text{pore},h\tau}$ of the discrete values $\{\tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1})\}_n$ and $\{\hat{p}_{\text{pore}}(\chi_h^n)\}_n$, respectively, define approximations of $\partial_t \boldsymbol{u}_{\varepsilon\eta}$ and $\hat{p}_{\text{pore}}(\chi_{\varepsilon\eta})$. The uniform stability estimate

$$\|\partial_t \hat{\boldsymbol{u}}_{t,h\tau}\|^2_{L^2(0,T;\boldsymbol{V})} + \|\partial_t \boldsymbol{u}_{h\tau}\|^2_{L^2(0,T;\boldsymbol{V})} + \|\partial_t \hat{p}_{\text{pore},h\tau}\|^2_{L^2(Q_T)} \leq C_{\zeta\eta}$$

guarantee the weak convergences

$$\begin{aligned}\partial_t \hat{\boldsymbol{u}}_{t,h\tau} &\rightharpoonup \partial_{tt}\boldsymbol{u}_{\varepsilon\eta}, && \text{in } L^2(0,T;\boldsymbol{V}),\\ \partial_t \boldsymbol{u}_{h\tau} &\rightharpoonup \partial_t \boldsymbol{u}_{\varepsilon\eta}, && \text{in } L^2(0,T;\boldsymbol{V}),\\ \partial_t \hat{p}_{\text{pore}}(\chi)_{h\tau} &\rightharpoonup \partial_t \hat{p}_{\text{pore}}(\chi_{\varepsilon\eta}), && \text{in } L^2(Q_T)\end{aligned}$$

up to subsequences, for $h, \tau \to 0$. Finally, one can identify (3.9) in the limit.

Step 5: Vanishing regularization in the mechanics equation. For each $\zeta, \eta > 0$, there exists a solution $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ to (3.5)–(3.6). For the limit $\zeta \to 0$, we employ compactness arguments similar to step 3. However, now the stability estimates ought to be independent of ζ. We show

$$\begin{aligned}&\|\boldsymbol{u}_{\varepsilon\eta}\|^2_{H^1(0,T;\boldsymbol{V})} + \|\chi_{\varepsilon\eta}\|^2_{L^\infty(0,T;Q)} + \|p_{\text{pore}}(\chi_{\varepsilon\eta})\|^2_{L^2(Q_T)}\\ &\quad + \left\|\hat{B}_\eta(\chi_{\varepsilon\eta})\right\|_{L^\infty(0,T;L^1(\Omega))} + \left\|\partial_t \hat{b}_\eta(\chi_{\varepsilon\eta})\right\|_{L^2(0,T;H^{-1}(\Omega))} \leq C,\end{aligned} \tag{3.10}$$

and

$$\|\partial_t \chi_{\varepsilon\eta}\|^2_{L^2(Q_T)} \leq C_\eta. \tag{3.11}$$

For (3.10), one can use $\boldsymbol{v} = \partial_t \boldsymbol{u}_{\varepsilon\eta}$ and $q = \partial_t \chi_{\varepsilon\eta}$ as test functions in (3.6) and (3.9). The coupling terms obviously do not match; but by using a binomial identity and the non-degeneracy condition (ND3), one can show that

$$\|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|^2_{L^2(0,T;\boldsymbol{V})} + \int_0^T \left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), \partial_t \chi_{\varepsilon\eta}\right\rangle + \alpha \int_0^T \left\langle \hat{s}_{\text{w}} \partial_t \chi_{\varepsilon\eta} - \partial_t \hat{p}_{\text{pore}}, \partial_t \boldsymbol{\nabla}\cdot\boldsymbol{u}_{\varepsilon\eta}\right\rangle \geq 0, \tag{3.12}$$

which effectively allows for dropping the coupling terms. With this, letting $\zeta \to 0$, one obtains for subsequences (denoted the same as before)

$$\begin{aligned}\boldsymbol{u}_{\varepsilon\eta} &\rightharpoonup \boldsymbol{u}_\eta && \text{in } L^2(0,T;\boldsymbol{V}),\\ \partial_t \boldsymbol{u}_{\varepsilon\eta} &\rightharpoonup \partial_t \boldsymbol{u}_\eta && \text{in } L^2(0,T;\boldsymbol{V}),\\ \zeta\partial_t \boldsymbol{u}_{\varepsilon\eta} &\to \boldsymbol{0} && \text{in } L^2(0,T;\boldsymbol{V}),\\ \chi_{\varepsilon\eta} &\rightharpoonup \chi_\eta && \text{in } L^\infty(0,T;Q),\\ \hat{p}_{\text{pore}}(\chi_{\varepsilon\eta}) &\rightharpoonup \hat{p}_{\text{pore}}(\chi_\eta) && \text{in } L^2(Q_T),\\ \hat{s}_{\text{w}}(\chi_{\varepsilon\eta})\partial_t \boldsymbol{\nabla}\cdot\boldsymbol{u}_{\varepsilon\eta} &\rightharpoonup \hat{s}_{\text{w}}(\chi_\eta)\partial_t \boldsymbol{\nabla}\cdot\boldsymbol{u}_\eta && \text{in } L^2(Q_T),\\ \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}) &\rightharpoonup \partial_t \hat{b}_\eta(\chi_\eta) && \text{in } L^2(0,T;Q^\star).\end{aligned}$$

Finally, it is straightforward to see that the limit $(\boldsymbol{u}_\eta, \chi_\eta)$ is weak solution of (3.5)–(3.6) for $\zeta = 0$.

We underline, that for showing (3.12), the time-continuous character of the variational problem is required. It is not obvious how to use a similar strategy on time-discrete level. Therefore, step 5 has been performed separately from step 3 and 4.

Step 6: Vanishing regularization in the flow equation. In the presence of fluid or solid grain compressibility in the original formulation, i.e., $c_\mathrm{w} > 0$ or $\frac{1}{N} > 0$, respectively, this final step is obsolete. Otherwise, we consider the limit process $\eta \to 0$ for the sequence of solutions $\{(\boldsymbol{u}_\eta, \chi_\eta)\}_\eta$, derived in step 5. The overall idea is the same as in step 5, namely to obtain estimates that are uniform wrt. η and to use compactness arguments. Referring to (3.10), the following estimate is uniform in η

$$\begin{aligned}\|\boldsymbol{u}_\eta\|_{H^1(0,T;\boldsymbol{V})} + \|\chi_\eta\|_{L^\infty(0,T;H_0^1(\Omega))} + \|\hat{p}_\mathrm{pore}(\chi_\eta)\|_{L^2(Q_T)}& \\ + \left\|\hat{B}_\eta(\chi_\eta)\right\|_{L^\infty(0,T;L^1(\Omega))} + \left\|\partial_t \hat{b}_\eta(\chi_\eta)\right\|_{L^2(0,T;H^{-1}(\Omega))} &\leq C.\end{aligned} \tag{3.13}$$

For estimating $\partial_t \chi_\eta$, we first show that the time derivative of the mechanics equation (3.7) is well-defined for $\zeta = 0$, i.e., it holds for all $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V})$ that

$$\int_0^T a(\partial_t \boldsymbol{u}_\eta, \boldsymbol{v})\, dt - \int_0^T \alpha \left\langle \partial_t \hat{p}_\mathrm{pore}(\chi_\eta), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle\, dt = \int_0^T \left\langle \partial_t \boldsymbol{f}_\mathrm{ext}, \boldsymbol{v} \right\rangle\, dt. \tag{3.14}$$

Since $\|\partial_t \chi_\eta\| \lesssim \|\partial_t \hat{p}_\mathrm{pore}(\chi_\eta)\|$, the uniform stability for $\partial_t \chi_\eta$ follows by an inf-sup argument, (3.14), and the stability bound (3.13). Due to the lack of a suitable bound on $\partial_{tt} \boldsymbol{u}_{\varepsilon\eta}$ in step 5, this approach only works for $\zeta = 0$. Standard compactness arguments allow for extracting subsequences (again denotes as before) such that for $\eta \to 0$ it holds that

$$\begin{aligned} \boldsymbol{u}_\eta &\rightharpoonup \boldsymbol{u} && \text{in } L^2(0,T;\boldsymbol{V}),\\ \chi_\eta &\rightharpoonup \chi && \text{in } L^\infty(0,T;Q),\\ \hat{p}_\mathrm{pore}(\chi_\eta) &\rightharpoonup \hat{p}_\mathrm{pore}(\chi) && \text{in } L^2(Q_T),\\ \hat{s}_\mathrm{w}(\chi_\eta)\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_\eta &\rightharpoonup \hat{s}_\mathrm{w}(\chi)\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u} && \text{in } L^2(Q_T),\\ \partial_t \hat{b}_\eta(\chi_\eta) &\rightharpoonup \partial_t \hat{b}(\chi) && \text{in } L^2(0,T;Q^\star). \end{aligned}$$

Ultimately, $(\boldsymbol{u}, \chi)$ can be identified as a weak solution to the unsaturated poroelasticity model in the sense of Definition 3.1. This finishes the proof of Theorem 3.2.

4 Step 1: Physical regularization – secondary consolidation and enhanced grain compressibility

We introduce a physical regularization of the weak formulation (3.1)–(3.2) by enhancing both the mechanics and the flow equations. We allow for secondary consolidation, which effectively incorporates a linear visco-elasticity contribution in the mechanics equations of the form $a(\partial_t \boldsymbol{u}, \boldsymbol{v})$. Additionally, we assume non-vanishing grain compressibility by regularizing $\hat{b}$. Specifically, we let $\zeta > 0$ and $\eta > 0$ be two regularization parameters and analyze the behavior of the solution when passing them to zero.

Motivated by the physical example (2.6), for $\eta > 0$, define the regularization $\hat{b}_\eta$ of $\hat{b}$ by

$$\hat{b}_\eta(\chi) := \hat{b}(\chi) + \eta \int_0^{\hat{p}_\mathrm{w}(\chi)} s_\mathrm{w}(p) p'_\mathrm{pore}(p)\, dp,$$

i.e., $\hat{b}_\eta$ has the same structure as $\hat{b}$, but with $\frac{1}{N} + \eta$ replacing $\frac{1}{N}$. Refering to Section 3.2, the function $\hat{b}_\eta$ still satisfies (A1). Additionally, a uniform growth condition holds

(A1*) There exists a $\hat{b}_{\chi,\mathrm{m}} > 0$ s.t. $\hat{b}_{\chi,\mathrm{m}}\|\chi_1 - \chi_2\|^2 \leq \left\langle \hat{b}_\eta(\chi_1) - \hat{b}_\eta(\chi_2), \chi_1 - \chi_2 \right\rangle$ for all $\chi_1, \chi_2 \in L^2(Q_T)$,

cf. also Section A. In the subsequent discussion, a growth condition for $\hat{b}_\eta$ (or $\hat{b}$) of type (A1*) will be required in order to to utilize strong compactness arguments. If $\min\left\{c_\mathrm{w}, \frac{1}{N}\right\} > 0$ in (2.6) holds, the growth condition (A1*) is fulfilled even for $\eta = 0$, and the regularization of the flow equation actually is not necessary, cf. Step 6 in Section 9. In this context, we emphasize that (ND3) also holds for $\hat{b}_\eta$ as $\hat{b}_\eta' \geq \hat{b}'$.

Also (A8) can be adapted for the regularization $\hat{b}_\eta$. With $\bar{b}_\eta := \hat{b}_\eta \circ \left(\frac{\hat{p}_\mathrm{pore}}{\hat{s}_\mathrm{w}}\right)^{-1}$, we let $\hat{B}_\eta$ and $\bar{B}_\eta$ be the Legendre transformations of $\hat{b}_\eta$ and $\bar{b}_\eta$, respectively, defined by

$$\hat{B}_\eta(z) := \int_0^z (\hat{b}_\eta(z) - \hat{b}_\eta(s))\, ds \geq 0, \tag{4.1}$$

$$\bar{B}_\eta(z) := \int_0^z (\bar{b}_\eta(z) - \bar{b}_\eta(s))\, ds \geq 0. \tag{4.2}$$

(A8*) There exists a $\eta_0 > 0$ and $C_0 > 0$, not depending on η_0, such that

$$\|\boldsymbol{u}_0\|_{\boldsymbol{V}}^2 + \|\boldsymbol{\nabla}\chi_0\|^2 + \left\|\hat{B}_\eta\left(\chi_0\right)\right)\right\|_{L^1(\Omega)} + \left\|\bar{B}_\eta\left(\frac{\hat{p}_\mathrm{pore}(\chi_0)}{\hat{s}_\mathrm{w}(\chi_0)}\right)\right\|_{L^1(\Omega)} \leq C_0$$

for all $\eta \in (0, \eta_0)$. Without loss of generality, we assume C_0 in (A8) and (A8*) to be the same.

For a non-degenerate initial condition χ_0, the additional terms in $\hat{B}_\eta$ and $\bar{B}_\eta$ can be essentially bounded by $\eta\|\chi_0\|^2$, which itself is bounded by (A8).

We introduce the notion of a weak solution of the doubly regularized unsaturated poroelasticity model.

Definition 4.1 (Weak solution of the doubly regularized model)**.** *For $\zeta > 0$ and $\eta > 0$, we call $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta}) \in L^2(0,T;\boldsymbol{V}) \times L^2(0,T;Q)$ a weak solution of the doubly regularized unsaturated poroelasticity model if it satisfies:*

(W1)$_{\zeta\eta}$ *$\hat{p}_\mathrm{pore}(\chi_{\varepsilon\eta}) \in L^2(Q_T)$, $\hat{s}_\mathrm{w}(\chi_{\varepsilon\eta}) \in L^\infty(Q_T)$.*

(W2)$_{\zeta\eta}$ *$\hat{b}_\eta(\chi_{\varepsilon\eta}) \in L^\infty(0,T;L^1(\Omega))$ and $\partial_t\hat{b}_\eta(\chi_{\varepsilon\eta}) \in L^2(0,T;Q^\star)$ such that*

$$\int_0^T \left\langle \partial_t\hat{b}_\eta(\chi_{\varepsilon\eta}), q \right\rangle dt + \int_0^T \left\langle \hat{b}_\eta(\chi_{\varepsilon\eta}) - \hat{b}_\eta(\chi_0), \partial_t q \right\rangle dt = 0,$$

for all $q \in L^2(0,T;Q)$ with $\partial_t q \in L^1(0,T;L^\infty(\Omega))$ and $q(T) = 0$.

(W3)$_{\zeta\eta}$ *$\partial_t\boldsymbol{u}_{\varepsilon\eta} \in L^2(0,T;\boldsymbol{V})$ such that*

$$\int_0^T a(\partial_t\boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v})\, dt + \int_0^T a(\boldsymbol{u}_{\varepsilon\eta} - \boldsymbol{u}_0, \partial_t\boldsymbol{v})\, dt = 0,$$

for all $\boldsymbol{v} \in H^1(0,T;\boldsymbol{V})$ with $\boldsymbol{v}(T) = \boldsymbol{0}$.

(W4)$_{\zeta\eta}$ *$(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ satisfies the variational equations*

$$\int_0^T \Big[\zeta a(\partial_t\boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) + a(\boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) - \alpha\left\langle \hat{p}_\mathrm{pore}(\chi_{\varepsilon\eta}), \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle\Big]\, dt = \int_0^T \left\langle \boldsymbol{f}_\mathrm{ext}, \boldsymbol{v}\right\rangle\, dt, \tag{4.3}$$

$$\int_0^T \Big[\left\langle \partial_t\hat{b}_\eta(\chi_{\varepsilon\eta}), q\right\rangle + \alpha\left\langle \hat{s}_\mathrm{w}(\chi_{\varepsilon\eta})\partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u}_{\varepsilon\eta}, q\right\rangle + \left\langle \kappa_\mathrm{abs}\boldsymbol{\nabla}\chi_{\varepsilon\eta}, \boldsymbol{\nabla}q\right\rangle\Big]\, dt = \int_0^T \left\langle h_\mathrm{ext}, q\right\rangle\, dt, \tag{4.4}$$

for all $(\boldsymbol{v}, q) \in L^2(0,T;\boldsymbol{V}) \times L^2(0,T;Q)$.

Furthermore, we call $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ *a weak solution with increased regularity for the doubly regularized unsaturated poroelasticity model if it satisfies* $(\mathrm{W1})_{\zeta\eta}$*–*$(\mathrm{W4})_{\zeta\eta}$ *and:*

$(\mathrm{W5})_{\zeta\eta}$ $\boldsymbol{u}_{\varepsilon\eta} \in H^2(0,T;\boldsymbol{V})$ *and* $\partial_t \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta}) \in L^2(Q_T)$.

$(\mathrm{W6})_{\zeta\eta}$ *It holds*

$$\int_0^T \Big[\zeta a(\partial_{tt}\boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) + a(\partial_t \boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) - \alpha \left\langle \partial_t \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta}), \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle\Big]\, dt = \int_0^T \left\langle \partial_t \boldsymbol{f}_{\mathrm{ext}}, \boldsymbol{v}\right\rangle\, dt, \quad (4.5)$$

for all $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V})$*, given that* $\boldsymbol{f}_{\mathrm{ext}} \in H^1(0,T;\boldsymbol{V}^\star)$.

We will later separately consider $\zeta \to 0$ and $\eta \to 0$. Therefore, we give the definition of a weak solution for the simply regularized unsaturated poroelasticity model, obtained for $\eta > 0$ and $\zeta = 0$.

Definition 4.2 (Weak solution of the simply regularized model)**.** *For* $\eta > 0$*, we call* $(\boldsymbol{u}_\eta, \chi_\eta)$ *a weak solution of the simply regularized unsaturated poroelasticity model if it satisfies* $(\mathrm{W1})_{\zeta\eta}$*–*$(\mathrm{W4})_{\zeta\eta}$ *for* $\zeta = 0$.

To distinguish between the equations satisfied by the weak solution of a doubly regularized model and the one of the simply regularized one, where $\epsilon_v = 0$, we use the notations $(\mathrm{W1})_\eta$–$(\mathrm{W4})_\eta$.

Lemma 4.3 (Existence of a weak solution to the doubly regularized model)**.** *Let* $\zeta > 0$ *and* $\eta > 0$ *be given. Under the assumptions* (A0)*–*(A9) *and* (ND1) *there exists a weak solution to the doubly regularized unsaturated poroelasticity model, in the sense of Definition 4.1.*

Proof. The assertion follows from steps 2–3. □

Lemma 4.4 (Existence of a weak solution with increased regularity for the doubly regularized model)**.** *Let* $\zeta > 0$ *and* $\eta > 0$ *be given. Under the assumptions* (A0)*–*(A9) *and the non-degeneracy conditions* (ND1)*–*(ND2)*, the doubly regularized unsaturated poroelasticity model has a weak solution with increased regularity, in the sense of Definition 4.1.*

Proof. The assertion follows from steps 2–4. □

Lemma 4.5 (Existence of a weak solution for the simply regularized model)**.** *Let* $\eta > 0$ *be given. Under the assumptions* (A0)*–*(A9) *and the non-degeneracy conditions* (ND1)*–*(ND3)*, the doubly regularized unsaturated poroelasticity model has a weak solution with increased regularity, in the sense of Definition 4.2.*

Proof. The assertion follows from step 5. □

5 Step 2: Implicit Euler non-linear FEM-TPFA discretization

The next two sections, identified with steps 2 and 3, are providing the proof of Lemma 4.3. To this aim, we employ the implicit Euler time stepping method, whereas for the spatial discretization of the mechanics equation (4.3) a conforming Galerkin finite element method is used. For the flow equation (4.4), the spatial discretization can be interpreted in various ways. It can be viewed as cell-centered finite volume method utilizing a two point flux approximation (TPFA), the simplest approximation one can consider, but it can also be interpreted as lowest order mixed finite element method with inexact quadrature allowing for lumping [65]. In this section, we show the existence of a fully discrete solution. We start with introducing the notations used in the discretization.

5.1 Finite volume and finite element notation

We use standard notations in the finite volume literature, see e.g. [61, 62]. In particular, we introduce notation for elements, faces, their measures, transmissibilities etc. We assume that the domain Ω is polygonal such that it can be discretized by an admissible mesh, as introduced by [64].

Definition 5.1 (Admissible mesh $\mathcal{T}$)**.** *Let $\mathcal{T}$ be a regular mesh of Ω with mesh size h, consisting of simplices in 2D or 3D, or convex quadrilaterals in 2D and convex hexahedrals in 3D. Furthermore, we introduce the following terminology:*

- *$K \in \mathcal{T}$ denotes a single element.*
- *$\mathcal{N}(K) := \{L \in \mathcal{T} \,|\, L \neq K,\ \bar{L} \cap \bar{K} \neq \emptyset\}$ denotes the set of neighboring elements of $K \in \mathcal{T}$.*
- *$\mathcal{E}$ denotes the set of all faces, i.e., boundaries of all elements; let $\mathcal{E}_K$ denote the faces of a single element $K \in \mathcal{T}$; let $\mathcal{E}_{\text{ext}}$ denote the faces lying on the boundary $\partial\Omega$.*
- *$K|L \in \mathcal{E}$ denotes the face between two neighboring elements $K, L \in \mathcal{T}$.*
- *$\{x_K\}_{K\in\mathcal{T}}$ is such that for all $K \in \mathcal{T}, L \in \mathcal{N}(K)$ the connecting line between x_K and x_L is perpendicular to $K|L$.*
- *$d_{K,\sigma}$ denotes the distance between center of K and $\sigma \in \mathcal{E}_K$;*

$$d_\sigma = \begin{cases} d_{K,\sigma} + d_{L,\sigma}, & K \in \mathcal{T},\ L \in \mathcal{N}(K),\ \sigma = K|L, \\ d_{K,\sigma}, & \sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K. \end{cases}$$

- *$\tau_\sigma = |\sigma|/d_\sigma$ denotes the transmissibility through $\sigma \in \mathcal{E}$.*

Assume there holds the regularity property: there exists a constant $C > 0$ such that

$$\sum_{\substack{L\in\mathcal{N}(K)\\ \sigma=K|L}} |\sigma| d_\sigma \leq C|K| \quad \textit{for all } K \in \mathcal{T}.$$

We introduce a dual grid $\mathcal{T}^\star$ with diamonds as elements. It will be used for the approximation of heterogeneous permeability fields. Additionally, it will be utilized within the proof.

Definition 5.2 (Dual grid to $\mathcal{T}$)**.** *Let $\mathcal{T}$ be an admissible mesh, cf. Definition 5.1. For each face $K|L \in \mathcal{E}$, $K \in \mathcal{T}$, $L \in \mathcal{N}(K)$, define a prism $P_{K|L} \subset \Omega$ with x_K, x_L and the vertices of $K|L$ as vertices. For all $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$, $K \in \mathcal{T}$ define $P_\sigma \subset \Omega$ to be the prism with x_K and the vertices of σ as vertices. By construction, $\mathcal{T}^\star := \{P_\sigma\}_{\sigma\in\mathcal{E}}$ defines a partition of Ω.*

Figure 1 displays a two-dimensional, admissible mesh and its auxiliary, dual grid.

The final discrete scheme is written in variational form. Given an admissible mesh $\mathcal{T}$, we introduce the discrete function spaces and implicitly their bases

$$\begin{aligned} \boldsymbol{V}_h &= \text{span}\,\{\boldsymbol{v}_{h,i}\}_{i\in\{1,\ldots,d_{\text{V}}\}}, \\ Q_h &= \text{span}\,\{q_{h,j}\}_{j\in\{1,\ldots,d_{\text{Q}}\}}, \end{aligned}$$

providing spaces for the discrete displacement and pressure, respectively. For the analysis below, we assume that the discrete function spaces to satisfy the following conditions:

(D1) Q_h is the space of all piecewise constant functions ($\mathbb{P}_0$) on $\mathcal{T}$ and the basis $\{q_{h,j}\}_j$ is equal to the indicator functions of all single elements. Note $Q_h \not\subset Q$.

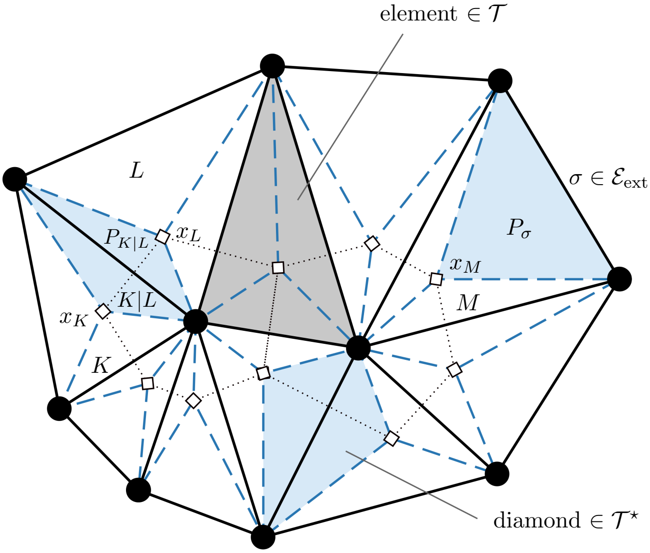

element $\in \mathcal{T}$
L
$\sigma \in \mathcal{E}_{\text{ext}}$
P_σ
$P_{K|L}$
x_L
x_M
$K|L$
M
x_K
K
diamond $\in \mathcal{T}^\star$

Figure 1: Admissible mesh $\mathcal{T}$ (consisting of elements) in two dimensions, together with the corresponding dual grid $\mathcal{T}^\star$ (consisting of diamonds).

(D2) $\boldsymbol{V}_h \subset \boldsymbol{V}$ such that $\boldsymbol{V}_h \times Q_h$ is inf-sup stable regarding the bilinear form

$$\boldsymbol{V}_h \times Q_h \to \mathbb{R}, \quad (\boldsymbol{v}_h, q_h) \mapsto \langle q_h, \boldsymbol{\nabla} \cdot \boldsymbol{v}_h \rangle.$$

In more detail, there exists a constant $\gamma_{\text{is}} = C_{\Omega,\text{is}}^{-1} > 0$ (independent of h), such that

$$\inf_{0 \neq q_h \in Q_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_h} \frac{\langle q_h, \boldsymbol{\nabla} \cdot \boldsymbol{v}_h \rangle}{\|q_h\| \, \|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \geq \gamma_{\text{is}}. \tag{5.1}$$

In the analysis, (D1) will allow for intuitively handling non-linearities in the pressure variable easily. Assumption (D2) will allow for using standard inf-sup arguments. In two dimensions, one can use piecewise quadratic elements for $\boldsymbol{V}_h$. In three dimensions, a practical choice is less trivial, cf. [66] for a thorough discussion.

In the analysis, we require the notion of a discrete $H^1(\Omega)$ norm for piecewise constant functions in Q_h, see also [62].

Definition 5.3 (Discrete $H^1(\Omega)$ norms on Q_h)**.** *Let $q_h \in Q_h$. We define*

$$\|q_h\|_{1,\mathcal{T}} := \left(\sum_{\sigma \in \mathcal{E}} \tau_\sigma \, \delta_\sigma(q_h)^2 \right)^{\frac{1}{2}},$$

where

$$\delta_\sigma q_h := \begin{cases} \left| q_h|_K - q_h|_L \right|, & K \in \mathcal{T}, \; L \in \mathcal{N}(K), \; \sigma = K|L, \\ \left| q_h|_K \right|, & \sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K. \end{cases}$$

In the same sense, given a uniformly positive field $\omega \in C(\Omega)$, a scaled inner product of discrete gradients is defined by

$$\langle \boldsymbol{\nabla}_h \chi_h, \boldsymbol{\nabla}_h q_h \rangle_\omega := \sum_{K \in \mathcal{T}} \sum_{L \in \mathcal{N}(K)} \tau_{K|L} \, \{\omega\}_{K|L} \, \delta_{K|L}(\chi_h) \, \delta_{K|L}(q_h) + \sum_{\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K} \tau_{K,\sigma} \, \{\omega\}_\sigma \, \chi_h|_K \, q_h|_K.$$

where the the weight ω evaluated at faces is approximated as weighted average incorporating the neighboring elements, i.e., utilizing the dual mesh $\mathcal{T}^\star$ to $\mathcal{T}$ it is

$$\{\omega\}_\sigma := \frac{1}{|P_\sigma|}\int_{P_\sigma} \omega(x)\,dx, \quad \sigma \in \mathcal{E}.$$

A norm $\|\cdot\|_{1,\mathcal{T},\omega} := \langle \boldsymbol{\nabla}_h \cdot, \boldsymbol{\nabla}_h \cdot\rangle_\omega^{1/2}$ is naturally induced.

A discrete Poincaré inequality can be showed for $\|\cdot\|_{1,\mathcal{T}}$, introducing a discrete Poincaré constant $C_{\Omega,\mathrm{DP}} > 0$ such that

$$\|q_h\| \le C_{\Omega,\mathrm{DP}} \|q_h\|_{1,\mathcal{T}} \qquad \text{for all } q_h \in Q_h,$$

cf. Lemma B.1; similarly also for $\|\cdot\|_{1,\mathcal{T},\omega}$.

5.2 Approximation of source terms and initial conditions

Let $0 = t_0 < t_1 < ... < t_N = T$ define a partition of the time interval $(0,T)$ with constant time step size $\tau = t_n - t_{n-1} = T/N$, $n, N \in \mathbb{N}$. We interpolate the source terms at discrete time steps. Let

$$\boldsymbol{f}_{\mathrm{ext}}^n := \frac{1}{\tau}\int_{t_{n-1}}^{t_n} \boldsymbol{f}_{\mathrm{ext}}(t)\,dt,$$
$$h_{\mathrm{ext}}^n := \frac{1}{\tau}\int_{t_{n-1}}^{t_n} h_{\mathrm{ext}}(t)\,dt.$$

Discrete initial conditions are chosen to imitate the compatibility assumption (A9). Let $\chi_h^0 \in Q_h$ be defined by the piecewise constant projection of χ^0, i.e., on $K \in \mathcal{T}$, we define

$$\chi_{h|_K}^0 := \frac{1}{|K|}\int_K \chi^0\,dx.$$

As $\chi_0 \in L^2(\Omega)$, cf. (A8*), it follows by classical approximation theory for $h \to 0$

$$\chi_h^0 \to \chi_0 \text{ in } L^2(\Omega),$$

and it holds that $\|\chi_h^0\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}} \le C\|\chi_0\|_1$ for some constant $C > 0$, cf., e.g., [64]. Furthermore, since $\hat{p}_{\mathrm{pore}} \in C(\mathbb{R})$, cf. (A3), and $\hat{p}_{\mathrm{pore}}(\chi_0) \in L^2(\Omega)$, cf. (A8*), it follows for $h \to 0$

$$\hat{p}_{\mathrm{pore}}(\chi_h^0) \to \hat{p}_{\mathrm{pore}}(\chi_0) \text{ in } L^2(\Omega),$$

similarly for $\{\hat{B}_\eta\left(\chi_h^0\right)\}_h$ and $\left\{\bar{B}_\eta\left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h^0)}{\hat{s}_{\mathrm{w}}(\chi_h^0)}\right)\right\}_h$. Then in order to satisfy (A9) in a discrete sense, we define $\boldsymbol{u}_h^0 \in \boldsymbol{V}_h$ to be the unique element in $\boldsymbol{V}_h$, satisfying

$$a(\boldsymbol{u}_h^0, \boldsymbol{v}_h) - \alpha\left\langle \hat{p}_{\mathrm{pore}}(\chi_h^0), \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\right\rangle = \left\langle \boldsymbol{f}_{\mathrm{ext}}(0), \boldsymbol{v}_h\right\rangle, \quad \text{for all } \boldsymbol{v}_h \in \boldsymbol{V}_h. \tag{5.2}$$

Using standard finite element techniques and (A9), it holds that

$$\|\boldsymbol{u}_0 - \boldsymbol{u}_h^0\|_{\boldsymbol{V}} \le 2\inf_{\boldsymbol{v}_h \in \boldsymbol{V}_h} \|\boldsymbol{u}_0 - \boldsymbol{v}_h\|_{\boldsymbol{V}} + \frac{\alpha}{K_{\mathrm{dr}}}\|\hat{p}_{\mathrm{pore}}(\chi_0) - \hat{p}_{\mathrm{pore}}(\chi_h^0)\|.$$

Hence, by classical approximation theory and the imposed regularity (A8*) it follows for $h \to 0$

$$\boldsymbol{u}_h^0 \to \boldsymbol{u}_0 \text{ in } \boldsymbol{V}.$$

All in all, due to the convergence, (A8*) also applies on discrete level.

(A8*)$_h$ For bounded $\eta > 0$, there exists a constant $C_0 > 0$ (wlog. the same as in (A8)) such that

$$\|\boldsymbol{u}_h^0\|_{\boldsymbol{V}}^2 + \|\chi_h^0\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 + \left\|\hat{B}_\eta\left(\chi_h^0\right)\right\|_{L^1(\Omega)} + \left\|\bar{B}_\eta\left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h^0)}{\hat{s}_{\mathrm{w}}(\chi_h^0)}\right)\right\|_{L^1(\Omega)}. \le C_0$$

5.3 Approximation of the evolutionary problem

The discretization of (4.3)–(4.4) is defined by the Galerkin method combined with the standard implicit Euler time discretization: for $n \geq 1$, given $(\boldsymbol{u}_h^{n-1}, \chi_h^{n-1}) \in \boldsymbol{V}_h \times Q_h$, find $(\boldsymbol{u}_h^n, \chi_h^n) \in \boldsymbol{V}_h \times Q_h$ satisfying for all $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h \times Q_h$

$$\zeta\tau^{-1} a(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) + a(\boldsymbol{u}_h^n, \boldsymbol{v}_h) - \alpha\langle \hat{p}_{\mathrm{pore}}(\chi_h^n), \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\rangle \;=\; \langle \boldsymbol{f}_{\mathrm{ext}}^n, \boldsymbol{v}_h\rangle, \tag{5.3}$$

$$\begin{aligned}\langle \hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^{n-1}), q_h\rangle + \alpha\langle \hat{s}_{\mathrm{w}}(\chi_h^n)\boldsymbol{\nabla}\cdot(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}), q_h\rangle \\ + \tau\langle \boldsymbol{\nabla}_h\chi_h^n, \boldsymbol{\nabla}_h q_h\rangle_{\kappa_{\mathrm{abs}}} \;=\; \tau\langle h_{\mathrm{ext}}^n, q_h\rangle.\end{aligned} \tag{5.4}$$

Lemma 5.4 (Existence of a discrete solution)**.** *Let* $n \geq 1$. (A0)–(A9), (ND1)*, and* (D1)–(D2) *hold true. Then there exists a discrete solution* $(\boldsymbol{u}_h^n, \chi_h^n) \in \boldsymbol{V}_h \times Q_h$ *satisfying* (5.3)–(5.4)*, and*

$$\left\| \bar{B}_\eta\left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h^n)}{\hat{s}_{\mathrm{w}}\left(\chi_h^n\right)}\right)\right\|_{L^1(\Omega)} + \|\boldsymbol{u}_h^n\|_{\boldsymbol{V}}^2 < \infty \quad \textit{for all } n \geq 1. \tag{5.5}$$

Proof. The proof is by induction. We present only the general step, since the proof for $n = 1$ is similar. We employ a corollary of Brouwer's fixed point theorem, cf. Lemma B.4, to show the existence of a solution of a non-linear algebraic system, which is equivalent to (5.3)–(5.4).

Introduction of a pressure-reduced algebraic problem. We introduce an isomorphism between the discrete function space corresponding to the fluid pressure χ and a suitable coefficient vector space

$$\chi_h : \mathbb{R}^{d_{\mathrm{Q}}} \to Q_h, \;\; \boldsymbol{\beta} \mapsto \sum_j \left(\tfrac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}}\right)^{-1}(\beta_j)\, q_{h,j}.$$

Due to (A4), χ_h is well-defined. Similarly, let

$$\boldsymbol{u}_h : \mathbb{R}^{d_{\mathrm{V}}} \to \boldsymbol{V}_h, \;\; \boldsymbol{\alpha} \mapsto \sum_i \alpha_i \boldsymbol{v}_{h,i}.$$

For given $\boldsymbol{\beta} \in \mathbb{R}^{d_{\mathrm{Q}}}$, define $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\boldsymbol{\beta}) \in \mathbb{R}^{d_{\mathrm{V}}}$ to be the unique solution to: find $\boldsymbol{\alpha} \in \mathbb{R}^{d_{\mathrm{V}}}$ such that

$$\begin{aligned}&\zeta\tau^{-1} a(\boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) + a(\boldsymbol{u}_h(\boldsymbol{\alpha}), \boldsymbol{v}_h) \\ &\quad = \langle \boldsymbol{f}^n, \boldsymbol{v}_h\rangle + \alpha\left\langle \hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta})), \boldsymbol{\nabla}\cdot\boldsymbol{v}_h\right\rangle, \text{ for all } \boldsymbol{v}_h \in \boldsymbol{V}_h.\end{aligned}$$

Finally, we define $\boldsymbol{F} : \mathbb{R}^{d_{\mathrm{Q}}} \to \mathbb{R}^{d_{\mathrm{Q}}}$ by

$$\begin{aligned}F_j(\boldsymbol{\beta}) = &\left\langle \hat{b}_\eta(\chi_h(\boldsymbol{\beta})) - \hat{b}_\eta(\chi_h^{n-1}), q_{h,j}\right\rangle + \alpha\left\langle \hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))\boldsymbol{\nabla}\cdot(\boldsymbol{u}_h(\boldsymbol{\alpha}(\boldsymbol{\beta})) - \boldsymbol{u}_h^{n-1}), q_{h,j}\right\rangle \\ &+ \tau\left\langle \boldsymbol{\nabla}_h\chi_h(\boldsymbol{\beta}), \boldsymbol{\nabla}_h q_{h,j}\right\rangle_{\kappa_{\mathrm{abs}}} - \tau\left\langle h_{\mathrm{ext}}^n, q_{h,j}\right\rangle, \quad j \in \{1, ..., d_{\mathrm{Q}}\}.\end{aligned}$$

We note, the existence of a discrete solution of Eq. (5.3)–(5.4) is equivalent to the existence of $\boldsymbol{\beta} \in \mathbb{R}^{d_{\mathrm{Q}}}$, satisfying $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{0}$. To prove the existence of a zero of $\boldsymbol{F}$, we employ Lemma B.4; we consider the expression

$$\begin{aligned}\langle \boldsymbol{F}(\boldsymbol{\beta}), \boldsymbol{\beta}\rangle = &\left\langle \hat{b}_\eta(\chi_h(\boldsymbol{\beta})) - \hat{b}_\eta(\chi_h^{n-1}), \frac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))}\right\rangle \\ &+ \alpha\left\langle \boldsymbol{\nabla}\cdot(\boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1}), \hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))\right\rangle \\ &+ \tau\left\langle \boldsymbol{\nabla}_h\chi(\boldsymbol{\beta}), \boldsymbol{\nabla}_h\tfrac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))}\right\rangle \\ &- \tau\left\langle h_{\mathrm{ext}}^n, \tfrac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))}\right\rangle \\ =: &\; T_1 + T_2 + T_3 + T_4.\end{aligned} \tag{5.6}$$

where we used

$$\sum_{j=1}^{d_{\mathrm{Q}}} \beta_j q_{h,j} = \frac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))}.$$

and dropped the explicit dependence of $\boldsymbol{\alpha}$ on $\boldsymbol{\beta}$. We discuss the terms $T_1, ..., T_4$ separately.

Discussion of T_1. Using (A4), we define $\bar{b}_\eta := \hat{b}_\eta \circ \left(\frac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}}\right)^{-1} : \mathbb{R} \to \mathbb{R}$ and its Legendre transformation, cf. (4.2). Finally, using standard properties of the Legendre transformation of non-decreasing functions, cf. Lemma B.12, we obtain for term T_1

$$T_1 \geq \left\| \bar{B}_\eta \left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))} \right) \right\|_{L^1(\Omega)} - \left\| \bar{B}_\eta \left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}^{n-1}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}^{n-1}))} \right) \right\|_{L^1(\Omega)},$$

where $\boldsymbol{\beta}^{n-1} \in \mathbb{R}^{d_{\mathrm{Q}}}$ such that $\chi_h^{n-1} = \chi_h(\boldsymbol{\beta}^{n-1})$.

Discussion of T_2. From the definition of $\boldsymbol{\alpha}$, under the use of a binomial identity, the Cauchy-Schwarz inequality and Young's inequality, the coupling term T_2 becomes

$$\begin{aligned}
T_2 &= \alpha \left\langle \boldsymbol{\nabla} \cdot (\boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1}), \hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta})) \right\rangle \\
&= \zeta \tau^{-1} \left\| \boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2 + \tfrac{1}{2} \left\| \boldsymbol{u}_h(\boldsymbol{\alpha}) \right\|_{\boldsymbol{V}}^2 + \tfrac{1}{2} \left\| \boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2 \\
&\quad - \tfrac{1}{2} \left\| \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2 - \left\langle \boldsymbol{f}^n, \boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1} \right\rangle \\
&\geq \zeta \tau^{-1} \left\| \boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2 + \tfrac{1}{2} \left\| \boldsymbol{u}_h(\boldsymbol{\alpha}) \right\|_{\boldsymbol{V}}^2 + \tfrac{1}{4} \left\| \boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2 \\
&\quad - \tfrac{1}{2} \left\| \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2 - \left\| \boldsymbol{f}^n \right\|_{\boldsymbol{V}^\star}^2.
\end{aligned}$$

Discussion of T_3. By the mean value theorem and (A4), the diffusion term T_3 can be estimated from below

$$T_3 \geq c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}} \tau \|\chi_h(\boldsymbol{\beta})\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2.$$

Discussion of T_4. Employing the definition of $h_{\mathrm{ext}} = (h, w_{\mathrm{N}})$, the non-degeneracy condition (ND1), a discrete trace inequality, cf. Lemma B.2, together with a discrete Poincaré inequality (introducing $C_{\Omega,\mathrm{DP}}$), cf. Lemma B.1, we obtain

$$\begin{aligned}
&\left\langle h_{\mathrm{ext}}^n, \tfrac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))} \right\rangle \\
&\quad \leq \left\| \tfrac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))\chi_h(\boldsymbol{\beta})} \right\|_\infty \left(\|h^n\| \, \|\chi_h(\boldsymbol{\beta})\| + \|w_{\mathrm{N}}^n\|_{L^2(\Gamma_{\mathrm{N}}^{\mathrm{f}})} \, \|\chi_h(\boldsymbol{\beta})\|_{L^2(\Gamma_{\mathrm{N}}^{\mathrm{f}})} \right) \\
&\quad \leq C\left(C_{\mathrm{ND},1}, C_{\mathrm{tr}}, C_{\Omega,\mathrm{DP}}\right) \|h_{\mathrm{ext}}^n\|_{L^2(\Omega)\times L^2(\Gamma_{\mathrm{N}}^{\mathrm{f}})} \, \|\chi_h(\boldsymbol{\beta})\|_{1,\mathcal{T}}
\end{aligned}$$

for a constant $C\left(C_{\mathrm{ND},1}, C_{\mathrm{tr}}, C_{\Omega,\mathrm{DP}}\right) > 0$ Hence, by (A6) and Young's inequality, for the term T_4 it holds that

$$T_4 \leq \frac{C\left(C_{\mathrm{ND},1}, C_{\mathrm{tr}}, C_{\Omega,\mathrm{DP}}\right)^2}{2 c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}} \kappa_{\mathrm{m,abs}}} \tau \left\| h_{\mathrm{ext}}^n \right\|_{L^2(\Omega)\times L^2(\Gamma_{\mathrm{N}}^{\mathrm{f}})}^2 + \frac{c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}}}{2} \tau \|\chi_h(\boldsymbol{\beta})\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2.$$

Combination of all results. By inserting the estimates for T_1, T_2, T_3, and T_4, (5.6) becomes

$$\begin{aligned}\langle \boldsymbol{F}(\boldsymbol{\beta}), \boldsymbol{\beta} \rangle \geq \Bigg(& \left\| \bar{B}_\eta \left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))}{\hat{s}_{\mathrm{w}}(\chi_h(\boldsymbol{\beta}))} \right) \right\|_{L^1(\Omega)} + \frac{c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}}}{2} \tau \|\chi_h(\boldsymbol{\beta})\|^2_{1,\mathcal{T},\kappa_{\mathrm{abs}}} \\ & + \frac{1}{4} \|\boldsymbol{u}_h(\boldsymbol{\alpha})\|^2_{\boldsymbol{V}} + \frac{1}{2}\zeta\tau^{-1} \left\|\boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1}\right\|^2_{\boldsymbol{V}} + \frac{1}{4} \left\|\boldsymbol{u}_h(\boldsymbol{\alpha}) - \boldsymbol{u}_h^{n-1}\right\|^2_{\boldsymbol{V}} \Bigg) \\ & - \Bigg(\left\| \bar{B}_\eta \left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h^{n-1})}{\hat{s}_{\mathrm{w}}(\chi_h^{n-1})} \right) \right\|_{L^1(\Omega)} + \frac{1}{2} \left\|\boldsymbol{u}_h^{n-1}\right\|^2_{\boldsymbol{V}} + \frac{5}{4} \|\boldsymbol{f}^n\|^2_{\boldsymbol{V}^\star} \\ & + \frac{C\left(C_{\mathrm{ND},1}, C_{\mathrm{tr}}, C_{\Omega,\mathrm{DP}}\right)^2}{2 c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}} \kappa_{\mathrm{m,abs}}} \tau \left\|h^n_{\mathrm{ext}}\right\|_{L^2(\Omega)\times L^2(\Gamma_{\mathrm{N}}^{\mathrm{f}})} \Bigg). \end{aligned} \tag{5.7}$$

Finally, since $\| \cdot \|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}$ defines a norm on Q_h and (5.5) holds by induction for $n-1$ if $n \geq 2$ or from (A8*) for $n = 1$, by a corollary of Brouwer's fixed point theorem, cf. Lemma B.4, there exists a $\boldsymbol{\beta} \in \mathbb{R}^{d_{\mathrm{Q}}}$ such that $\boldsymbol{F}(\boldsymbol{\beta}) = \boldsymbol{0}$, which implies existence of a solution. The bound (5.5) for n follows immediately from (5.7). □

6 Step 3: Limit $h, \tau \to 0$

In the following, we show that the fully-discrete FEM-TPFA discretization, introduced in the previous section, converges to a weak solution of the doubly regularized unsaturated poroelasticity model, i.e., we prove Lemma 4.3. The proof follows the steps: 1) derive stability results for the fully discrete approximation; 2) define suitable approximations a.e. in time using interpolation; 3) deduce stability for those as well; 4) relative compactness arguments are performed yielding a well-defined limit for $h, \tau \to 0$; 5) the limit is showed to be a weak solution of the doubly regularized model. Throughout the entire section, we assume (A0)–(A9) and (ND1) hold true.

6.1 Stability estimates for the fully-discrete approximation

Lemma 6.1 (Stability estimate for the primary variables)**.** *Let $\tau < \frac{1}{8}$. There exists a constant $C^{(1)} > 0$ (independent of h, τ, ζ, η), such that*

$$\begin{aligned} \zeta \sum_n \tau^{-1} \left\|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\|^2_{\boldsymbol{V}} + \sup_n \|\boldsymbol{u}_h^n\|^2_{\boldsymbol{V}} + \sum_n \|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\|^2_{\boldsymbol{V}} + \sum_{n=1}^{N} \tau \|\chi_h^n\|^2_{1,\mathcal{T}} \\ \leq C^{(1)} \Big(C_0, C_{\mathrm{ND},1} \|h_{\mathrm{ext}}\|_{L^2(0,T;Q^\star)}, \|\boldsymbol{f}_{\mathrm{ext}}\|_{H^1(0,T;\boldsymbol{V}^\star)} \Big), \end{aligned}$$

where C_0 and $C_{\mathrm{ND},1}$ are defined in $(\mathrm{A8}^\star)_h$ *and* (ND1)*, respectively.*

Proof. The proof follows essentially the same steps as in the proof of Lemma 5.4. Therefore, we are quick on similar steps. We consider the reduced displacement-pressure formulation (5.3)–(5.4). We choose $\boldsymbol{v}_h = \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}$ and $q_h = \frac{\hat{p}_{\mathrm{pore}}(\chi_h^n)}{\hat{s}_{\mathrm{w}}(\chi_h^n)}$ as test functions and sum the two equations; note that the second is well-defined as $\hat{s}_{\mathrm{w}}(\chi) > 0$ for all $\chi \in \mathbb{R}$, by (A2). We obtain

$$\begin{aligned} & \zeta\tau^{-1} \left\|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\|^2_{\boldsymbol{V}} + a(\boldsymbol{u}_h^n, \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}) \\ & \quad + \left\langle \hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^{n-1}), \frac{\hat{p}_{\mathrm{pore}}(\chi_h^n)}{\hat{s}_{\mathrm{w}}(\chi_h^n)} \right\rangle + \tau \left\langle \boldsymbol{\nabla}_h \chi_h^n, \boldsymbol{\nabla}_h \frac{\hat{p}_{\mathrm{pore}}(\chi_h^n)}{\hat{s}_{\mathrm{w}}(\chi_h^n)} \right\rangle_{\kappa_{\mathrm{abs}}} \\ & \quad = \left\langle \boldsymbol{f}^n_{\mathrm{ext}}, \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\rangle + \tau \left\langle h^n_{\mathrm{ext}}, \frac{\hat{p}_{\mathrm{pore}}(\chi_h^n)}{\hat{s}_{\mathrm{w}}(\chi_h^n)} \right\rangle. \end{aligned}$$

On the left hand side, we employ the binomial identity (B.2), the Legendre transformation, $\bar{B}_\eta$, of $\bar{b}_\eta = \hat{b}_\eta \circ \left(\frac{\hat{p}_\mathrm{pore}}{\hat{s}_\mathrm{w}}\right)^{-1}$, cf. (4.2) and Lemma B.12, and the uniform increase of $\frac{\hat{p}_\mathrm{pore}}{\hat{s}_\mathrm{w}}$, cf., (A4). It holds that

$$\begin{aligned}
&\zeta\tau^{-1}\left\|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\|_{\boldsymbol{V}}^2 + \frac{1}{2}\left(\|\boldsymbol{u}_h^n\|_{\boldsymbol{V}}^2 - \|\boldsymbol{u}_h^{n-1}\|_{\boldsymbol{V}}^2 + \|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\|_{\boldsymbol{V}}^2\right) \\
&\quad + \left\|\bar{B}_\eta\left(\frac{\hat{p}_\mathrm{pore}(\chi_h^n)}{\hat{s}_\mathrm{w}(\chi_h^n)}\right)\right\|_{L^1(\Omega)} - \left\|\bar{B}_\eta\left(\frac{\hat{p}_\mathrm{pore}(\chi_h^{n-1})}{\hat{s}_\mathrm{w}(\chi_h^{n-1})}\right)\right\|_{L^1(\Omega)} + c_{\hat{p}_\mathrm{pore}/\hat{s}_\mathrm{w}}\tau\|\chi_h^n\|_{1,\mathcal{T},\kappa_\mathrm{abs}}^2 \\
&\quad \leq \left\langle \boldsymbol{f}_\mathrm{ext}^n, \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\rangle + \left\langle h_\mathrm{ext}^n, \frac{\hat{p}_\mathrm{pore}(\chi_h^n)}{\hat{s}_\mathrm{w}(\chi_h^n)}\right\rangle.
\end{aligned}$$

Summing over the time steps 1 to N and rearranging terms, yields

$$\begin{aligned}
&\zeta\sum_n \tau^{-1}\left\|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\|_{\boldsymbol{V}}^2 + \frac{1}{2}\|\boldsymbol{u}_h^N\|_{\boldsymbol{V}}^2 + \frac{1}{2}\sum_{n=1}^N \|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\|_{\boldsymbol{V}}^2 \\
&\quad + \left\|\bar{B}_\eta\left(\frac{\hat{p}_\mathrm{pore}(\chi_h^N)}{\hat{s}_\mathrm{w}(\chi_h^N)}\right)\right\|_{L^1(\Omega)} + c_{\hat{p}_\mathrm{pore}/\hat{s}_\mathrm{w}}\sum_{n=1}^N \tau\|\chi_h^n\|_{1,\mathcal{T},\kappa_\mathrm{abs}}^2 \\
&\quad \leq \frac{1}{2}\|\boldsymbol{u}_h^0\|_{\boldsymbol{V}}^2 + \left\|\bar{B}_\eta\left(\frac{\hat{p}_\mathrm{pore}(\chi_h^0)}{\hat{s}_\mathrm{w}(\chi_h^0)}\right)\right\|_{L^1(\Omega)} \\
&\quad + \sum_{n=1}^N \left\langle \boldsymbol{f}_\mathrm{ext}^n, \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\rangle + \sum_{n=1}^N \tau\left\langle h_\mathrm{ext}^n, \frac{\hat{p}_\mathrm{pore}(\chi_h^n)}{\hat{s}_\mathrm{w}(\chi_h^n)}\right\rangle.
\end{aligned}$$

It remains to discuss the last two terms on the right hand side. For the first of them, we employ summation by parts, cf. Lemma B.6, as well as the Cauchy-Schwarz inequality and Young's inequality:

$$\begin{aligned}
&\sum_{n=1}^N \left\langle \boldsymbol{f}_\mathrm{ext}^n, \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\rangle \\
&\quad = \left\langle \boldsymbol{f}_\mathrm{ext}^N, \boldsymbol{u}_h^N\right\rangle - \left\langle \boldsymbol{f}_\mathrm{ext}^1, \boldsymbol{u}_h^0\right\rangle - \sum_{n=1}^{N-1}\left\langle \boldsymbol{f}_\mathrm{ext}^{n+1} - \boldsymbol{f}_\mathrm{ext}^n, \boldsymbol{u}_h^n\right\rangle \\
&\quad \leq \|\boldsymbol{f}_\mathrm{ext}^N\|_{\boldsymbol{V}^\star}^2 + \frac{1}{4}\|\boldsymbol{u}_h^N\|_{\boldsymbol{V}}^2 + \frac{1}{2}\|\boldsymbol{f}_\mathrm{ext}^1\|_{\boldsymbol{V}^\star}^2 + \frac{1}{2}\|\boldsymbol{u}_h^0\|_{\boldsymbol{V}}^2 + \sum_{n=1}^N \tau^{-1}\left\|\boldsymbol{f}_\mathrm{ext}^n - \boldsymbol{f}_\mathrm{ext}^{n-1}\right\|_{\boldsymbol{V}^\star}^2 + \sum_{n=1}^N \tau\|\boldsymbol{u}_h^n\|_{\boldsymbol{V}}^2.
\end{aligned}$$

The second term is estimated as in the discussion of T_4 within the proof of Lemma 5.4. We obtain

$$\begin{aligned}
&\sum_{n=1}^N \tau\left\langle h_\mathrm{ext}^n, \frac{\hat{p}_\mathrm{pore}(\chi_h^n)}{\hat{s}_\mathrm{w}(\chi_h^n)}\right\rangle \\
&\quad \leq \frac{C\left(C_{\mathrm{ND},1}, C_\mathrm{tr}, C_{\Omega,\mathrm{DP}}\right)^2}{2c_{\hat{p}_\mathrm{pore}/\hat{s}_\mathrm{w}}\kappa_\mathrm{m,abs}}\sum_{n=1}^N \tau\left\|h_\mathrm{ext}^n\right\|_{L^2(\Omega)^2\times L^2(\Gamma_\mathrm{N}^\mathrm{f})} + \frac{c_{\hat{p}_\mathrm{pore}/\hat{s}_\mathrm{w}}}{2}\sum_{n=1}^N \tau\|\chi_h^n\|_{1,\mathcal{T},\kappa_\mathrm{abs}}^2.
\end{aligned}$$

Altogether, after rearranging terms, we obtain

$$\begin{aligned}
&\frac{\zeta}{2}\sum_n \tau^{-1}\left\|\boldsymbol{u}_h^n-\boldsymbol{u}_h^{n-1}\right\|_{\boldsymbol{V}}^2+\frac{1}{4}\|\boldsymbol{u}_h^N\|_{\boldsymbol{V}}^2+\frac{1}{4}\sum_{n=1}^N\|\boldsymbol{u}_h^n-\boldsymbol{u}_h^{n-1}\|_{\boldsymbol{V}}^2\\
&\quad+\left\|\bar{B}_\eta\left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h^N)}{\hat{s}_{\mathrm{w}}(\chi_h^N)}\right)\right\|_{L^1(\Omega)}+\frac{c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}}}{2}\sum_{n=1}^N\tau\|\chi_h^n\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2\\
&\quad\le\|\boldsymbol{u}_h^0\|_{\boldsymbol{V}}^2+\left\|\bar{B}_\eta\left(\frac{\hat{p}_{\mathrm{pore}}(\chi_h^0)}{\hat{s}_{\mathrm{w}}(\chi_h^0)}\right)\right\|_{L^1(\Omega)}+\frac{C\left(C_{\mathrm{ND},1},C_{\mathrm{tr}},C_{\Omega,\mathrm{DP}}\right)^2}{2c_{\hat{p}_{\mathrm{pore}}/\hat{s}_{\mathrm{w}}}\kappa_{\mathrm{m,abs}}}\sum_{n=1}^N\tau\left\|h_{\mathrm{ext}}^n\right\|_{L^2(\Omega)\times L^2(\Gamma_{\mathrm{N}}^{\mathrm{f}})}^2\\
&\quad+\|\boldsymbol{f}_{\mathrm{ext}}^N\|_{\boldsymbol{V}^\star}^2+\|\boldsymbol{f}_{\mathrm{ext}}^1\|_{\boldsymbol{V}^\star}^2+\sum_{n=1}^N\tau^{-1}\left\|\boldsymbol{f}_{\mathrm{ext}}^n-\boldsymbol{f}_{\mathrm{ext}}^{n-1}\right\|_{\boldsymbol{V}^\star}^2+\sum_{n=1}^N\tau\|\boldsymbol{f}_{\mathrm{ext}}^n\|_{\boldsymbol{V}^\star}^2+2\sum_{n=1}^N\tau\|\boldsymbol{u}_h^n\|_{\boldsymbol{V}}^2.
\end{aligned}$$

Finally, the last term on the right hand side can be controlled after applying a discrete Grönwall inequality, cf. Lemma B.7, using that $2\tau<\frac{1}{4}$. The thesis follows from the assumptions on the regularity of the source terms (A7) (together with a Sobolev embedding) and initial data (A8*). □

Lemma 6.2 (Stability for the Kirchhoff pressure)**.** *There exists a constant* $C_{\zeta\eta}^{(2)}>0$ *(independent of* h,τ*) such that*

$$b_{\chi,m}\sum_{n=1}^N\tau^{-1}\left\|\chi_h^n-\chi_h^{n-1}\right\|^2+\left\|\chi_h^N\right\|_{1,\mathcal{T}}^2+\sum_{n=1}^N\left\|\chi_h^n-\chi_h^{n-1}\right\|_{1,\mathcal{T}}^2\le C_{\zeta\eta}^{(2)}\left(C_0,\frac{1+\zeta^{-1}}{b_{\chi,\mathrm{m}}}C^{(1)}\right),$$

where $C^{(1)}$ *is the stability constant from Lemma 6.1,* $b_{\chi,\mathrm{m}}$ *is from the growing condition* (A1*)*, and* C_0 *is the bound in* $(\mathrm{A8}^\star)_h$*.*

Proof. We choose $q_h=\chi_h^n-\chi_h^{n-1}$ in (5.4). By using the binomial identity (B.2) for the diffusion term, we obtain

$$\begin{aligned}
&\left\langle\hat{b}_\eta(\chi_h^n)-\hat{b}_\eta(\chi_h^{n-1}),\chi_h^n-\chi_h^{n-1}\right\rangle+\alpha\left\langle\hat{s}_{\mathrm{w}}(\chi_h^n)\boldsymbol{\nabla}\cdot(\boldsymbol{u}_h^n-\boldsymbol{u}_h^{n-1}),\chi_h^n-\chi_h^{n-1}\right\rangle\\
&\quad+\frac{\tau}{2}\left(\|\chi_h^n\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2-\left\|\chi_h^{n-1}\right\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2+\left\|\chi_h^n-\chi_h^{n-1}\right\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2\right)\\
&\quad=\tau\left\langle h_{\mathrm{ext}}^n,\chi_h^n-\chi_h^{n-1}\right\rangle.
\end{aligned}$$

Dividing by τ and summing over time steps 1 to N, yields

$$\begin{aligned}
&\sum_{n=1}^N\tau^{-1}\left\langle\hat{b}_\eta(\chi_h^n)-\hat{b}_\eta(\chi_h^{n-1}),\chi_h^n-\chi_h^{n-1}\right\rangle+\frac{1}{2}\left\|\chi_h^N\right\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2+\frac{1}{2}\sum_{n=1}^N\left\|\chi_h^n-\chi_h^{n-1}\right\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 \qquad (6.1)\\
&\quad=\frac{1}{2}\left\|\chi_h^0\right\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2+\sum_{n=1}^N\left\langle h_{\mathrm{ext}}^n,\chi_h^n-\chi_h^{n-1}\right\rangle-\alpha\sum_{n=1}^N\tau^{-1}\left\langle\hat{s}_{\mathrm{w}}(\chi_h^n)\boldsymbol{\nabla}\cdot(\boldsymbol{u}_h^n-\boldsymbol{u}_h^{n-1}),\chi_h^n-\chi_h^{n-1}\right\rangle.
\end{aligned}$$

We discuss some of the terms above separately. Employing the growth condition (A1*), yields for the first term on the left hand side of (6.1)

$$\sum_{n=1}^N\tau^{-1}\left\langle\hat{b}_\eta(\chi_h^n)-\hat{b}_\eta(\chi_h^{n-1}),\chi_h^n-\chi_h^{n-1}\right\rangle\ge b_{\chi,\mathrm{m}}\sum_{n=1}^N\tau^{-1}\left\|\chi_h^n-\chi_h^{n-1}\right\|^2.$$

By employing the Cauchy-Schwarz inequality and Young's inequality, we get for the second term on the right hand side of (6.1)

$$\sum_{n=1}^N\left\langle h_{\mathrm{ext}}^n,\chi_h^n-\chi_h^{n-1}\right\rangle\le\frac{b_{\chi,\mathrm{m}}}{2}\sum_{n=1}^N\tau^{-1}\left\|\chi_h^n-\chi_h^{n-1}\right\|^2+\frac{1}{2b_{\chi,\mathrm{m}}}\sum_{n=1}^N\tau\left\|h_{\mathrm{ext}}^n\right\|_{Q^\star}^2.$$

Similarly, for the last term on the right hand side of (6.1), we get

$$\alpha \sum_{n=1}^{N} \tau^{-1} \langle \hat{s}_{\mathrm{w}}(\chi_h^n) \boldsymbol{\nabla} \cdot (\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}), \chi_h^n - \chi_h^{n-1} \rangle \\ \leq \frac{b_{\chi,\mathrm{m}}}{4} \sum_{n=1}^{N} \tau^{-1} \|\chi_h^n - \chi_h^{n-1}\|^2 + \frac{\alpha^2}{b_{\chi,\mathrm{m}}} \sum_{n=1}^{N} \tau^{-1} \|\boldsymbol{\nabla} \cdot (\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1})\|^2.$$

All in all, (6.1) becomes

$$\frac{b_{\chi,m}}{4} \sum_{n=1}^{N} \tau^{-1} \|\chi_h^n - \chi_h^{n-1}\|^2 + \frac{1}{2} \|\chi_h^N\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 + \frac{1}{2} \sum_{n=1}^{N} \|\chi_h^n - \chi_h^{n-1}\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 \\ \leq \frac{1}{2} \|\chi_h^0\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 + \frac{1}{2b_{\chi,\mathrm{m}}} \sum_{n=1}^{N} \tau \|h_{\mathrm{ext}}^n\|_{Q^\star}^2 + \frac{\alpha^2}{b_{\chi,\mathrm{m}}} \sum_{n=1}^{N} \tau^{-1} \|\boldsymbol{\nabla} \cdot (\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1})\|^2.$$

Finally, the first term on the right hand side is bounded by $(\mathrm{A8}^\star)_h$, whereas the last term can be bounded by employing Lemma 6.1. On the left hand side, we employ (A6). □

Lemma 6.3 (Stability for the Legendre transformation $\hat{B}_\eta$ of $\hat{b}_\eta$)**.** *Let $\hat{B}_\eta(z)$ denote the Legendre transformation of $\hat{b}$, cf. (4.1). There exists a constant $C_\zeta^{(3)} > 0$ (independent of h, τ, η), such that*

$$\sup_n \left\| \hat{B}_\eta(\chi_h^n) \right\|_{L^1(\Omega)} \leq C_\zeta^{(3)} \left(C_0, C^{(1)} \left(1 + \zeta^{-1}\right) \right),$$

where $C^{(1)}$ is the stability constant from Lemma 6.1, and C_0 is the bound in $(\mathrm{A8}^\star)_h$.

Proof. Testing (5.4) with $q_h = \chi_h^n$ and employing the properties of the Legendre transformation $\hat{B}_\eta$, cf. Lemma B.12, yields for all n

$$\left\| \hat{B}_\eta(\chi_h^n) \right\|_{L^1(\Omega)} - \left\| \hat{B}_\eta(\chi_h^{n-1}) \right\|_{L^1(\Omega)} + \tau \left\| \chi_h^n \right\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 \\ \leq \tau \left\langle h_{\mathrm{ext}}^n, \chi_h^n \right\rangle - \alpha \left\langle \hat{s}_{\mathrm{w}}(\chi_h^n) \boldsymbol{\nabla} \cdot (\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}), \chi_h^n \right\rangle.$$

For the first term on the right hand side, we employ a similar bound as in the discussion of T_4 within the proof of Lemma 5.4; for the second term, we employ the Cauchy-Schwarz inequality, a discrete Poincaré inequality (introducing $C_{\Omega,\mathrm{DP}}$), and (A6). We obtain

$$\left\| \hat{B}_\eta(\chi_h^n) \right\|_{L^1(\Omega)} - \left\| \hat{B}_\eta(\chi_h^{n-1}) \right\|_{L^1(\Omega)} + \frac{\tau}{2} \|\chi_h^n\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 \\ \leq \frac{C\left(C_{\mathrm{ND},1}, C_{\mathrm{tr}}, C_{\Omega,\mathrm{DP}}\right)^2}{\kappa_{\mathrm{m,abs}}} \tau \|h_{\mathrm{ext}}^n\|_{Q^\star}^2 + \frac{C_{\Omega,\mathrm{DP}}}{\kappa_{\mathrm{m,abs}}} \frac{\alpha^2}{K_{\mathrm{dr}}} \tau^{-1} \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2.$$

Finally, summing over time steps 1 to N and employing Lemma 6.1 and (A7) proves the assertion. □

Lemma 6.4 (Stability for the pore pressure)**.** *There exists a constant $C^{(4)} > 0$ (independent of h, τ, ζ, η), such that*

$$\sum_{n=1}^{N} \tau \|\hat{p}_{\mathrm{pore}}(\chi_h^n)\|^2 \leq C^{(4)} \left(C^{(1)} \right),$$

where $C^{(1)}$ is the stability constant from Lemma 6.1.

Proof. We utilize a standard inf-sup argument (introducing $C_{\Omega,\mathrm{is}}$), cf. Lemma B.11. Due to (D2), there exists a $\boldsymbol{v}_h \in \boldsymbol{V}_h$ such that

$$\|\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))\|^2 = \alpha \left\langle \hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta})), \boldsymbol{\nabla} \cdot \boldsymbol{v}_h \right\rangle, \qquad \|\boldsymbol{v}_h\|_{\boldsymbol{V}} \leq C_{\Omega,\mathrm{is}} \|\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))\|,$$

Employing the mechanics equation (4.3), we obtain

$$\|\hat{p}_{\mathrm{pore}}(\chi_h(\boldsymbol{\beta}))\| \leq C_{\Omega,\mathrm{is}} \left(\zeta \tau^{-1} \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}} + \|\boldsymbol{u}_h^n\|_{\boldsymbol{V}} + \|\boldsymbol{f}_{\mathrm{ext}}^n\|_{\boldsymbol{V}^\star} \right),$$

and hence,

$$\sum_{n=1}^N \tau \|\hat{p}_{\mathrm{pore}}(\chi_h^n)\|^2 \leq 3C_{\Omega,\mathrm{is}}^2 \left(\zeta^2 \sum_{n=1}^N \tau^{-1} \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2 + \sum_{n=1}^N \tau \|\boldsymbol{u}_h^n\|_{\boldsymbol{V}}^2 + \sum_{n=1}^N \tau \|\boldsymbol{f}_{\mathrm{ext}}^n\|_{\boldsymbol{V}^\star}^2 \right).$$

Finally, the assertion follows from Lemma 6.1, assuming wlog. ζ is bounded from above. □

Lemma 6.5 (Stability for the temporal change of $\hat{b}$)**.** *There exists a constant $C_\zeta^{(5)} > 0$ (independent of h, τ, η), such that*

$$\sup_{\{q_h^n\}_n \subset Q_h \setminus \{0\}} \frac{\sum_{n=1}^N \tau \left\langle \frac{\hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^{n-1})}{\tau}, q_h^n \right\rangle}{\left(\sum_{n=1}^N \tau \|q_h^n\|_{1,\mathcal{T}}^2 \right)^{1/2}} \leq C_\zeta^{(5)} \left(C^{(1)} \left(1 + \zeta^{-1} \right) \right),$$

where $C^{(1)}$ is the stability constant from Lemma 6.1.

Proof. Let $\{q_h^n\}_n \subset Q_h \setminus \{0\}$ be an arbitrary sequence of test functions. Employ q_h^n as test function for (5.4). Summing over time steps 1 to N and applying the Cauchy-Schwarz inequality, yields

$$\begin{aligned} &\sum_{n=1}^N \tau \left\langle \frac{\hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^{n-1})}{\tau}, q_h^n \right\rangle \\ &\leq \left(\frac{\alpha^2}{K_{\mathrm{dr}}} \sum_{n=1}^N \tau^{-1} \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\|_{\boldsymbol{V}}^2 \right)^{1/2} \left(\sum_{n=1}^N \tau \|q_h^n\|^2 \right)^{1/2} \\ &\quad + \left(\sum_{n=1}^N \tau \|\chi_h^n\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 \right)^{1/2} \left(\sum_{n=1}^N \tau \|q_h^n\|_{1,\mathcal{T},\kappa_{\mathrm{abs}}}^2 \right)^{1/2} \\ &\quad + (1 + C_{\mathrm{tr}}) C_{\Omega,\mathrm{DP}} \left(\sum_{n=1}^N \tau \|h_{\mathrm{ext}}^n\|_{Q^\star}^2 \right)^{1/2} \left(\sum_{n=1}^N \tau \|q_h^n\|_{1,\mathcal{T}}^2 \right)^{1/2}. \end{aligned}$$

For the last term, we employed a discrete trace inequality, cf. Lemma B.2, and a discrete Poincaré inequality, cf. Lemma B.1. Finally, utilizing a discrete Poincaré inequality for the first term on the right hand side, (A6), and employing Lemma 6.1, we prove the assertion with $C_\zeta^{(5)} := 3\sqrt{C^{(1)}} \left(C_{\Omega,\mathrm{DP}} \frac{\alpha}{\zeta^{1/2} K_{\mathrm{dr}}^{1/2}} + \kappa_{\mathrm{M,abs}}^{1/2} + (1 + C_{\mathrm{tr}}) C_{\Omega,\mathrm{DP}} \right)$. □

6.2 Stability estimates for interpolants in time

Utilizing the discrete-in-time approximations $(\boldsymbol{u}_h^n, \chi_h^n)_n$, defined by (5.3)–(5.4), we define continuous-in-time approximations on $(0, T]$ by piecewise constant interpolation

$$\begin{aligned} \bar{\boldsymbol{u}}_{h\tau}(t) &:= \boldsymbol{u}_h^n,\ t \in (t_{n-1}, t_n], \\ \bar{\chi}_{h\tau}(t) &:= \chi_h^n,\ t \in (t_{n-1}, t_n], \end{aligned}$$

and by piecewise linear interpolation

$$\hat{\boldsymbol{u}}_{h\tau}(t) := \boldsymbol{u}_h^{n-1} + \frac{t-t_{n-1}}{\tau}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}),\ t \in (t_{n-1}, t_n], \tag{6.2}$$

$$\hat{\chi}_{h\tau}(t) := \chi_h^{n-1} + \frac{t-t_{n-1}}{\tau}(\chi_h^n - \chi_h^{n-1}),\ \ t \in (t_{n-1}, t_n]. \tag{6.3}$$

We deduce stability for the interpolants from the stability of the fully discrete approximation.

Lemma 6.6 (Stability estimate for time interpolants of the mechanical displacement)**.** *For all $h, \tau > 0$ and $\hat{\tau} \in [0, \tau)$ it holds that*

$$\zeta \int_0^T \|\partial_t \hat{\boldsymbol{u}}_{h\tau}\|_{\boldsymbol{V}}^2 \, dt + \|\bar{\boldsymbol{u}}_{h\tau}\|_{L^\infty(0,T;\boldsymbol{V})}^2 \le C^{(1)}, \tag{6.4}$$

$$\int_0^{T-\hat{\tau}} \|\bar{\boldsymbol{u}}_{h\tau}(t+\hat{\tau}) - \bar{\boldsymbol{u}}_{h\tau}(t)\|_{\boldsymbol{V}}^2 \, dt \le C^{(1)}\hat{\tau}, \tag{6.5}$$

$$\|\bar{\boldsymbol{u}}_{h\tau} - \hat{\boldsymbol{u}}_{h\tau}\|_{L^2(Q_T)}^2 \le C^{(1)}\tau, \tag{6.6}$$

where $C^{(1)}$ is the stability constant from Lemma 6.1.

Proof. The assertion (6.4) follows directly from Lemma 6.1 by definition of the interpolants. Similarly, by definition of the piecewise constant in time interpolation, it holds that

$$\begin{aligned}
&\int_0^{T-\hat{\tau}} \|\bar{\boldsymbol{u}}_{h\tau}(t+\hat{\tau}) - \bar{\boldsymbol{u}}_{h\tau}(t)\|_{\boldsymbol{V}}^2 \, dt \\
&= \sum_{n=1}^{N-1} \int_{t_{n-1}}^{t_n} \|\bar{\boldsymbol{u}}_{h\tau}(t+\hat{\tau}) - \bar{\boldsymbol{u}}_{h\tau}(t)\|_{\boldsymbol{V}}^2 \, dt + \int_{t_{N-1}}^{t_N-\hat{\tau}} \|\bar{\boldsymbol{u}}_{h\tau}(t+\hat{\tau}) - \bar{\boldsymbol{u}}_{h\tau}(t)\|_{\boldsymbol{V}}^2 \, dt \\
&= \sum_{n=1}^{N-1} \int_{t_n-\hat{\tau}}^{t_n} \|\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n\|_{\boldsymbol{V}}^2 \, dt \\
&= \hat{\tau} \sum_{n=1}^{N} \|\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n\|_{\boldsymbol{V}}^2.
\end{aligned}$$

We obtain (6.5) from Lemma 6.1. By definition of the piecewise constant and piecewise linear interpolation, it holds that

$$\begin{aligned}
\|\bar{\boldsymbol{u}}_{h\tau} - \hat{\boldsymbol{u}}_{h\tau}\|_{L^2(Q_T)}^2 &= \sum_{n=1}^{N} \int_{t_{n-1}}^{t^n} \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} - \tfrac{t-t_{n-1}}{\tau}\left(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right) \right\|^2 \\
&= \frac{1}{3}\tau \sum_{n=1}^{N} \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\|^2.
\end{aligned}$$

We conclude (6.6). □

Analogously, we conclude stability for the interpolants of the Kirchhoff pressure.

Lemma 6.7 (Stability estimate for time interpolants of the Kirchhoff pressure)**.** *For all $h, \tau > 0$ and $\hat{\tau} \in [0, \tau)$ it holds that*

$$\begin{aligned}
\int_0^T \|\bar{\chi}_{h\tau}(t)\|_{1,\mathcal{T}}^2 \, dt &\le C^{(1)}, \\
b_{\chi,\mathrm{m}} \|\partial_t \hat{\chi}_{h\tau}\|_{L^2(Q_T)}^2 + \|\bar{\chi}_{h\tau}\|_{L^\infty(0,T;L^2(\Omega))}^2 &\le C_{\zeta\eta}^{(2)}, \\
\int_0^{T-\hat{\tau}} \|\bar{\chi}_{h\tau}(t+\hat{\tau}) - \bar{\chi}_{h\tau}(t)\|^2 \, dt &\le C_{\Omega,\mathrm{DP}}^2 C_{\zeta\eta}^{(2)} \hat{\tau}, \\
\|\bar{\chi}_{h\tau} - \hat{\chi}_{h\tau}\|_{L^2(Q_T)}^2 &\le C_{\Omega,\mathrm{DP}}^2 C_{\zeta\eta}^{(2)} \tau,
\end{aligned}$$

where $C^{(1)}$ and $C^{(2)}_{\zeta\eta}$ are the stability constants from Lemma 6.1 and Lemma 6.2, respectively, and $C_{\Omega,\mathrm{DP}}$ is the discrete Poincaré constant, cf. Lemma B.1.

Proof. The proof is analogous to the proof Lemma 6.6. For the last two estimates in the assertion, a discrete Poincaré inequality, cf. Lemma B.1, has to be applied before utilizing the stability bound on $\sum_{n=1}^{N} \|\chi_h^n - \chi_h^{n-1}\|_{1,\mathcal{T}}^2$ from Lemma 6.2. □

Similarly, by definition of the piecewise constant interpolation, we deduce stability for some of the non-linearities used in the model.

Lemma 6.8 (Stability estimates for non-linearities evaluated in interpolants)**.** *It holds that*

$$\left\| \hat{B}_\eta(\bar{\chi}_{h\tau}) \right\|_{L^\infty(0,T;L^1(\Omega))} \leq C_\zeta^{(3)},$$
$$\|\hat{p}_{\mathrm{pore}}(\bar{\chi}_{h\tau})\|^2_{L^2(Q_T)} \leq C^{(4)},$$

where $C_\zeta^{(3)}$ and $C^{(4)}$ are the stability constants from Lemma 6.3 and Lemma 6.4, respectively.

Lemma 6.9 (Stability estimate for the temporal change of $\hat{b}$)**.** *For*

$$\bar{\lambda}_{h\tau}(t) := \frac{\hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^{n-1})}{\tau} \quad t \in (t_{n-1}, t_n]$$

it holds that

$$\|\bar{\lambda}_{h\tau}\|_{L^2(0,T;H^{-1}(\Omega)} \leq C_{\Omega,\mathrm{P}}^{1/2} C_\zeta^{(5)},$$

where $C_\zeta^{(5)}$ is the stability constant from Lemma 6.5, and $C_{\Omega,\mathrm{P}}$ is a Poincaré constant.

Proof. Let $q \in L^2(0,T;Q)$. We define a piecewise constant interpolation in both space and time, and only time by

$$\begin{aligned} q_h^n(x,t) &:= \tfrac{1}{\tau} \int_{t_{n-1}}^{t_n} \frac{1}{|K|} \int_K q \, dx \, dt, && (x,t) \in K \times (t_{n-1}, t_n],\ K \in \mathcal{T}, \\ q^n(x,t) &:= \tfrac{1}{\tau} \int_{t_{n-1}}^{t_n} q \, dt, && (x,t) \in \Omega \times (t_{n-1}, t_n]. \end{aligned}$$

Then by Lemma 6.5 it holds that

$$\int_0^T \langle \bar{\lambda}_{h\tau}, q \rangle = \sum_{n=1}^{N} \tau \left\langle \frac{\hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^{n-1})}{\tau}, q_h^n \right\rangle \leq C_\zeta^{(5)} \left(\sum_{n=1}^{N} \tau \|q_h^n\|_{1,\mathcal{T}}^2 \right)^{1/2}.$$

By Lemma B.3, a (continuous) Poincaré inequality (introducing $C_{\Omega,\mathrm{P}}$), analogous to Lemma B.1, the triangle inequality and the Cauchy-Schwarz inequality, it holds that

$$\begin{aligned} \sum_{n=1}^{N} \tau \|q_h^n\|_{1,\mathcal{T}}^2 &\leq C_{\Omega,\mathrm{P}} \sum_{n=1}^{N} \tau \|\boldsymbol{\nabla} q^n\|^2 \\ &= C_{\Omega,\mathrm{P}} \sum_{n=1}^{N} \tau \left\| \tau^{-1} \int_{t_{n-1}}^{t_n} \boldsymbol{\nabla} q \, dt \right\|^2 \\ &\leq C_{\Omega,\mathrm{P}} \sum_{n=1}^{N} \tau^{-1} \left(\int_{t_{n-1}}^{t_n} \|\boldsymbol{\nabla} q\| \, dt \right)^2 \\ &\leq C_{\Omega,\mathrm{P}} \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \|\boldsymbol{\nabla} q\|^2 \, dt \\ &= C_{\Omega,\mathrm{P}} \|q\|^2_{L^2(0,T;H_0^1(\Omega))}, \end{aligned}$$

which concludes the proof. □

6.3 Relative (weak) compactness for the limit $h, \tau \to 0$

We utilize the stability results from the previous section to conclude relative compactness. We deduce limits for the interpolants which eventually converge towards a weak solution of the doubly regularized unsaturated poroelasticity model, i.e., it fulfils $(\mathrm{W1})_{\zeta\eta}$–$(\mathrm{W4})_{\zeta\eta}$.

Lemma 6.10 (Convergence of the mechanical displacement)**.** *We can extract subsequences of $\{\bar{\boldsymbol{u}}_{h\tau}\}_{h,\tau}$ and $\{\hat{\boldsymbol{u}}_{h\tau}\}_{h,\tau}$ (still denoted like the original sequences), and there exists $\boldsymbol{u}_{\varepsilon\eta} \in L^\infty(0,T;\boldsymbol{V})$ with $\partial_t \boldsymbol{u}_{\varepsilon\eta} \in L^2(0,T;\boldsymbol{V})$ such that for $h, \tau \to 0$*

$$\begin{aligned} \bar{\boldsymbol{u}}_{h\tau} &\rightharpoonup \boldsymbol{u}_{\varepsilon\eta} \quad \textit{in } L^\infty(0,T;\boldsymbol{V}), && (6.7)\\ \bar{\boldsymbol{u}}_{h\tau} &\to \boldsymbol{u}_{\varepsilon\eta} \quad \textit{in } L^2(Q_T), && (6.8)\\ \hat{\boldsymbol{u}}_{h\tau} &\rightharpoonup \boldsymbol{u}_{\varepsilon\eta} \quad \textit{in } L^2(0,T;\boldsymbol{V}), && (6.9)\\ \partial_t\hat{\boldsymbol{u}}_{h\tau} &\rightharpoonup \partial_t\boldsymbol{u}_{\varepsilon\eta} \ \textit{in } L^2(0,T;\boldsymbol{V}). && (6.10) \end{aligned}$$

Proof. By the Eberlein-Šmulian theorem, cf. Lemma B.8, and Lemma 6.6, we obtain directly (6.7). For (6.8), we employ a relaxed Aubin-Lions-Simon type compactness result for Bochner spaces, cf. Lemma B.9, together with Lemma 6.6. Furthermore, by the Eberlein-Šmulian theorem, cf. Lemma B.8, and Lemma 6.6, there exists a $\hat{\boldsymbol{u}} \in L^2(0,T,\boldsymbol{V})$ such that up to a subsequence

$$\begin{aligned} \hat{\boldsymbol{u}}_{h\tau} &\rightharpoonup \hat{\boldsymbol{u}} \quad \text{in } L^2(0,T;\boldsymbol{V}),\\ \partial_t\hat{\boldsymbol{u}}_{h\tau} &\rightharpoonup \partial_t\hat{\boldsymbol{u}} \text{ in } L^2(0,T;\boldsymbol{V}). \end{aligned}$$

We can identify $\hat{\boldsymbol{u}} = \boldsymbol{u}_{\varepsilon\eta}$ as follows. Employing the triangle inequality and Lemma 6.6, yields

$$\begin{aligned} \|\hat{\boldsymbol{u}}_{h\tau} - \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(Q_T)} &\leq \|\hat{\boldsymbol{u}}_{h\tau} - \bar{\boldsymbol{u}}_{h\tau}\|_{L^2(Q_T)} + \|\bar{\boldsymbol{u}}_{h\tau} - \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(Q_T)}\\ &\leq C^{(1)}\tau + \|\bar{\boldsymbol{u}}_{h\tau} - \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(Q_T)}, \end{aligned}$$

which converges to zero for $h, \tau \to 0$. This concludes the proof. □

In order to discuss the limit of the pressure, we utilize techniques employed in the finite volume literature [61]. We define a piecewise constant discrete gradient of $\bar{\chi}_{h\tau}$ utilizing the dual grid $\mathcal{T}^\star$, cf. Definition 5.2,

$$\left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau} := \begin{cases} d\frac{\chi^n_{h|L} - \chi^n_{h|K}}{d_{K|L}}\boldsymbol{n}_{K|L}, & (x,t) \in P_\sigma \times (t_{n-1}, t_n], \quad K \in \mathcal{T},\ L \in \mathcal{N}(L),\ \sigma = K|L,\\ d\frac{\chi^n_{h|K}}{d_{\sigma,K}}\boldsymbol{n}_{\sigma,K}, & (x,t) \in P_\sigma \times (t_{n-1}, t_n], \quad \sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K, \end{cases}$$

where $\boldsymbol{n}_{K|L}$ denotes the outward normal on $K|L \in \mathcal{E}$, pointing towards L; and $\boldsymbol{n}_{\sigma,K}$ denotes the outward normal on $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$, pointing towards K.

Lemma 6.11 (Convergence of the Kirchhoff pressure)**.** *We can extract a subsequence of $\{\bar{\chi}_{h\tau}\}_{h,\tau}$ (still denoted like the original sequences), and there exists $\chi_{\varepsilon\eta} \in H^1(0,T;Q)$ such that*

$$\begin{aligned} \bar{\chi}_{h\tau} &\to \chi_{\varepsilon\eta} \quad \textit{in } L^2(Q_T), && (6.11)\\ \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau} &\rightharpoonup \boldsymbol{\nabla}\chi_{\varepsilon\eta} \ \textit{in } L^2(Q_T), && (6.12)\\ \partial_t\hat{\chi}_{h\tau} &\rightharpoonup \partial_t\chi_{\varepsilon\eta} \ \textit{in } L^2(Q_T). && (6.13) \end{aligned}$$

Proof. Let $\hat{h} \in \mathbb{R}^d$ and $\Omega_{\hat{h}} := \{x \in \Omega \,|\, x + \hat{h} \in \Omega\}$. Using Lemma 4 from [62], for all $q_h \in Q_h$ it holds that

$$\int_{\Omega_{\hat{h}}} \left\| q_h(x+\hat{h}) - q_h(x) \right\|^2 \, dx \leq C\,\|q_h\|^2_{1,\mathcal{T}}\,|\hat{h}|\left(|\hat{h}| + |\Omega|\right)$$

for some $C > 0$. Hence, we obtain

$$\begin{aligned}\int_0^T \int_{\Omega_{\hat{h}}} \left\| \bar{\chi}_{h\tau}(x+\hat{h}) - \bar{\chi}_{h\tau}(x) \right\|^2 \, dx \, dt &= \sum_{n=1}^{N} \tau \int_{\Omega_{\hat{h}}} \left\| \chi_h^n(x+\hat{h}) - \chi_h^n(x) \right\|^2 \, dx \\ &\leq |\hat{h}| \left(|\hat{h}| + |\Omega| \right) \sum_{n=1}^{N} \tau \, \|\chi_h^n\|_{1,\mathcal{T}}^2 .\end{aligned}$$

Consequently, by Lemma 6.7, $\bar{\chi}_{h\tau}$ satisfies a translation property in space and time wrt. $L^2(Q_T)$. We conclude by the Riesz-Frechet-Kolmogorov compactness criterion, cf. Lemma B.10, that there exists a $\chi_{\varepsilon\eta} \in L^2(Q_T)$ satisfying (6.11).

By definition of $\left(\overline{\nabla\chi}\right)_{h\tau}$ and the geometrical identity $|P_\sigma| = d^{-1}|\sigma| d_\sigma$, it holds that

$$\begin{aligned}&\left\| \left(\overline{\nabla\chi}\right)_{h\tau} \right\|_{L^2(Q_T)}^2 \\ &= \sum_{n=1}^{N} \tau \sum_{\sigma\in\mathcal{E}} \int_{P_\sigma} \left| \left(\overline{\nabla\chi}\right)_{h\tau} \right|^2 \, dx \\ &= \sum_{n=1}^{N} \tau \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} |P_{K|L}| d^2 \frac{\left| \chi_{h|_K}^n - \chi_{h|_L}^n \right|^2}{d_{K|L}^2} + \sum_{n=1}^{N} \tau \sum_{\sigma\in\mathcal{E}_{\text{ext}}\cap\mathcal{E}_K} |P_\sigma| d^2 \frac{\left| \chi_{h|_K}^n \right|^2}{d_{\sigma,K}^2} \\ &= d \sum_{n=1}^{N} \tau \sum_{\sigma\in\mathcal{E}} \tau_\sigma \left| \delta_\sigma(\chi_h^n) \right|^2 \\ &= d \int_0^T \|\bar{\chi}_{h\tau}\|_{1,\mathcal{T}}^2 \, dt,\end{aligned}$$

which is uniformly bounded by Lemma 6.7. Hence, by the Eberlein-Šmulian theorem, cf. Lemma B.8, there exist a $\boldsymbol{g}_\chi \in L^2(Q_T)$ such that (up to a subsequence)

$$\left(\overline{\nabla\chi}\right)_{h\tau} \rightharpoonup \boldsymbol{g}_\chi \text{ in } L^2(Q_T).$$

It remains to show that $\boldsymbol{g}_\chi = \nabla\chi_{\varepsilon\eta}$ in the sense of distributions, i.e.,

$$\int_0^T \langle \boldsymbol{g}_\chi, \boldsymbol{\varphi} \rangle \, dt + \int_0^T \langle \chi_{\varepsilon\eta}, \nabla \cdot \boldsymbol{\varphi} \rangle \, dt = 0 \quad \text{for all } \boldsymbol{\varphi} \in C^\infty(Q_T)^d .$$

For that, we follow an argument in [61]. Let $\boldsymbol{\varphi} \in C^\infty(Q_T)^d$. As

$$\begin{aligned}\int_0^T \langle \left(\overline{\nabla\chi}\right)_{h\tau}, \boldsymbol{\varphi} \rangle \, dt &\to \int_0^T \langle \boldsymbol{g}_\chi, \boldsymbol{\varphi} \rangle \, dt, \quad \text{and} \\ \int_0^T \langle \bar{\chi}_{h\tau}, \nabla \cdot \boldsymbol{\varphi} \rangle \, dt &\to \int_0^T \langle \chi_{\varepsilon\eta}, \nabla \cdot \boldsymbol{\varphi} \rangle \, dt\end{aligned}$$

for $h, \tau \to 0$, it suffices to show that

$$\int_0^T \langle \left(\overline{\nabla\chi}\right)_{h\tau}, \boldsymbol{\varphi} \rangle \, dt + \int_0^T \langle \bar{\chi}_{h\tau}, \nabla \cdot \boldsymbol{\varphi} \rangle \, dt \to 0.$$

By definition of $\left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau}$ and the construction of $\mathcal{T}^\star$ with $\frac{d}{d_\sigma} = \frac{|\sigma|}{|P_\sigma|}$ for all $\sigma \in \mathcal{E}$, it holds that

$$\begin{aligned}
\int_0^T \left\langle \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau}, \boldsymbol{\varphi} \right\rangle dt &= \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{\sigma\in\mathcal{E}} \int_{P_\sigma} \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau} \cdot \boldsymbol{\varphi} \, dx \, dt \\
&= \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} d \frac{\chi^n_{h|_L} - \chi^n_{h|_K}}{d_{K|L}} \int_{P_{K|L}} \boldsymbol{\varphi} \cdot \boldsymbol{n}_{K|L} \, dx \, dt \\
&\quad + \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} d \frac{\chi^n_{h|_K}}{d_{\sigma,K}} \int_{P_\sigma} \boldsymbol{\varphi} \cdot \boldsymbol{n}_{\sigma,K} \, dx \, dt \\
&= \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} |\sigma| \left(\chi^n_{h|_L} - \chi^n_{h|_K}\right) \frac{1}{|P_{K|L}|} \int_{P_{K|L}} \boldsymbol{\varphi} \cdot \boldsymbol{n}_{K|L} \, dx \, dt \\
&\quad + \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} |\sigma| \chi^n_{h|_K} \frac{1}{|P_\sigma|} \int_{P_\sigma} \boldsymbol{\varphi} \cdot \boldsymbol{n}_{\sigma,K} \, dx \, dt.
\end{aligned}$$

On the other hand, since $\bar{\chi}_{h\tau}$ is constant and hence continuous within each $K \in \mathcal{T}$, it holds that

$$\begin{aligned}
&\int_0^T \left\langle \bar{\chi}_{h\tau}, \boldsymbol{\nabla} \cdot \boldsymbol{\varphi} \right\rangle \, dt \\
&= \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \sum_{\sigma\in\mathcal{E}} \int_{P_\sigma} \bar{\chi}_{h\tau} \boldsymbol{\nabla} \cdot \boldsymbol{\varphi} \, dx \, dt \\
&= \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \left[\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \left(\chi^n_{h|_K} \int_{P_{K|L}\cap K} \boldsymbol{\nabla} \cdot \boldsymbol{\varphi} \, dx + \chi^n_{h|_L} \int_{P_{K|L}\cap L} \boldsymbol{\nabla} \cdot \boldsymbol{\varphi} \, dx \right) \right. \\
&\qquad \left. + \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} \chi^n_{h|_K} \int_{P_\sigma\cap K} \boldsymbol{\nabla} \cdot \boldsymbol{\varphi} \, dx \right] dt \\
&= -\sum_{n=1}^N \int_{t_{n-1}}^{t_n} \left[\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \left(\chi^n_{h|_L} - \chi^n_{h|_K}\right) \int_{K|L} \boldsymbol{\varphi} \cdot \boldsymbol{n}_{K|L} \, ds \right. \\
&\qquad \left. + \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} \chi^n_{h|_K} \int_\sigma \boldsymbol{\varphi} \cdot \boldsymbol{n}_{\sigma,K} \, ds \right] dt.
\end{aligned}$$

As $\boldsymbol{\varphi} \in C^\infty(Q_T)^d$ is smooth, there exists a constant $C > 0$ such that

$$\left| \frac{1}{\tau} \int_{t_{n-1}}^{t_n} \frac{1}{|P_\sigma|} \int_{P_\sigma} \boldsymbol{\varphi} \cdot \boldsymbol{n}_\sigma \, dx \, dt - \frac{1}{\tau} \int_{t_{n-1}}^{t_n} \frac{1}{|\sigma|} \int_\sigma \boldsymbol{\varphi} \cdot \boldsymbol{n}_\sigma \, ds \, dt \right| \le Ch.$$

By abuse of notation, we used $\boldsymbol{n}_\sigma$ for both $\boldsymbol{n}_{K|L}$ and $\boldsymbol{n}_{\sigma,K}$. After all, together with the Cauchy-Schwarz inequality, it holds that

$$\begin{aligned}
&\left| \int_0^T \left\langle \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau}, \boldsymbol{\varphi} \right\rangle dt + \int_0^T \int_\Omega \bar{\chi}_{h\tau} \boldsymbol{\nabla} \cdot \boldsymbol{\varphi} \, dx \, dt \right| \\
&\le Ch \sum_{n=1}^N \tau \left(\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} |\sigma| \left| \chi^n_{h|_L} - \chi^n_{h|_K} \right| + \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} |\sigma| \left| \chi^n_{h|_K} \right| \right) \\
&\le Ch \left(\sum_{n=1}^N \tau \left\| \chi^n_h \right\|^2_{1,\mathcal{T}} \right)^{1/2} \left(\sum_{n=1}^N \tau \sum_{\sigma\in\mathcal{E}} |\sigma| d_\sigma \right)^{1/2}.
\end{aligned}$$

By Lemma 6.7 and the regularity assumption on $\mathcal{T}$, convergence towards 0 follows for $h, \tau \to 0$. This concludes the proof of (6.12).

The proof of (6.13) is standard and follows mainly from the stability results in Lemma 6.7 and the Eberlein-Šmulian theorem, cf. Lemma B.8. This concludes the proof. □

The main purpose of the double regularization has been the aim to get control over the non-linear coupling terms, and eventually establish convergence.

Lemma 6.12 (Convergence of the coupling terms)**.** *We can extract a subsequence of* $\{\bar{\chi}_{h\tau}\}_{h,\tau}$ *(still denoted like the original sequences) such that*

$$\hat{p}_{\mathrm{pore}}(\bar{\chi}_{h\tau}) \rightharpoonup \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta}) \qquad \text{in } L^2(Q_T), \tag{6.14}$$

$$\hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau})\partial_t \boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}}_{h\tau} \rightharpoonup \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})\partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u}_{\varepsilon\eta} \qquad \text{in } L^2(Q_T). \tag{6.15}$$

Proof. By the Eberlein-Šmulian theorem, cf. Lemma B.8, and Lemma 6.8, we can extract a subsequence of $\{\bar{\chi}_{h\tau}\}_{h,\tau}$ (still denoted $\{\bar{\chi}_{h\tau}\}_{h,\tau}$), and there exists a $\hat{p} \in L^2(Q_T)$ such that

$$\hat{p}_{\mathrm{pore}}(\bar{\chi}_{h\tau}) \rightharpoonup \hat{p} \text{ in } L^2(Q_T).$$

We can identify $\hat{p} = \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta})$ as follows. From Lemma 6.11, we have $\bar{\chi}_{h\tau} \to \chi_{\varepsilon\eta}$ a.e. on Q_T for a subsequence (still denoted $\{\bar{\chi}_{h\tau}\}_{h,\tau}$). As $\hat{p}_{\mathrm{pore}}$ is continuous by (A3), it holds that $\hat{p}_{\mathrm{pore}}(\bar{\chi}_{h\tau}) \to \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta})$ a.e. on Q_T. This concludes (6.14).

The convergence property (6.15) follows from the convergence properties of the single contributions. Let $q \in L^2(Q_T)$; it holds that $\hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau})q \to \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})q$ in $L^2(Q_T)$ (up to a subsequence). Indeed, by Lemma 6.11, we have $\bar{\chi}_{h\tau} \to \chi_{\varepsilon\eta}$ a.e. on Q_T (up to a subsequence); due to (A2), it holds that $\hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau})q \to \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})q$ a.e. on Q_T and $|\hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau})q| \leq |q|$ a.e.; hence, by the dominated convergence theorem $\hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau})q \to \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})q$ in $L^2(Q_T)$. In particular, it holds that $\hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})q \in L^2(\Omega)$. Moreover from Lemma 6.10, we have $\partial_t\boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}}_{h\tau} \rightharpoonup \partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u}_{\varepsilon\eta}$ in $L^2(Q_T)$. Altogether, we obtain

$$\begin{aligned}
&|\langle \hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau})\partial_t\boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}}_{h\tau} - \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})\partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u}_{\varepsilon\eta}, q\rangle| \\
&\quad \leq |\langle (\hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau}) - \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta}))\,\partial_t\boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}}_{h\tau}, q\rangle| + |\langle \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})\left(\partial_t\boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}}_{h\tau} - \partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u}_{\varepsilon\eta}\right), q\rangle| \\
&\quad \leq \|\hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau})q - \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})q\|\,\|\partial_t\boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}}_{h\tau}\| + |\langle \partial_t\boldsymbol{\nabla}\cdot\hat{\boldsymbol{u}}_{h\tau} - \partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u}_{\varepsilon\eta}, \hat{s}_{\mathrm{w}}(\chi_{\varepsilon\eta})q\rangle|\,,
\end{aligned}$$

which converges towards 0 for $h, \tau \to 0$, due to strong and weak convergence of the single components. □

Lemma 6.13 (Initial conditions for the fluid flow)**.** *It holds that*

$$\bar{\lambda}_{h\tau} \rightharpoonup \partial_t\hat{b}_\eta(\chi_{\varepsilon\eta}) \text{ in } L^2(0,T;Q^\star) \tag{6.16}$$

(up to a subsequence), where $\partial_t\hat{b}_\eta(\chi_{\varepsilon\eta}) \in L^2(0,T;Q^\star)$ *is understood in the sense of* (W2)$_{\zeta\eta}$.

Proof. By definition of the Legendre transformation $\hat{B}$ and its properties, cf. Lemma B.12, it holds that

$$|\hat{b}_\eta(x)| \leq \delta\hat{B}_\eta(x) + \sup_{|y|\leq\delta^{-1}} |\hat{b}_\eta(y)|,$$

for all $\delta > 0$. Since $\hat{B}_\eta(\bar{\chi}_{h\tau}) \in L^\infty(0,T;L^1(\Omega))$ is uniformly bounded by Lemma 6.8, and $\hat{b}_\eta$ is continuous by (A1)*, it holds that $\|\hat{b}_\eta(\bar{\chi}_{h\tau})\|_{L^\infty(0,T;L^1(\Omega))}$ is uniformly bounded. Hence, by the Eberlein-Šmulian theorem, cf. Lemma B.8, we can extract a subsequence of $\{\bar{\chi}_{h\tau}\}_{h,\tau}$ (still denoted $\{\bar{\chi}_{h\tau}\}_{h,\tau}$), and there exists a $\hat{b}_\chi \in L^\infty(0,T;L^1(\Omega))$ such that

$$\hat{b}_\eta(\bar{\chi}_{h\tau}) \rightharpoonup \hat{b}_\chi \text{ in } L^\infty(0,T;L^1(\Omega)).$$

As $\hat{b}$ is continuous by (A1), and $\bar{\chi}_{h\tau} \to \chi_{\varepsilon\eta}$ in $L^2(Q_T)$ (up to a subsequence) by Lemma 6.11, it holds that $\hat{b}_\eta(\bar{\chi}_{h\tau}) \to \hat{b}_\eta(\chi_{\varepsilon\eta})$ a.e. on Q_T (up to a subsequence). We conclude $\hat{b}_\chi = \hat{b}_\eta(\chi_{\varepsilon\eta})$, which proves

$$\hat{b}_\eta(\bar{\chi}_{h\tau}) \rightharpoonup \hat{b}_\eta(\chi_{\varepsilon\eta}) \text{ in } L^\infty(0,T;L^1(\Omega)). \tag{6.17}$$

By the Eberlein-Šmulian theorem, cf. Lemma B.8, and Lemma 6.9, we can extract a subsequence of $\{\bar{\chi}_{h\tau}\}_{h,\tau}$ (still denoted $\{\bar{\chi}_{h\tau}\}_{h,\tau}$), and there exists a $\hat{b}_t \in L^2(0,T;Q^\star)$ such that

$$\bar{\lambda}_{h\tau} \rightharpoonup \hat{b}_t \text{ in } L^2(0,T;Q^\star).$$

It remains to show that $\hat{b}_t = \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta})$ in the sense of $(\mathrm{W2})_{\zeta\eta}$. For this, we follow arguments by [53] as follows. Let $q \in L^2(0,T;Q)$ with $\partial_t q \in L^1(0,T;L^\infty(\Omega))$ and $q(T)=0$. Due to (6.17) it holds that

$$\int_0^T \left\langle \hat{b}_\eta(\chi_h^0) - \hat{b}_\eta(\chi_0), \partial_t q \right\rangle \, dt \to 0,$$

for $h,\tau \to 0$. Thus, it suffices to show that

$$\int_0^T \langle \bar{\lambda}_{h\tau}, q\rangle \, dt + \int_0^T \left\langle \hat{b}_\eta(\bar{\chi}_{h\tau}) - \hat{b}_\eta(\chi_h^0), \partial_t q \right\rangle \, dt \to 0,$$

for $h,\tau \to 0$. By definition of $\bar{\lambda}_{h\tau}$, after applying summation by parts, cf. Lemma B.6, we obtain

$$\begin{aligned}
&\int_0^T \langle \bar{\lambda}_{h\tau}, q\rangle \, dt \\
&\quad = \sum_{n=1}^N \left\langle \hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^{n-1}), \tau^{-1}\int_{t_{n-1}}^{t_n} q \, dt \right\rangle \\
&\quad = \left\langle \hat{b}_\eta(\chi_h^N), \tau^{-1}\int_{T-\tau}^{T} q \, dt \right\rangle - \left\langle \hat{b}_\eta(\chi_h^0), \tau^{-1}\int_0^\tau q \, dt \right\rangle \\
&\qquad - \sum_{n=1}^{N-1} \left\langle \hat{b}_\eta(\chi_h^n), \tau^{-1}\int_{t_n}^{t_{n+1}} q \, dt - \tau^{-1}\int_{t_{n-1}}^{t_n} q \, dt \right\rangle \\
&\quad = \left\langle \hat{b}_\eta(\chi_h^N) - \hat{b}_\eta(\chi_h^0), \tau^{-1}\int_{T-\tau}^{T} q \, dt \right\rangle \\
&\qquad - \sum_{n=1}^{N-1} \int_{t_{n-1}}^{t_n} \left\langle \hat{b}_\eta(\chi_h^n) - \hat{b}_\eta(\chi_h^0), \frac{\tau^{-1}\int_{t_n}^{t_{n+1}} q \, dt - \tau^{-1}\int_{t_{n-1}}^{t_n} q \, dt}{\tau} \right\rangle \, d\tilde{t} \\
&\quad \to 0 - \int_0^T \left\langle \hat{b}_\eta(\chi_{\varepsilon\eta}) - \hat{b}_\eta(\chi_0), \partial_t q \right\rangle \, dt,
\end{aligned}$$

for $h,\tau \to 0$, due to the smoothness of q and the convergence properties of $\hat{b}_\eta(\bar{\chi}_{h\tau})$. This concludes the proof. □

Lemma 6.14 (Initial conditions for the mechanical displacement)**.** *The limit* $\boldsymbol{u}_{\varepsilon\eta} \in H^1(0,T;\boldsymbol{V})$ *from Lemma 6.10 satisfies* $(\mathrm{W3})_{\zeta\eta}$.

Proof. Let $\boldsymbol{v} \in H^1(0,T;\boldsymbol{V})$ with $\boldsymbol{v}(T) = \boldsymbol{0}$. We obtain, using the same calculations as in the proof of Lemma 6.13,

$$\int_0^T a(\partial_t \hat{\boldsymbol{u}}_{h\tau}, \boldsymbol{v}) \, dt = a\left(\boldsymbol{u}_h^N - \boldsymbol{u}_h^0, \tau^{-1}\int_{T-\tau}^T \boldsymbol{v} \, dt\right) - \int_0^{T-\tau} a\left(\bar{\boldsymbol{u}}_{h\tau} - \boldsymbol{u}_h^0, \partial_t \hat{\boldsymbol{v}}_{h\tau}\right),$$

where

$$\hat{\boldsymbol{v}}_{h\tau}(t) = \tau^{-1}\int_{t_{n-2}}^{t_{n-1}} \boldsymbol{v}\,dt + \frac{t-t_{n-1}}{\tau}\left(\tau^{-1}\int_{t_{n-1}}^{t_n} \boldsymbol{v}\,dt - \tau^{-1}\int_{t_{n-2}}^{t_{n-1}} \boldsymbol{v}\,dt\right), \quad t\in(t_{n-1},t_n].$$

By construction of $\boldsymbol{u}_h^0$ it holds that $\boldsymbol{u}_h^0 \rightharpoonup \boldsymbol{u}_0$ in $L^2(0,T;\boldsymbol{V})$. Furthermore, by Lemma 6.10, it holds that $\bar{\boldsymbol{u}}_{h\tau} \rightharpoonup \boldsymbol{u}_{\varepsilon\eta}$ in $L^2(0,T;\boldsymbol{V})$ and $\partial_t\hat{\boldsymbol{u}}_{h\tau} \rightharpoonup \partial_t\boldsymbol{u}_{\varepsilon\eta}$ in $L^2(0,T;\boldsymbol{V})$ (up to subsequences). Hence, for $h,\tau\to 0$, we obtain

$$\int_0^T a(\partial_t \boldsymbol{u}_{\varepsilon\eta},\boldsymbol{v})\,dt = -\int_0^T a\left(\boldsymbol{u}_{\varepsilon\eta}-\boldsymbol{u}_0,\partial_t\boldsymbol{v}\right),$$

and thereby (W3)$_{\zeta\eta}$. □

6.4 Identifying a weak solution for $h,\tau\to 0$

Finally, we show the limit $(\boldsymbol{u}_{\varepsilon\eta},\chi_{\varepsilon\eta})$, introduced in the previous section, is a weak solution of the doubly regularized unsaturated poroelasticity model, cf. Definition 4.1.

Lemma 6.15 (Limit satisfies (W1)$_{\zeta\eta}$–(W4)$_{\zeta\eta}$)**.** *The limit $(\boldsymbol{u}_{\varepsilon\eta},\chi_{\varepsilon\eta})$ introduced in the previous section is a weak solution to the doubly regularized unsaturated poroelasticity model, cf. Definition 4.1.*

Proof. The limit $(\boldsymbol{u}_{\varepsilon\eta},\chi_{\varepsilon\eta})$ satisfies (W1)$_{\zeta\eta}$–(W3)$_{\zeta\eta}$ by Lemma 6.10, Lemma 6.11 Lemma 6.12, and Lemma 6.14. It remains to show (W4)$_{\zeta\eta}$, i.e., that $(\boldsymbol{u}_{\varepsilon\eta},\chi_{\varepsilon\eta})$ satisfies the balance equations (4.3)–(4.4). We first consider sufficiently smooth test functions and then use a density argument. Let $(\boldsymbol{v},q)\in L^2(0,T;\boldsymbol{V}\cap C^\infty(\Omega)^d)\times L^2(0,T;Q\cap C^\infty(\Omega))$. For given mesh $\mathcal{T}$, we define spatial projection and interpolation operators, respectively, by

$$\Pi_{\boldsymbol{V}_h} : \boldsymbol{V}\cap C^\infty(\Omega)\to \boldsymbol{V}_h, \text{ s.t. } \langle \Pi_{\boldsymbol{V}_h}\boldsymbol{v},\boldsymbol{v}_h\rangle = \langle \boldsymbol{v},\boldsymbol{v}_h\rangle \text{ for all } \boldsymbol{v}_h\in\boldsymbol{V}_h, \tag{6.18}$$

$$\mathcal{I}_{Q_h} : Q\cap C^\infty(\Omega)\to Q_h, \text{ s.t. } \mathcal{I}_{Q_h}q_{|K} = q(x_K) \text{ for all } K\in\mathcal{T}. \tag{6.19}$$

Using those, we define piecewise-constant-in-time interpolants of $(\boldsymbol{v},q)$

$$\bar{\boldsymbol{v}}_{h\tau}(t) := \boldsymbol{v}_h^n,\ t\in(t_{n-1},t_n], \qquad \boldsymbol{v}_h^n := \Pi_{\boldsymbol{V}_h}\boldsymbol{v}^n, \qquad \boldsymbol{v}^n := \tau^{-1}\int_{t_{n-1}}^{t_n}\boldsymbol{v}\,dt, \tag{6.20}$$

$$\bar{q}_{h\tau}(t) := q_h^n,\ t\in(t_{n-1},t_n], \qquad q_h^n := \mathcal{I}_{Q_h}q^n, \qquad q^n := \tau^{-1}\int_{t_{n-1}}^{t_n} q\,dt. \tag{6.21}$$

Similarly, let

$$\begin{aligned}\bar{\boldsymbol{f}}_{\mathrm{ext},\tau}(t) &:= \boldsymbol{f}_{\mathrm{ext}}^n,\ t\in(t_{n-1},t_n],\\ \bar{h}_{\mathrm{ext},\tau}(t) &:= h_{\mathrm{ext}}^n,\ t\in(t_{n-1},t_n].\end{aligned}$$

Combining classical results, based on the assumed regularity (A7), for $h,\tau\to 0$ it holds that

$$\begin{aligned}\bar{\boldsymbol{v}}_{h\tau} &\to \boldsymbol{v} && \text{in } L^2(0,T;\boldsymbol{V}),\\ \bar{q}_{h\tau} &\to q && \text{in } L^2(0,T;Q),\\ \bar{\boldsymbol{f}}_{\mathrm{ext},\tau} &\to \boldsymbol{f}_{\mathrm{ext}} && \text{in } L^2(0,T;\boldsymbol{V}^\star),\\ \bar{h}_{\mathrm{ext},\tau} &\to h_{\mathrm{ext}} && \text{in } L^2(0,T;Q^\star).\end{aligned}$$

We choose $\boldsymbol{v}_h = \boldsymbol{v}_h^n$ and $q_h = q_h^n$ as test functions in (5.3)–(5.4), multiply both equations with τ and sum over all time steps 1 to N; we obtain

$$\int_0^T \Big[\lambda_{\mathrm{v}} \left\langle \partial_t \boldsymbol{\nabla}\cdot \hat{\boldsymbol{u}}_{h\tau}, \boldsymbol{\nabla}\cdot \bar{\boldsymbol{v}}_{h\tau}\right\rangle + a(\bar{\boldsymbol{u}}_{h\tau}, \bar{\boldsymbol{v}}_{h\tau}) - \alpha \left\langle \hat{p}_{\mathrm{pore}}(\bar{\chi}_{h\tau}), \boldsymbol{\nabla}\cdot \bar{\boldsymbol{v}}_{h\tau}\right\rangle \Big]\, dt = \int_0^T \left\langle \bar{\boldsymbol{f}}_{\mathrm{ext},\tau}, \bar{\boldsymbol{v}}_{h\tau}\right\rangle\, dt, \tag{6.22}$$

$$\int_0^T \Big[\left\langle \bar{\lambda}_{h\tau}, \bar{q}_{h\tau}\right\rangle + \alpha \left\langle \hat{s}_{\mathrm{w}}(\bar{\chi}_{h\tau}) \partial_t \boldsymbol{\nabla}\cdot \hat{\boldsymbol{u}}_{h\tau}, \bar{q}_{h\tau}\right\rangle + \left\langle \boldsymbol{\nabla}_h \bar{\chi}_{h\tau}, \boldsymbol{\nabla}_h \bar{q}_{h\tau}\right\rangle_{\kappa_{\mathrm{abs}}}\Big]\, dt = \int_0^T \left\langle \bar{h}_{\mathrm{ext},\tau}, \bar{q}_{h\tau}\right\rangle\, dt. \tag{6.23}$$

For most terms we can apply the fact that the product of weakly and strongly convergent sequences converge to the product of their limits. The only term needing discussion is the diffusion term in the flow equation. For this, we follow an argument by [61].

By definition of the continuous extension of the discrete gradient $\left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau}$, it holds that

$$\begin{aligned}
&\int_0^T \left\langle \boldsymbol{\nabla}_h \bar{\chi}_{h\tau}, \boldsymbol{\nabla}_h \bar{q}_{h\tau}\right\rangle_{\kappa_{\mathrm{abs}}}\, dt \\
&\quad = \sum_{n=1}^N \tau \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K|L} \{\kappa_{\mathrm{abs}}\}_{K|L} \left(\chi^n_{h|_K} - \chi^n_{h|_L}\right) \left(q^n(x_K) - q^n(x_L)\right) \\
&\qquad + \sum_{n=1}^N \tau \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} \tau_\sigma \{\kappa_{\mathrm{abs}}\}_\sigma \chi^n_{h|_K}\, q^n(x_K) \\
&\quad = \sum_{n=1}^N \tau \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} |P_{K|L}| \{\kappa_{\mathrm{abs}}\}_{K|L} \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau|_{P_{K|L}\times(t_{n-1},t_n]}} \cdot \boldsymbol{n}_{L|K}\, \tfrac{1}{d_{K|L}} \left(q^n(x_K) - q^n(x_L)\right) \\
&\qquad + \sum_{n=1}^N \tau \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} |P_\sigma| \{\kappa_{\mathrm{abs}}\}_\sigma \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau|_{P_\sigma\times(t_{n-1},t_n]}} \cdot (-\boldsymbol{n}_{\sigma,K})\, \tfrac{1}{d_{\sigma,K}} q^n(x_K).
\end{aligned}$$

By the mean value theorem, there exists an $x_{K|L} \in P_{K|L}$ on the line between x_K and x_L, and an $x_\sigma \in P_\sigma$ on the line between x_K and the closest point of x_K on σ such that

$$\begin{aligned}
\tfrac{1}{d_{K|L}} \left(q^n(x_K) - q^n(x_L)\right) &= \boldsymbol{\nabla} q^n(x_{K|L}) \cdot \boldsymbol{n}_{L|K}, \\
\tfrac{1}{d_{\sigma,K}} q^n(x_K) &= \boldsymbol{\nabla} q^n(x_\sigma) \cdot \left(-\boldsymbol{n}_{\sigma,K}\right).
\end{aligned}$$

Due to identical alignment of the discrete gradients, it holds that

$$\begin{aligned}
&\int_0^T \left\langle \boldsymbol{\nabla}_h \bar{\chi}_{h\tau}, \boldsymbol{\nabla}_h \bar{q}_{h\tau}\right\rangle_{\kappa_{\mathrm{abs}}}\, dt \\
&\quad = \sum_{n=1}^N \tau \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} |P_{K|L}| \{\kappa_{\mathrm{abs}}\}_{K|L} \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau|_{P_{K|L}\times(t_{n-1},t_n]}} \cdot \boldsymbol{\nabla} q^n(x_{K|L}) \\
&\qquad + \sum_{n=1}^N \tau \sum_{\sigma\in\mathcal{E}_{\mathrm{ext}}\cap\mathcal{E}_K} |P_\sigma| \{\kappa_{\mathrm{abs}}\}_\sigma \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau|_{P_\sigma\times(t_{n-1},t_n]}} \cdot \boldsymbol{\nabla} q(x_\sigma).
\end{aligned}$$

We define the piecewise constant functions

$$\begin{aligned}
\left(\overline{\boldsymbol{\nabla} q}\right)_{h\tau}(x,t) &= \boldsymbol{\nabla} q^n(x_\sigma), \quad (x,t) \in P_\sigma \times (t_{n-1}, t_n],\ \sigma\in\mathcal{E}, \\
\{\kappa_{\mathrm{abs}}\}_{\mathcal{T}}(x) &= \{\kappa_{\mathrm{abs}}\}_\sigma, \qquad x\in P_\sigma,\ \sigma\in\mathcal{E}.
\end{aligned}$$

We obtain for $h, \tau \to 0$

$$\begin{aligned}
&\int_0^T \langle \boldsymbol{\nabla}_h \bar{\chi}_{h\tau}, \boldsymbol{\nabla}_h \bar{q}_{h\tau} \rangle_{\kappa_{\text{abs}}} \, dt \\
&\quad = \int_0^T \int_\Omega \{\kappa_{\text{abs}}\}_{\mathcal{T}} \left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau} \cdot \left(\overline{\boldsymbol{\nabla} q}\right)_{h\tau} \, dx \, dt \to \int_0^T \int_\Omega \kappa_{\text{abs}} \boldsymbol{\nabla}\chi_{\varepsilon\eta} \cdot \boldsymbol{\nabla} q \, dx \, dt.
\end{aligned}$$

Indeed, due to sufficient regularity, it holds that $\left(\overline{\boldsymbol{\nabla} q}\right)_{h\tau} \to \boldsymbol{\nabla} q$ a.e., and also in $L^2(Q_T)$ by the dominated convergence theorem. Furthermore, it holds that $\{\kappa_{\text{abs}}\}_{\mathcal{T}} \to \kappa_{\text{abs}}$ in $L^\infty(Q_T)$, and by Lemma 6.11, it holds that $\left(\overline{\boldsymbol{\nabla}\chi}\right)_{h\tau} \rightharpoonup \boldsymbol{\nabla}\chi_{\varepsilon\eta}$ in $L^2(Q_T)$. That suffices to discuss the product.

All in all, together with the convergence properties of the test functions $\bar{\boldsymbol{v}}_{h\tau}$, $\bar{q}_{h\tau}$, the source terms $\bar{\boldsymbol{f}}_{\text{ext},\tau}$, $\bar{h}_{\text{ext},\tau}$, and the interpolants for the fully discrete approximations (cf. Lemma 6.10, Lemma 6.11, Lemma 6.12 and Lemma 6.13), we conclude that (6.22)–(6.23) converges to (4.3)–(4.4), evaluated in $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ and tested with $(\boldsymbol{v}, q) \in L^2(0,T; \boldsymbol{V} \cap C^\infty(\Omega)^d) \times L^2(0,T; Q \cap C^\infty(\Omega))$. Finally, a density argument yields the final result. □

7 Step 4: Increased regularity in a non-degenerate case

In the following, further stability estimates for the fully-discrete problem are derived, allowing for showing that the limit $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ introduced in the previous section also satisfies $(\text{W5})_{\zeta\eta}$ and $(\text{W6})_{\zeta\eta}$, i.e., we prove Lemma 4.4. For this, non-degeneracy assumptions are required. For compact presentation throughout the entire section, we assume (A0)–(A9) and (ND1)–(ND2) hold true, and we define $\boldsymbol{u}_h^{-1} := \boldsymbol{u}_h^0$.

7.1 Improved stability estimates for fully-discrete approximation

Lemma 7.1 (Improved stability estimate for the structural velocity)**.** *There exists a constant $C_{\zeta\eta}^{(6)} > 0$ (independent of h, τ), satisfying*

$$\begin{aligned}
\zeta \sup_n \left\| \tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}) \right\|_{\boldsymbol{V}}^2 &+ \sum_{n=1}^N \tau^{-1} \| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \|_{\boldsymbol{V}}^2 \\
&+ \zeta \sum_{n=1}^N \left\| \tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}) - \tau^{-1}(\boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2}) \right\|_{\boldsymbol{V}}^2 \\
&+ \sum_{n=1}^N \tau^{-1} \| \hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1}) \|^2 \\
&\leq C_{\zeta\eta}^{(6)} \left(\| \partial_t \boldsymbol{f}_{\text{ext}} \|_{L^2(Q_T)}^2, \frac{C_{\text{ND},2}}{b_{\chi,\text{m}}} C_{\zeta\eta}^{(2)} \right),
\end{aligned}$$

where $C_{\zeta\eta}^{(2)}$ is the stability constant from Lemma 6.2, $C_{\text{ND},2}$ comes from the non-degeneracy condition (ND2)*, and $b_{\chi,\text{m}}$ comes from the growth condition* (A1*)*.*

Proof. First we observe, that the compatibility condition for the initial conditions (5.2) is equivalent to the mechanics equation (5.3) for $n = 0$, since $\boldsymbol{u}_h^0 - \boldsymbol{u}_h^{-1} = \boldsymbol{0}$. This allows for considering the difference of the mechanics equation (5.3) at time steps n and $n-1$, $n \geq 1$,

$$\begin{aligned}
&\zeta a\left(\tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}) - (\boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2}), \boldsymbol{v}_h\right) + a(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) \\
&\qquad - \alpha \langle \hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1}), \boldsymbol{\nabla} \cdot \boldsymbol{v}_h \rangle = \langle \boldsymbol{f}_{\text{ext}}^n - \boldsymbol{f}_{\text{ext}}^{n-1}, \boldsymbol{v}_h \rangle \quad \text{for all } \boldsymbol{v}_h \in \boldsymbol{V}_h.
\end{aligned}$$

By testing with $\boldsymbol{v}_h = \tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1})$ and using the binomial identity (B.2), we obtain

$$\begin{aligned}&\frac{\zeta}{2}\Big(\left\|\tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1})\right\|_{\boldsymbol{V}}^2 - \left\|\tau^{-1}(\boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2})\right\|_{\boldsymbol{V}}^2 \\ &\quad + \left\|\tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}) - \tau^{-1}(\boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2})\right\|_{\boldsymbol{V}}^2 \Big) + \tau^{-1}\|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\|_{\boldsymbol{V}}^2 \\ &\quad = \tau^{-1}\left\langle \boldsymbol{f}_{\text{ext}}^n - \boldsymbol{f}_{\text{ext}}^{n-1}, \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\rangle + \alpha\tau^{-1}\left\langle \hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1}), \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right\rangle.\end{aligned}$$

Summing over $n \in \{1, ..., N\}$, yields after applying the Cauchy-Schwarz inequality and Young's inequality for the right hand side terms

$$\begin{aligned}&\frac{\zeta}{2}\left\|\tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1})\right\|_{\boldsymbol{V}}^2 + \frac{1}{2}\sum_{n=1}^{N}\tau^{-1}\|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\|_{\boldsymbol{V}}^2 \\ &\quad + \frac{1}{2}\sum_{n=1}^{N}\left\|\tau^{-1}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}) - \tau^{-1}(\boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2})\right\|_{\boldsymbol{V}}^2 \\ &\quad \leq \sum_{n=1}^{N}\tau^{-1}\left\|\boldsymbol{f}_{\text{ext}}^n - \boldsymbol{f}_{\text{ext}}^{n-1}\right\|_{\boldsymbol{V}^\star}^2 + \frac{\alpha^2}{K_{\text{dr}}}\sum_{n=1}^{N}\tau^{-1}\left\|\hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1})\right\|^2. \qquad (7.1)\end{aligned}$$

Due to (ND2), $\hat{p}_{\text{pore}} = \hat{p}_{\text{pore}}(\chi)$ is Lipschitz continuous. Therefore, by Lemma 6.2 it holds that

$$\sum_{n=1}^{N}\tau^{-1}\|\hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1})\|^2 \leq C_{\text{ND},2}^2\frac{C_{\zeta\eta}^{(2)}}{b_{\chi,\text{m}}},$$

which together with (7.1) concludes the proof. □

Lemma 7.2 (Consequence for the structural acceleration). *There exists a constant* $C_{\zeta\eta}^{(7)} > 0$ *(independent of* h, τ*), such that*

$$\sum_{n=1}^{N}\tau\left\|\frac{\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}}{\tau^2}\right\|_{\boldsymbol{V}}^2 \leq C_{\zeta\eta}^{(7)}\left(\zeta^{-2}C_{\zeta\eta}^{(6)}\right),$$

where $C_{\zeta\eta}^{(6)}$ *is the stability constant from Lemma 7.1.*

Proof. Let $\{\boldsymbol{v}_h^n\}_n \subset \boldsymbol{V}_h \setminus \{\boldsymbol{0}\}$ be an arbitrary sequence of test functions. Consider the difference of (5.3) at n and $n-1$, $n \geq 1$; it holds that

$$\begin{aligned}&\tau^{-1}\zeta a\left(\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}, \boldsymbol{v}_h^n\right) \\ &\quad = \left\langle \boldsymbol{f}_{\text{ext}}^n - \boldsymbol{f}_{\text{ext}}^{n-1}, \boldsymbol{v}_h^n\right\rangle - a\left(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h^n\right) + \alpha\left\langle \hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1}), \boldsymbol{\nabla}\cdot\boldsymbol{v}_h^n\right\rangle.\end{aligned}$$

Summing over $n \in \{1, ..., N\}$, applying the Cauchy-Schwarz inequality and Lemma 7.1, yields

$$\sup_{\{\boldsymbol{v}_h^n\}_n \subset \boldsymbol{V}_h\setminus\{\boldsymbol{0}\}}\frac{\zeta\sum_{n=1}^{N}\tau^{-1}a(\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}, \boldsymbol{v}_h^n)}{\left(\sum_{n=1}^{N}\tau\|\boldsymbol{v}_h^n\|_{\boldsymbol{V}}^2\right)^{1/2}} \leq 3\sqrt{C_{\zeta\eta}^{(6)}}.$$

As $\|\cdot\|_{\boldsymbol{V}}^2 = a(\cdot,\cdot)$, we obtain equivalence of norms, which concludes the proof. □

7.2 Improved stability estimates for interpolants in time

We define piecewise linear interpolations of the discrete structural velocities and the pore pressure. For $t \in (t_{n-1}, t_n]$, $n \geq 1$, let

$$\hat{\boldsymbol{u}}_{t,h\tau}(t) := \frac{\boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2}}{\tau} + \frac{t - t_{n-1}}{\tau} \frac{\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}}{\tau}, \tag{7.2}$$

$$\hat{p}_{\mathrm{pore},h\tau}(t) := \hat{p}_{\mathrm{pore}}(\chi_h^{n-1}) + \frac{t - t_{n-1}}{\tau} \left(\hat{p}_{\mathrm{pore}}(\chi_h^n) - \hat{p}_{\mathrm{pore}}(\chi_h^{n-1}) \right). \tag{7.3}$$

Note that $\partial_t \hat{\boldsymbol{u}}_{h\tau}$ defines the piecewise constant analog of $\hat{\boldsymbol{u}}_{t,h\tau}$. Stability bounds are obtained as direct consequence of Lemma 7.1 and Lemma 7.2.

Lemma 7.3 (Stability estimate for interpolations of the structural velocity)**.** *Let* $\hat{\boldsymbol{u}}_{h\tau}$ *and* $\hat{\boldsymbol{u}}_{t,h\tau}$*, as defined by* (6.2) *and* (7.2)*. For all* $h, \tau > 0$ *and* $\hat{\tau} \in (0, \tau)$*, it holds that*

$$\|\partial_t \hat{\boldsymbol{u}}_{h\tau}\|^2_{L^2(0,T;\boldsymbol{V})} \leq C^{(6)}_{\zeta\eta}, \tag{7.4}$$

$$\int_0^{T-\hat{\tau}} \|\partial_t \hat{\boldsymbol{u}}_{h\tau}(t + \hat{\tau}) - \partial_t \hat{\boldsymbol{u}}_{h\tau}(t)\|^2 \, dt \leq C^2_{\Omega,\mathrm{PK}} C^{(6)}_{\zeta\eta} \hat{\tau}, \tag{7.5}$$

$$\|\hat{\boldsymbol{u}}_{t,h\tau}\|^2_{L^2(0,T;\boldsymbol{V})} \leq 2C^{(6)}_{\zeta\eta}, \tag{7.6}$$

$$\|\hat{\boldsymbol{u}}_{t,h\tau} - \partial_t \hat{\boldsymbol{u}}_{h\tau}\|^2_{L^2(Q_T)} \leq \frac{C^2_{\Omega,\mathrm{PK}} C^{(7)}_{\zeta\eta}}{\zeta} \tau^2, \tag{7.7}$$

$$\|\partial_t \hat{\boldsymbol{u}}_{t,h\tau}\|^2_{L^2(0,T;\boldsymbol{V})} \leq \frac{C^{(7)}_{\zeta\eta}}{\zeta}, \tag{7.8}$$

where $C^{(6)}_{\zeta\eta}$ *and* $C^{(7)}_{\zeta\eta}$ *are the stability constants from Lemma 7.1 and Lemma 7.2, respectively, and* $C_{\Omega,\mathrm{PK}}$ *is the product of the Poincaré and the Korn constants.*

Proof. By construction, it holds that

$$\|\partial_t \hat{\boldsymbol{u}}_{h\tau}\|^2_{L^2(0,T;\boldsymbol{V})} = \sum_{n=1}^N \tau^{-1} \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\|^2_{\boldsymbol{V}}.$$

Hence, (7.4) follows directly from Lemma 7.1. The time-translation property (7.5) follows from the fact that $\partial_t \hat{\boldsymbol{u}}_{h\tau}$ is piecewise constant. Analogous to the proof of Lemma 6.6, one can show

$$\int_0^{T-\tau} \|\partial_t \hat{\boldsymbol{u}}_{h\tau}(t + \tau) - \partial_t \hat{\boldsymbol{u}}_{h\tau}(t)\|^2 \, dt = \hat{\tau} \sum_{n=1}^N \left\| \tau^{-1} \left(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right) - \tau^{-1} \left(\boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2} \right) \right\|^2.$$

Finally, after using a Poincaré inequality and Korn's inequality, (7.5) follows from Lemma 7.1.

In order to show (7.6), we expand the integral over the time interval. By definition of $\hat{\boldsymbol{u}}_{t,h\tau}$, it holds that

$$\begin{aligned}
&\|\hat{\boldsymbol{u}}_{t,h\tau}\|^2_{L^2(0,T;\boldsymbol{V})} \\
&\quad = \sum_{n=1}^N \tau^{-2} \int_{t_{n-1}}^{t_n} \left\| \left(\boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2} \right) + \frac{t - t_{n-1}}{\tau} \left(\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2} \right) \right\|^2_{\boldsymbol{V}} dt \\
&\quad \leq 2 \sum_{n=2}^N \tau^{-2} \int_{t_{n-1}}^{t_n} \left(\frac{t - t_n}{\tau} \right)^2 \left\| \boldsymbol{u}_h^{n-1} - \boldsymbol{u}_h^{n-2} \right\|^2_{\boldsymbol{V}} dt \\
&\qquad + 2 \sum_{n=2}^N \tau^{-2} \int_{t_{n-1}}^{t_n} \left(\frac{t - t_{n-1}}{\tau} \right)^2 \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\|^2_{\boldsymbol{V}} dt \\
&\quad \leq \frac{4}{3} \sum_{n=1}^N \tau^{-1} \left\| \boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1} \right\|^2_{\boldsymbol{V}}.
\end{aligned}$$

Hence, (7.6) follows by Lemma 7.1. In order to show (7.7), we again expand the integral over the time interval. By definition of $\hat{\boldsymbol{u}}_{t,h\tau}$ and $\hat{\boldsymbol{u}}_{h\tau}$, it holds that

$$\begin{aligned}
&\|\hat{\boldsymbol{u}}_{t,h\tau} - \partial_t \hat{\boldsymbol{u}}_{h\tau}\|^2_{L^2(Q_T)} \\
&\quad = \sum_{1=2}^{N} \int_{t_{n-1}}^{t_n} \left(1 - \frac{t - t_{n-1}}{\tau}\right)^2 \tau^{-2} \left\|\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}\right\|^2 \, dt \\
&\quad = \frac{1}{3} \sum_{n=1}^{N} \tau^{-1} \left\|\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}\right\|^2 .
\end{aligned}$$

Hence, after employing a Poincaré inequality and Korn's inequality, (7.7) follows from Lemma 7.2. Finally, (7.8) follows directly from Lemma 7.2, since

$$\|\partial_t \hat{\boldsymbol{u}}_{t,h\tau}\|^2_{L^2(0,T;\boldsymbol{V})} = \sum_{n=1}^{N} \tau \left\| \frac{\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}}{\tau^2} \right\|^2_{\boldsymbol{V}} .$$

□

Lemma 7.4 (Stability result for the interpolation of the pore pressure)**.** *For* $\hat{p}_{\text{pore},h\tau}$ *defined in* (7.3)*. It holds that*

$$\begin{aligned}
\|\partial_t \hat{p}_{\text{pore},h\tau}\|^2_{L^2(Q_T)} &\leq C^{(6)}_{\zeta\eta}, \\
\|\hat{p}_{\text{pore},h\tau} - \hat{p}_{\text{pore}}(\bar{\chi}_{h\tau})\|^2_{L^2(Q_T)} &\leq C^{(6)}_{\zeta\eta} \tau^2,
\end{aligned}$$

where $C^{(6)}_{\zeta\eta}$ *is the stability constant from Lemma 7.1.*

Proof. By construction, it holds that

$$\begin{aligned}
\|\partial_t \hat{p}_{\text{pore},h\tau}\|^2_{L^2(Q_T)} &= \sum_{n=1}^{N} \tau^{-1} \|\hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1})\|^2, \qquad \text{and} \\
\|\hat{p}_{\text{pore},h\tau} - \hat{p}_{\text{pore}}(\bar{\chi}_{h\tau})\|^2_{L^2(Q_T)} &= \sum_{n=1}^{N} \frac{\tau}{3} \|\hat{p}_{\text{pore}}(\chi_h^n) - \hat{p}_{\text{pore}}(\chi_h^{n-1})\|^2,
\end{aligned}$$

where the second result follows by expanding time integration. Hence, the assertion follows directly from the stability result for the discrete time derivative of the pore pressure, cf. Lemma 7.1. □

7.3 More relative (weak) compactness for $h, \tau \to 0$

The previous stability results allow for analyzing the limit in relation to $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$.

Lemma 7.5 (Convergence of the structural velocity and acceleration)**.** *We can extract subsequences of* $\{\hat{\boldsymbol{u}}_{h\tau}\}_{h,\tau}$ *and* $\{\hat{\boldsymbol{u}}_{t,h\tau}\}_{h,\tau}$ *(still denoted like the original sequences) such that* $\partial_t \boldsymbol{u}_{\varepsilon\eta}, \partial_{tt} \boldsymbol{u}_{\varepsilon\eta} \in L^2(0,T;\boldsymbol{V})$ *and*

$$\partial_t \hat{\boldsymbol{u}}_{h\tau} \rightharpoonup \partial_t \boldsymbol{u}_{\varepsilon\eta}, \quad \text{in } L^2(0,T;\boldsymbol{V}), \tag{7.9}$$

$$\partial_t \hat{\boldsymbol{u}}_{t,h\tau} \rightharpoonup \partial_{tt} \boldsymbol{u}_{\varepsilon\eta}, \quad \text{in } L^2(0,T;\boldsymbol{V}). \tag{7.10}$$

Proof. The convergence result (7.9) follows from the stability result (7.4), the Eberlein-Šmulian theorem, cf. Lemma B.8, and the fact that $\hat{\boldsymbol{u}}_{h\tau} \rightharpoonup \boldsymbol{u}_{\varepsilon\eta}$ in $L^2(0,T;\boldsymbol{V})$, cf. Lemma 6.10. Furthermore, due to the additional translation property (7.5), by employing a relaxed Aubin-Lions-Simon type compactness result for Bochner spaces, cf. Lemma B.9, we can extract a further subsequence (still denoted the same)

$$\partial_t \hat{\boldsymbol{u}}_{h\tau} \to \partial_t \boldsymbol{u}_{\varepsilon\eta}, \quad \text{in } L^2(Q_T). \tag{7.11}$$

Using the stability result (7.8), by the Eberlein-Šmulian theorem, cf. Lemma B.8, we can extract a subsequence (still denoted the same) such that $\partial_t \hat{\boldsymbol{u}}_{t,h\tau} \rightharpoonup \boldsymbol{u}_{tt}$ in $L^2(0,T;\boldsymbol{V})$ for some $\boldsymbol{u}_{tt} \in L^2(0,T;\boldsymbol{V})$. It holds that $\boldsymbol{u}_{tt} = \partial_{tt}\boldsymbol{u}_{\varepsilon\eta}$ if also $\hat{\boldsymbol{u}}_{t,h\tau} \rightharpoonup \partial_t \boldsymbol{u}_{\varepsilon\eta}$ in $L^2(0,T;\boldsymbol{V})$. From the stability result (7.6), and the Eberlein-Šmulian theorem, cf. Lemma B.8, there exists a $\boldsymbol{u}_t \in L^2(0,T;\boldsymbol{V})$ such that $\hat{\boldsymbol{u}}_{t,h\tau} \rightharpoonup \boldsymbol{u}_t$ in $L^2(0,T;\boldsymbol{V})$ (up to a subsequence). Employing the triangle inequality, yields

$$\|\hat{\boldsymbol{u}}_{t,h\tau} - \partial_t \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(Q_T)} \leq \|\hat{\boldsymbol{u}}_{t,h\tau} - \partial_t \hat{\boldsymbol{u}}_{h\tau}\|_{L^2(Q_T)} + \|\partial_t \hat{\boldsymbol{u}}_{h\tau} - \partial_t \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(Q_T)}.$$

Hence, due to (7.7) and (7.11), it holds that $\boldsymbol{u}_t = \partial_t \boldsymbol{u}_{\varepsilon\eta}$, and consequently $\boldsymbol{u}_{tt} = \partial_{tt}\boldsymbol{u}_{\varepsilon\eta}$, concluding the proof. □

Lemma 7.6 (Convergence of the time derivative of the pore pressure). *There exists a subsequence of $\{\hat{p}_{\mathrm{pore},h\tau}\}_{h,\tau}$ (still denoted $\{\hat{p}_{\mathrm{pore},h\tau}\}_{h,\tau}$) satisfying*

$$\partial_t \hat{p}_{\mathrm{pore},h\tau} \rightharpoonup \partial_t \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta}), \ \text{in } L^2(Q_T).$$

Proof. By Lemma 6.11, we have $\bar{\chi}_{h\tau} \to \chi_{\varepsilon\eta}$ in $L^2(Q_T)$ (up to a subsequence). Hence, also $\hat{p}_{\mathrm{pore}}(\bar{\chi}_{h\tau}) \to \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta})$ in $L^2(Q_T)$ (up to a subsequence). From Lemma 7.4, it follows $\hat{p}_{\mathrm{pore},h\tau} \to \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta})$ and $\partial_t \hat{p}_{\mathrm{pore},h\tau} \rightharpoonup p_t$ for some $p_t \in L^2(Q_T)$ (up to a subsequence). Consequently, $p_t = \partial_t \hat{p}_{\mathrm{pore}}(\chi_{\varepsilon\eta})$, which concludes the proof. □

7.4 Identifying a weak solution with increased regularity for $h, \tau \to 0$

Finally, we show the limit $(\boldsymbol{u}_{\varepsilon\eta}, \chi)$, derived in Section 6.3, also satisfies (W5)$_{\zeta\eta}$–(W6)$_{\zeta\eta}$, i.e., $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ is a weak solution with increased regularity for the doubly regularized unsaturated poroelasticity model, cf. Definition 4.1.

Lemma 7.7 (Limit satisfies (W1)$_{\zeta\eta}$–(W6)$_{\zeta\eta}$). *The limit $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$, derived in Section 6.3, is a weak solution with increased regularity for the doubly regularized unsaturated poroelasticity model, cf. Definition 4.1.*

Proof. The limit $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ satisfies (W1)$_{\zeta\eta}$–(W4)$_{\zeta\eta}$ by Lemma 9.7. Furthermore, (W5)$_{\zeta\eta}$ follows directly from Lemma 7.5 and Lemma 7.6. In order to show (W6)$_{\zeta\eta}$, let $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V} \cap C^\infty(\Omega)^d)$. We utilize $\bar{\boldsymbol{v}}_{h\tau}$ and $\boldsymbol{v}_h^n$, as introduced in (6.18) and (6.20), respectively; again it holds that

$$\bar{\boldsymbol{v}}_{h\tau} \to \boldsymbol{v} \quad \text{in } L^2(0,T;\boldsymbol{V}). \tag{7.12}$$

We consider the difference of the mechanics equation (5.3) at time steps n and $n-1$, $n \geq 1$, tested with $\boldsymbol{v}_h = \boldsymbol{v}_h^n$; we obtain

$$\begin{aligned} &\zeta\tau^{-1} a(\boldsymbol{u}_h^n - 2\boldsymbol{u}_h^{n-1} + \boldsymbol{u}_h^{n-2}, \boldsymbol{v}_h^n) + a(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h^n) \\ &\quad - \alpha \left\langle \hat{p}_{\mathrm{pore}}(\chi_h^n) - \hat{p}_{\mathrm{pore}}(\chi_h^{n-1}), \boldsymbol{\nabla} \cdot \boldsymbol{v}_h^n \right\rangle = \left\langle \boldsymbol{f}_{\mathrm{ext}}^n - \boldsymbol{f}_{\mathrm{ext}}^{n-1}, \boldsymbol{v}_h^n \right\rangle. \end{aligned}$$

Summing over $n \in \{1, ..., N\}$, and employing the definitions of $\bar{\boldsymbol{v}}_{h\tau}$, $\hat{\boldsymbol{u}}_{t,h\tau}$, $\hat{\boldsymbol{u}}_{h\tau}$, and $\hat{p}_{\mathrm{pore},h\tau}$, yields

$$\int_0^T \Big[\zeta a(\partial_t \hat{\boldsymbol{u}}_{t,h\tau}, \bar{\boldsymbol{v}}_{h\tau}) + a(\partial_t \hat{\boldsymbol{u}}_{h\tau}, \bar{\boldsymbol{v}}_{h\tau}) - \alpha \left\langle \partial_t \hat{p}_{\mathrm{pore},h\tau}, \boldsymbol{\nabla} \cdot \bar{\boldsymbol{v}}_{h\tau} \right\rangle \Big] \, dt = \int_0^T \Big\langle \partial_t \hat{\boldsymbol{f}}_\tau, \bar{\boldsymbol{v}}_{h\tau} \Big\rangle \, dt, \tag{7.13}$$

where $\hat{\boldsymbol{f}}_\tau$ denotes the piecewise linear interpolation of the discrete values $\{\boldsymbol{f}_{\text{ext}}^n\}_n$

$$\hat{\boldsymbol{f}}_{\text{ext},\tau}(t) := \boldsymbol{f}_{\text{ext}}^{n-1} + \frac{t-t_{n-1}}{\tau}\left(\boldsymbol{f}_{\text{ext}}^n - \boldsymbol{f}_{\text{ext}}^{n-1}\right), \quad t \in (t_{n-1}, t_n].$$

It holds $\hat{\boldsymbol{f}}_{\text{ext},\tau} \to \boldsymbol{f}_{\text{ext}}$ in $L^2(0,T;\boldsymbol{V}^\star)$ and also $\partial_t \hat{\boldsymbol{f}}_{\text{ext},\tau} \rightharpoonup \partial_t \boldsymbol{f}_{\text{ext}}$ in $L^2(0,T;\boldsymbol{V}^\star)$, for $\tau \to 0$. Hence, together with the weak convergence properties of $\hat{\boldsymbol{u}}_{t,h\tau}$, $\hat{\boldsymbol{u}}_{h\tau}$ and $\hat{p}_{\text{pore},h\tau}$, cf. Lemma 7.5 and Lemma 7.6, and the strong convergence properties of the test function $\bar{\boldsymbol{v}}_{h\tau}$, cf. (7.12), we conclude that

$$\int_0^T \Big[\zeta a(\partial_{tt}\boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) + a(\partial_t \boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) - \alpha \left\langle \partial_t \hat{p}_{\text{pore}}(\chi_{\varepsilon\eta}), \boldsymbol{\nabla}\cdot\boldsymbol{v}\right\rangle\Big]\, dt = \int_0^T \left\langle \partial_t \boldsymbol{f}_{\text{ext}}, \boldsymbol{v}\right\rangle\, dt,$$

for all $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V}\cap C^\infty(\Omega)^d)$. A density argument yields the final result. □

8 Step 5: Limit $\zeta \to 0$

In this section, we prove Lemma 4.5, i.e., the existence of weak solution to the simply regularized unsaturated poroelasticity model, cf. Definition 4.2. For this we utilize the fact that under the assumptions of Lemma 4.5, there exists weak solution, $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$, with increased regularity for the doubly regularized unsaturated poroelasticity model, cf. Definition 4.1. We show that $\{(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})\}_\zeta$ has a limit for $\zeta \to 0$, which is a weak solution to the simply regularized unsaturated poroelasticity model, i.e., it satisfies $(\text{W1})_\eta$–$(\text{W4})_\eta$ for $\zeta = 0$. For this, we employ compactness arguments. The central uniform stability bound is derived utilizing $(\text{W6})_{\zeta\eta}$ and the non-degeneracy condition (ND3). Throughout the entire section, we assume (A0)–(A9) and (ND1)–(ND3) hold true.

8.1 Stability estimates independent of ζ

The key ingredients for the subsequent discussion are stability estimates, which are independent of ζ. In Section 6.2, some derived stability results are independent of ζ; they remain true for weak limits. In particular, there exists a constant $C = C\left(C^{(1)}, C^{(4)}\right) > 0$ (independent of $\zeta > 0$ and $\eta > 0$), such that

$$\|\boldsymbol{u}_{\varepsilon\eta}\|_{L^\infty(0,T;\boldsymbol{V})}^2 + \|p_{\text{pore}}(\chi_{\varepsilon\eta})\|_{L^2(Q_T)}^2 \le C, \tag{8.1}$$

where Lemma 6.6 and Lemma 6.10 yield stability for the displacement, and Lemma 6.8 and Lemma 6.12 yield stability for the pore pressure. Further stability bounds can be obtained by exploiting the continuous nature of the balance equations and the time derivative of the mechanics equation. The following stability estimate is the essential step.

Lemma 8.1 (Stability for the primal variables)**.** *There exists a constant $C^{(8)} > 0$ (independent of ζ and η), such that*

$$\begin{aligned}\zeta\, \|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|_{L^\infty(0,T;\boldsymbol{V})}^2 + \|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(0,T;\boldsymbol{V})}^2 + \|\boldsymbol{\nabla}\chi_{\varepsilon\eta}\|_{L^\infty(0,T;L^2(\Omega))}^2 \\ \le C^{(8)}\left(C_0, \|\partial_t \boldsymbol{f}_{\text{ext}}\|_{L^2(0,T;\boldsymbol{V}^\star)}^2, \|h_{\text{ext}}\|_{H^1(0,T;Q^\star)}^2\right),\end{aligned}$$

where C_0 comes from (A8*)*.*

Proof. Consider the flow equation (4.4) and the mechanics equation differentiated in time (4.5), tested with $q = \partial_t \chi_{\varepsilon\eta}$ and $\boldsymbol{v} = \partial_t \boldsymbol{u}_{\varepsilon\eta}$, respectively. Summing both equation yields

$$\begin{aligned} &\zeta \int_0^T a(\partial_{tt}\boldsymbol{u}_{\varepsilon\eta}, \partial_t \boldsymbol{u}_{\varepsilon\eta})\, dt + \int_0^T \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}, \boldsymbol{\nabla} \partial_t \chi_{\varepsilon\eta} \rangle\, dt \\ &\quad + \|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|^2_{L^2(0,T;\boldsymbol{V})} + \int_0^T \left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), \partial_t \chi_{\varepsilon\eta} \right\rangle + \alpha \int_0^T \langle \hat{s}_{\mathrm{w}} \partial_t \chi_{\varepsilon\eta} - \partial_t \hat{p}_{\mathrm{pore}}, \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta} \rangle \\ &\quad = \int_0^T \langle \partial_t \boldsymbol{f}_{\mathrm{ext}}, \partial_t \boldsymbol{u}_{\varepsilon\eta} \rangle\, dt + \int_0^T \langle h_{\mathrm{ext}}, \partial_t \chi_{\varepsilon\eta} \rangle\, dt. \end{aligned} \tag{8.2}$$

We discuss the individual terms separately. For the first two terms on the left hand side of (8.2), we employ the fundamental theorem of calculus

$$\begin{aligned} &\zeta \int_0^T a(\partial_{tt}\boldsymbol{u}_{\varepsilon\eta}, \partial_t \boldsymbol{u}_{\varepsilon\eta})\, dt + \int_0^T \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}, \boldsymbol{\nabla} \partial_t \chi_{\varepsilon\eta} \rangle\, dt \\ &\quad = \frac{\zeta}{2} \|\partial_t \boldsymbol{u}_{\varepsilon\eta}(T)\|^2_{L^2(0,T;\boldsymbol{V})} + \frac{1}{2} \left(\langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T) \rangle - \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0) \rangle \right), \end{aligned}$$

where we used that $\partial_t \boldsymbol{u}_{\varepsilon\eta}(0) = \boldsymbol{0}$, following from the temporal derivative of the mechanics equations (4.5) and the compatibility condition for the initial conditions (A9).

For the remaining terms on the left hand side of (8.2), we employ the fact that $\hat{b}_\eta$ is increasing with $\hat{b}'_\eta \geq \hat{b}'$, that $a(\boldsymbol{v}, \boldsymbol{v}) \geq K_{\mathrm{dr}} \|\boldsymbol{\nabla} \cdot \boldsymbol{v}\|^2$ for all $\boldsymbol{v} \in \boldsymbol{V}$ with $K_{\mathrm{dr}} = \frac{2\mu}{d} + \lambda$, and (ND3). Starting with a binomial identity, we obtain

$$\begin{aligned} &\|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|^2_{L^2(0,T;\boldsymbol{V})} + \int_0^T \left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), \partial_t \chi_{\varepsilon\eta} \right\rangle + \alpha \int_0^T \langle \hat{s}_{\mathrm{w}} \partial_t \chi_{\varepsilon\eta} - \partial_t \hat{p}_{\mathrm{pore}}, \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta} \rangle \\ &\quad = \|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|^2_{L^2(0,T;\boldsymbol{V})} - \frac{\alpha^2}{4} \int_0^T \int_\Omega \left(\frac{s_{\mathrm{w}}(\chi_{\varepsilon\eta})}{\hat{p}'_{\mathrm{pore}}(\chi_{\varepsilon\eta})} - 1 \right)^2 \frac{\left(\hat{p}'_{\mathrm{pore}}(\chi_{\varepsilon\eta})\right)^2}{\hat{b}'_\eta(\chi_{\varepsilon\eta})} \left| \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta} \right|^2\, dx\, dt \\ &\qquad + \int_0^T \int_\Omega \left[\left(\partial_t \hat{b}_\eta \partial_t \chi_{\varepsilon\eta} \right)^{1/2} + \frac{\alpha}{2} \left(\hat{s}_{\mathrm{w}} \partial_t \chi_{\varepsilon\eta} - \partial_t \hat{p}_{\mathrm{pore}} \right) \left(\partial_t \hat{b}_\eta \partial_t \chi_{\varepsilon\eta} \right)^{-1/2} \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta} \right]^2 dx\, dt \\ &\quad \geq (1 - C_{\mathrm{ND},3}) \|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|^2_{L^2(0,T;\boldsymbol{V})}. \end{aligned}$$

For the first term on the right hand side of (8.2), we apply the Cauchy-Schwarz inequality and Young's inequality

$$\int_0^T \langle \partial_t \boldsymbol{f}_{\mathrm{ext}}, \partial_t \boldsymbol{u}_{\varepsilon\eta} \rangle\, dt \leq \frac{1}{2(1 - C_{\mathrm{ND},3})} \|\partial_t \boldsymbol{f}_{\mathrm{ext}}\|^2_{L^2(0,T;\boldsymbol{V}^\star)} + \frac{1 - C_{\mathrm{ND},3}}{2} \|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|^2_{L^2(0,T;\boldsymbol{V})}.$$

For the second term on the right hand side of (8.2), we apply integration by parts, a Cauchy-Schwarz inequality and Young's inequality, a Poincaré inequality (introducing the Poincaré constant $C_{\Omega,\mathrm{P}}$) and a Sobolev embedding (introducing the constant $C_{\mathrm{T,Sob}}$), as well as (A6). All in all, we obtain

$$\begin{aligned}&\int_0^T \langle h_{\mathrm{ext}}, \partial_t \chi_{\varepsilon\eta}\rangle \, dt \\ &\quad = \langle h_{\mathrm{ext}}(T), \chi_{\varepsilon\eta}(T)\rangle - \langle h_{\mathrm{ext}}(0), \chi(0)\rangle - \int_0^T \langle \partial_t h_{\mathrm{ext}}, \chi_{\varepsilon\eta}\rangle \, dt \\ &\quad \leq \frac{C_{\Omega,\mathrm{P}}^2}{\kappa_{\mathrm{m,abs}}} \left(\|h_{\mathrm{ext}}(T)\|^2 + \|h_{\mathrm{ext}}(0)\|^2 + \|\partial_t h_{\mathrm{ext}}\|_{L^2(Q_T)}^2 \right) \\ &\qquad + \frac{\kappa_{\mathrm{m,abs}}}{4C_{\Omega,\mathrm{P}}^2} \left(\|\chi_{\varepsilon\eta}(T)\|^2 + \|\chi_{\varepsilon\eta}(0)\|^2 + \|\chi_{\varepsilon\eta}\|_{L^2(Q_T)}^2 \right) \\ &\quad \leq \frac{3\left(C_{\mathrm{T,Sob}} C_{\Omega,\mathrm{P}}\right)^2}{\kappa_{\mathrm{m,abs}}} \|h_{\mathrm{ext}}\|_{H^1(0,T;Q^\star)}^2 \\ &\qquad + \frac{1}{4}\bigg(\langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T)\rangle + \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0)\rangle + \int_0^T \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}, \boldsymbol{\nabla} \chi_{\varepsilon\eta}\rangle \, dt \bigg).\end{aligned}$$

Altogether, (8.2) becomes

$$\begin{aligned}&\frac{\zeta}{2} \|\partial_t \boldsymbol{u}_{\varepsilon\eta}(T)\|_{L^2(0,T;\boldsymbol{V})}^2 + \frac{1}{4} \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T)\rangle + \frac{1 - C_{\mathrm{ND},3}}{2} \|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(0,T;\boldsymbol{V})}^2 \\ &\quad \leq \frac{3}{4} \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0)\rangle + \frac{1}{2(1 - C_{\mathrm{ND},3})} \|\partial_t \boldsymbol{f}_{\mathrm{ext}}\|_{L^2(0,T;\boldsymbol{V}^\star)}^2 \\ &\qquad + \frac{3\left(C_{\mathrm{T,Sob}} C_{\Omega,\mathrm{P}}\right)^2}{\kappa_{\mathrm{m,abs}}} \|h_{\mathrm{ext}}\|_{H^1(0,T;Q^\star)}^2 + \frac{1}{4}\int_0^T \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}, \boldsymbol{\nabla} \chi_{\varepsilon\eta}\rangle \, dt.\end{aligned}$$

Applying a Grönwall inequality proves the assertion under the given assumptions. □

The last stability estimate allows for deriving further stability estimates.

Lemma 8.2 (Stability for the Legendre transformation of $\hat{b}_\eta$)**.** *There exists a constant $C^{(9)} > 0$ (independent of ζ, η), such that*

$$\left\| \hat{B}_\eta(\chi_{\varepsilon\eta}) \right\|_{L^\infty(0,T;L^1(\Omega))} \leq C^{(9)}\left(C_0, C^{(8)}\right),$$

where $C^{(8)}$ is the stability constant from Lemma 8.1, and C_0 is the stability constant from (A8*).

Proof. Testing the flow equation (4.4) with $q = \chi_{\varepsilon\eta}$, yields

$$\int_0^T \left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), \chi_{\varepsilon\eta} \right\rangle dt + \int_0^T \|\boldsymbol{\nabla} \chi_{\varepsilon\eta}\|_{\kappa_{\mathrm{abs}}}^2 \, dt = \int_0^T \langle h_{\mathrm{ext}}, \chi_{\varepsilon\eta}\rangle \, dt - \alpha \int_0^T \langle s_{\mathrm{w}} \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta}\rangle \, dt.$$

For the first term on the left hand side, we apply an identity for Legendre transformations, cf. [53],

$$\int_0^T \left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), \chi_{\varepsilon\eta} \right\rangle dt = \left\| \hat{B}_\eta(\chi_{\varepsilon\eta}(T)) \right\|_{L^1(\Omega)} - \left\| \hat{B}_\eta(\chi_0) \right\|_{L^1(\Omega)},$$

where $\hat{B}_\eta$ is the Legendre transformation for $\hat{b}_\eta$. On the right hand side, we apply the Cauchy-Schwarz inequality, Young's inequality, a Poincaré inequality (introducing $C_{\Omega,\mathrm{P}}$) and (A6), and obtain

$$\begin{aligned}&\left\| \hat{B}_\eta(\chi_{\varepsilon\eta}(T)) \right\|_{L^1(\Omega)} + \frac{1}{2}\int_0^T \|\boldsymbol{\nabla} \chi_{\varepsilon\eta}\|_{\kappa_{\mathrm{abs}}}^2 \, dt \\ &\quad \leq \left\| \hat{B}_\eta(\chi_0) \right\|_{L^1(\Omega)} + \frac{C_{\Omega,\mathrm{P}}^2}{\kappa_{\mathrm{m,abs}}} \left(\|h_{\mathrm{ext}}\|_{L^2(0,T;Q^\star)}^2 + \alpha^2 \|\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(Q_T)}^2 \right).\end{aligned}$$

Finally, the thesis follows from Lemma 8.1. □

Lemma 8.3 (Stability for the temporal change of $\hat{b}_\eta$)**.** *There exists a constant $C^{(10)} > 0$ (independent of ζ, η), such that*

$$\sup_{0 \neq q \in L^2(0,T;Q)} \frac{\int_0^T \left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), q \right\rangle \, dt}{\|\boldsymbol{\nabla} q\|_{L^2(Q_T)}} \leq C^{(10)} \left(C^{(8)} \right),$$

where $C^{(8)}$ is the stability constant from Lemma 8.1.

Proof. The proof is analog to the proof of Lemma 6.5. However, this time, we exploit

$$\|\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(Q_T)} \leq \frac{1}{K_{\mathrm{dr}}^{1/2}} \|\partial_t \boldsymbol{u}_{\varepsilon\eta}\|_{L^2(0,T;\boldsymbol{V})} \leq \frac{C^{(8)}}{K_{\mathrm{dr}}^{1/2}}$$

by Lemma 8.1. Thus, we drop the dependence on ζ. □

We will require to show strong convergence of the Kirchhoff pressure. Having that in mind, we conclude with a stability estimate for $\partial_t \chi_{\varepsilon\eta}$. We note, this is the only stability estimate in this section, requiring the regularizing growth condition (A1*).

Lemma 8.4 (Stability estimate for the temporal change of the Kirchhoff pressure)**.** *There exists a constant $C_\eta^{(11)} > 0$ (independent of ζ), such that*

$$\|\partial_t \chi_{\varepsilon\eta}\|^2_{L^2(Q_T)} \leq C_\eta^{(11)} \left(b_{\chi,\mathrm{m}}^{-1} C_0, b_{\chi,\mathrm{m}}^{-2} C^{(8)} \right),$$

where $C^{(8)}$ is the stability constant from Lemma 8.1, $b_{\chi,\mathrm{m}}$ is from (A1*)*, and C_0 is from* (A8*)*.*

Proof. We repeat parts of the proof of Lemma 8.1. We test the flow equation (4.4) with $q = \partial_t \chi_{\varepsilon\eta}$ and apply (A1*) and the Cauchy-Schwarz inequality; we obtain

$$\begin{aligned}
& b_{\chi,\mathrm{m}} \|\partial_t \chi_{\varepsilon\eta}\|^2_{L^2(Q_T)} + \frac{1}{2} \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T) \rangle \\
& \leq \int_0^T \left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), \partial_t \chi_{\varepsilon\eta} \right\rangle dt + \frac{1}{2} \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(T) \rangle \\
& = \frac{1}{2} \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0) \rangle + \int_0^T \left(\langle h_{\mathrm{ext}}, \partial_t \chi_{\varepsilon\eta} \rangle - \alpha \langle \hat{s}_{\mathrm{w}} \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta}, \partial_t \chi_{\varepsilon\eta} \rangle \right) dt \\
& \leq \frac{1}{2} \langle \kappa_{\mathrm{abs}} \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0), \boldsymbol{\nabla} \chi_{\varepsilon\eta}(0) \rangle + \frac{1}{b_{\chi,\mathrm{m}}} \left(\|h_{\mathrm{ext}}\|^2_{L^2(0,T;Q^\star)} + \alpha^2 \|\partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta}\|^2_{L^2(Q_T)} \right) \\
& \quad + \frac{b_{\chi,\mathrm{m}}}{2} \|\partial_t \chi_{\varepsilon\eta}\|^2_{L^2(Q_T)}.
\end{aligned}$$

After rearranging terms, applying the regularity of the data, and applying from Lemma 8.1, the assertion follows. □

8.2 Relative (weak) compactness for $\zeta \to 0$

We utilize the stability results from the previous section to conclude relative compactness.

Lemma 8.5 (Convergence of the primary variables)**.** *We can extract subsequences of $\{\boldsymbol{u}_{\varepsilon\eta}\}_\zeta$ and $\{\chi_{\varepsilon\eta}\}_\zeta$ (still denoted like the original sequences), and there exist $\boldsymbol{u}_\eta \in H^1(0,T;\boldsymbol{V})$ and $\chi_\eta \in H^1(0,T;L^2(\Omega)) \cap L^\infty(0,T;Q)$ such that for $\zeta \to 0$*

$$\boldsymbol{u}_{\varepsilon\eta} \rightharpoonup \boldsymbol{u}_\eta \quad \text{in } H^1(0,T;\boldsymbol{V}), \tag{8.3}$$

$$\zeta \partial_t \boldsymbol{u}_{\varepsilon\eta} \to \boldsymbol{0} \quad \text{in } L^2(0,T;\boldsymbol{V}), \tag{8.4}$$

$$\chi_{\varepsilon\eta} \to \chi_\eta \quad \text{in } L^2(Q_T), \tag{8.5}$$

$$\chi_{\varepsilon\eta} \rightharpoonup \chi_\eta \quad \text{in } L^\infty(0,T;Q), \tag{8.6}$$

$$\partial_t \chi_{\varepsilon\eta} \rightharpoonup \partial_t \chi_\eta \text{ in } L^2(Q_T). \tag{8.7}$$

Proof. The proof follows standard arguments based on the Eberlein-Šmulian theorem, cf. Lemma B.8, the Aubin-Lions lemma, cf. Lemma B.9, and the stability results for $\boldsymbol{u}_{\varepsilon\eta}$, cf. Lemma 8.1 and (8.1), as well as the stability results for $\chi_{\varepsilon\eta}$, cf. Lemma 8.1 and Lemma 8.4. In particular, for (8.4), we employ the uniform stability result from Lemma 8.1; for all fixed $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V})$ it holds that

$$\left| \int_0^T \zeta a(\partial_t \boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) \, dt \right| \leq \zeta C^{(8)} \|\boldsymbol{v}\|_{L^2(0,T;\boldsymbol{V})} \to 0 \quad \text{for } \zeta \to 0.$$

□

Lemma 8.6 (Convergence of the coupling terms)**.** *Up to subsequences it holds for* $\zeta \to 0$

$$\hat{p}_{\text{pore}}(\chi_{\varepsilon\eta}) \rightharpoonup \hat{p}_{\text{pore}}(\chi_\eta) \qquad \textit{in } L^2(Q_T), \tag{8.8}$$

$$\hat{s}_{\text{w}}(\chi_{\varepsilon\eta}) \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta} \rightharpoonup \hat{s}_{\text{w}}(\chi_\eta) \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_\eta \quad \textit{in } L^2(Q_T). \tag{8.9}$$

Proof. The proof is analogous to the proof of Lemma 6.12. Essentially, first, one has to utilize stability estimates together with the Eberlein-Šmulian theorem, cf. Lemma B.8; second, continuity properties of the non-linearities have to be employed together with the convergence of $\{\boldsymbol{u}_{\varepsilon\eta}\}_\zeta$ and $\{\chi_{\varepsilon\eta}\}_\zeta$, cf. Lemma 8.5. We note that for (8.8) the stability result (8.1) has to be utilized. □

Lemma 8.7 (Initial conditions for the fluid flow)**.** *Up to subsequences it holds for* $\zeta \to 0$

$$\partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}) \rightharpoonup \partial_t \hat{b}_\eta(\chi_\eta) \;\; \textit{in } L^2(0,T;Q^\star), \tag{8.10}$$

where $\partial_t \hat{b}_\eta(\chi_\eta) \in L^2(0,T;Q^\star)$ *is understood in the sense of* $(\text{W2})_\eta$*.*

Proof. The proof is analogous to the proof of Lemma 6.13. By Lemma 8.3 and the Eberlein-Šmulian theorem, cf. Lemma B.8, there exists a $b_t \in L^2(0,T;Q^\star)$ such that $\partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}) \rightharpoonup b_t$ in $L^2(0,T;Q^\star)$ (up to a subsequence). We can identify $b_t = \partial_t b(\chi_\eta)$ by showing $(\text{W2})_\eta$. For this we utilize $(\text{W2})_{\zeta\eta}$. For $q \in L^2(0,T;Q)$ with $\partial_t q \in L^1(0,T:L^\infty(\Omega))$ and $q(T) = 0$, it holds that

$$\int_0^T \left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), q \right\rangle \, dt = \int_0^T \left\langle \hat{b}_\eta(\chi_0) - \hat{b}_\eta(\chi_{\varepsilon\eta}), \partial_t q \right\rangle \, dt.$$

The assertion follows immediately if

$$\hat{b}_\eta(\chi_{\varepsilon\eta}) \to \hat{b}_\eta(\chi_\eta) \quad \text{in } L^\infty(0,T;L^1(\Omega)) \tag{8.11}$$

(up to a subsequence). And indeed, by the uniform boundedness of the Legendre transformation, $\|\hat{B}_\eta(\chi_\eta)\|_{L^\infty(0,T;L^1(\Omega))}$, there exists $b_\chi \in L^\infty(0,T;L^1(\Omega))$ such that $\hat{b}_\eta(\chi_{\varepsilon\eta}) \rightharpoonup b_\chi$ in $L^\infty(0,T;L^1(\Omega))$. Using the strong convergence of $\{\chi_\eta\}_\eta$ and the dominated convergence theorem, we can identify $b_\chi = \hat{b}_\eta(\chi_\eta)$, and thus (8.11). □

Lemma 8.8 (Initial conditions of the mechanical displacement)**.** $\partial_t \boldsymbol{u}_\eta \in L^2(0,T;\boldsymbol{V})$ *satisfies* $(\text{W3})_\eta$*.*

Proof. Using the uniform stability bound for $\{\partial_t \boldsymbol{u}_{\varepsilon\eta}\}_\zeta$ by Lemma 8.1 and the weak convergence $\boldsymbol{u}_{\varepsilon\eta} \rightharpoonup \boldsymbol{u}_\eta$ in $L^2(0,T;\boldsymbol{V})$ (up to a subsequence) by Lemma 8.5, standard compactness arguments yield $\partial_t \boldsymbol{u}_{\varepsilon\eta} \rightharpoonup \partial_t \boldsymbol{u}_\eta$ in $L^2(0,T;\boldsymbol{V})$ (up to a subsequence). Hence, $\zeta \to 0$ of $(\text{W3})_{\zeta\eta}$ yields $(\text{W3})_\eta$ immediately. □

8.3 Identifying a weak solution for $\zeta \to 0$

Finally, we show the limit $(\boldsymbol{u}_\eta, \chi_\eta)$, introduced above, is a weak solution of the simply regularized unsaturated poroelasticity model.

Lemma 8.9 (Limit satisfies (W1)$_\eta$–(W4)$_\eta$)**.** *The limit $(\boldsymbol{u}_\eta, \chi_\eta)$, derived in Section 8.2, is a weak solution of the simply regularized unsaturated poroelasticity model, cf. Definition 4.2.*

Proof. The limit $(\boldsymbol{u}_\eta, \chi_\eta)$ satisfies (W1)$_\eta$–(W3)$_\eta$ by Lemma 8.5, Lemma 8.6, Lemma 8.7, and Lemma 8.8. It remains to show (W4)$_\eta$, i.e., that $(\boldsymbol{u}_\eta, \chi_\eta)$ satisfies the balance equations (4.3)–(4.4) for $\zeta = 0$. By definition, the sequence $(\boldsymbol{u}_{\varepsilon\eta}, \chi_{\varepsilon\eta})$ satisfies (W4)$_{\zeta\eta}$ for $\zeta > 0$, i.e., it holds for all $(\boldsymbol{v}, q) \in L^2(0,T;\boldsymbol{V}) \cap L^2(0,T;Q)$

$$\int_0^T \left[\zeta a(\partial_t \boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) + a(\boldsymbol{u}_{\varepsilon\eta}, \boldsymbol{v}) - \alpha \left\langle \hat{p}_\mathrm{pore}(\chi_{\varepsilon\eta}), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle\right] dt = \int_0^T \left\langle \boldsymbol{f}_\mathrm{ext}, \boldsymbol{v} \right\rangle \, dt,$$

$$\int_0^T \left[\left\langle \partial_t \hat{b}_\eta(\chi_{\varepsilon\eta}), q \right\rangle + \alpha \left\langle \hat{s}_\mathrm{w}(\chi_{\varepsilon\eta}) \partial_t \boldsymbol{\nabla} \cdot \boldsymbol{u}_{\varepsilon\eta}, q \right\rangle + \left\langle \kappa_\mathrm{abs} \boldsymbol{\nabla} \chi_{\varepsilon\eta}, \boldsymbol{\nabla} q \right\rangle \right] dt = \int_0^T \left\langle h_\mathrm{ext}, q \right\rangle \, dt.$$

Utilizing the weak convergence results, cf. Lemma 8.5 and Lemma 8.6, (W4)$_\eta$ follows directly for $\zeta \to 0$. □

Remark 8.10 (Existence of a weak solution for compressible system)**.** *If compressibility is present either for the fluid or the solid grains, the regularizing property* (A1*) *is fulfilled for $\eta = 0$. For instance, for b as in* (2.6)*, the equivalent pore pressure and the van Genuchten-Mualem model, it holds that $b_{\chi,\mathrm{m}} = \phi_0 c_\mathrm{w} + \frac{1}{N}$, cf. Appendix A. Consequently, the limit $(\boldsymbol{u}_\eta, \chi_\eta)$ in Lemma 8.9 is also well-defined for $\eta = 0$. In particular, it is a weak solution of* (2.15)–(2.22)*, cf. Definition 3.1.*

9 Step 6: Limit $\eta \to 0$ in the incompressible case

In this section, we show the main result, Theorem 3.2, for the more demanding case of an incompressible fluid and incompressible solid grains. Otherwise, by Remark 8.10 the main result of this paper follows already. In the incompressible case, b as in (2.6) is monotone but with $\hat{b}' = 0$ on a part of the domain with non-zero measure. Under the use of regularization with $\eta > 0$, it holds that $b_{\chi,\mathrm{m}} = \eta$. In the following, we prove that the limit of $\{(\boldsymbol{u}_\eta, \chi_\eta)\}_\eta$ for $\eta \to 0$ exists, and that it is a weak solution of (2.15)–(2.22) according to Definition 3.1. Throughout the entire section, we assume (A0)–(A9) and (ND1)–(ND3) hold true.

9.1 Stability estimates independent of η

In Section 8, almost all stability bounds have been independent of η. To summarize, there exists a constant $C > 0$ (independent of η) such that

$$\begin{aligned} &\|\boldsymbol{u}_\eta\|_{H^1(0,T;\boldsymbol{V})} + \|\chi_\eta\|_{L^\infty(0,T;H_0^1(\Omega))} + \|\hat{p}_\mathrm{pore}(\chi_\eta)\|_{L^2(Q_T)} \\ &\quad + \left\|\hat{B}_\eta(\chi_\eta)\right\|_{L^\infty(0,T;L^1(\Omega))} + \left\|\partial_t \hat{b}_\eta(\chi_\eta)\right\|_{L^2(0,T;H^{-1}(\Omega))} \leq C. \end{aligned} \tag{9.1}$$

The only bound depending on η is the stability of $\partial_t \chi_\eta$, cf. Lemma 8.4. We recall, there exists a constant $C_\eta > 0$, depending on η, satisfying

$$\|\partial_t \chi_\eta\|_{L^2(Q_T)} \leq C_\eta. \tag{9.2}$$

In order to conclude that $(\boldsymbol{u}_\eta, \chi_\eta)$ converges towards a weak solution of the unsaturated poroelasticity model, it will be sufficient to replace the stability result (9.2) by a uniform stability estimate. The remaining discussion for $\eta \to 0$ can be done along the lines of Section 8.2–8.3.

In the following, we prove a uniform stability bound replacing (9.2) in two steps. We show that the temporal derivative of the mechanics equation, i.e., (W5)$_{\zeta\eta}$ for $\zeta = 0$, is well-defined; and then we use an inf-sup argument and the uniform stability estimate (9.1).

Lemma 9.1 (Temporal derivative of the mechanics equation)**.** *It holds for all* $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V})$

$$\int_0^T a(\partial_t \boldsymbol{u}_\eta, \boldsymbol{v})\, dt - \int_0^T \alpha \left\langle \partial_t \hat{p}_{\mathrm{pore}}(\chi_\eta), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle\, dt = \int_0^T \left\langle \partial_t \boldsymbol{f}_{\mathrm{ext}}, \boldsymbol{v} \right\rangle\, dt. \tag{9.3}$$

Proof. First, we argue that the mechanics equation (3.1) holds pointwise on $[0,T]$. Let $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V}) \cap C^\infty(0,T;\boldsymbol{V})$. By Lemma 8.9, it holds that

$$\int_0^T a(\boldsymbol{u}_\eta, \boldsymbol{v})\, dt - \int_0^T \alpha \left\langle \hat{p}_{\mathrm{pore}}(\chi_\eta), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle\, dt = \int_0^T \left\langle \boldsymbol{f}_{\mathrm{ext}}, \boldsymbol{v} \right\rangle\, dt.$$

By the fundamental lemma of calculus of variations it follows a.e. on $[0,T]$

$$a(\boldsymbol{u}_\eta, \boldsymbol{v}) - \alpha \left\langle \hat{p}_{\mathrm{pore}}(\chi_\eta), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle = \left\langle \boldsymbol{f}_{\mathrm{ext}}, \boldsymbol{v} \right\rangle, \quad \text{for all } \boldsymbol{v} \in \boldsymbol{V}. \tag{9.4}$$

Applying a standard embedding for Bochner spaces [67], we can assume wlog. that $\boldsymbol{u}_\eta \in C(0,T;\boldsymbol{V})$ and $\hat{p}_{\mathrm{pore}}(\chi_\eta) \in C(0,T;L^2(\Omega))$, as $\partial_t \boldsymbol{u}_\eta \in L^2(0,T;\boldsymbol{V})$ and $\partial_t \hat{p}_{\mathrm{pore}}(\chi_\eta) \in L^2(Q_T)$ by (9.2) and assumption (ND2). Hence, (9.4) holds pointwise on $[0,T]$.

Now we show (9.3). Let $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V}) \cap C^\infty(0,T;\boldsymbol{V})$. By Lemma 8.9, it holds that

$$\int_0^T a(\boldsymbol{u}_\eta, \partial_t \boldsymbol{v})\, dt - \alpha \int_0^T \left\langle \hat{p}_{\mathrm{pore}}(\chi_\eta), \boldsymbol{\nabla} \cdot \partial_t \boldsymbol{v} \right\rangle\, dt = \int_0^T \left\langle \boldsymbol{f}_{\mathrm{ext}}, \partial_t \boldsymbol{v} \right\rangle\, dt.$$

Since $\partial_t \boldsymbol{u}_\eta \in L^2(0,T;\boldsymbol{V})$, $\partial_t \hat{p}_{\mathrm{pore}}(\chi_\eta) \in L^2(Q_T)$ and $\partial_t \boldsymbol{f}_{\mathrm{ext}} \in L^2(0,T;\boldsymbol{V}^\star)$, integration by parts is well-defined. Together with (9.4), we obtain

$$\int_0^T a(\partial_t \boldsymbol{u}_\eta, \boldsymbol{v})\, dt - \alpha \int_0^T \left\langle \partial_t \hat{p}_{\mathrm{pore}}(\chi_\eta), \boldsymbol{\nabla} \cdot \boldsymbol{v} \right\rangle\, dt = \int_0^T \left\langle \partial_t \boldsymbol{f}_{\mathrm{ext}}, \boldsymbol{v} \right\rangle\, dt.$$

The assertion follows after applying a density argument allowing for arbitrary test functions in $L^2(0,T;\boldsymbol{V})$ in (9.3). □

Lemma 9.2 (Stability estimate for the temporal derivative of the Kirchhoff pressure)**.** *There exists a constant* $C^{(12)} > 0$ *(independent of* η*) such that*

$$\|\partial_t \chi_\eta\|_{L^2(Q_T)} \leq C^{(12)}.$$

Proof. We show that $\|\partial_t \hat{p}_{\mathrm{pore}}(\chi_\eta)\|_{L^2(Q_T)}$ is uniformly bounded. The assertion follows then from assumption (ND2), as

$$\|\partial_t \chi_\eta\|_{L^2(Q_T)} \leq C_{\mathrm{ND},2} \left\| \partial_t \hat{p}_{\mathrm{pore}}(\chi_\eta) \right\|_{L^2(Q_T)}.$$

By Lemma 9.1, the time derivative of the mechanics equations is well-defined, cf. (9.3). Using a standard inf-sup argument (introducing the constant $C_{\Omega,\mathrm{is}}$), cf. Lemma B.11, it follows from (9.3) that

$$\left\| \partial_t \hat{p}_{\mathrm{pore}}(\chi_\eta) \right\|_{L^2(Q_T)} \leq C_{\Omega,\mathrm{is}} \left(\|\partial_t \boldsymbol{u}_\eta\|_{L^2(0,T;\boldsymbol{V})} + \|\partial_t \boldsymbol{f}_{\mathrm{ext}}\|_{L^2(0,T;\boldsymbol{V}^\star)} \right).$$

Since $\|\partial_t \boldsymbol{u}_\eta\|_{L^2(0,T;\boldsymbol{V})}$ is uniformly bounded by (9.1), $\|\partial_t \hat{p}_{\mathrm{pore}}(\chi_\eta)\|_{L^2(Q_T)}$ is uniformly bounded, which concludes the proof. □

9.2 Relative (weak) compactness for $\eta \to 0$

Using the same line of argumentation used in Section 8.2, we can discuss the limit process $\eta \to 0$.

Lemma 9.3 (Convergence of the primary variables)**.** *We can extract subsequences of* $\{\boldsymbol{u}_\eta\}_\eta$ *and* $\{\chi_\eta\}_\eta$ *(still denoted like the original sequences), and there exist* $\boldsymbol{u} \in H^1(0,T;\boldsymbol{V})$ *and* $\chi \in H^1(0,T;L^2(\Omega)) \cap L^\infty(0,T;Q)$ *such that for* $\eta \to 0$

$$\begin{aligned} \boldsymbol{u}_\eta &\rightharpoonup \boldsymbol{u} && \text{in } H^1(0,T;\boldsymbol{V}),\\ \chi_\eta &\to \chi && \text{in } L^2(Q_T),\\ \chi_\eta &\rightharpoonup \chi && \text{in } L^\infty(0,T;Q),\\ \partial_t\chi_\eta &\rightharpoonup \partial_t\chi && \text{in } L^2(Q_T). \end{aligned}$$

Proof. The proof is analog to the proofs of Lemma 8.5. □

Lemma 9.4 (Convergence of the coupling terms)**.** *Up to subsequences it holds for* $\eta \to 0$ *that*

$$\begin{aligned} \hat{p}_{\mathrm{pore}}(\chi_\eta) &\rightharpoonup \hat{p}_{\mathrm{pore}}(\chi) && \text{in } L^2(Q_T),\\ \hat{s}_{\mathrm{w}}(\chi_\eta)\partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u}_\eta &\rightharpoonup \hat{s}_{\mathrm{w}}(\chi)\partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u} && \text{in } L^2(Q_T). \end{aligned}$$

Proof. The proof is analog to the proof of Lemma 8.6. □

Lemma 9.5 (Initial conditions for the fluid flow)**.** *Up to subsequences it holds that*

$$\partial_t\hat{b}_\eta(\chi_\eta) \rightharpoonup \partial_t\hat{b}(\chi) \text{ in } L^2(0,T;Q^\star),$$

where $\partial_t\hat{b}(\chi) \in L^2(0,T;Q^\star)$ *is understood in the sense of* (W2)*.*

Proof. The proof is analog to the proof of Lemma 8.7. We only stress that due to construction of $\hat{b}_\eta$, one can show that if $\chi_\eta \to \chi$ in $L^2(Q_T)$, it also holds

$$\begin{aligned} \hat{b}_\eta(\chi_0) &\rightharpoonup \hat{b}(\chi_0) && \text{in } L^\infty(0,T;L^1(\Omega)),\\ \hat{b}_\eta(\chi_\eta) &\rightharpoonup \hat{b}(\chi) && \text{in } L^\infty(0,T;L^1(\Omega)), \end{aligned}$$

for $\eta \to 0$. Hence, (W2) can be deduced from $(\mathrm{W2})_\eta$ for $\eta \to 0$. □

Lemma 9.6 (Initial conditions of the mechanical displacement)**.** $\partial_t\boldsymbol{\nabla}\cdot\boldsymbol{u} \in L^2(Q_T)$ *satisfies* (W3)*.*

Proof. The proof is almost identical to the proof of Lemma 8.8. Standard compactness arguments and $(\mathrm{W3})_\eta$ yield

$$\int_0^T a(\partial_t\boldsymbol{u},\boldsymbol{v})\,dt + \int_0^T a(\boldsymbol{u}-\boldsymbol{u}_0,\partial_t\boldsymbol{v})\,dt = 0$$

for all $\boldsymbol{v} \in H^1(0,T;\boldsymbol{V})$ with $\boldsymbol{v}(T) = \boldsymbol{0}$. Hence, $\boldsymbol{u}(0) = \boldsymbol{u}_0$ in $\boldsymbol{V}$; note that $\boldsymbol{u} \in C(0,T;\boldsymbol{V})$ by a Sobolev embedding. Therefore also $\boldsymbol{\nabla}\cdot\boldsymbol{u}(0) = \boldsymbol{\nabla}\cdot\boldsymbol{u}_0$ in $L^2(\Omega)$, which yields (W3). □

9.3 Identifying a weak solution for $\eta \to 0$

Finally, we prove the existence of a weak solution to the unsaturated poroelasticity model.

Lemma 9.7 (Limit satisfies (W1)–(W4))**.** *The limit* $(\boldsymbol{u},\chi)$ *is a weak solution of* (2.15)–(2.22)*, cf. Definition 3.1.*

Proof. The proof follows directly from the convergence results in Lemma 9.5 and Lemma 9.6 together with the validity of the regularized problem (4.3)–(4.4) for $\zeta = 0$. □

A Feasibility of assumptions

The analysis of this paper allows for arbitrary constitutive laws for b, p_{pore}, s_{w} and κ_{rel}, as long as they satisfy the conditions (A0)–(A4), (ND1)–(ND3) and (A1*). In the following, we demonstrate the feasibility of those conditions for a prominent choice of models. Let b as derived by [4]

$$b(p_{\text{w}}) = \phi_0 s_{\text{w}}(p_{\text{w}}) + c_{\text{w}}\phi_0 \int_0^{p_{\text{w}}} s_{\text{w}}(p)\,dp + \frac{1}{N}\int_0^{p_{\text{w}}} s_{\text{w}}(p) p_{\text{pore}}'(p)\,dp,$$

with p_{pore} chosen as equivalent pore pressure [5]

$$p_{\text{pore}}(p_{\text{w}}) = \int_0^{p_{\text{w}}} s_{\text{w}}(p)\,dp,$$

and the hydraulic properties s_{w} and κ_{rel} given by the van Genuchten-Mualem relations [63, 68]

$$s_{\text{w}}(p_{\text{w}}) = \begin{cases} [1+(-\alpha_{\text{vG}}p_{\text{w}})^{n_{\text{vG}}}]^{-m_{\text{vG}}}, & p_{\text{w}} \leq 0, \\ 1, & p_{\text{w}} \geq 0, \end{cases}$$

$$\kappa_{\text{rel}}(s_{\text{w}}) = \sqrt{s_{\text{w}}}\left[1-\left(1-s_{\text{w}}^{\frac{1}{m_{\text{vG}}}}\right)^{m_{\text{vG}}}\right]^2.$$

where $m_{\text{vG}} \in (0,1)$, $n_{\text{vG}} = (1-m_{\text{vG}})^{-1}$, and $\alpha_{\text{vG}} > 0$ are constant fitting parameters.

A.1 Checking (A0)

By definition, it holds that $s_{\text{w}}(p_{\text{w}}) > 0$ for all $p_{\text{w}} \in \mathbb{R}$ and $\kappa_{\text{rel}}(s_{\text{w}}) > 0$ for all $s_{\text{w}} > 0$. Hence, (A0) is satisfied for the van Genuchten-Mualem relations.

A.2 Checking (A1)–(A4) and (A1*)

By definition, it follows directly, that s_{w} is differentiable with a non-negative and uniformly bounded derivative s_{w}', i.e., s_{w} satisfies (A2). Furthermore, $p_{\text{pore}}'(p_{\text{w}}) = s_{\text{w}}(p_{\text{w}})$, and hence, p_{pore} satisfies (A3). We therefore only focus on (A1), (A1*) and (A4).

(A1) Monotonicity of $\hat{b}$. The function $\hat{b} = \hat{b}(\chi)$ is non-decreasing since

$$\hat{b}'(\chi) = c_{\text{w}}\phi_0 \frac{\hat{s}_{\text{w}}(\chi)}{\hat{\kappa}_{\text{rel}}(\chi)} + \phi_0 \frac{s_{\text{w}}'(\hat{p}_{\text{w}}(\chi))}{\hat{\kappa}_{\text{rel}}(\chi)} + \frac{1}{N}\frac{\hat{s}_{\text{w}}(\chi)^2}{\hat{\kappa}_{\text{rel}}(\chi)} \geq 0. \tag{A.1}$$

(A1*) Regularizing property of $\hat{b}_\eta$. As $\hat{b}_\eta$ is essentially equal to $\hat{b}$ but with enhanced Biot Modulus, $\hat{b}_\eta$ essentially satisfies (A1) with

$$\hat{b}_\eta'(\chi) = c_{\text{w}}\phi_0 \frac{\hat{s}_{\text{w}}(\chi)}{\hat{\kappa}_{\text{rel}}(\chi)} + \phi_0 \frac{s_{\text{w}}'(\hat{p}_{\text{w}}(\chi))}{\hat{\kappa}_{\text{rel}}(\chi)} + \left(\frac{1}{N}+\eta\right)\frac{\hat{s}_{\text{w}}(\chi)^2}{\hat{\kappa}_{\text{rel}}(\chi)} \geq 0.$$

In particular, it holds that

$$\langle \hat{b}(\chi_1) - \hat{b}(\chi_2), \chi_1 - \chi_2 \rangle \geq \left(c_{\text{w}}\phi_0 \left\| \tfrac{\kappa_{\text{rel}}}{s_{\text{w}}} \right\|_\infty^{-1} + \left(\frac{1}{N}+\eta\right) \left\| \tfrac{\kappa_{\text{rel}}}{s_{\text{w}}^2} \right\|_\infty^{-1} \right) \|\chi_1 - \chi_2\|^2.$$

By l'Hôspital's rule (note $0 < m_{\text{vG}} < 1$) it holds that

$$\lim_{s_{\text{w}}\to 0} \left(\frac{\kappa_{\text{rel}}(s_{\text{w}})}{s_{\text{w}}}\right)^2 = \lim_{s_{\text{w}}\to 0} 4\left[1-\left(1-s_{\text{w}}^{\frac{1}{m_{\text{vG}}}}\right)^{m_{\text{vG}}}\right]^3 \left(1-s_{\text{w}}^{\frac{1}{m_{\text{vG}}}}\right)^{m_{\text{vG}}-1} s_{\text{w}}^{1/m_{\text{vG}}-1} = 0.$$

and

$$\lim_{s_{\mathrm{w}}\to 0} \left(\frac{\kappa_{\mathrm{rel}}(s_{\mathrm{w}})}{s_{\mathrm{w}}^2}\right)^{2/3} = \lim_{s_{\mathrm{w}}\to 0} \frac{4}{3}\left[1-(1-s_{\mathrm{w}}^{\frac{1}{m_{\mathrm{vG}}}})^{m_{\mathrm{vG}}}\right]^{1/3}\left(1-s_{\mathrm{w}}^{\frac{1}{m_{\mathrm{vG}}}}\right)^{m_{\mathrm{vG}}-1} s_{\mathrm{w}}^{1/m_{\mathrm{vG}}-1} = 0.$$

Hence, there exists a generic constant $c > 0$, such that

$$\frac{\kappa_{\mathrm{rel}}(p_{\mathrm{w}})}{s_{\mathrm{w}}(p_{\mathrm{w}})} \in (0, c], \tag{A.2}$$

$$\frac{\kappa_{\mathrm{rel}}(p_{\mathrm{w}})}{s_{\mathrm{w}}(p_{\mathrm{w}})^2} \in (0, c]. \tag{A.3}$$

After all, it follows, for $\eta > 0$, $\hat{b}_\eta$ satisfies (A1*). Furthermore, in the compressible case $\max\{c_{\mathrm{w}}, \frac{1}{N}\} > 0$, also $\hat{b}$ satisfies (A1*), cf. Remark 8.10.

(A4) Uniform growth of $\frac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}}$. For all $p_{\mathrm{w}} \in \mathbb{R}$, it holds that

$$\frac{d}{dp_{\mathrm{w}}}\left(\frac{p_{\mathrm{pore}}}{s_{\mathrm{w}}}\right) = 1 - \frac{p_{\mathrm{pore}}(p_{\mathrm{w}})s_{\mathrm{w}}'(p_{\mathrm{w}})}{s_{\mathrm{w}}(p_{\mathrm{w}})^2} \geq 1,$$

$$\chi'(p_{\mathrm{w}}) = \kappa_{\mathrm{rel}}(s_{\mathrm{w}}(p_{\mathrm{w}})) \leq 1.$$

Hence, by using the chain rule, $\frac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}}$ satisfies the uniform growth condition (A4) with

$$\frac{d}{d\chi}\left(\frac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}}\right) \geq 1.$$

A.3 Checking (ND1)–(ND2)

We demonstrate, that (ND1)–(ND2) hold assuming $s_{\mathrm{w}} \geq s_{\mathrm{min}}$ for some minimal saturation value $s_{\mathrm{min}} > 0$. It holds that

$$\frac{\hat{p}_{\mathrm{pore}}}{\hat{s}_{\mathrm{w}}\chi} \sim \frac{1}{\hat{\kappa}_{\mathrm{rel}}} \quad \text{for} \quad \chi \to -\infty.$$

Under above assumption, one can assume that $\hat{\kappa}_{\mathrm{rel}} \geq \kappa_{\mathrm{min}} > 0$, such that (ND1) holds. Furthermore,

$$\hat{p}_{\mathrm{pore}}'(\chi) = \frac{\hat{s}_{\mathrm{w}}}{\hat{\kappa}_{\mathrm{rel}}}.$$

By (A.2), $\hat{p}_{\mathrm{pore}}'(\chi)$ is bounded from below by a constant independent of χ. Assuming $s_{\mathrm{w}} \geq s_{\mathrm{min}}$ for some minimal saturation value, also an upper bound is given. After all, (ND2) holds.

A.4 Discussion of (ND3)

The condition (ND3) is equivalent with

$$\left(\frac{\hat{s}_{\mathrm{w}}(\chi)}{\hat{p}_{\mathrm{pore}}'(\chi)} - 1\right)^{-2} \frac{\hat{b}'(\chi)}{\left(\hat{p}_{\mathrm{pore}}'(\chi)\right)^2} > \frac{\alpha^2}{4K_{\mathrm{dr}}} \qquad \text{for all } \chi \in \mathbb{R}. \tag{A.4}$$

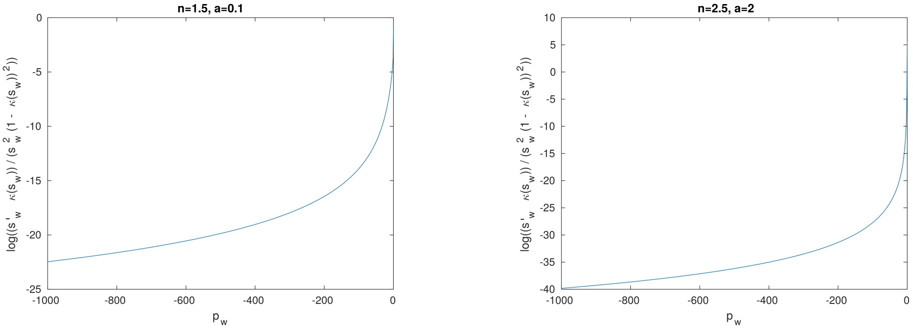
First, we note that in the fully saturated regime condition, (ND3) is fulfilled since

$$\frac{\hat{s}_{\mathrm{w}}(\chi)}{\hat{p}_{\mathrm{pore}}'(\chi)} = 1, \quad \text{for all } \chi \geq 0.$$

For the combination of the specific choices of b, s_w and κ_rel condition (A.4) becomes

$$(\kappa_\mathrm{rel}(s_\mathrm{w})-1)^{-2}\left(\phi_0 c_\mathrm{w}\frac{\kappa_\mathrm{rel}(s_\mathrm{w})}{s_\mathrm{w}}+\phi_0 s_\mathrm{w}'\frac{\kappa_\mathrm{rel}(s_\mathrm{w})}{s_\mathrm{w}^2}+\frac{1}{N}\kappa_\mathrm{rel}(s_\mathrm{w})\right)>\frac{\alpha^2}{4K_\mathrm{dr}}\qquad \text{for all } s_\mathrm{w}<1.$$

We consider the more demanding case, the incompressible case with $c_\mathrm{w}=\frac{1}{N}=0$. The expression $\frac{s_\mathrm{w}'\kappa_\mathrm{rel}(s_\mathrm{w})}{s_\mathrm{w}^2(1-\kappa_\mathrm{rel}(s_\mathrm{w}))^2}$ is increasing in p_w, see Figure 2 for two examples. Hence, there exists a minimal saturation value s_min such that (A.4) holds in the regime $s_\mathrm{w}\in[s_\mathrm{min},1]$. This value will depend on ϕ_0, α_vG, n_vG, α and K_dr. Assuming $\phi_0=0.1$ and $\alpha=1$, we compute s_min for a set of realistic parameters, see Table 1. We observe, that the range of admissible saturation values becomes larger, the stiffer the system. Furthermore, for all parameters, s_min is relatively small. Hence, we can expect (ND3) to hold for geotechnical applications, for which K_dr is typically large.



(a) $n_\mathrm{vG}=1.5$, $\alpha_\mathrm{vG}=0.1$

(b) $n_\mathrm{vG}=2.5$, $\alpha_\mathrm{vG}=2$

Figure 2: Increasing behavior of $\frac{s_\mathrm{w}'\kappa_\mathrm{rel}(s_\mathrm{w})}{s_\mathrm{w}^2(1-\kappa_\mathrm{rel}(s_\mathrm{w}))^2}$ in the unsaturated regime.

α_vG	n_vG	s_min for $K_\mathrm{dr}=10^5$	s_min for $K_\mathrm{dr}=10^8$	s_min for $K_\mathrm{dr}=10^{11}$
0.1	1.5	0.26	0.10	0.04
2	1.5	0.17	0.07	0.03
0.1	2	0.08	0.02	0.004
2	2	0.04	0.009	0.002
0.1	2.5	0.03	0.004	0.0006
2	2.5	0.01	0.002	0.0003

Table 1: Minimal allowed saturation values for a set of realistic model parameters, assuming $\alpha=1$.

B Useful results from literature

Lemma B.1 (Discrete Poincaré inequality [62]). *Let $\mathcal{T}$ be an admissible mesh, cf. Definition 5.1, and u a piecewise constant function. Then there exists a constant $C_{\Omega,\mathrm{DP}}\in(0,\mathrm{diam}(\Omega)]$ such that*

$$\|u\|_{L^2(\Omega)}\le C_{\Omega,\mathrm{DP}}\|u\|_{1,\mathcal{T}},$$

where $\|\cdot\|_{1,\mathcal{T}}$ denotes the discrete $H_0^1(\Omega)$ norm, cf. Definition 5.3.

Lemma B.2 (Discrete trace inequality [64]). *Let $\mathcal{T}$ be an admissible mesh, cf. Definition 5.1, and u a piecewise constant function. Let $\gamma(u)$ denote the trace of u, defined by $\gamma(u) = u_K$ on $\sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K$, $K \in \mathcal{T}$. Then there exists a constant $C_{\text{tr}} > 0$ such that*

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq C_{\text{tr}} \left(\|u\|_{1,\mathcal{T}} + \|u\|_{L^2(\Omega)} \right),$$

where $\|\cdot\|_{1,\mathcal{T}}$ denotes the discrete $H_0^1(\Omega)$ norm, cf. Definition 5.3.

Lemma B.3 (Stability of discrete gradients [62]). *Let $\mathcal{T}$ be an admissible mesh of some domain Ω, cf. Definition 5.1, and $u \in H_0^1(\Omega)$. Define a piecewise constant function $\tilde{u}$ by*

$$\tilde{u}(x) := \frac{1}{|K|} \int_K u(x)\, dx, \quad x \in K \in \mathcal{T}.$$

Then there exists a constant $C > 0$ (independent of h for regular meshes) such that

$$\|\tilde{u}\|_{1,\mathcal{T}} \leq C \|u\|_{H^1(\Omega)}.$$

Lemma B.4 (Corollary of Brouwer's fixed point theorem [69]). *Let $\langle \cdot, \cdot \rangle$ denote the standard $\mathbb{R}^d$ scalar product and let $\boldsymbol{F} : \mathbb{R}^d \to \mathbb{R}^d$ be a continuous function, satisfying*

$$\langle \boldsymbol{F}(\boldsymbol{x}), \boldsymbol{x} \rangle \geq 0 \tag{B.1}$$

for all $\boldsymbol{x} \in \mathbb{R}^d$ with $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq M$ for some fixed $M \in \mathbb{R}_+$. Then there exists a $\boldsymbol{x}^\star \in \mathbb{R}^d$ with $\langle \boldsymbol{x}^\star, \boldsymbol{x}^\star \rangle \leq M$ and $\boldsymbol{F}(\boldsymbol{x}^\star) = 0$.

Lemma B.5 (Binomial identity). *For $a, b \in \mathbb{R}$ it holds that*

$$a(a-b) = \frac{1}{2} \left(a^2 + (a-b)^2 - b^2 \right). \tag{B.2}$$

Lemma B.6 (Summation by parts). *Given two sequences $(a_k)_{k \in \mathbb{N}_0}$, $(b_k)_{k \in \mathbb{N}_0} \subset \mathbb{R}$, for all $N \in \mathbb{N}$ it holds that*

$$\sum_{n=1}^{N} a_n (b_n - b_{n-1}) = a_N b_N - a_1 b_0 - \sum_{n=1}^{N-1} b_n (a_{n+1} - a_n).$$

Lemma B.7 (Discrete Grönwall inequality [70]). *Let $(a_n)_n \subset \mathbb{R}_+$, $(\lambda_n)_n \subset \mathbb{R}_+$, $B \geq 0$. Assume for all $n \in \mathbb{N}$ it holds that*

$$a_n \leq B + \sum_{k=0}^{n-1} \lambda_k a_k.$$

Then it follows

$$a_n \leq B \prod_{k=0}^{n-1} (1 + \lambda_k).$$

In particular, if $\lambda_k = \frac{\lambda T}{N}$ for all $k \in \mathbb{N}$ for some $\lambda, T \in \mathbb{R}_+$ and $N \in \mathbb{N}$, it holds that

$$a_N \leq B \exp(\lambda T).$$

Lemma B.8 (Eberlein-Šmulian theorem [69]). *Assume that B is a reflexive Banach space and let $\{x_n\}_n \subset B$ be a bounded sequence in B. Then there exists a subsequence $\{x_{n_k}\}_k$ that converges weakly in B.*

Lemma B.9 (Relaxed Aubin-Lions lemma [71]). *Let $\{f_n\}_n \subset L^p(0,T;B)$, $1 \leq p < \infty$, B a Banach space. $\{f_n\}_n$ is relatively compact in $L^p(0,T;B)$ if the following two are fulfilled:*

- *$\{f_n\}_n$ is uniformly bounded in $L^p(0,T;X)$, for $X \subset B$ with compact embedding.*
- *$\int_\tau^T \|f_n(t) - f_n(t-\tau)\|_B^p \, dt \leq \mathcal{O}(\tau)$, as $\tau \to 0$.*

For the second property it is sufficient that $\{\partial_t f_n\}_n$ is uniformly bounded in $L^p(0,T;B)$.

Lemma B.10 (Riesz-Frechet-Kolmogorov compactness criterion [72]). *Let F be a bounded set in $L^p(\mathbb{R}^N)$ with $1 \leq p < \infty$, $N \in \mathbb{N}$. Assume that*

$$\lim_{|h|\to 0} \|f(\cdot + h) - f(\cdot)\|_{L^p(\mathbb{R}^N)} = 0 \quad \textit{uniformly in } f \in F.$$

Then the closure of $F|_\Omega := \{f : \Omega \to \mathbb{R} \,|\, f \in F\}$ is compact for any measurable set $\Omega \subset \mathbb{R}^N$ with finite measure.

Lemma B.11 (Standard inf-sup argument [66]). *Let V and Q be Hilbert spaces, and let B be a linear continuous operator from V to Q'. Denote by B^t the transposed operator of B. Then, the following two statements are equivalent:*

- *B^t is bounding, i.e., there exists a $\gamma > 0$ such that $\left\|B^t q\right\|_{V'} \geq \gamma \left\|q\right\|_Q$ for all $q \in Q$.*
- *There exists a $L_B \in \mathcal{L}(Q', V)$ such that $B(L_B(\xi)) = \xi$ for all $\xi \in Q'$ with $\|L_b\| = \dfrac{1}{\gamma} =: C_{\Omega,\text{is}}$.*

Lemma B.12 (Properties of the Legendre transformation [53]). *Given $b : \mathbb{R} \to \mathbb{R}$ continuous and non-decreasing , we define its Legendre transformation*

$$B(z) := \int_0^z (b(z) - b(s))\, ds \geq 0.$$

It holds for all $x, y \in \mathbb{R}$ and for all $\delta > 0$

$$\begin{aligned} 0 &\leq B(x), \\ B(x) - B(y) &\leq (b(x) - b(y))\, x, \\ |b(x)| &\leq \delta\, B(x) + \sup_{|y|\leq \delta^{-1}} |b(y)|\,. \end{aligned}$$

Acknowledgments

The authors would like to thank Florin A. Radu and Willi Jäger for their useful discussions.

References

[1] R. De Boer, *Theory of porous media: highlights in historical development and current state.* Springer Science & Business Media, 2000.

[2] K. v. Terzaghi, “The shearing resistance of saturated soils and the angle between the planes of shear,” in *First international conference on soil mechanics, 1936*, vol. 1, pp. 54–59, 1936.

[3] M. Biot, “General theory of three-dimensional consolidation,” *Journal of applied physics*, vol. 12, no. 2, pp. 155–164, 1941.

[4] R. Lewis and B. Schrefler, *The finite element method in the static and dynamic deformation and consolidation of porous media*. Numerical methods in engineering, John Wiley, 1998.

[5] O. Coussy, *Poromechanics*. Wiley, 2004.

[6] A. Szymkiewicz, *Modelling water flow in unsaturated porous media: accounting for nonlinear permeability and material heterogeneity*. Springer Science & Business Media, 2012.

[7] J. M. Nordbotten and M. A. Celia, *Geological storage of CO2: modeling approaches for large-scale simulation*. John Wiley & Sons, 2011.

[8] J.-L. Auriault and E. Sanchez-Palencia, "Etude de comportment macroscopique d'un milieu poreux sature deformable," *Journal de Mécanique*, vol. 16, pp. 575–603, 1977.

[9] A. Zenisek, "The existence and uniquencess theorem in Biot's consolidation theory," *Aplikace matematiky*, vol. 29, no. 3, pp. 194–211, 1984.

[10] R. Showalter, "Diffusion in Poro-Elastic Media," *Journal of Mathematical Analysis and Applications*, vol. 251, no. 1, pp. 310 – 340, 2000.

[11] M. Ferronato, N. Castelletto, and G. Gaolati, "A fully coupled 3-D mixed finite element model of Biot consolidation," *Journal of Computational Physics*, vol. 229, no. 12, pp. 4813 – 4830, 2010.

[12] J. B. Haga, H. Osnes, and H. P. Langtangen, "On the causes of pressure oscillations in low-permeable and low-compressible porous media," *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 36, no. 12, pp. 1507–1522, 2012.

[13] M. Wheeler, G. Xue, and I. Yotov, "Coupling multipoint flux mixed finite element methods with continuous Galerkin methods for poroelasticity," *Computational Geosciences*, vol. 18, no. 1, pp. 57–75, 2014.

[14] J. M. Nordbotten, "Stable cell-centered finite volume discretization for Biot equations," *SIAM Journal on Numerical Analysis*, vol. 54, no. 2, pp. 942–968, 2016.

[15] C. Rodrigo, F. Gaspar, X. Hu, and L. Zikatanov, "Stability and monotonicity for some discretizations of the Biot's consolidation model," *Computer Methods in Applied Mechanics and Engineering*, vol. 298, pp. 183 – 204, 2016.

[16] J. A. White, N. Castelletto, and H. A. Tchelepi, "Block-partitioned solvers for coupled poromechanics: A unified framework," *Computer Methods in Applied Mechanics and Engineering*, vol. 303, pp. 55 – 74, 2016.

[17] N. Castelletto, H. Hajibeygi, and H. A. Tchelepi, "Multiscale finite-element method for linear elastic geomechanics," *Journal of Computational Physics*, vol. 331, pp. 337 – 356, 2017.

[18] J. Lee, K. Mardal, and R. Winther, "Parameter-Robust Discretization and Preconditioning of Biot's Consolidation Model," *SIAM Journal on Scientific Computing*, vol. 39, no. 1, pp. 1–24, 2017.

[19] J. Kim, H. Tchelepi, and R. Juanes, "Stability and convergence of sequential methods for coupled flow and geomechanics: Fixed-stress and fixed-strain splits," *Computer Methods in Applied Mechanics and Engineering*, vol. 200, no. 13, pp. 1591 – 1606, 2011.

[20] A. Mikelić and M. F. Wheeler, "Convergence of iterative coupling for coupled flow and geomechanics," *Computational Geosciences*, vol. 17, no. 3, pp. 455–461, 2013.

[21] J. W. Both, M. Borregales, J. M. Nordbotten, K. Kumar, and F. A. Radu, "Robust fixed stress splitting for Biot's equations in heterogeneous media," *Applied Mathematics Letters*, vol. 68, pp. 101 – 108, 2017.

[22] F. J. Gaspar and C. Rodrigo, "On the fixed-stress split scheme as smoother in multigrid methods for coupling flow and geomechanics," *Computer Methods in Applied Mechanics and Engineering*, vol. 326, pp. 526 – 540, 2017.

[23] M. A. Borregales, K. Kumar, J. M. Nordbotten, and F. A. Radu, "Iterative solvers for Biot model under small and large deformation," *arxiv e-prints*, 2019. arXiv:1905.12996 [math.NA].

[24] E. Storvik, J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu, "On the optimization of the fixed-stress splitting for biot's equations," *International Journal for Numerical Methods in Engineering*, vol. 120, no. 2, pp. 179–194, 2019.

[25] K. Kumar, S. Matculevich, J. Nordbotten, and S. Repin, "Guaranteed and computable bounds of approximation errors for the semi-discrete Biot problem," *arxiv e-prints*, 2018. arXiv:1808.08036 [math.NA].

[26] E. Ahmed, F. A. Radu, and J. M. Nordbotten, "Adaptive poromechanics computations based on a posteriori error estimates for fully mixed formulations of Biot's consolidation model," *Computer Methods in Applied Mechanics and Engineering*, vol. 347, pp. 264 – 294, 2019.

[27] E. Ahmed, J. M. Nordbotten, and F. A. Radu, "Adaptive asynchronous time-stepping, stopping criteria, and a posteriori error estimates for fixed-stress iterative schemes for coupled poromechanics problems," *Journal of Computational and Applied Mathematics*, vol. 364, p. 112312, 2020.

[28] A. Mikelić and M. F. Wheeler, "Theory of the dynamic Biot-Allard equations and their link to the quasi-static Biot system," *Journal of Mathematical Physics*, vol. 53, no. 12, p. 123702, 2012.

[29] R. E. Showalter, "Poroelastic filtration coupled to Stokes flow," *Lecture Notes in Pure and Applied Mathematics*, vol. 242, p. 229, 2005.

[30] I. Ambartsumyan, V. J. Ervin, T. Nguyen, and I. Yotov, "A nonlinear Stokes-Biot model for the interaction of a non-Newtonian fluid with poroelastic media," *arxiv e-prints*, 2018. arXiv:1803.00947 [math.NA].

[31] I. Ambartsumyan, E. Khattatov, I. Yotov, and P. Zunino, "A lagrange multiplier method for a Stokes–Biot fluid–poroelastic structure interaction model," *Numerische Mathematik*, vol. 140, no. 2, pp. 513–553, 2018.

[32] A. Tavakoli and M. Ferronato, "On existence-uniqueness of the solution in a nonlinear Biot's model," *Appl. Math*, vol. 7, no. 1, pp. 333–341, 2013.

[33] L. Bociu, G. Guidoboni, R. Sacco, and J. T. Webster, "Analysis of nonlinear poro-elastic and poro-visco-elastic models," *Archive for Rational Mechanics and Analysis*, vol. 222, no. 3, pp. 1445–1519, 2016.

[34] A. Mikelić, M. F. Wheeler, and T. Wick, "Phase-field modeling of a fluid-driven fracture in a poroelastic medium," *Computational Geosciences*, vol. 19, no. 6, pp. 1171–1195, 2015.

[35] V. Girault, K. Kumar, and M. F. Wheeler, “Convergence of iterative coupling of geomechanics with flow in a fractured poroelastic medium,” *Computational Geosciences*, vol. 20, no. 5, pp. 997–1011, 2016.

[36] R. L. Berge, I. Berre, E. Keilegavlen, J. M. Nordbotten, and B. Wohlmuth, “Finite volume discretization for poroelastic media with fractures modeled by contact mechanics,” *arxiv e-prints*, 2019. arXiv:1904.11916 [math.NA].

[37] E. Ucar, E. Keilegavlen, I. Berre, and J. M. Nordbotten, “A finite-volume discretization for deformation of fractured media,” *Computational Geosciences*, vol. 22, no. 4, pp. 993–1007, 2018.

[38] J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu, “The gradient flow structures of thermo-poro-visco-elastic processes in porous media,” *arxiv e-prints*, 2019. arXiv:1907.03134 [math.NA].

[39] M. Borregales, F. A. Radu, K. Kumar, and J. M. Nordbotten, “Robust iterative schemes for non-linear poromechanics,” *Computational Geosciences*, vol. 22, no. 4, pp. 1021–1038, 2018.

[40] C. J. Van Duijn and A. Mikelic, “Mathematical Theory of Nonlinear Single-Phase Poroelasticity,” 2019.

[41] C. van Duijn, A. Mikelić, M. F. Wheeler, and T. Wick, “Thermoporoelasticity via homogenization: Modeling and formal two-scale expansions,” *International Journal of Engineering Science*, vol. 138, pp. 1 – 25, 2019.

[42] M. K. Brun, E. Ahmed, J. M. Nordbotten, and F. A. Radu, “Well-posedness of the fully coupled quasi-static thermo-poroelastic equations with nonlinear convective transport,” *Journal of Mathematical Analysis and Applications*, vol. 471, no. 1, pp. 239 – 266, 2019.

[43] M. Kirkesæther Brun, E. Ahmed, I. Berre, J. M. Nordbotten, and F. A. Radu, “Monolithic and splitting based solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport,” *arxiv e-prints*, 2019. arXiv:1902.05783 [math.NA].

[44] J. Kim, “Unconditionally stable sequential schemes for all-way coupled thermoporomechanics: Undrained-adiabatic and extended fixed-stress splits,” *Computer Methods in Applied Mechanics and Engineering*, vol. 341, pp. 93 – 112, 2018.

[45] Q. Hong, J. Kraus, M. Lymbery, and F. Philo, “Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelasticity models,” *Numerical Linear Algebra with Applications*, vol. 26, no. 4, p. 2242, 2019.

[46] Q. Hong, J. Kraus, M. Lymbery, and M. Fanett Wheeler, “Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems,” *arxiv e-prints*, 2018. arXiv:1812.11809 [math.NA].

[47] J. Lee, E. Piersanti, K. Mardal, and M. Rognes, “A Mixed Finite Element Method for Nearly Incompressible Multiple-Network Poroelasticity,” *SIAM Journal on Scientific Computing*, vol. 41, no. 2, pp. 722–747, 2019.

[48] R. Showalter and N. Su, “Partially saturated flow in a poroelastic medium,” *Discrete and Continuous Dynamical Systems - Series B*, vol. 1, no. 4, pp. 403–420, 2001.

[49] J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu, “Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media,” *Computers & Mathematics with Applications*, vol. 77, no. 6, pp. 1479–1502, 2019.

[50] J. Kim, H. A. Tchelepi, and R. Juanes, “Rigorous Coupling of Geomechanics and Multiphase Flow with Strong Capillarity,” *Society of Petroleum Engineers*, 2013.

[51] B. Jha and R. Juanes, “Coupled multiphase flow and poromechanics: A computational model of pore pressure effects on fault slip and earthquake triggering,” *Water Resources Research*, vol. 50, no. 5, pp. 3776–3808, 2014.

[52] Q. M. Bui, D. Osei-Kuffuor, N. Castelletto, and J. A. White, “A Scalable Multigrid Reduction Framework for Multiphase Poromechanics of Heterogeneous Media,” *arxiv e-prints*, 2019. arXiv:1904.05960 [math.NA].

[53] H. Wilhelm Alt and S. Luckhaus, “Quasilinear elliptic-parabolic differential equations,” *Mathematische Zeitschrift*, vol. 183, no. 3, pp. 311–341, 1983.

[54] N. Castelletto, S. Klevtsov, H. Hajibeygi, and H. A. Tchelepi, “Multiscale two-stage solver for Biot’s poroelasticity equations in subsurface media,” *Computational Geosciences*, vol. 23, no. 2, pp. 207–224, 2019.

[55] R. Eymard, M. Gutnic, and D. Hilhorst, “The finite volume method for Richards equation,” *Computational Geosciences*, vol. 3, no. 3-4, pp. 259–294, 1999.

[56] R. A. Klausen, F. A. Radu, and G. T. Eigestad, “Convergence of MPFA on triangulations and for Richards’ equation,” *International Journal for Numerical Methods in Fluids*, vol. 58, no. 12, pp. 1327–1351, 2008.

[57] C. Cancès and C. Guichard, “Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations,” *Mathematics of Computation*, vol. 85, no. 298, pp. 549–580, 2016.

[58] A. Ait Hammou Oulhaj, C. Cancès, and C. Chainais-Hillairet, “Numerical analysis of a nonlinearly stable and positive control volume finite element scheme for Richards equation with anisotropy,” *ESAIM: Mathematical Modelling & Numerical Analysis*, vol. 52, no. 4, 2018.

[59] T. Arbogast and M. Wheeler, “A Nonlinear Mixed Finite Element Method for a Degenerate Parabolic Equation Arising in Flow in Porous Media,” *SIAM Journal on Numerical Analysis*, vol. 33, no. 4, pp. 1669–1687, 1996.

[60] F. A. Radu, I. S. Pop, and P. Knabner, “Error estimates for a mixed finite element discretization of some degenerate parabolic equations,” *Numerische Mathematik*, vol. 109, no. 2, pp. 285–311, 2008.

[61] B. Saad and M. Saad, “Study of full implicit petroleum engineering finite-volume scheme for compressible two-phase flow in porous media,” *SIAM Journal on Numerical Analysis*, vol. 51, no. 1, pp. 716–741, 2013.

[62] R. Eymard, T. Gallouët, and R. Herbin, “Convergence of finite volume schemes for semilinear convection diffusion equations,” *Numerische Mathematik*, vol. 82, no. 1, pp. 91–116, 1999.

[63] M.Th. van Genuchten, “A closed-form equation for predicting the hydraulic conductivity of unsaturated soils,” *Soil Science Society of America Journal*, vol. 44(5), pp. 892–898, 1980.

[64] R. Eymard, T. Gallouët, and R. Herbin, “Finite volume methods,” vol. 7, pp. 713 – 1018, 2000.

[65] Baranger, Jacques, Maitre, Jean-François, and Oudin, Fabienne, “Connection between finite volume and mixed finite element methods,” *ESAIM: M2AN*, vol. 30, no. 4, pp. 445–465, 1996.

[66] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications.* Springer Series in Computational Mathematics, Springer, 2013.

[67] L. Evans, *Partial Differential Equations.* Graduate studies in mathematics, American Mathematical Society, 2010.

[68] Y. Mualem, “A new model for predicting the hydraulic conductivity of unsaturated porous media,” *Water Resources Research*, vol. 12, no. 3, pp. 513–522.

[69] P. G. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications.* Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2013.

[70] D. S. Clark, “Short proof of a discrete gronwall inequality,” *Discrete Applied Mathematics*, vol. 16, no. 3, pp. 279 – 281, 1987.

[71] J. Simon, “Compact sets in the space $L^p(0,T;B)$,” *Annali di Matematica Pura ed Applicata*, vol. 146, no. 1, pp. 65–96, 1986.

[72] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations.* Springer Science & Business Media, 2010.

Paper E

Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media

Both, J.W., Kumar, K., Nordbotten, J.M., and Radu, F.A.

Computers & Mathematics with Applications **77(6)** (2019), pg. 1479–1502.

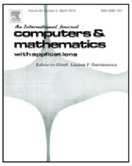
doi:10.1016/j.camwa.2018.07.033

Computers and Mathematics with Applications 77 (2019) 1479–1502

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa



Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media

Jakub Wiktor Both [a,*], Kundan Kumar [a], Jan Martin Nordbotten [a,b], Florin Adrian Radu [a]

[a] Department of Mathematics, University of Bergen, Bergen, Norway
[b] Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA

ARTICLE INFO

Article history:
Available online 10 August 2018

Keywords:
Nonlinear poroelasticity
Partially saturated porous media
Iterative coupling
L-scheme
Fixed-stress splitting
Anderson acceleration

ABSTRACT

In this paper, we study the robust linearization of nonlinear poromechanics of unsaturated materials. The model of interest couples the Richards equation with linear elasticity equations, generalizing the classical Biot equations. In practice a monolithic solver is not always available, defining the requirement for a linearization scheme to allow the use of separate simulators. It is not met by the classical Newton method. We propose three different linearization schemes incorporating the fixed-stress splitting scheme, coupled with an L-scheme, Modified Picard and Newton linearization of the flow equations. All schemes allow the efficient and robust decoupling of mechanics and flow equations. In particular, the simplest scheme, the Fixed-Stress-L-scheme, employs solely constant diagonal stabilization, has low cost per iteration, and is very robust. Under mild, physical assumptions, it is theoretically shown to be a contraction. Due to possible break-down or slow convergence of all considered splitting schemes, Anderson acceleration is applied as post-processing. Based on a special case, we justify theoretically the general ability of the Anderson acceleration to effectively accelerate convergence and stabilize the underlying scheme, allowing even non-contractive fixed-point iterations to converge. To our knowledge, this is the first theoretical indication of this kind. Theoretical findings are confirmed by numerical results. In particular, Anderson acceleration has been demonstrated to be very effective for the considered Picard-type methods. Finally, the Fixed-Stress-Newton scheme combined with Anderson acceleration shows the best performance among the splitting schemes.

1. Introduction

The coupling of fluid flow and mechanical deformation in unsaturated porous media is relevant for many applications, ranging from modeling rainfall-induced land subsidence or levee failure to understanding the swelling and drying-shrinkage of wooden or cement-based materials. Assuming linear elastic behavior, the process can be modeled by coupling the Richards equation with quasi-static linear elasticity equations, generalizing the classical Biot equations [1]. In this work, we consider the equivalent pore pressure [2], which allows a thermodynamically stable formulation [3].

For the numerical simulation of large scale applications, the solution of linear problems is typically the computationally most expensive component. For nonlinear problems, the dominating cost is determined by both the linearization scheme

* Corresponding author.
E-mail addresses: jakub.both@uib.no (J.W. Both), kundan.kumar@uib.no (K. Kumar), jan.nordbotten@uib.no (J.M. Nordbotten), florin.radu@uib.no (F.A. Radu).

https://doi.org/10.1016/j.camwa.2018.07.033

and the solver technology. Commonly, Newton's method is the first choice linearization scheme. However, for the nonlinear Biot equations, the monolithic Newton method requires the solver technology to solve saddle point problems coupling mechanics and flow equations. Additionally, in practice, constitutive laws employed in the model might be not Lipschitz continuous [4]. Thus, the arising systems are possibly ill-conditioned and require an advanced, monolithic simulator. As the latter might be not available, the goal of this work is to develop a linearization scheme, which is robust and allows the use of decoupled simulators for mechanics and flow equations. For this purpose, we adopt closely related concepts for the linear Biot equations and the Richards equation.

For the numerical solution of the linear Biot equations, splitting schemes are widely used; either as iterative solvers [5] or as preconditioners [6]. In particular, the fixed-stress splitting scheme has aroused much interest, being unconditionally stable in the sense of a von Neumann analysis [7] and being a global contraction [8–10]. As iterative solver, the scheme has been extended in various ways; e.g., it can be rewritten to a parallel-in-time solver [11], a multi-scale version allowing separate grids for the mechanics and flow problem has been developed [12], and the concept has been extended to nonlinear multi-phase flow coupled with linear elasticity [3]. In the context of monolithic solvers, the scheme has been applied as preconditioner for Krylov subspace methods [13–16] and as smoother for multigrid methods [17]. All in all, the scheme defines a promising strategy to decouple mechanics and flow equations.

For the linearization of the Richards equation, the standard Newton method has to be used with care, since the Richards equation is a degenerate elliptic–parabolic equation, modeling saturated/unsaturated flow, and additionally material laws might be only Hölder continuous. Various problem-specific alternatives have been developed in the literature. We want to point out two particular, simple linearization schemes; the L-scheme and the Modified Picard method. The L-scheme [18], employs diagonal stabilization for monotone, Lipschitz continuous nonlinearities. Global convergence has been rigorously proven for several porous media applications [19–21]; in particular also for the Richards equation [22]. The L-scheme can be also applied for Hölder continuous problems [21,23]. Furthermore, for the Richards equation it can be used to define a robust, linear domain decomposition method [24]. The L-scheme linearization has been coupled with the fixed-stress splitting scheme for nonlinear Biot equations with linear coupling [25]. Less robust, but in some cases more efficient is the Modified Picard method [26], which is a linearization scheme of the Richards equation, employing a first order Taylor approximation for the saturation and still allowing a Hölder continuous permeability.

In this paper, we combine the fixed-stress splitting scheme separately with the L-scheme, the Modified Picard method and Newton's method. The resulting schemes decouple and linearize simultaneously the mechanics and flow equations, utilizing only a single loop and allowing separate simulators. We show theoretically linear convergence of the Fixed-Stress-L-scheme, assuming non-vanishing residual saturation, permeability and porosity and an inf–sup stable discretization. However, the theoretical convergence rate might deteriorate in unfavorable situations, leading to either slow convergence or even stagnation in practice. As remedy for this, we apply Anderson acceleration.

Anderson acceleration has been originally introduced by [27] in order to accelerate fixed point iterations in electronic structure computation. It has been successfully applied in various other fields; in particular, we would like to highlight its use for the modified Picard iteration [28]. Reusing previous iterations to approximate directional derivatives, it can be related to a multi-secant quasi-Newton method [29] and to preconditioned GMRES for linear problems [30]. For nonlinear problems, it can be interpreted as a preconditioned nonlinear GMRES. Being a post-processing, it can be combined with splitting methods, still allowing separate simulators unlike preconditioned monolithic solvers. So far, theoretical results in the literature guarantee only convergence for contractive fixed-point iterations [31]. Furthermore, those results do not guarantee actual acceleration. However, in practice, Anderson acceleration may be observed to even possibly recover convergence for diverging methods. In this work, based on a special case, we justify theoretically the ability of the Anderson acceleration to effectively accelerate convergence of contractive fixed-point iterations and moreover stabilize non-contractive fixed-point iterations. To our knowledge, this is the first theoretical indication of this kind. Instead of Anderson acceleration, other stabilization techniques could be applied as, e.g., adaptive step size control, adaptive time stepping or the combination of a Picard-type method with a Newton-type method, following ideas by [32]. These concepts have not been considered in the scope of this work.

We present numerical results confirming the theoretical findings of this work. Indeed, the Fixed-Stress-L-scheme is more robust than the modifications employing Newton's method and the Modified Picard method. Moreover, convergence of the Picard-type methods can be accelerated significantly by the Anderson acceleration. When applied to initially diverging methods, convergence can be reliably recovered. These results are shown for two discretizations (i) the lowest-order discretization, employing constant, lowest-order Raviart–Thomas and linear finite elements for pressure, flux and displacement, respectively, and (ii) an unconditionally inf–sup stable discretization, equal to the previous one, besides using quadratic finite elements for the displacement. The first discretization is only conditionally stable and does not satisfy the assumptions of the theoretical convergence result for the Fixed-Stress-L-scheme. Nevertheless, in our numerical examples, the choice of the lowest-order discretization or the unconditionally stable discretization does not influence the performance of the linearization schemes.

The main, new contributions of this work are:

- We propose three linearization schemes incorporating the fixed-stress splitting scheme, coupled with an L-scheme, Modified Picard and Newton linearization of the flow. All schemes allow the efficient and robust decoupling of mechanics and flow equations. For the simplest scheme, the Fixed-Stress-L-scheme, we show theoretical convergence, assuming non-vanishing residual saturation, permeability and porosity and an inf–sup stable discretization, cf. Theorem 3.

- Based on a special case, we justify theoretically the general ability of the Anderson acceleration to effectively accelerate convergence and stabilize the underlying scheme, allowing even non-contractive fixed-point iterations to converge, cf. Section 7.3.
- The combination of the proposed linearization schemes and Anderson acceleration is demonstrated numerically to be robust and efficient. In particular, Anderson acceleration allows the schemes to converge in challenging situations even if the plain linearization schemes diverge. The Fixed-Stress-Newton method coupled with Anderson acceleration shows best performance among the splitting schemes, cf. Section 8.

The paper is organized as follows. In Section 2, the mathematical model is explained. In Section 3, a three-field discretization is introduced, employing conforming finite elements and mixed finite element for the mechanics and flow equations, respectively. In Section 4, we recall the monolithic Newton method and introduce the three splitting schemes, simultaneously linearizing and decoupling the mechanics and flow equations. In Section 5, convergence is proved for the Fixed-Stress-L-scheme. In Section 6, Anderson acceleration is recalled, and in Section 7, the ability of the Anderson acceleration to effectively accelerate convergence and increase robustness is discussed theoretically. In Section 8, numerical results are presented, illustrating in particular the increase of robustness via Anderson acceleration. The work is closed with concluding remarks in Section 9.

2. Mathematical model — Nonlinear Biot's equations coupling Richards equation and linear elasticity

We consider a nonlinear extension of the classical, linear Biot equations modeling flow in deformable porous media under possibly both fully and partially saturated conditions. For this, we assume:

(A1) The bulk material is linearly elastic and deforms solely under infinitesimal deformations.
(A2) There exists two fluid phases — one active and one passive phase (standard assumption for the Richards equation).
(A3) The active fluid phase is incompressible and corresponding fluxes are described by Darcy's law.
(A4) Mechanical inertia effects are negligible allowing to consider the quasi-static balance of linear momentum.

We model the medium at initial conditions by a reference configuration $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$. Due to the limitation to infinitesimal deformations, the domain of the primary fields is approximated by Ω on the entire time interval of interest $(0,T)$, with final time $T>0$.

Finally, the governing equations describing coupled fluid flow and mechanical deformation of a porous medium with mechanical displacement $\boldsymbol{u}$, fluid pressure p_w and volumetric flux $\boldsymbol{q}_\mathrm{w}$ as primary variables are given by

$$\partial_t(\phi s_\mathrm{w}) + \nabla\cdot\boldsymbol{q}_\mathrm{w} = 0, \tag{1}$$

$$\boldsymbol{q}_\mathrm{w} + k_\mathrm{w}(s_\mathrm{w})(\nabla p_\mathrm{w} - \rho_\mathrm{w}\boldsymbol{g}) = \boldsymbol{0}, \tag{2}$$

$$-\nabla\cdot[2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\nabla\cdot\boldsymbol{u}\boldsymbol{I} - \alpha p_\mathrm{E}(p_\mathrm{w})\boldsymbol{I}] = \rho_\mathrm{b}\boldsymbol{g}, \tag{3}$$

where ϕ denotes variable porosity, s_w denotes fluid saturation, k_w denotes fluid-dependent mobility, ρ_w and ρ_b denote fluid and bulk density, respectively, $\boldsymbol{g}$ is the gravitational acceleration, $\boldsymbol{\varepsilon}(\boldsymbol{u})$ and $\nabla\cdot\boldsymbol{u}$ denote the linear strain and the volumetric deformation, respectively, μ and λ denote the Lamé parameters, α is the Biot coefficient and p_E denotes the pore pressure. In the following, we comment briefly on the single components of the mathematical model and refer to [2] for a detailed derivation.

Eq. (1): For the fluid flow, an active and a passive fluid phase are assumed, cf. (A2). In other words, the passive phase responds instantaneously to the active phase and therefore has a constant pressure. The behavior of the active fluid phase is governed by mass conservation, equivalent to volume conservation for an incompressible fluid. The volume is given by the product of porosity ϕ and saturation s_w. As in linear poroelasticity, the porosity is assumed to change linearly with volumetric deformation $\nabla\cdot\boldsymbol{u}$ and pore pressure p_E by

$$\phi(\boldsymbol{u},p_\mathrm{w}) = \phi_0 + \alpha\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_0) + \frac{1}{N}(p_\mathrm{E}(p_\mathrm{w}) - p_\mathrm{E}(p_{\mathrm{w},0})), \tag{4}$$

where ϕ_0, $\boldsymbol{u}_0$ and $p_{\mathrm{w},0}$ are the initial porosity, displacement and pressure, respectively, and α is the Biot coefficient and N is the Biot modulus. Eq. (4) is a byproduct of the thermodynamic derivation of the effective stress by Coussy [2]. Furthermore, the saturation $s_\mathrm{w} = s_\mathrm{w}(p)$ is assumed to be described by a material law $s_\mathrm{w} : \mathbb{R} \to (0,1]$, satisfying $s_\mathrm{w}(p)=1$, $p\geq 0$, and having a negative inverse $p_\mathrm{c} : (0,1] \to \mathbb{R}_+$, such that $s_\mathrm{w}(-p_\mathrm{c}(s)) = s$, $s\in(0,1]$. In the literature, the function p_c is often referred to as capillary pressure.

Eq. (2): The volumetric flux $\boldsymbol{q}_\mathrm{w}$ is assumed to be described by Darcy's law for multiphase flow. Here, the permeability scaled by the inverse of the viscosity is given by a material law $k_\mathrm{w} = k_\mathrm{w}(s_\mathrm{w})$. In practice, the material law can become Hölder continuous.

Eq. (3): The mechanical behavior is governed by balance of linear momentum under quasi-static conditions, combined with an effective stress formulation. Allowing only for small deformations, we employ the St. Venant Kirchhoff model for the effective stress, determining the total, poroelastic stress as $\boldsymbol{\sigma}^\mathrm{por}(\boldsymbol{u},p_\mathrm{w}) = 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\nabla\cdot\boldsymbol{u}\boldsymbol{I} - \alpha p_\mathrm{E}(p_\mathrm{w})\boldsymbol{I}$. As pore pressure, we use the equivalent pore pressure [2]

$$p_\mathrm{E}(p) = s_\mathrm{w}(p)p - \int_{s_\mathrm{w}(p)}^{1} p_\mathrm{c}(s)\,ds, \tag{5}$$

which takes into account interfacial effects. By construction it satisfies $dp_\mathrm{E} = s_\mathrm{w}(p)\,dp$. As body force we assume solely gravity, where for the sake of simplicity the bulk density ρ_b is assumed to be constant. In general it is a function of porosity and saturation. All in all, Eq. (3) acts as compatibility condition to be satisfied at each time.

Introducing two partitions $\Gamma_D^f \cup \Gamma_N^f = \Gamma_D^m \cup \Gamma_N^m = \partial\Omega$ of the boundary of Ω and the outer normal $\boldsymbol{n}$ on $\partial\Omega$, we assume boundary conditions and initial conditions

$$\begin{aligned}
p_\mathrm{w} &= p_{w,D} && \text{on } \Gamma_D^f \times (0,T), & \boldsymbol{u} &= \boldsymbol{u}_D && \text{on } \Gamma_D^m \times (0,T),\\
\boldsymbol{q}_\mathrm{w}\cdot\boldsymbol{n} &= \boldsymbol{q}_{w,N} && \text{on } \Gamma_N^f \times (0,T), & \boldsymbol{\sigma}^\mathrm{por}(\boldsymbol{u},p_\mathrm{w})\boldsymbol{n} &= \boldsymbol{\sigma}_n^\mathrm{por} && \text{on } \Gamma_N^m \times (0,T),\\
p_\mathrm{w} &= p_{\mathrm{w},0} && \text{in } \Omega \times \{0\}, & \boldsymbol{u} &= \boldsymbol{u}_0 && \text{in } \Omega \times \{0\}.
\end{aligned}$$

All in all, the nonlinear Biot equations (1)–(3) couple nonlinearly the Richards equation and linear elasticity equations. In the fully saturated regime ($p_\mathrm{w} \geq 0$), the model reduces locally to the classical, linear Biot equations for an incompressible fluid and compressible rock. We note, as long as the fluid saturation is not vanishing, the nonlinear Biot equations (1)–(3) are parabolic, unlike the degenerate elliptic–parabolic Richards equation, cf. Remark 5.

3. Finite element discretization

We discretize the Biot equations (1)–(3) in space by the finite element method. More precisely, given a regular triangulation $\mathcal{T}_h$ of the domain Ω, we employ linear/quadratic, constant and lowest-order Raviart–Thomas finite elements to approximate displacement, pressure and volumetric flux, respectively. For the sake of simplicity, in the following, we assume homogeneous boundary conditions on $\partial\Omega = \Gamma_D^m = \Gamma_D^f$. The corresponding discrete function spaces are then given by

$$\begin{aligned}
W_h &= \left\{ w_h \in L^2(\Omega) \,\middle|\, \forall T \in \mathcal{T}_h,\, w_h|_T \in \mathbb{P}_0 \right\},\\
\boldsymbol{Z}_h &= \left\{ \boldsymbol{z}_h \in H(\mathrm{div};\Omega) \,\middle|\, \forall T \in \mathcal{T}_h,\, \boldsymbol{z}_h|_T \in \mathbb{RT}_0 \right\},\\
\boldsymbol{V}_h &= \left\{ \boldsymbol{v}_h \in [H_0^1(\Omega)]^d \,\middle|\, \forall T \in \mathcal{T}_h,\, \boldsymbol{v}_h|_T \in [\mathbb{P}_k]^d \right\}, \qquad k \in \{1,2\},
\end{aligned}$$

where $\mathbb{P}_l$ denotes the space of scalar piecewise polynomial functions with polynomial degree $l \in \{0,1,2\}$, and $\mathbb{RT}_0 = \{\boldsymbol{x} \mapsto \boldsymbol{a} + b\boldsymbol{x} \mid \boldsymbol{a} \in \mathbb{R}^d,\ b \in \mathbb{R}\}$ denotes the space of lowest-order Raviart–Thomas elements. We note, the elements of $\boldsymbol{V}_h$ are zero on the boundary. For better distinction, in the remaining work, let $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$ denote $W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ for $k = 1$ and $k = 2$, respectively. We refer also to them as lowest-order discretization and unconditionally stable discretization.

Furthermore, for the temporal discretization, we use the implicit Euler method. For this, let a partition $\{t^n\}_n$ of the time interval $(0,T)$ with (constant) time step size $\tau = t^n - t^{n-1} > 0$ be given.

Then given initial data $(p,\boldsymbol{u})_h^0 \in W_h \times \boldsymbol{V}_h$, at each time step $n \geq 1$, the discrete problem reads: Given $(p,\boldsymbol{q},\boldsymbol{u})_h^{n-1} \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, find $(p,\boldsymbol{q},\boldsymbol{u})_h^n \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, satisfying for all $(w,\boldsymbol{z},\boldsymbol{v})_h \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$

$$\left\langle \phi^{n-1}(s_\mathrm{w}^n - s_\mathrm{w}^{n-1}), w_h \right\rangle + \alpha \left\langle s_\mathrm{w}^n \nabla\cdot(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}), w_h \right\rangle + \tfrac{1}{N}\left\langle s_\mathrm{w}^n(p_\mathrm{E}^n - p_\mathrm{E}^{n-1}), w_h \right\rangle + \tau \left\langle \nabla\cdot\boldsymbol{q}_h^n, w_h \right\rangle = 0, \tag{6}$$

$$\left\langle k_\mathrm{w}(s_\mathrm{w}^n)^{-1}\boldsymbol{q}_h^n, \boldsymbol{z}_h \right\rangle - \left\langle p_h^n, \nabla\cdot\boldsymbol{z}_h \right\rangle = \left\langle \rho_\mathrm{w}\boldsymbol{g}, \boldsymbol{z}_h \right\rangle, \tag{7}$$

$$2\mu\left\langle \boldsymbol{\varepsilon}(\boldsymbol{u}_h^n), \boldsymbol{\varepsilon}(\boldsymbol{v}_h) \right\rangle + \lambda \left\langle \nabla\cdot\boldsymbol{u}_h^n, \nabla\cdot\boldsymbol{v}_h \right\rangle - \alpha \left\langle p_\mathrm{E}^n, \nabla\cdot\boldsymbol{v}_h \right\rangle = \left\langle \rho_\mathrm{b}\boldsymbol{g}, \boldsymbol{v}_h \right\rangle, \tag{8}$$

where $s_\mathrm{w}^k = s_\mathrm{w}(p_h^k)$ and $p_\mathrm{E}^k = p_\mathrm{E}(p_h^k)$, $k \in \{n-1,n\}$ and $\phi^{n-1} = \phi(\boldsymbol{u}_h^{n-1}, p_h^{n-1})$. Here, $\langle\cdot,\cdot\rangle$ denotes the standard $L^2(\Omega)$ scalar product.

Remark 1 (*Volume Conservation*)**.** The discretization (6)–(8) is volume-conservative as by Eq. (4) it holds

$$\phi^n s_\mathrm{w}^n - \phi^{n-1}s_\mathrm{w}^{n-1} = \phi^{n-1}(s_\mathrm{w}^n - s_\mathrm{w}^{n-1}) + s_\mathrm{w}^n\left(\alpha\nabla\cdot(\boldsymbol{u}^n - \boldsymbol{u}^{n-1}) + \frac{1}{N}(p_\mathrm{E}^n - p_\mathrm{E}^{n-1})\right).$$

Remark 2 (*Stability*)**.** For the linear Biot equations, the discretization $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ is only conditionally stable [33]. It does not satisfy an inf–sup condition uniformly with respect to the physical parameters. In particular for small permeability, volumetric locking may occur. However, the discretization $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$ is unconditionally stable [34], satisfying the inf–sup condition uniformly for the physical parameters. As the linear Biot equations are only a special case of the nonlinear Biot equations (1)–(3), obtained for $p_\mathrm{w} \geq 0$ in Ω, stability properties of $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$ also translate to the nonlinear Biot equations.

4. Monolithic and decoupled linearization schemes

In the following, we consider four linearization schemes. First, we apply the monolithic Newton method, being commonly the first choice when linearizing a nonlinear problem. Second, we propose a linearization scheme, which employs constant diagonal stabilization in order to linearize and decouple simultaneously the mechanics and flow equations. Furthermore, we introduce two modifications of the latter method, utilizing both decoupling and first order Taylor approximations. For direct comparison, we formulate all schemes in incremental form.

Monolithic schemes vs. iterative operator splitting schemes for saddle point problems. Containing more information on coupling terms, a monolithic scheme for solving Biot's equations is per se more stable, whereas for iterative operator splitting schemes, stability is always an issue necessary to be checked. However, in contrast to robust splitting schemes, for a monolithic scheme, a fully coupled simulator with advanced solver technology is often required, able to handle the saddle point structure of the problem. For that, one possibility is to apply an iterative splitting scheme as preconditioner for a Krylov subspace method. In fact, this is more efficient and robust than applying the iterative solver itself, cf. e.g. [13] for the linear Biot equations. In case only separate simulators are available, the concept of preconditioning the coupled problem cannot be applied in the same sense. But we note that acceleration techniques as Anderson acceleration can be applied as post-processing to the iterative splitting schemes, acting as a preconditioned nonlinear GMRES solver applied to the coupled problem, cf. Section 6.

4.1. Notation of residuals

For the incremental formulation of the linearization schemes, we introduce naturally defined residuals of the coupled, discrete problem (6)–(8). Given data $(p,\boldsymbol{q},\boldsymbol{u})_h^{n-1} \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ for time step $n-1$, the residuals at time step n evaluated at some state $(p,\boldsymbol{q},\boldsymbol{u})_h \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ and tested with $(w,\boldsymbol{z},\boldsymbol{v})_h \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ are defined by

$$\begin{aligned} r_p^n((p,\boldsymbol{q},\boldsymbol{u})_h; w_h) &= -\Big(\left\langle \phi^{n-1}(s_\mathrm{w}(p_h) - s_\mathrm{w}^{n-1}), w_h \right\rangle + \alpha \left\langle s_\mathrm{w}(p_h) \nabla\cdot(\boldsymbol{u}_h - \boldsymbol{u}_h^{n-1}), w_h \right\rangle \\ &\qquad + \frac{1}{N} \left\langle s_\mathrm{w}(p_h)(p_\mathrm{E}(p_h) - p_\mathrm{E}^{n-1}), w_h \right\rangle + \tau \left\langle \nabla\cdot\boldsymbol{q}_h, w_h \right\rangle \Big), \\ r_q^n((p,\boldsymbol{q},\boldsymbol{u})_h; \boldsymbol{z}_h) &= \left\langle \rho_\mathrm{w}\boldsymbol{g}, \boldsymbol{z}_h \right\rangle - \Big(\left\langle k_\mathrm{w}(s_\mathrm{w}(p_h))^{-1}\boldsymbol{q}_h, \boldsymbol{z}_h \right\rangle - \left\langle p_h, \nabla\cdot\boldsymbol{z}_h \right\rangle \Big), \\ r_u^n((p,\boldsymbol{q},\boldsymbol{u})_h; \boldsymbol{v}_h) &= \left\langle \rho_\mathrm{b}\boldsymbol{g}, \boldsymbol{v}_h \right\rangle - \Big(2\mu \left\langle \boldsymbol{\varepsilon}(\boldsymbol{u}_h), \boldsymbol{\varepsilon}(\boldsymbol{v}_h) \right\rangle + \lambda \left\langle \nabla\cdot\boldsymbol{u}_h, \nabla\cdot\boldsymbol{v}_h \right\rangle - \alpha \left\langle p_\mathrm{E}(p_h), \nabla\cdot\boldsymbol{v}_h \right\rangle \Big). \end{aligned}$$

Shorter, given a sequence of approximations $(p,\boldsymbol{q},\boldsymbol{u})_h^{n,i} \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ of $(p,\boldsymbol{q},\boldsymbol{u})_h^n \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, for time steps n and iterations $i \in \mathbb{N}$, we define

$$\begin{aligned} r_p^{n,i}(w_h) &= r_p^n((p,\boldsymbol{q},\boldsymbol{u})_h^{n,i}; w_h), \\ r_q^{n,i}(\boldsymbol{z}_h) &= r_q^n((p,\boldsymbol{q},\boldsymbol{u})_h^{n,i}; \boldsymbol{z}_h), \\ r_u^{n,i}(\boldsymbol{v}_h) &= r_u^n((p,\boldsymbol{q},\boldsymbol{u})_h^{n,i}; \boldsymbol{v}_h), \\ r_u^{n,i/i-1}(\boldsymbol{v}_h) &= r_u^n((p,\boldsymbol{q})_h^{n,i}, \boldsymbol{u}_h^{n,i-1}; \boldsymbol{v}_h). \end{aligned}$$

4.2. Monolithic Newton's method

We apply the standard Newton method, linearizing the coupled, discrete problem (6)–(8) in a monolithic fashion.

Scheme. The monolithic Newton method reads: For each time step n, given the initial guess $(p,\boldsymbol{q},\boldsymbol{u})_h^{n,0} = (p,\boldsymbol{q},\boldsymbol{u})_h^{n-1}$, loop over the iterations $i \in \mathbb{N}$ until convergence is reached. Given data at the previous time step $n-1$ and iteration $i-1$, find the increments $\Delta(p,\boldsymbol{q},\boldsymbol{u})_h^{n,i} \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, satisfying the coupled, linear problem, for all $(w,\boldsymbol{z},\boldsymbol{v})_h \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$,

$$\left\langle \left(\phi^{n,i-1} \frac{\partial s_\mathrm{w}}{\partial p_\mathrm{w}}(p_h^{n,i-1}) + \frac{1}{N}(s_\mathrm{w}^{n,i-1})^2 \right) \Delta p_h^{n,i}, w_h \right\rangle + \alpha \left\langle s_\mathrm{w}^{n,i-1} \nabla\cdot\Delta\boldsymbol{u}_h^{n,i}, w_h \right\rangle + \tau \left\langle \nabla\cdot\Delta\boldsymbol{q}_h^{n,i}, w_h \right\rangle = r_p^{n,i-1}(w_h), \tag{9}$$

$$\left\langle k_\mathrm{w}(s_\mathrm{w}^{n,i-1})^{-1} \Delta\boldsymbol{q}_h^{n,i}, \boldsymbol{z}_h \right\rangle + \left\langle \left(\frac{\partial}{\partial p_\mathrm{w}} k_\mathrm{w}(s_\mathrm{w}) \Big|_{p_h^{n,i-1}} \right)^{-1} \boldsymbol{q}_h^{n,i-1} \Delta p_h^{n,i}, \boldsymbol{z}_h \right\rangle - \left\langle \Delta p_h^{n,i}, \nabla\cdot\boldsymbol{z}_h \right\rangle = r_q^{n,i-1}(\boldsymbol{z}_h), \tag{10}$$

$$2\mu \left\langle \boldsymbol{\varepsilon}(\Delta\boldsymbol{u}_h^{n,i}), \boldsymbol{\varepsilon}(\boldsymbol{v}_h) \right\rangle + \lambda \left\langle \nabla\cdot\Delta\boldsymbol{u}_h^{n,i}, \nabla\cdot\boldsymbol{v}_h \right\rangle - \alpha \left\langle s_\mathrm{w}^{n,i-1} \Delta p_h^{n,i}, \nabla\cdot\boldsymbol{v}_h \right\rangle = r_u^{n,i-1}(\boldsymbol{v}_h), \tag{11}$$

and set

$$(p,\boldsymbol{q},\boldsymbol{u})_h^{n,i} = (p,\boldsymbol{q},\boldsymbol{u})_h^{n,i-1} + \Delta(p,\boldsymbol{q},\boldsymbol{u})_h^{n,i}.$$

After convergence is reached at iteration N, set $(p,\boldsymbol{q},\boldsymbol{u})_h^n = (p,\boldsymbol{q},\boldsymbol{u})_h^{n,N}$.

Properties. Newton's method is known to be locally, quadratically convergent, which makes the method commonly the first choice linearization method. However, in general it is not robust and has the following drawbacks:

- In order to ensure convergence, the time step size has to be chosen sufficiently small depending on the mesh size. Then the initial guess is sufficiently close to the unknown solution.
- The need for a good initial guess can be relaxed by using step size control, allowing a bigger time step size. Anderson acceleration applied as post-processing can be interpreted as such, for more details, cf. Section 6.

- Bounded derivatives of constitutive laws have to be available. In practice, nonlinearities employed in the model (1)–(3) are not necessarily Lipschitz continuous; e.g., the relative permeability for soils. In particular, in the transition from the partially to the fully saturated regime, the derivative of the relative permeability modeled by the van Genuchten model [4] can be unbounded. Consequently, the Jacobian might become ill-conditioned.
- The coupled problem (9)–(11) has a saddle point structure. Hence, an advanced solver architecture is required, in particular when considering preconditioned Krylov subspace methods. In the context of Biot's equations, e.g., the application of a fixed-stress type solver or preconditioner [13] fixes this issue.

4.3. Fixed-Stress-L-scheme — A Picard-type simultaneous linearization and splitting

We propose a novel, robust linearization scheme for Eqs. (6)–(8). It is essentially a simultaneous application of the L-scheme linearization for the Richards equation, cf. e.g. [22], and the fixed-stress splitting scheme for the linear Biot equations, cf. e.g. [7]. Both follow the same concept, utilizing diagonal stabilization. In the following, we refer to the scheme as Fixed-Stress-L-scheme (FSL). As derived in Section 5, it can be interpreted as L-scheme linearization of the nonlinear Biot equations reduced to a pure pressure formulation or alternatively as nonlinear Gauss–Seidel-type solver, consisting of cheap iterations allowing separate, sophisticated simulators for the mechanical and flow subproblems. In Section 5, we show convergence of the Fixed-Stress-L-scheme under physical assumptions. Before defining the Fixed-Stress-L-scheme, we recall main ideas of both the L-scheme and the fixed-stress splitting scheme.

Main ideas of the L-scheme. The L-scheme is an inexact Newton's method, employing constant linearization for monotone and Lipschitz continuous terms. For remaining contributions Picard-type linearization is applied. Effectively, this approach is identical with applying a standard Picard iteration with additional diagonal stabilization. All in all, no explicit derivatives are required, easing the cost of the assembly at the price that only linear convergence can be expected. Under mild conditions, this concept has been rigorously proven to be globally convergent for various porous media applications, e.g., [20–22,25]. Moreover, as pointed out by [22], the resulting linear problem is expected to be significantly better conditioned than the corresponding linear problem obtained by Newton's method.

Regarding the nonlinear Biot equations, models for the saturation $s_{\mathrm{w}} = s_{\mathrm{w}}(p)$ are commonly non-decreasing and Lipschitz continuous. Hence, assuming that $\phi^{n,i-1} \geq 0$ on Ω, the above criteria apply to the saturation contribution in Eq. (9). The approximation of the saturation at iteration i is then given by

$$s_{\mathrm{w}}^{n} = s_{\mathrm{w}}(p_h^n) \approx s_{\mathrm{w}}(p_h^{n,i}) \approx s_{\mathrm{w}}(p_h^{n,i-1}) + L(p_h^{n,i} - p_h^{n,i-1}) = s_{\mathrm{w}}(p_h^{n,i-1}) + L\Delta p_h^{n,i}, \tag{12}$$

where $L \in \mathbb{R}_+$ is a sufficiently large tuning parameter, usually set equal to the Lipschitz constant L_{s} of s_{w}. As coupling terms are not monotone, Picard-type linearization is applied to the remaining contributions of Eqs. (6)–(8).

Main ideas of the fixed-stress splitting scheme. Considering the linear Biot equations, their linearization results in a saddle point problem, thus, requiring an advanced solver technology for efficient solution. For this purpose, physically motivated, robust, iterative splitting schemes are widely-used as, e.g., the fixed-stress splitting scheme, originally introduced by [5]. As it decouples mechanics and flow equations, separate simulators can be utilized for both subproblems, reducing the total complexity to solving simpler, better conditioned problems. The robust decoupling is accomplished via sufficient diagonal stabilization, introducing a tuning parameter β_{FS}, for which several values are suggested in the literature. Commonly, it is chosen as $\beta_{\mathrm{FS}} = \frac{\alpha^2}{K_{\mathrm{dr}}}$ with K_{dr} the effective bulk modulus. A physically motivated choice is $K_{\mathrm{dr}}^{\mathrm{phy}} = \frac{2\mu}{d} + \lambda$, yielding $\beta_{\mathrm{FS}}^{\mathrm{phy}} = \frac{\alpha^2}{K_{\mathrm{dr}}^{\mathrm{phy}}}$, which is not necessarily optimal. An optimal, theoretical choice is still an open research question [8,9,35]. Theoretical approaches suggest choosing only the half of $\beta_{\mathrm{FS}}^{\mathrm{phy}}$. However, also the domain and the boundary conditions seem to have an influence which has not been quantified in the literature, yet.

Scheme. We observe that applying the monolithic L-scheme, as just explained to the nonlinear, discrete Biot equations (6)–(8), results in a linear problem equivalent with that for single phase flow in heterogeneous media, for which the fixed-stress splitting scheme is an attractive solver [9]. Both schemes are realized via diagonal stabilization. Anticipating the dynamics to be mainly governed by the flow problem, cf. Assumption (A4), and the mechanics problem to be much simpler, a simultaneous application of the L-scheme and the fixed-stress splitting scheme yields an attractive linearization scheme incorporating the decoupling of flow and mechanics equations.

Written as iterative scheme in incremental form, the resulting Fixed-Stress-L-scheme reads: For each time step n, given the initial guess $(p, \boldsymbol{q}, \boldsymbol{u})_h^{n,0} = (p, \boldsymbol{q}, \boldsymbol{u})_h^{n-1}$, loop over the iterations $i \in \mathbb{N}$ until convergence is reached. For each iteration i, perform two steps:

1. Step: Set $L = L_{\mathrm{s}}$, the Lipschitz constant of s_{w}, and $\beta_{\mathrm{FS}} = \alpha^2/\left(\frac{2\mu}{d} + \lambda\right)$. Given $(p, \boldsymbol{q}, \boldsymbol{u})_h^{n,i-1}, (p, \boldsymbol{q}, \boldsymbol{u})_h^{n-1} \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, find the increments $\Delta(p, \boldsymbol{q})_h^{n,i} \in W_h \times \boldsymbol{Z}_h$, satisfying, for all $(w, \boldsymbol{z})_h \in W_h \times \boldsymbol{Z}_h$,

$$\left\langle \left(L + \tfrac{1}{N} + \beta_{\mathrm{FS}}\right) \Delta p_h^{n,i}, w_h \right\rangle + \tau \left\langle \nabla \cdot \Delta \boldsymbol{q}_h^{n,i}, w_h \right\rangle = r_p^{n,i-1}(w_h), \tag{13}$$

$$\left\langle k_{\mathrm{w}}(s_{\mathrm{w}}^{n,i-1})^{-1} \Delta \boldsymbol{q}_h^{n,i}, \boldsymbol{z}_h \right\rangle - \left\langle \Delta p_h^{n,i}, \nabla \cdot \boldsymbol{z}_h \right\rangle = r_q^{n,i-1}(\boldsymbol{z}_h), \tag{14}$$

and set

$$(p, \boldsymbol{q})_h^{n,i} = (p, \boldsymbol{q})_h^{n,i-1} + \Delta(p, \boldsymbol{q})_h^{n,i}.$$

2. Step: Given $((p, \boldsymbol{q})_h^{n,i}, \boldsymbol{u}_h^{n,i-1}) \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, find the increment $\Delta\boldsymbol{u}_h^{n,i} \in \boldsymbol{V}_h$, satisfying, for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$,

$$2\mu \left\langle \boldsymbol{\varepsilon}(\Delta\boldsymbol{u}_h^{n,i}), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\right\rangle + \lambda \left\langle \nabla \cdot \Delta\boldsymbol{u}_h^{n,i}, \nabla \cdot \boldsymbol{v}_h\right\rangle = r_u^{n,i/i-1}(\boldsymbol{v}_h), \tag{15}$$

and set

$$\boldsymbol{u}_h^{n,i} = \boldsymbol{u}_h^{n,i-1} + \Delta\boldsymbol{u}_h^{n,i}.$$

After convergence is reached at iteration N, set $(p, \boldsymbol{q}, \boldsymbol{u})_h^n = (p, \boldsymbol{q}, \boldsymbol{u})_h^{n,N}$.

Properties. The Fixed-Stress-L-scheme inherits its properties from the underlying methods. It does not require the evaluation of any derivatives, increasing the speed of the assembly process. It is very robust but guarantees only linear convergence, cf. Theorem 3. Furthermore, the Fixed-Stress-L-scheme requires the independent solution of the mechanical and flow equations, allowing to use separate simulators. In particular, the overall method utilizes a single loop in contrast to the Newton's method combined with a fixed-stress splitting scheme as iterative solver. Inheriting the block structure of the linear Biot equations, the implementation of the Fixed-Stress-L-scheme and its subsequent modifications follows the structure of splitting schemes for Biot equations in general, cf. e.g. [15,36].

4.4. Quasi-Newton modifications of the Fixed-Stress-L-scheme

The Fixed-Stress-L-scheme employs constant linearization for the fluid volume ϕs_w with respect to fluid pressure, utilizing an upper bound for the Lipschitz constant. In many practical situations, this approach is quite pessimistic. Recalling the assumption that the flow problem dominates the dynamics of the system, we expect the simultaneous application of the fixed-stress splitting scheme and more sophisticated flow linearizations to be only slightly less robust than the Fixed-Stress-L-scheme. Independent of the flow linearization, diagonal stabilization is added by the splitting scheme anyhow increasing the robustness. In the following, based on the derivation of the Fixed-Stress-L-scheme in Section 5, cf. Remark 3, we couple simultaneously a modified Picard method [26] and Newton's method with the fixed-stress splitting scheme yielding the Fixed-Stress-Modified-Picard method and the Fixed-Stress-Newton method, respectively. The modified Picard method, in particular, is a widely-used linearization scheme for the Richards equation and hence rises also interest for its use for the linearization of the discrete, nonlinear Biot equations (6)–(8).

Fixed-Stress-Modified-Picard method. Applied to the Richards equation, the modified Picard method employs a first order Taylor approximation as linearization for the saturation and a Picard-type linearization for the possibly Hölder continuous permeability. By employing a first order approximation of the fluid volume ϕs_w with respect to fluid pressure instead, and by coupling simultaneously with the fixed-stress splitting scheme, we obtain a linearization scheme for Eqs. (6)–(8). For later reference, we denote the resulting scheme by Fixed-Stress-Modified-Picard-scheme. It is essentially identical with the Fixed-Stress-L-scheme but with an iteration dependent L parameter and hence modified first fixed-stress step (1. Step). We exchange Eqs. (13)–(14) with

$$\left\langle \left(\phi^{n,i-1}\frac{\partial s_\mathrm{w}}{\partial p_\mathrm{w}}(p_h^{n,i-1}) + \left(\tfrac{1}{N} + \beta_\mathrm{FS}\right)(s_\mathrm{w}^{n,i-1})^2\right)\Delta p_h^{n,i}, w_h\right\rangle + \tau\left\langle \nabla\cdot\Delta\boldsymbol{q}_h^{n,i}, w_h\right\rangle = r_p^{n,i-1}(w_h), \tag{16}$$

$$\left\langle k_\mathrm{w}(s_\mathrm{w}^{n,i-1})^{-1}\Delta\boldsymbol{q}_h^{n,i}, \boldsymbol{z}_h\right\rangle - \left\langle \Delta p_h^{n,i}, \nabla\cdot\boldsymbol{z}_h\right\rangle = r_q^{n,i-1}(\boldsymbol{z}_h). \tag{17}$$

Fixed-Stress-Newton method. In case the permeability is Lipschitz continuous, the simultaneous application of the fixed-stress splitting scheme and linearization of the flow equations via Newton's method yields an attractive linearization scheme for Eqs. (6)–(8). For later reference, we denote the scheme by Fixed-Stress-Newton method. It is essentially identical with the Fixed-Stress-L-scheme but with an iteration dependent L parameter and additional contribution in Darcy's law and hence modified first fixed-stress step (1. Step). We exchange Eqs. (13)–(14) with

$$\left\langle \left(\phi^{n,i-1}\frac{\partial s_\mathrm{w}}{\partial p_\mathrm{w}}(p_h^{n,i-1}) + \left(\tfrac{1}{N} + \beta_\mathrm{FS}\right)(s_\mathrm{w}^{n,i-1})^2\right)\Delta p_h^{n,i}, w_h\right\rangle + \tau\left\langle \nabla\cdot\Delta\boldsymbol{q}_h^{n,i}, w_h\right\rangle = r_p^{n,i-1}(w_h), \tag{18}$$

$$\left\langle k_\mathrm{w}(s_\mathrm{w}^{n,i-1})^{-1}\Delta\boldsymbol{q}_h^{n,i}, \boldsymbol{z}_h\right\rangle + \left\langle \left(\frac{\partial}{\partial p_\mathrm{w}}k_\mathrm{w}(s_\mathrm{w})\Big|_{p_h^{n,i-1}}\right)^{-1}\boldsymbol{q}_h^{n,i-1}\Delta p_h^{n,i}, \boldsymbol{z}_h\right\rangle - \left\langle \Delta p_h^{n,i}, \nabla\cdot\boldsymbol{z}_h\right\rangle = r_q^{n,i-1}(\boldsymbol{z}_h). \tag{19}$$

We note that the Fixed-Stress-Newton method is also closely related to applying a single fixed-stress iteration as inexact solver for the linear problem (9)–(11) arising from Newton's method.

4.5. $L^2(\Omega)$-type stopping criterion

For the numerical examples in Section 8, we employ a combination of an absolute and a relative $L^2(\Omega)$-type stopping criterion, closely related to the standard algebraic $l^2(\mathbb{R})$-type criterion. Given tolerances $\varepsilon_\mathrm{a},\ \varepsilon_\mathrm{r} \in \mathbb{R}_+$, we denote an iteration as converged if it holds

$$\|\Delta p_h^{n,i}\|_{L^2(\Omega)} + \|\Delta \boldsymbol{q}_h^{n,i}\|_{L^2(\Omega)} + \|\Delta \boldsymbol{u}_h^{n,i}\|_{L^2(\Omega)} < \varepsilon_{\mathrm{a}}, \qquad \text{and} \qquad \frac{\|\Delta p_h^{n,i}\|_{L^2(\Omega)}}{\|p_h^{n,i}\|_{L^2(\Omega)}} + \frac{\|\Delta \boldsymbol{q}_h^{n,i}\|_{L^2(\Omega)}}{\|\boldsymbol{q}_h^{n,i}\|_{L^2(\Omega)}} + \frac{\|\Delta \boldsymbol{u}_h^{n,i}\|_{L^2(\Omega)}}{\|\boldsymbol{u}_h^{n,i}\|_{L^2(\Omega)}} < \varepsilon_{\mathrm{r}}.$$

5. Convergence theory for simultaneous linearization and splitting via the L-scheme

In the following, we show convergence of the Fixed-Stress-L-scheme (13)–(15) under mild, physical assumptions. For this purpose, we first formulate the nonlinear discrete problem (6)–(8) as an algebraic problem, reduce the problem to a pure pressure problem by exact inversion and apply the L-scheme as linearization identical to the Fixed-Stress-L-scheme (13)–(15). Convergence follows then from an abstract convergence result for the L-scheme. For simplicity, we assume vanishing initial data and a homogeneous and isotropic material.

Notation . Let $\mathbb{R}^{n_{\mathrm{p}}}$, $\mathbb{R}^{n_{\mathrm{q}}}$ and $\mathbb{R}^{n_{\mathrm{u}}}$ denote the coefficient vector spaces corresponding to the functional spaces W_h, $\boldsymbol{Z}_h$ and $\boldsymbol{V}_h$, respectively, given standard bases. Let $\langle\cdot,\cdot\rangle$ denote the classical l^2 vector scalar product on $\mathbb{R}^n$, $n \in \mathbb{N}$. Furthermore, for symmetric, positive definite matrices $\mathbf{M} \in \mathbb{R}^{n\times n}$, let the vector norm $\|\cdot\|_{\mathbf{M}}$ be defined by $\|\mathbf{v}\|_{\mathbf{M}}^2 = \langle \mathbf{M}\mathbf{v}, \mathbf{v}\rangle$, $\mathbf{v} \in \mathbb{R}^n$.

Algebraic formulation of the nonlinear, discrete Biot equations. Given finite element bases for $W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, the nonlinear, discrete Biot equations (6)–(8) translate to the algebraic equations

$$\mathbf{S}_{\mathrm{pp}}(\mathbf{p})\left(\mathbf{M}_{\mathrm{pp}}\boldsymbol{\phi}_0 + \alpha \mathbf{D}_{\mathrm{pu}}\mathbf{u} + \tfrac{1}{N}\mathbf{M}_{\mathrm{pp}}\mathbf{p}_{\mathrm{E}}(\mathbf{p})\right) + \tau \mathbf{D}_{\mathrm{pq}}\mathbf{q} = \mathbf{f}_{\mathrm{p}} \tag{20}$$

$$\mathbf{K}_{\mathrm{qq}}(\mathbf{p})^{-1}\mathbf{q} - \mathbf{D}_{\mathrm{pq}}^{\top}\mathbf{p} = \mathbf{f}_{\mathrm{q}} \tag{21}$$

$$\mathbf{A}_{\mathrm{uu}}\mathbf{u} - \alpha \mathbf{D}_{\mathrm{pu}}^{\top}\mathbf{p}_{\mathrm{E}}(\mathbf{p}) = \mathbf{f}_{\mathrm{u}}. \tag{22}$$

We omit the detailed definition of the finite element matrices and vectors used in Eqs. (20)–(22), as they are assembled in a standard way. We comment solely on their origin and their properties relevant for further discussion. For this purpose, let $(\mathbf{p}', \mathbf{q}', \mathbf{u}') \in \mathbb{R}^{n_{\mathrm{p}}} \times \mathbb{R}^{n_{\mathrm{q}}} \times \mathbb{R}^{n_{\mathrm{u}}}$, $(\mathbf{q}^\star, \mathbf{u}^\star) \in \mathbb{R}^{n_{\mathrm{q}}} \times \mathbb{R}^{n_{\mathrm{u}}}$ be arbitrary coefficient vectors corresponding to some $(p_h', \boldsymbol{q}_h', \boldsymbol{u}_h') \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, $(\boldsymbol{q}_h^\star, \boldsymbol{u}_h^\star) \in \boldsymbol{Z}_h \times \boldsymbol{V}_h$.

- Let $\mathbf{p} \in \mathbb{R}^{n_{\mathrm{p}}}$, $\mathbf{q} \in \mathbb{R}^{n_{\mathrm{q}}}$, $\mathbf{u} \in \mathbb{R}^{n_{\mathrm{u}}}$ denote the algebraic pressure, volumetric flux and displacement coefficient vectors corresponding to $(p, \boldsymbol{q}, \boldsymbol{u})_h^n \in W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ with respect to the chosen bases.
- Let $\mathbf{M}_{\mathrm{pp}} \in \mathbb{R}^{n_{\mathrm{p}}\times n_{\mathrm{p}}}$ be the natural mass matrix for the pressure variable incorporating local mesh information for $\mathcal{T}_h$ such that $\|\mathbf{p}'\|_{\mathbf{M}_{\mathrm{pp}}} = \|p_h'\|_{L^2(\Omega)}$.
- Let $\mathbf{S}_{\mathrm{pp}} : \mathbb{R}^{n_{\mathrm{p}}} \to \mathbb{R}^{n_{\mathrm{p}}\times n_{\mathrm{p}}}$ denote a diagonal matrix with element-wise saturation s_{w} on the diagonal, i.e., $\mathbf{S}_{\mathrm{pp}}(\mathbf{p}')_{kk} = s_{\mathrm{w}}(\mathbf{p}_k')$ for $k \in \{1, \ldots, n_{\mathrm{p}}\}$.
- Let $\mathbf{D}_{\mathrm{pu}} \in \mathbb{R}^{n_{\mathrm{p}}\times n_{\mathrm{u}}}$ and $\mathbf{D}_{\mathrm{pq}} \in \mathbb{R}^{n_{\mathrm{p}}\times n_{\mathrm{q}}}$ denote the matrices corresponding to the divergence operating on displacement and volumetric flux spaces, respectively, mapping into the pressure space, such that $\langle \mathbf{p}', \mathbf{D}_{\mathrm{pu}}\mathbf{u}'\rangle = \langle p_h', \nabla\cdot \boldsymbol{u}_h'\rangle$ and $\langle \mathbf{p}', \mathbf{D}_{\mathrm{pq}}\mathbf{q}'\rangle = \langle p_h', \nabla\cdot \boldsymbol{q}_h'\rangle$.
- Let $\mathbf{p}_{\mathrm{E}} : \mathbb{R}^{n_{\mathrm{p}}} \to \mathbb{R}^{n_{\mathrm{p}}}$ correspond to the element-wise equivalent pore pressure p_{E}, i.e., $\mathbf{p}_{\mathrm{E}}(\mathbf{p}')_k = p_{\mathrm{E}}(\mathbf{p}_k')$ for $k \in \{1, \ldots, n_{\mathrm{p}}\}$.
- Let $\boldsymbol{\phi}_0 \in \mathbb{R}^{n_{\mathrm{p}}}$ correspond to the element-wise porosity, such that the components of $\mathbf{M}_{\mathrm{pp}}\boldsymbol{\phi}_0 + \alpha\mathbf{D}_{\mathrm{pu}}\mathbf{u} + \frac{1}{N}\mathbf{M}_{\mathrm{pp}}\mathbf{p}_{\mathrm{E}}(\mathbf{p})$ correspond to the element-wise porosity of the deformed material scaled by the element size.
- Let $\mathbf{K}_{\mathrm{qq}}^{-1} : \mathbb{R}^{n_{\mathrm{p}}} \to \mathbb{R}^{n_{\mathrm{q}}\times n_{\mathrm{q}}}$ denote the volumetric flux mass matrix, weighted by the nonlinear permeability contribution $k_{\mathrm{w}}^{-1}(s_{\mathrm{w}})$ in Darcy's law, such that $\langle \mathbf{K}_{\mathrm{qq}}^{-1}(\mathbf{p}')\mathbf{q}', \mathbf{q}^\star\rangle = \langle k_{\mathrm{w}}^{-1}(s_{\mathrm{w}}(p_h'))\boldsymbol{q}_h', \boldsymbol{q}_h^\star\rangle$.
- Let $\mathbf{A}_{\mathrm{uu}} \in \mathbb{R}^{n_{\mathrm{u}}\times n_{\mathrm{u}}}$ denote the stiffness matrix, corresponding to the linear elasticity equations, such that $\langle \mathbf{A}_{\mathrm{uu}}\mathbf{u}', \mathbf{u}^\star\rangle = 2\mu\langle \boldsymbol{\varepsilon}(\boldsymbol{u}_h'), \boldsymbol{\varepsilon}(\boldsymbol{u}_h^\star)\rangle + \lambda\langle \nabla\cdot\boldsymbol{u}_h', \nabla\cdot\boldsymbol{u}_h^\star\rangle$.
- $\mathbf{f}_{\mathrm{p}} \in \mathbb{R}^{n_{\mathrm{p}}}$, $\mathbf{f}_{\mathrm{q}} \in \mathbb{R}^{n_{\mathrm{q}}}$ and $\mathbf{f}_{\mathrm{u}} \in \mathbb{R}^{n_{\mathrm{u}}}$ incorporate solution independent contributions as volume effects and Neumann boundary conditions and data at the previous time step. Furthermore, let local mesh information be incorporated.

Compact formulation of the algebraic problem. First, we define the porosity of the deformed material

$$\boldsymbol{\phi}(\mathbf{p}, \mathbf{u}) = \mathbf{M}_{\mathrm{pp}}\boldsymbol{\phi}_0 + \alpha\mathbf{D}_{\mathrm{pu}}\mathbf{u} + \tfrac{1}{N}\mathbf{M}_{\mathrm{pp}}\mathbf{p}_{\mathrm{E}}(\mathbf{p}).$$

Then given data at the previous time step and the corresponding coefficient vectors $(\mathbf{p}^{n-1}, \mathbf{u}^{n-1}) \in \mathbb{R}^{n_{\mathrm{p}}} \times \mathbb{R}^{n_{\mathrm{p}}}$, it holds

$$\boldsymbol{\phi}(\mathbf{p}, \mathbf{u}) = \boldsymbol{\phi}(\mathbf{p}^{n-1}, \mathbf{u}^{n-1}) + \alpha\mathbf{D}_{\mathrm{pu}}(\mathbf{u} - \mathbf{u}^{n-1}) + \tfrac{1}{N}\mathbf{M}_{\mathrm{pp}}(\mathbf{p}_{\mathrm{E}}(\mathbf{p}) - \mathbf{p}_{\mathrm{E}}(\mathbf{p}^{n-1})) \tag{23}$$

We rewrite the displacement contribution by inverting the mechanics equation (22). The equation holds for each time step n and $\mathbf{f}_{\mathrm{u}}$ is constant in time, as only Dirichlet boundary conditions are applied, cf. Section 3. Hence, it holds

$$\mathbf{A}_{\mathrm{uu}}(\mathbf{u} - \mathbf{u}^{n-1}) - \alpha\mathbf{D}_{\mathrm{pu}}^{\top}(\mathbf{p}_{\mathrm{E}}(\mathbf{p}) - \mathbf{p}_{\mathrm{E}}(\mathbf{p}^{n-1})) = \mathbf{0}.$$

Inverting and inserting in Eq. (23), allows us to write the porosity as function of pressure only

$$\boldsymbol{\phi}(\mathbf{p}) = \boldsymbol{\phi}(\mathbf{p},\mathbf{u}) = \boldsymbol{\phi}(\mathbf{p}^{n-1}) + \left(\alpha^2 \mathbf{D}_{\mathrm{pu}}\mathbf{A}_{\mathrm{uu}}^{-1}\mathbf{D}_{\mathrm{pu}}^{\top} + \tfrac{1}{N}\mathbf{M}_{\mathrm{pp}}\right)(\mathbf{p}_{\mathrm{E}}(\mathbf{p}) - \mathbf{p}_{\mathrm{E}}(\mathbf{p}^{n-1})). \tag{24}$$

Next, we define the abbreviations

$$\mathbf{b}(\mathbf{p}) = \mathbf{S}_{\mathrm{pp}}(\mathbf{p})\boldsymbol{\phi}(\mathbf{p}), \qquad \mathbf{D} = \mathbf{D}_{\mathrm{pq}}, \qquad \mathbf{K}(\mathbf{p}) = \mathbf{K}_{\mathrm{qq}}(\mathbf{p}). \tag{25}$$

Finally, by inverting exactly Eq. (21) with respect to $\mathbf{q}$, and inserting that together with above abbreviations into Eq. (20), we obtain an equivalent, reduced problem for $\mathbf{p}$ in compact form

$$\mathbf{b}(\mathbf{p}) + \tau\mathbf{D}\mathbf{K}(\mathbf{p})\left(\mathbf{f}_{\mathrm{q}} + \mathbf{D}^{\top}\mathbf{p}\right) = \mathbf{f}_{\mathrm{p}}. \tag{26}$$

L-scheme linearization. We linearize the abstract problem (26) using the L-scheme, introducing a sequence $\{\mathbf{p}^i\}_i \subset \mathbb{R}^{n_{\mathrm{p}}}$ approximating the exact solution $\mathbf{p} \in \mathbb{R}^{n_{\mathrm{p}}}$. Given a user-defined parameter $L \in \mathbb{R}_+$, we set $\mathbf{L}_{\mathrm{pp}} = L\mathbf{M}_{\mathrm{pp}}$. Then given an initial guess $\mathbf{p}^0 \in \mathbb{R}^{n_{\mathrm{p}}}$, the scheme is defined as follows: Loop over the iterations $i \in \mathbb{N}$ until convergence is reached. At iteration i, given data $\mathbf{p}^{i-1} \in \mathbb{R}^{n_{\mathrm{p}}}$, find $\mathbf{p}^i \in \mathbb{R}^{n_{\mathrm{p}}}$ solving the linear problem

$$\mathbf{L}_{\mathrm{pp}}(\mathbf{p}^i - \mathbf{p}^{i-1}) + \mathbf{b}(\mathbf{p}^{i-1}) + \tau\mathbf{D}\mathbf{K}(\mathbf{p}^{i-1})\left(\mathbf{f}_{\mathrm{q}} + \mathbf{D}^{\top}\mathbf{p}^i\right) = \mathbf{f}_{\mathrm{p}}. \tag{27}$$

Lemma 1 (*Convergence of the L-Scheme*). *Assume* (26) *and* (27) *both have unique solutions* $\mathbf{p} \in \mathbb{R}^{n_{\mathrm{p}}}$ *and* $\mathbf{p}^i \in \mathbb{R}^{n_{\mathrm{p}}}$*, respectively. Furthermore, let the following assumptions be satisfied:*

(L1) *There exists a constant* $L_{\mathrm{b}} \in \mathbb{R}_+$ *satisfying* $\|\mathbf{b}(\mathbf{p}) - \mathbf{b}(\tilde{\mathbf{p}})\|^2_{\mathbf{M}_{\mathrm{pp}}^{-1}} \leq L_{\mathrm{b}}\langle \mathbf{b}(\mathbf{p}) - \mathbf{b}(\tilde{\mathbf{p}}), \mathbf{p} - \tilde{\mathbf{p}}\rangle$ *for all* $\mathbf{p}, \tilde{\mathbf{p}} \in \mathbb{R}^{n_{\mathrm{p}}}$*, i.e.,* $\mathbf{b}$ *is in some sense monotonically increasing and Lipschitz continuous.*

(L2) *There exist constants* $k_{\mathrm{m}}, k_{\mathrm{M}} \in \mathbb{R}_+$ *satisfying* $k_{\mathrm{m}}\|\mathbf{q}\|^2_{\mathbf{M}_{\mathrm{qq}}^{-1}} \leq \langle \mathbf{K}(\mathbf{p})\mathbf{q}, \mathbf{q}\rangle \leq k_{\mathrm{M}}\|\mathbf{q}\|^2_{\mathbf{M}_{\mathrm{qq}}^{-1}}$ *for all* $\mathbf{p} \in \mathbb{R}^{n_{\mathrm{p}}}$, $\mathbf{q} \in \mathbb{R}^{n_{\mathrm{q}}}$*. Furthermore, there exists a constant* L_{K} *satisfying* $\|(\mathbf{K}(\mathbf{p}) - \mathbf{K}(\tilde{\mathbf{p}}))\mathbf{M}_{\mathrm{qq}}\|_{\mathbf{M}_{\mathrm{qq}},\infty} \leq L_{\mathrm{K}}\|\mathbf{b}(\mathbf{p}) - \mathbf{b}(\tilde{\mathbf{p}})\|_{\mathbf{M}_{\mathrm{pp}}^{-1}}$ *for all* $\mathbf{p}, \tilde{\mathbf{p}} \in \mathbb{R}^{n_{\mathrm{p}}}$*, i.e.* $\mathbf{K}$ *is in some sense Lipschitz continuous. Here,* $\mathbf{M}_{\mathrm{qq}} \in \mathbb{R}^{n_{\mathrm{q}} \times n_{\mathrm{q}}}$ *is the natural mass matrix for the flux variable satisfying* $\|\mathbf{q}\|_{\mathbf{M}_{\mathrm{qq}}} = \|\boldsymbol{q}_h\|_{L^2(\Omega)}$ *for* $\boldsymbol{q}_h \in \boldsymbol{Z}_h$ *and corresponding coefficient vector* $\mathbf{q} \in \mathbb{R}^{n_{\mathrm{q}}}$*. Furthermore, the subordinate matrix norm* $\|\cdot\|_{\mathbf{M}_{\mathrm{qq}},\infty}$ *is defined by* $\|\mathbf{K}\|_{\mathbf{M}_{\mathrm{qq}},\infty} = \sup_{\mathbf{q}\neq 0} \|\mathbf{K}\mathbf{q}\|_{\mathbf{M}_{\mathrm{qq}}} / \|\mathbf{q}\|_{\infty}$, $\mathbf{K} \in \mathbb{R}^{n_{\mathrm{q}} \times n_{\mathrm{q}}}$.

(L3) *There exists a constant* $q_\infty \in \mathbb{R}_+$ *satisfying* $\|\mathbf{M}_{\mathrm{qq}}^{-1}\mathbf{f}_{\mathrm{q}} + \mathbf{D}^{\top}\mathbf{p}\|_\infty \leq q_\infty$ *for the solution of problem* (26).

If the parameter L *and the time step size* τ *are chosen such that* $\frac{2}{L_{\mathrm{b}}} - \frac{1}{L} - \tau\frac{q_\infty^2 L_{\mathrm{K}}^2}{2k_{\mathrm{m}}} \geq 0$*, for a Poincaré constant* $C_\Omega > 0$*, it holds*

$$\|\mathbf{p}^i - \mathbf{p}\|^2_{\mathbf{M}_{\mathrm{pp}}} \leq \frac{L}{L + \tau k_{\mathrm{m}} C_\Omega^2}\|\mathbf{p}^{i-1} - \mathbf{p}\|^2_{\mathbf{M}_{\mathrm{pp}}}.$$

The proof of Lemma 1 is given in Appendix A. The proof is essentially the same as for the Richards equation by [22] but formulated in a slightly more general framework. Assumptions (L1)–(L2) are generalized versions of assumptions made in [22], adapted to the possible global dependence of each component of $\mathbf{b} = \mathbf{b}(\mathbf{p})$ on $\mathbf{p}$.

Consequence for the Fixed-Stress-L-scheme. In the context of the poroelasticity problem (26), the L-scheme (27) is equivalent with the Fixed-Stress-L-scheme (13)–(15), revealing the close connection between the fixed-stress splitting scheme and the L-scheme. Therefore, we check Assumptions (L1)–(L3) of Lemma 1 particularly for Eq. (26) in order to analyze the Fixed-Stress-L-scheme. We make the following physical assumptions:

(F1) With the varying porosity $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{p})$ as defined in Eq. (24), let $\mathbf{P}_{\phi\geq 0} = \{\mathbf{p} \in \mathbb{R}^{n_{\mathrm{p}}} \,|\, \boldsymbol{\phi}(\mathbf{p}) \in [0,1] \text{ component-wise}\}$ denote the space of all pressures leading to physical deformations.

(F2) Let the saturation model $s_{\mathrm{w}} : \mathbb{R} \to [0,1]$ have a bounded derivative and assume a non-vanishing residual saturation $0 < s_{\mathrm{w,res}} = \inf_{\mathbf{p}\in\mathbf{P}_{\phi\geq 0},\ i\in\{1,\ldots,n_{\mathrm{p}}\}} s(\mathbf{p}_i)$.

(F3) Let the material law $k_{\mathrm{w}} = k_{\mathrm{w}}(s_{\mathrm{w}}) : [0,1] \to \mathbb{R}$ be Lipschitz continuous and assume there exist constants $k_{\mathrm{w,m}}, k_{\mathrm{w,M}} \in \mathbb{R}_+$ satisfying $k_{\mathrm{w,m}} \leq k_{\mathrm{w}}(s_{\mathrm{w}}(\mathbf{p}_i)) \leq k_{\mathrm{w,M}}$ for all $\mathbf{p} \in \mathbf{P}_{\phi\geq 0}$, $i \in \{1,\ldots,n_{\mathrm{p}}\}$.

(F4) There exists a constant $q_\infty \in \mathbb{R}_+$ satisfying $\|\mathbf{M}_{\mathrm{qq}}^{-1}\mathbf{f}_{\mathrm{q}} + \mathbf{D}^{\top}\mathbf{p}\|_\infty \leq q_\infty$ for the solution of problem (26), i.e., fluxes are essentially bounded.

Assumption (F1) is physical. Assumptions (F2)–(F4) are standard assumptions generally accepted for the numerical analysis of the Richards equation, cf. e.g. [37–40]. In particular, if the latter assumptions are not satisfied, the Richards equation as model for flow in partially saturated porous media has to be questioned. Under these physical assumptions, (L1)–(L3) are satisfied.

Lemma 2 (*Assumptions for L-Scheme satisfied*). *Let Assumptions (F1)–(F2) be satisfied. Furthermore, assume* $W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ *yields an inf–sup stable discretization. Then Assumption (L1) is satisfied, in the sense, that there exists a constant* $L_b \in \mathbb{R}_+$*, satisfying for* *all* $\mathbf{p}, \tilde{\mathbf{p}} \in \mathbf{P}_{\phi\geq 0}$

$$\|\mathbf{b}(\mathbf{p}) - \mathbf{b}(\tilde{\mathbf{p}})\|^2_{\mathbf{M}_{\mathrm{pp}}^{-1}} \leq L_{\mathrm{b}} \langle \mathbf{b}(\mathbf{p}) - \mathbf{b}(\tilde{\mathbf{p}}), \mathbf{p} - \tilde{\mathbf{p}} \rangle. \tag{28}$$

Furthermore, let Assumptions (F1)–(F3) be satisfied. Then, Assumption (L2) is satisfied, in the sense, that there exists a constant $L_{\mathrm{K}} \in \mathbb{R}_+$*, satisfying for all* $\mathbf{p}, \tilde{\mathbf{p}} \in \mathbf{P}_{\phi\geq 0}$*,*

$$\|(\mathbf{K}(\mathbf{p}) - \mathbf{K}(\tilde{\mathbf{p}}))\mathbf{M}_{\mathrm{qq}}\|_{\mathbf{M}_{\mathrm{qq}},\infty} \leq L_{\mathrm{K}} \|\mathbf{b}(\mathbf{p}) - \mathbf{b}(\tilde{\mathbf{p}})\|_{\mathbf{M}_{\mathrm{pp}}^{-1}}. \tag{29}$$

Moreover, there exist constants $k_{\mathrm{m}}, k_{\mathrm{M}} \in \mathbb{R}_+$*, satisfying for all* $\mathbf{p} \in \mathbb{R}^{n_{\mathrm{p}}}$*,* $\mathbf{q} \in \mathbb{R}^{n_{\mathrm{q}}}$

$$k_{\mathrm{m}} \|\mathbf{q}\|^2_{\mathbf{M}_{\mathrm{qq}}^{-1}} \leq \langle \mathbf{K}(\mathbf{p})\mathbf{q}, \mathbf{q} \rangle \leq k_{\mathrm{M}} \|\mathbf{q}\|^2_{\mathbf{M}_{\mathrm{qq}}^{-1}}. \tag{30}$$

The proof of Lemma 2 is given in Appendix B. All in all, under the assumptions of non-vanishing residual saturation, permeability, and porosity, the L-scheme (27) converges, which follows from Lemma 1. Hence, also the Fixed-Stress-L-scheme (13)–(15) converges.

Theorem 3. *Let Assumptions (F1)–(F4) be satisfied. Furthermore, assume* $W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ *yields an inf–sup stable discretization. Let* $\mathbf{p} \in \mathbb{R}^{n_{\mathrm{p}}}$ *and* $\mathbf{p}^i \in \mathbb{R}^{n_{\mathrm{p}}}$ *be the solutions of the nonlinear problem (26) and the L-scheme (27), respectively. Assume they are unique. Let the initial guess* $\mathbf{p}^0 \in \mathbb{R}^{n_{\mathrm{p}}}$ *satisfy* $\mathcal{B}_{\mathbf{p}}(\|\mathbf{p}^0 - \mathbf{p}\|_{\mathbf{M}_{\mathrm{pp}}}) \subset \mathbf{P}_{\phi\geq 0}$*, where* $\mathcal{B}_{\mathbf{p}}(r) \subset \mathbb{R}^{n_{\mathrm{p}}}$ *denotes the sphere with center* $\mathbf{p}$ *and radius* $r > 0$*. Let* L *and* τ *be chosen such that* $\frac{1}{L_{\mathrm{b}}} - \frac{1}{2L} - \tau \frac{q_\infty^2 L_{\mathrm{K}}^2}{2k_{\mathrm{m}}} \geq 0$*. Then the L-scheme (27) converges linearly with mesh-independent convergence rate* $\sqrt{\frac{L}{L+\tau k_{\mathrm{m}} C_\Omega^2}}$*. Furthermore, by induction, each iterate is a physical solution* $\{\mathbf{p}^i\}_i \subset \mathbf{P}_{\phi\geq 0}$*.*

Remark 3 (*Choice of L*)**.** The Jacobian of **b**

$$\mathbf{D}_{\mathbf{b}}(\mathbf{p}) = \begin{bmatrix} s'(\mathbf{p}_1)\boldsymbol{\phi}_1(\mathbf{p}) & & \\ & \ddots & \\ & & s'(\mathbf{p}_{n_{\mathrm{p}}})\boldsymbol{\phi}_{n_{\mathrm{p}}}(\mathbf{p}) \end{bmatrix} + \alpha^2 \mathbf{S}_{\mathrm{pp}}(\mathbf{p}) \mathbf{D}_{\mathrm{pu}} \mathbf{A}_{\mathrm{uu}}^{-1} \mathbf{D}_{\mathrm{pu}}^\top \mathbf{S}_{\mathrm{pp}}(\mathbf{p})^\top + \frac{1}{N} \mathbf{S}_{\mathrm{pp}}(\mathbf{p}) \mathbf{M}_{\mathrm{pp}} \mathbf{S}_{\mathrm{pp}}(\mathbf{p})^\top \tag{31}$$

justifies the choice of the tuning parameters L and β_{FS} for the Fixed-Stress-L-scheme, cf. Section 4.3, being an approximation of the Jacobian. Assuming the worst case scenario, the eigenvalues of all three contributions are maximized yielding an *a priori choice*. This pessimistic choice slows down potential convergence but increases robustness. From the proof of Theorem 3, it follows that local optimization would be sufficient, yielding an optimal but solution-dependent tuning parameter. In this spirit, Eq. (31) also provides the basis for the modification of the tuning parameter used for both the Fixed-Stress-Modified-Picard method and the Fixed-Stress-Newton method, cf. Section 4.4.

Remark 4 (*Limitations of the Fixed-Stress-L-Scheme*)**.** Based on Theorem 3, we expect the convergence of the Fixed-Stress-L-scheme (13)–(15) to deteriorate for either too large time steps or too large Lipschitz constants for the constitutive laws s_{w} and k_{w}. This applies in particular if the constitutive laws are only Hölder continuous. Furthermore, given the parameter L is sufficiently large and the time step size sufficiently small, theoretical convergence of the Fixed-Stress-L-scheme is guaranteed. However, in practice, numerical round-off errors might lead to stagnation.

Remark 5 (*Parabolic Character of the Nonlinear Biot Equations*)**.** The Richards equation itself is a degenerate elliptic–parabolic equation due to possible development of fully saturated regions. However, from Eq. (31) it follows, that this type of degeneracy is not adopted by the nonlinear Biot equations (1)–(3) and by corresponding stable discretizations (6)–(8). Independent of the mesh size, the derivative of the fluid volume ϕs_{w} with respect to fluid pressure is not vanishing, as long as the fluid saturation is not vanishing. This observation is consistent with considerations by [41] on the classical, linear Biot equations. We note for weak coupling of mechanics and flow equations, numerically the parabolic character might be effectively lost, making the original two-way coupled problem essentially equivalent to the Richards equation, one-way coupled with the linear elasticity equations.

6. Acceleration and stabilization by Anderson acceleration

The Fixed-Stress-L-scheme is expected to be a linearly convergent fixed-point iteration with the convergence rate depending on the tuning parameter. Its Quasi-Newton modifications, cf. Section 4.4, employ a less conservative choice for the tuning parameter with the risk of failing convergence. Consequently, we are concerned with two issues — slow convergence and robustness with respect to the tuning parameter.

All presented linearization schemes in Section 4 can be interpreted as fixed-point iterations $\mathbf{x}^i = \mathcal{FP}(\mathbf{x}^{i-1}) = \mathbf{x}^{i-1} + \Delta\mathcal{FP}(\mathbf{x}^{i-1})$, where $\mathbf{x}^i$ denotes the algebraic vector associated with $(p, \boldsymbol{q}, \boldsymbol{u})_h^{n,i}$ and $\Delta\mathcal{FP}(\mathbf{x}^{i-1})$ is the actual, computed increment within the linearization scheme. For fixed-point iterations in general, Anderson acceleration [27] has been demonstrated on several occasions to be a suitable method to accelerate convergence. Furthermore, due to its relation to preconditioned, nonlinear GMRES [30], we also expect Anderson acceleration to increase robustness with respect to the tuning parameter for the considered linearization schemes. Both properties are justified by theoretical considerations in Section 7 and demonstrated numerically in Section 8.

Scheme. The main idea of the Anderson acceleration applied to a fixed-point iteration is to utilize previous iterates and mix their contributions in order to obtain a new iterate. The method is applied as post-processing, not interacting with the underlying fixed-point iteration. In the following, we denote AA(m) the Anderson acceleration reusing $m + 1$ previous iterations, such that AA(0) is identical to the original fixed point iteration. We can apply AA(m) to post-process the presented linearization schemes. In compact notation, the scheme reads:

Algorithm 1 (AA(m) accelerated $\mathcal{FP}$)

```
Given: $\mathcal{FP}$, $\mathbf{x}^0$
for $i$=1,2, ..., until convergence do
    Define depth $m_i = \min\{i-1, m\}$
    Define matrix of increments $\mathbf{F}_i = \left[\Delta\mathcal{FP}(\mathbf{x}^{i-m_i-1}), \dots, \Delta\mathcal{FP}(\mathbf{x}^{i-1})\right]$
    Minimize $\|\mathbf{F}_i\boldsymbol{\alpha}\|_2$ wrt. $\boldsymbol{\alpha} \in \mathbb{R}^{m_i+1}$ s.t. $\sum_k \alpha_k = 1$
    Define next iterate $\mathbf{x}^i = \sum_{k=0}^{m_i} \alpha_k \mathcal{FP}(\mathbf{x}^{k+i-m_i-1})$
end for
```

Properties. For the implementation, we follow Walker and Peng [30], using an equivalent, unconstrained minimization problem formulation in Step 3 of the loop. The resulting problem becomes better conditioned, relatively small and cheap. It is solved via a QR decomposition employing Householder transformations and subsequent inversion, using a matrix–vector multiplication and backward substitution. Given the depth m and problem size n ($m \ll n$), the total, algorithmic complexity of one Anderson acceleration iteration (without the evaluation of the fixed-point iteration) is of order $\mathcal{O}(2nm^2 + 2nm + m^2) = \mathcal{O}(nm^2)$, dominated by the cost of the QR decomposition. Additionally, the price for the storage of the vectors $[\Delta\mathcal{FP}(\mathbf{x}^{i-m-1}), \dots, \Delta\mathcal{FP}(\mathbf{x}^{i-1})]$ and $[\mathcal{FP}(\mathbf{x}^{i-m-1}), \dots, \mathcal{FP}(\mathbf{x}^{i-1})]$ has to be paid, similar to GMRES. For the numerical examples in Section 8, we employ direct solvers. Hence, compared to the application of the plain linearization schemes, the additional cost for Anderson acceleration with small depth is insignificant.

Moreover, as post-processing Anderson acceleration does not modify the character of the underlying method, i.e., a decoupled character remains unchanged. In particular, in contrast to classical preconditioning, no monolithic simulator is required. Hence, all in all, Anderson acceleration is an attractive method in order to accelerate splitting schemes.

In many practical applications, effective acceleration can be observed. Though, there is no general, theoretical guarantee for the Anderson acceleration to accelerate convergence of an underlying, convergent fixed-point iteration. Theoretically, even divergence is possible [30]. In the literature, so far, theoretical convergence results are solely known for contractive fixed-point iterations [31]. For nonlinear problems, AA(m) is guaranteed locally r-linearly convergent with theoretical convergence rate not larger than the original contraction constant if the coefficients $\boldsymbol{\alpha}$ remain bounded. Without assumptions on $\boldsymbol{\alpha}$, AA(1) converges globally, q-linearly in case the contraction constant is sufficiently small. After all, both results only guarantee the lack of deterioration but not actual acceleration.

For a special, linear case, in Section 7, we show global convergence and theoretical acceleration for a variant of AA(1), fortifying the potential of Anderson acceleration. In particular, Corollary 5 predicts the ability of the Anderson acceleration to increase robustness, allowing non-contractive fixed-point iterations to converge. Although the structure of the considered linearization schemes does not satisfy the structure of the special linear case (symmetric and linear), this motivates to apply AA(m) also to accelerate possibly diverging Newton-like methods with the risk of losing potential, quadratic convergence, as well as slow Picard type methods.

7. Theoretical contraction and acceleration for the restarted Anderson acceleration

For a special linear case, we prove global convergence of a restarted version of the Anderson acceleration. In particular, convergence for non-contractive fixed-point iterations and effective acceleration for a class of contractive fixed-point iterations is shown. We note that the results cannot be transferred to the application of Anderson acceleration for fixed-stress type linearization schemes, but they indicate theoretically the benefit of its application.

7.1. Restarted Anderson acceleration

The original Anderson acceleration AA(m) constantly utilizes the full set of m previous iterates. By defining the depth $m_i^\star = \min\{i - 1 \bmod m + 1, m\}$ in the first step of Algorithm 1 and apart from that following the remaining steps, we define a restarted version AA*(m) of AA(m), closer related to GMRES(m). In words, in each iteration we update the set of considered iterates by the most current iterate. And in case the number of iterates becomes $m + 1$, we flush the memory and restart filling it again. In particular, for $m = 1$, the algorithm reads:

Algorithm 2 **(AA*(1) accelerated $\mathcal{FP}$)**

```
Given: $\mathcal{FP}$, $\mathbf{x}^0$
for $i$=0,2,4,..., until convergence do
    Set $\mathbf{x}^{i+1} = \mathcal{FP}(\mathbf{x}^i)$
    Minimize $\left\|\Delta\mathcal{FP}(\mathbf{x}^{i+1}) + \alpha^{i+1}(\Delta\mathcal{FP}(\mathbf{x}^i) - \Delta\mathcal{FP}(\mathbf{x}^{i+1}))\right\|_2$ wrt. $\alpha^{(i+1)} \in \mathbb{R}$
    Set $\mathbf{x}^{i+2} = \mathcal{FP}(\mathbf{x}^{i+1}) + \alpha^{(i+1)}(\mathcal{FP}(\mathbf{x}^i) - \mathcal{FP}(\mathbf{x}^{i+1}))$
end for
```

From [31], it follows directly, that for $\mathcal{FP}$, a linear contraction, AA*(1) converges globally with convergence rate at most equal the contraction constant of $\mathcal{FP}$. In the following, we extend the result to a special class of non-contractive fixed-point iterations.

7.2. Convergence results

For the convergence results, cf. Lemma 4 and Corollaries 5, 6, we make the following assumptions:

(C1) $\mathcal{FP}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ defines the Richardson iteration for $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $n > 1$, $\mathbf{b} \in \mathbb{R}^n$.
(C2) $\mathbf{A}$ is symmetric, and hence, $\mathbf{A}$ is orthogonally diagonalizable and there exists an orthogonal basis of eigenvectors $\{\mathbf{v}_j\}_j$ and a corresponding set of eigenvalues $\{\lambda_j\}_j$ satisfying $\mathbf{A}\mathbf{v}_j = \lambda_j\mathbf{v}_j$.
(C3) There exists a unique $\mathbf{x}^\star$ such that $\mathcal{FP}(\mathbf{x}^\star) = \mathbf{x}^\star$, i.e., $\mathbf{I} - \mathbf{A}$ is invertible.
(C4) The initial iterate $\mathbf{x}^0$ is chosen such that the initial error $\mathbf{x}^0 - \mathbf{x}^\star \in \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$, where $\mathbf{v}_1, \mathbf{v}_2$ are two orthogonal eigenvectors of $\mathbf{A}$. To avoid a trivial case, we assume $\lambda_1, \lambda_2 \neq 0$.

Then we are able to relate the errors between iterations of AA*(1), allowing to prove further convergence and acceleration results, cf. Corollaries 5 and 6. All in all, the proof employs solely elementary calculations. However, as we are not aware of a general result of same type in the literature, we present the proof.

Lemma 4 (*Error Propagation*)**.** *Let the Assumptions (C1)–(C4) be satisfied and let $\{\mathbf{x}^i\}_i$ define the sequence defined by AA*(1) applied to $\mathcal{FP}$. Furthermore, let $\mathbf{e}^i = \mathbf{x}^i - \mathbf{x}^\star$ denote the error. Then it holds*

$$\|\mathbf{e}^{i+4}\| \leq r(\lambda_1, \lambda_2)\|\mathbf{e}^i\|, \quad i = 0, 4, 8, 12, \ldots$$

for

$$r(\lambda_1, \lambda_2) = \frac{\lambda_1^2\lambda_2^2(\lambda_2 - \lambda_1)^2}{(|\lambda_1(\lambda_1 - 1)| + |\lambda_2(\lambda_2 - 1)|)^2}.$$

The proof of Lemma 4 is given in Appendix C. Based on the contraction result, we are finally able to show convergence and actual acceleration of AA*(1).

Corollary 5 (*AA*(1) Converges for Non-Contractive $\mathcal{FP}$*)**.** *Let the Assumptions (C1)–(C4) be satisfied. Let* $\mathbf{A}$ *be positive definite with at most one eigenvalue among $\{\lambda_1, \lambda_2\}$ larger than 1 and none equal to 1. Then AA*(1) converges for the underlying non-contractive fixed-point iteration, cf. Assumption (C1).*

Proof. Due to symmetry, it is sufficient, to consider solely $(\lambda_1, \lambda_2) \in (\mathbb{R}_+ \setminus \{1\}) \times (0, 1)$. For $\lambda_1 < 1$, the result follows immediately from Corollary 6. Let $\lambda_1 > 1$. It holds $r(1, \lambda_2) = 1$ for all $\lambda_2 \in (0, 1)$ and $\partial_1 r(\lambda_1, \lambda_2) < 0$ for all $(\lambda_1, \lambda_2) \in (1, \infty) \times (0, 1)$. Thus, it follows directly that $r(\lambda_1, \lambda_2) < 1$ for all $(\lambda_1, \lambda_2) \in (1, \infty) \times (0, 1)$. □

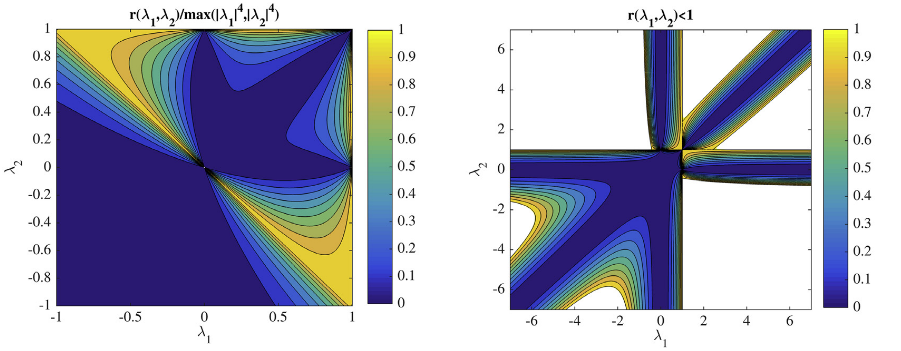
Corollary 6 (*AA*(1) Accelerates Contractive $\mathcal{FP}$*)**.** *Let the Assumptions (C1)–(C4) be satisfied. Let $\rho(\mathbf{A}) < 1$, where $\rho(\mathbf{A})$ denotes the spectral radius of* $\mathbf{A}$*. Then it holds $r(\lambda_1, \lambda_2) < \rho(\mathbf{A})^4$ if $\lambda_1 \neq -\lambda_2$, and $r(\lambda_1, \lambda_2) = \rho(\mathbf{A})^4$ otherwise . Consequently, AA*(1) is effectively accelerating the underlying fixed-point iteration, cf. Assumption (C1).*

Proof. By plotting $r(\lambda_1, \lambda_2)/\max\left\{|\lambda_1|^4, |\lambda_2|^4\right\}$, we demonstrate $r(\lambda_1, \lambda_2) \leq \max\left\{|\lambda_1|^4, |\lambda_2|^4\right\} = \rho(\mathbf{A})^4$ for $(\lambda_1, \lambda_2) \in [-1, 1] \times [-1, 1]$, cf. Fig. 1(a). □

7.3. Discussion

We make the following comments:

- The convergence result in Corollary 5 deals only with positive definite matrices. In Fig. 1(b), eigenvalue pairs $(\lambda_1, \lambda_2) \in \mathbb{R} \times \mathbb{R}$ are displayed satisfying $r(\lambda_1, \lambda_2) < 1$ and therefore guaranteeing AA*(1) to converge. In particular, AA*(1) converges also for matrices with two eigenvalues larger than 1 with relatively close distance to each other.



(a) Acceleration rate for AA*(1) compared to theoretical contraction rate of Richardson iteration for contractive 2×2 matrices.

(b) The (λ_1, λ_2)-convergence plane for AA*(1) applied to 2×2 linear problems, characterizing which matrices guarantee AA* to define a uniform contraction.

Fig. 1. Acceleration and convergence factor for the restarted AA*(1).

- In practice, we do not experience AA*(1) or AA(1) to fail as long as Assumption (C4) is valid and $|\lambda_1| < 1$ or $|\lambda_2| < 1$. This observation extends also to arbitrarily large decompositions of $\mathbf{e}^0$ as long as at most one eigenvalue of $\mathbf{A}$ satisfies $|\lambda_j| > 1$. Based on similar observations, we state the following claim: If $|\lambda_j| > 1$ for exactly m eigenvalues $\{\lambda_j\}_j$, then AA(m) converges for arbitrary $\mathbf{e}^0$. We note that the worst case approach used in order to prove Lemma 4 cannot be applied to prove the general claim. It can be verified numerically that in general the eigenvalues of the error propagation matrix (C.10) can be larger than 1 even if $\rho(\mathbf{A}) < 1$.
- From Fig. 1(b), it follows, the closer the eigenvalues to 1, the slower the convergence of AA*(1). This is consistent with the interpretation of Anderson acceleration as secant method. The Richardson iteration does only damp slowly directions corresponding to eigenvalues close to 1. Hence, a directional derivative in these directions cannot be approximated well, purely based on the iterations of fixed-point iterations. Quite contrary to directions corresponding to small or large eigenvalues relative to 1.
- The theoretical convergence result has been obtained from a worst case analysis. Practical convergence rates might be lower than predicted, depending on the weights of the initial error.
- Convergence of AA(m) is not guaranteed to be monotone, when applied for non-contractive fixed-point iterations.

8. Numerical results — Performance study

In this section, we consider three numerical examples with increasing complexity. In all examples, we compare the linearization schemes, presented in Section 4, coupled with Anderson acceleration. In particular, we confirm numerically the parabolic character of the nonlinear Biot equations, cf. Remark 5, the convergence result for the Fixed-Stress-L-scheme, cf. Theorem 3, as well as the acceleration and stabilization properties of the Anderson acceleration, cf. Section 7. All numerical results have been obtained using the software environment DUNE [42–44], where linear systems are solved with a direct solver.

8.1. Test case I — Injection in a 2D homogeneous medium with Lipschitz continuous constitutive laws

We consider a two-dimensional, homogeneous, unsaturated porous medium $(-1, 1) \times (0, 1) \subset \mathbb{R}^2$, in which a fluid is injected at the top $(-0.2, 0.2) \times \{1\}$, cf. Fig. 2. Due to the symmetry of the problem, we consider only the right half $\Omega = (0, 1) \times (0, 1)$, discretized by 50×50 regular quadrilaterals. As initial condition, we choose a constant displacement $\boldsymbol{u}(0) = \mathbf{0}$ and pressure field $p_\mathrm{w}(0) = p_0$, satisfying the stationary version of the continuous problem (1)–(3). In order to avoid inconsistent initial data, we ramp the injection at the top with inflow rate $q_\mathrm{inflow}(t) = q^\star \times \min\{t^2, 1.0\}$ for given $q^\star \in \mathbb{R}$. Apart from the inflow at the top, we consider no flow at the remaining boundaries, no normal displacement at left, right and bottom boundary and no stress on the top. The boundary conditions are displayed in Fig. 2.

For the spatial discretization, we consider both variants of $W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, the lowest-order discretization $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and the unconditionally stable discretization $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$.

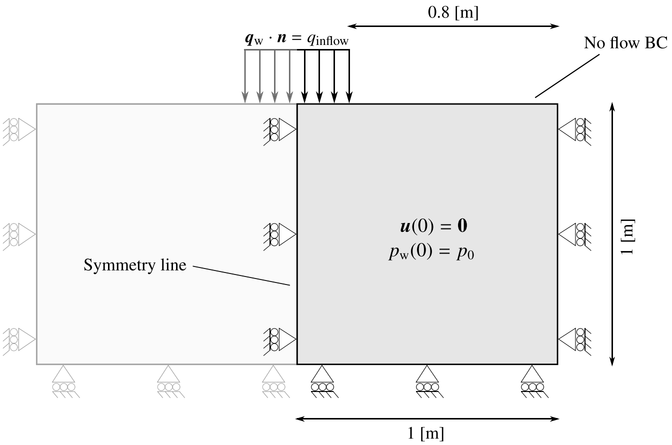

0.8 [m]
$\boldsymbol{q}_{\mathrm{w}} \cdot \boldsymbol{n} = q_{\mathrm{inflow}}$
No flow BC
$\boldsymbol{u}(0) = \boldsymbol{0}$
$p_{\mathrm{w}}(0) = p_0$
Symmetry line
1 [m]
1 [m]

Fig. 2. Domain Ω and boundary conditions for test cases I and II.

Table 1
Parameters employed for test cases I, II and III. Top: Physical model parameters. Bottom: Numerical parameters.

Parameter	Variable [unit]	Test case I (Section 8.1)	Test case II (Section 8.2)	Test case III (Section 8.3)
Young's modulus	E [Pa]	3e1	3e1	1e6
Poisson's ratio	ν [–]	0.2	0.2	0.3
Initial pressure	p_0 [Pa]	−7.78	−15.3	*hydrostatic*
Initial porosity	ϕ_0 [–]	0.2	0.2	0.2
Inverse of air suction	a_{vG} [Pa^{-1}]	0.1844	0.627	1e−4
Pore size distribution	n_{vG} [–]	3.0	1.4	0.7^{-1}
Abs. permeability	k_{abs} [m^2]	3e−2	3e−2	5e−13
Fluid viscosity	μ_{w} [Pa·s]	1.0	1.0	1e−3
Fluid density	ρ_{w} [kg/m^3]	1e3	1e3	1e3
Bulk density	ρ_{b} [kg/m^3]	[-]	[-]	1.8e3
Gravitational acc.	$\boldsymbol{g}$ [m/s^2]	0.0	0.0	9.81
Biot coefficient	α [–]	0.1 |0.5| 1.0	0.1 |0.5| 1.0	1.0
Biot modulus	N [Pa]	∞	∞	∞
Maximal inflow rate	$q^{\star}$ [m^2/s]	−1.25	−0.175	[-]
Final time	T [s]	1.0	1.0	86400 (= 10 [days])
Time step size	τ [s]	1e−1	1e−1	3600 (= 1 [hours])
Absolute tolerance	ε_{a}	1e−8	1e−8	1e−3
Relative tolerance	ε_{r}	1e−8	1e−8	1e−6

Physical and numerical parameters. For the constitutive laws, governing saturation and permeability, we use the van Genuchten–Mualem model [4], defining

$$s_{\mathrm{w}}(p_{\mathrm{w}}) = \begin{cases} \left(1 + (-a_{\mathrm{vG}} p_{\mathrm{w}})^{n_{\mathrm{vG}}}\right)^{-\frac{n_{\mathrm{vG}}-1}{n_{\mathrm{vG}}}} & , p_{\mathrm{w}} \le 0, \\ 1 & , \text{else}, \end{cases} \quad k_{\mathrm{w}}(s_{\mathrm{w}}) = \frac{k_{\mathrm{abs}}}{\mu_{\mathrm{w}}} \sqrt{s_{\mathrm{w}}} \left(1 - \left(1 - s_{\mathrm{w}}^{\frac{n_{\mathrm{vG}}}{n_{\mathrm{vG}}-1}}\right)^{\frac{n_{\mathrm{vG}}-1}{n_{\mathrm{vG}}}}\right)^2, \; s_{\mathrm{w}} \in [0, 1],$$

where a_{vG} and n_{vG} are model parameters associated to the inverse of the air suction value and pore size distribution, respectively, k_{abs} is the intrinsic absolute permeability and μ_{w} is the dynamic fluid viscosity.

Values chosen for model parameters and numerical parameters are displayed in Table 1. The parameters have been chosen such that the initial saturation is $s_{\mathrm{w},0} = 0.4$ in Ω and a region of full saturation ($s_{\mathrm{w}} = 1$) is developed after seven time steps. Furthermore, the constitutive laws for saturation and permeability are Lipschitz continuous with respect to pressure (with $L_s = 0.12$). We consider three different, realistic values for the Biot coefficient α, all relevant for applications with unsaturated, deformable materials [2]. They control the coupling strength and thereby whether the Richards equation or the nonlinear coupling terms determine the character of the numerical difficulties. The simulation result at final time $t = 1$ for strong coupling ($\alpha = 1.0$) and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ is illustrated exemplarily in Fig. 3.

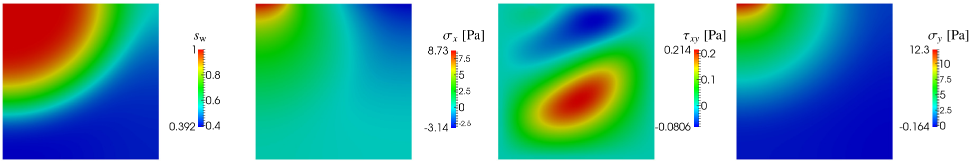


Fig. 3. Simulation results for test case I (Lipschitz continuous permeability) solved with $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ for $\alpha = 1.0$: Saturation, normal stresses σ_x, σ_y and shear stress τ_{xy} at time $t = 1$.

Table 2
Abbreviations for methods (top) and additional stabilizations (bottom).

Abbreviation	Explanation
Newton	Monolithic Newton's method
FS-Newton	Fixed-Stress-Newton method
FS-MP	Fixed-Stress-Modified-Picard method
FSL	Fixed-Stress-L-scheme with $L = L_s$, $\beta_{FS} = \beta_{FS}^{phy}$
FSL/2	Fixed-Stress-L-scheme with $L = \frac{1}{2}L_s$, $\beta_{FS} = \frac{1}{2}\beta_{FS}^{phy}$
AA(m)	Anderson acceleration with $m + 1$ reused iterations

Table 3
Performance for test case I with different coupling strengths ($\alpha = 0.1,\ 0.5,\ 1.0$), for $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$. Average number of (nonlinear) iterations per time step for Newton's method, the Fixed-Stress-Newton method, the Fixed-Stress-Modified-Picard method and the Fixed-Stress-L-scheme; both plain and coupled with Anderson acceleration for different depths ($m = 1,\ 3,\ 5,\ 10$). Minimal numbers per linearization type and Biot coefficient are in bold. Numbers of iterations equal for $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$, if not noted differently.

$\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ / $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$															
Linearization	Newton			FS-Newton			FS-MP			FSL			FSL/2		
Biot coeff. α	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0
AA(0)	**5.3**	**5.1**	**5.0**	**6.0**	8.3	10.6	18.2	18.2	16.7	23.2	21.2	18.9	46.8	41.4	41.1
AA(1)	6.1	6.0	6.0	6.2	**7.6**	8.9	15.8	15.5	15.7	21.2	19.7	17.7	17.4	17.3	17.3
AA(3)	7.4	7.4	7.5	7.4	7.7	8.5	13.4	13.6	13.5	16.1	15.3	15.0	14.3	14.5	14.7
AA(5)	8.3	8.1	8.2	7.9	7.9[a]	**8.4**	13.1	12.8	12.5	14.9	14.6	14.3	13.3	13.5	13.6
AA(10)	–	–	–	–	–	–	**12.8**	**12.5**	**12.3**	**14.4**	**14.3**	**14.1**	**13.3**	**13.1**	**13.4**

[a] 8.0 for $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$.

Performance of linearization schemes. We consider the four linearization schemes introduced in Section 4, coupled with Anderson acceleration as post-processing. Furthermore, motivated by the theoretical convergence result, cf. Theorem 3, we employ the Fixed-Stress-L-scheme with half sized stabilization parameter. Abbreviations used in this section are introduced in Table 2.

We use the average number of iterations per time step as measure for performance, cf. Table 3, which is a reasonable measure due to the insignificant, additional cost for the application of Anderson acceleration for small depth. In particular, we disregard the use of CPU time as performance measure due to a not finely-tuned implementation. We just note, that a single iteration of a splitting method is significantly faster than a single monolithic Newton iteration.

First of all, despite the requirement of Theorem 3 for an inf–sup stable discretization, the number of iterations per time step for $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$ is virtually the same. Only for a single setting a slight, but not significant difference can be observed. Next, all plain linearization schemes (AA(0)) succeed to converge for all three coupling strengths. This is consistent with Remark 5, demonstrating that the nonlinear Biot equations do not adopt the degeneracy of the Richards equation and remain parabolic in a fully saturated regime. Not surprisingly, the monolithic Newton method requires fewest iterations. So at first impression, it seems to be the preferred method. However, as stressed above, on larger scale, a fixed-stress type iterative solver or another advanced monolithic solver is required for efficient solution independent of the coupling strength, i.e., additional costs are hidden. On the other hand, the remaining linearization schemes allow separate simulators from the beginning. As the Fixed-Stress-L-scheme does not utilize an exact evaluation of derivatives, the Fixed-Stress-Newton method and the Fixed-Stress-Modified-Picard method perform better for all three coupling strengths. Solely the performance of the Fixed-Stress-Newton method shows weak dependence on the coupling strength, having the character of Newton's method and a Picard-type method for weak and strong coupling, respectively. The remaining methods show improved convergence behavior for increasing coupling strength, due to the decreasing numerical complexity of the problem itself following from Remark 5.

When applying Anderson acceleration, we observe that Anderson acceleration slows down the convergence of the monolithic Newton method, which is consistent with considerations in Section 6. In contrast, Anderson acceleration speeds

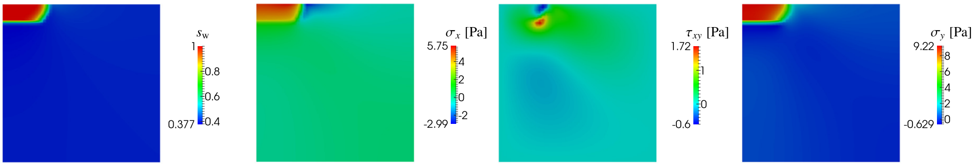


Fig. 4. Simulation results for test case II (Hölder continuous permeability) solved with $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ for $\alpha = 1.0$: Saturation, normal stresses σ_x, σ_y and shear stress τ_{xy} at time $t = 1$.

up significantly the convergence of the Picard-type methods (Fixed-Stress-L-scheme and the variation FSL/2, and Fixed-Stress-Modified-Picard). We recall the insignificant, additional cost for the application of Anderson acceleration for small depth. Largest acceleration effect can be seen for largest considered depth. For the Fixed-Stress-Newton method, the effect of Anderson acceleration depends on the numerical character of the problem. This is due to the fact, that for weak coupling, the method is essentially identical with Newton's method.

Regarding the Fixed-Stress-L-scheme, according to Theorem 3, optimally, the diagonal stabilization parameter has to be chosen as small as possible. However, smaller values do not necessarily lead to faster convergence, as can be observed by comparing the plain Fixed-Stress-L-scheme and the plain FSL/2-scheme. Yet when utilizing Anderson acceleration, robustness with respect to the tuning parameter is increased, and eventually the FSL/2-scheme converges faster than the Fixed-Stress-L-scheme. In particular, it performs as good as the Fixed-Stress-Modified-Picard method.

Without presenting any detailed numerical results, when repeating the test case with lower time step size, we observe a small quantitative change for the average number of iterations (e.g., on the order of 20% savings in iterations per time step for $\tau = 0.01$). Though the total number of iterations might increase. The qualitative discussion of the merits of the different nonlinear solvers remains largely unchanged.

All in all, the theory has been confirmed. The Fixed-Stress-L-scheme converges despite the simple linearization approach even for the lowest-order discretization $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and Anderson acceleration is able to accelerate Picard-type schemes. Moreover, the latter has been shown to stabilize the Fixed-Stress-L-scheme. It allows to choose a smaller tuning parameter, leading to improved convergence behavior. Considering the cost per iteration, despite some additional iterations, we finally recommend the use of the Fixed-Stress-Newton method with Anderson acceleration with low depth. It is cheap and allows separate simulators. For strongly coupled problems or in the absence of exact derivatives, the Fixed-Stress-L-scheme with small tuning parameters is an attractive alternative to the Fixed-Stress-Newton method.

8.2. Test case II — Injection in homogeneous 2D medium with Hölder continuous permeability

In the following, we reveal limitations of the considered linearization schemes. Moreover, we demonstrate the stabilization property of Anderson acceleration, allowing non-convergent methods to converge. For this purpose, we repeat test case I with modified physical parameters. In particular, we choose the saturation to be Lipschitz continuous with same Lipschitz constant as in test case I. In contrast, the permeability is chosen to be only Hölder continuous. Hence, the derivative becomes unbounded in the transition between partial and full saturation, causing potential trouble for the Newton-type methods. Again, we choose the initial pressure and the maximal inflow rate such that $s_{w,0} = 0.4$ and a region of full saturation ($s_w = 1$) is developed after seven time steps. The simulation result at final time $t = 1$ for strong coupling ($\alpha = 1.0$) and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ is illustrated in Fig. 4.

As mentioned in Remark 4, due to lack of regularity for the permeability, each of the considered methods faces difficulties. For Newton-type methods (Newton, Fixed-Stress-Newton), the derivative of the permeability is evaluated, which might be unbounded. Effectively, for the Fixed-Stress-L-scheme, this also means that $L_K \to \infty$ or in practice L_K becomes very large. Hence, by Theorem 3, the time step size has to be chosen sufficiently small or possibly L has to be chosen larger to guarantee convergence. We note that for chosen initial saturation the permeability is significantly lower than for test case I. Consequently, the theoretical convergence rate for the plain Fixed-Stress-L-scheme (13)–(15) deteriorates. Due to round off errors stagnation is possible.

Performance of linearization schemes. The average number of iterations per time step is presented in Table 4, both for $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$. This time we observe slightly different results for the different discretizations. But overall, for both discretizations, the results follow same trends. Again, the lack of inf–sup stability does not show to be a serious issue, cf. Theorem 3. In contrast to test case I, not all plain linearization schemes (AA(0)) converge. For weak coupling, all Newton-like methods (Newton, Fixed-Stress-Newton) diverge with the Fixed-Stress-Newton method being slightly more robust due to added fixed-stress stabilization. The Fixed-Stress-L-scheme stagnates and shows to be slightly more robust than the Newton-type methods. The Fixed-Stress-Modified-Picard method is least robust and stagnates already after three time steps. For strong coupling, all methods converge, which is consistent with Remark 5. If convergent, the schemes sorted by required number of iterations are the monolithic Newton method, the Fixed-Stress-Newton method, the Fixed-Stress-Modified-Picard and the Fixed-Stress-L-scheme, meeting our expectations.

Table 4
Performance for test case II with different coupling strengths ($\alpha = 0.1,\ 0.5,\ 1.0$), for $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ and $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$. Average number of (nonlinear) iterations per time step for Newton's method, the Fixed-Stress-Newton method, the Fixed-Stress-Modified-Picard method and the Fixed-Stress-L-scheme; both plain and coupled with Anderson acceleration for different depths ($m = 1,\ 3,\ 5,\ 10$). Minimal numbers per linearization type and Biot coefficient are in bold. Failing linearization due to stagnation at time step n is marked by $\rightarrow$ [n]. Failing linearization due to divergence at time step n is marked by $\nearrow$ [n].

Linearization	Newton			FS-Newton			FS-MP			FSL			FSL/2		
Biot coeff. α	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0
$\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$															
AA(0)	↗ [8]	**8.5**	**8.1**	↗ [9]	13.2	19.1	→ [3]	36.9	55.0	→ [9]	126.9	134.9	→ [8]	→ [9]	→ [10]
AA(1)	**10.7**	9.4	→ [8]	**11.0**	**11.8**	14.6	45.2	34.2	33.8	133.6	84.0	83.2	→ [9]	68.5	65.1
AA(3)	17.2	11.7	→ [8]	15.6	12.1	**13.0**	30.5	26.9	28.1	68.3	54.3	56.9	48.4	37.9	35.5
AA(5)	24.8	13.9	→ [8]	23.3	13.1	13.2	29.2	24.7	23.5	62.4	48.7	44.9	43.4	34.8	32.7
AA(10)	33.3	18.4	→ [8]	43.0	14.7	13.8	**29.8**	**23.5**	**23.5**	**52.6**	**42.6**	**42.5**	**39.3**	**31.8**	**29.2**
$\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$															
AA(0)	↗ [8]	**8.4**	**8.7**	↗ [9]	13.2	18.9	→ [3]	36.8	53.9	→ [9]	126.9	133.1	→ [8]	→ [9]	→ [10]
AA(1)	**10.7**	9.4	12.9	**11.1**	**11.7**	14.5	45.2	34.2	33.9	133.6	84.4	85.2	→ [9]	65.6	64.3
AA(3)	16.2	11.8	→ [8]	15.2	12.2	**12.9**	30.5	27.0	27.7	68.3	55.3	53.6	48.4	37.8	35.5
AA(5)	18.5	14.0	→ [8]	45.6	13.2	13.1	**29.1**	24.7	25.2	61.5	49.2	46.4	43.4	35.0	32.4
AA(10)	36.7	17.8	→ [8]	↗ [9]	14.8	13.9	30.1	**23.7**	**23.4**	**52.3**	**42.7**	**41.9**	**39.3**	**32.1**	**28.8**

By utilizing Anderson acceleration, convergence can be observed for all coupling strengths and all linearization schemes besides Newton's method for $\alpha = 1$. In particular, all previously failing schemes converge. This confirms the possible increase of robustness by Anderson acceleration, postulated in Section 7. Similar observations as before are made for the splitting schemes under Anderson acceleration. All in all, the theory has been confirmed.

As before, for increasing depth, the performance of Newton's method deteriorates. For strong coupling stagnation is observed. For weak coupling, for several time steps practical stagnation is observed with eventual convergence after a very large number of iterations. This is consistent with the fact that Anderson acceleration can also lead to divergence for increasing depth [30]. Hence, Anderson acceleration has to be applied carefully for the monolithic Newton method.

Motivated by test case I, we apply the Fixed-Stress-L-scheme with a decreased tuning parameter. For this test case, the plain FSL/2-scheme fails for all coupling strengths. As the FSL/2-scheme is *a priori* less robust as the Fixed-Stress-L-scheme, this has been expected. Utilizing Anderson acceleration, the FSL/2-scheme eventually converges. In particular, convergence is always faster than for the corresponding Fixed-Stress-L-scheme. This again demonstrates the ability of the Anderson acceleration to increase robustness and to relax assumptions for practical convergence.

According to the theory for the Fixed-Stress-L-scheme, a larger tuning parameter or a lower time step size could enable convergence, e.g., for $\alpha = 0.1$, AA(0). However, we do not consider those strategies here, as they lead to worse convergence rates and utilizing Anderson acceleration should be anyhow preferred.

Concerning the best splitting method, we again recommend the use of the Fixed-Stress-Newton method combined with Anderson acceleration with low depth. It is cheap, robust and allows separate simulators.

8.3. Test case III — Unsteady seepage flow through a 2D homogeneous levee

Finally, we consider a test case challenging all linearization schemes, in particular Newton's method. The purpose of this example is again the demonstration of the ability of Anderson acceleration to recover reliably convergence for the splitting methods. We consider unsteady seepage flow through a simple, two-dimensional, homogeneous levee, enforced by a flood. The levee consists of a lower and an upper part (lower 5 [m] and upper 10 [m], respectively), cf. Fig. 5. Initially, the water table lies at the interface between lower and upper parts. The initial fluid pressure is a hydrostatic pressure with $p = 0$ at the water table. The reference configuration, defined by the domain, is initially already consolidated under the influence of gravity. As $\boldsymbol{u}$ is the deviation of the reference configuration, effectively, no gravity is applied in the mechanics equation, but only in the flow equations.

Over time, on the left hand side of the levee, the water table rises with constant speed for four days and remains constant for the next six days, defining $h(t) = 2t$ [m/days], $t \leq 4$ [days], and $h(t) = 8$ [m], 4 [days] $\leq t \leq 10$ [days]. Below $h(t)$ on the left, a hydrostatic pressure boundary condition is applied. On the right side, we apply approximate seepage face boundary conditions, based on the previous time step; i.e., given a fully saturated cell at the previous time step, a pressure boundary condition $p = 0$ is applied on corresponding boundary for the next time step, otherwise a no-flow boundary condition is applied for the volumetric flux. On the remaining boundary, no-flow boundary conditions are applied for all time. For the mechanics, no displacement in normal direction is assumed on the boundary of the lower part of the levee. On the boundary of the upper part and the interface, zero effective stress is applied. The boundary conditions are visualized in Fig. 5.

Physical and numerical parameters. The domain is discretized by a regular, unstructured, simplicial mesh with approximately 67,000 elements and 201,000 nodes. The test cases in Sections 8.1 and 8.2 have shown that the choice of the discretizations $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ or $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$ does not influence the performance of the linearization schemes. Hence, for this test case, for computational reasons, we present the results only for the lowest-order discretization $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$.

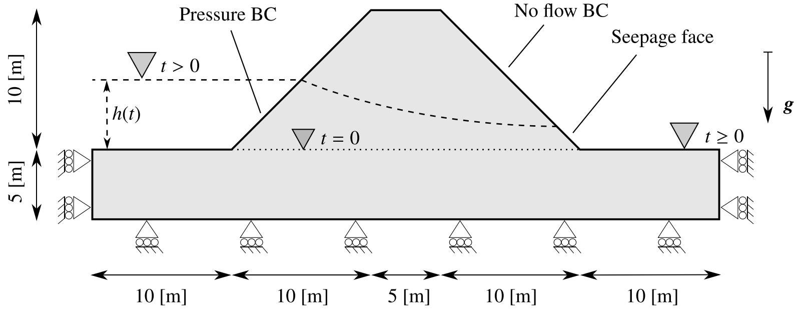


Fig. 5. Domain, boundary and initial conditions for test case III.

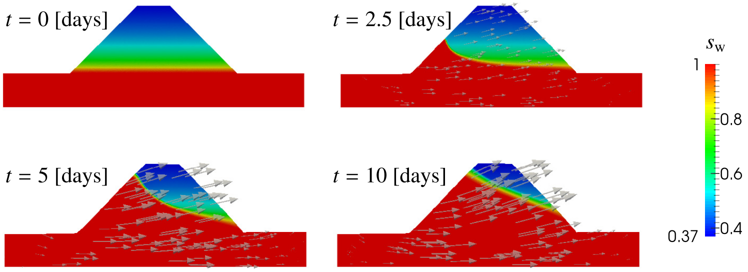


Fig. 6. Simulation result for test case III. Saturation for the deformed material at time $t = 0$ [days], 2.5 [days], 5 [days], 10 [days]. Direction and intensity of deformation also indicated by arrows.

Compared to the previous test cases, we employ more realistic material parameters. Values chosen for model parameters and numerical parameters are displayed in Table 1. We note, the resulting permeability is only Hölder continuous. The saturation history and deformation at four times is displayed in Fig. 6. We observe steep saturation gradients during the flooding. Furthermore, both consolidation and swelling can be observed. All in all, the levee is pushed to the right.

Performance of linearization schemes. We consider the same linearization schemes as in the previous test cases, all but FSL, i.e., the Fixed-Stress-L-scheme with $L = L_s$ and $\beta_{FS} = \beta_{FS}^{phy}$. Based on previous observations, we expect FSL/2 coupled with Anderson acceleration to be more efficient than FSL. The average number of iterations per time step is presented in Table 5. First of all, we observe that all plain linearization schemes fail in the same phase of the simulation (after around 50 time steps). The reason for that lies mainly in the steep saturation gradients. As before, Anderson acceleration can remedy the failure of convergence. However, for this test case, the simple combination of Newton's method and Anderson acceleration does not converge for any considered depth. Indeed, Newton's method combined with AA(1) is not convergent, and for increasing depth the robustness decreases again, which is consistent with observations from the previous test cases. For the remaining linearization schemes convergence can be recovered. In particular, the Fixed-Stress-Newton method combined with AA(1) converges with the least amount of iterations. The Picard-type methods are slower, but show again more robustness with respect to increasing depth, whereas the Fixed-Stress-Newton method diverges eventually for $m = 10$. Here, the Picard type methods require at least depth $m = 3$ for successful convergence. After all, we conclude that the diagonal stabilization is essential for the success of the linearization schemes. The stabilization is added via both the fixed-stress splitting scheme and the L-scheme. Consequently, we expect also the monolithic Newton method to be convergent when adding sufficient diagonal stabilization.

9. Concluding remarks

In this paper, we have proposed three different linearization schemes for nonlinear poromechanics of unsaturated materials. All schemes incorporate the fixed-stress splitting scheme and allow the efficient and robust decoupling of mechanics and flow equations. In particular, the simplest scheme, the Fixed-Stress-L-scheme, employs solely constant diagonal stabilization. It has been derived as L-scheme linearization of the Biot equations reduced to a pure pressure formulation. Under mild, physical assumptions, also needed for the mathematical model to be valid, it has been rigorously

Table 5
Performance for test case III. Average number of (nonlinear) iterations per time step for Newton's method, the Fixed-Stress-Newton method, the Fixed-Stress-Modified-Picard method and the Fixed-Stress-L-scheme; both plain and coupled with Anderson acceleration for different depths ($m = 1,\ 3,\ 5,\ 10$). Minimal numbers per linearization type are in bold. Failing linearization due to stagnation at time step n is marked by → [n]. Failing linearization due to divergence at time step n is marked by ↗ [n].

$\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$				
Linearization	Newton	FS-Newton	FS-MP	FSL/2
AA(0)	↗ [56]	→ [57]	→ [48]	→ [48]
AA(1)	→ [**165**]	**10.2**	→ [86]	→ [73]
AA(3)	→ [90]	11.4	18.1	33.2
AA(5)	→ [87]	10.6	16.9	30.2
AA(10)	→ [87]	→ [87]	**16.0**	**28.3**

shown to be a contraction. This has been also verified numerically. In particular, the numerical examples have shown that the choice of the discretizations $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_1$ or $\mathbb{P}_0 \times \mathbb{RT}_0 \times \mathbb{P}_2$, and by inference, the inf–sup properties of the discretizations, do not influence the performance of the linearization schemes. Exploiting the derivation of the Fixed-Stress-L-scheme allows modifications including first order Taylor approximations. In this way, we have introduced the Fixed-Stress-Modified-Picard and the Fixed-Stress-Newton method.

The derivation of the Fixed-Stress-L-scheme provides two particular side products. First, it reveals the close relation of the L-scheme and the fixed-stress splitting scheme. Second, the nonlinear Biot equations can be shown to be parabolic in the pressure variable. This holds in particular in the fully saturated regime unlike for the Richards equation.

The theoretical convergence rate of the Fixed-Stress-L-scheme might deteriorate for unfavorable situations, leading to slow convergence or even stagnation in practice. Similarly, the Fixed-Stress-Modified-Picard and Fixed-Stress-Newton methods are prone to diverge for Hölder continuous nonlinearities. In order to accelerate or recover convergence, we apply Anderson acceleration, which is a post-processing, maintaining the decoupled character of the underlying splitting methods. The general increase of robustness and acceleration of convergence via the Anderson acceleration has been justified theoretically considering a special linear case. To our knowledge, this is the first theoretical indication of this kind, considering non-contractive fixed-point iterations.

In practice, Anderson acceleration has shown to be very effective for the considered Picard-type methods, confirming the theoretical considerations. We note also that Anderson acceleration can possibly be used to recover convergence for a diverging Newton method. However, no increase in speed can be expected with increasing depth, but rather performance deterioration. Hence, Anderson acceleration should not be used per se, but an adaptive choice should be made. After all the standard Newton method has to be used with care. Instead, we recommend to use splitting schemes with Anderson acceleration as they have been demonstrated to be very robust, even for Hölder continuous nonlinearities. In particular, the Fixed-Stress-Newton method together with the Anderson acceleration showed best performance. However, it also requires the most fine tuning regarding the depth for the Anderson acceleration. If derivatives are not available, we recommend the combination of the Fixed-Stress-L-scheme with a decreased tuning parameter. Without Anderson acceleration, convergence might not be guaranteed. Including it, does not only recover but it also significantly accelerates convergence. This is interesting, as the optimal tuning parameter is not necessarily known *a priori* and can be more safely approached under the use of Anderson acceleration.

As outlook, with focus on large scale applications, the performance of the linearization schemes should be analyzed under the use of parallel, iterative solvers; in particular, as due to added stabilization, the arising linear systems are expected to be better conditioned than for the monolithic Newton method. Additionally, Anderson acceleration should be further studied in the context of possibly non-contractive fixed point iterations. Examples are (i) the linearization of degenerate problems including Hölder continuities, which are known to be difficult to solve [23], and (ii) numerical schemes employing a tuning parameter. Based on the numerical results in this paper, the approach seems very promising.

Acknowledgments

This work was partially supported by the Norwegian Academy of Science and Letters and Statoil through the VISTA AdaSim project #6367 and by the Norwegian Academy of Science and Letters through NFR project 250223.

Appendix A. Convergence proof of abstract L-scheme

We present the proof of Lemma 1 showing convergence of the L-scheme (27) as linearization for Eq. (26). The proof is essentially the same as given by [22], but now written for an algebraic problem.

Proof of Lemma 1. Let $\mathbf{e}_{\mathbf{p}}^{i} = \mathbf{p}^{i} - \mathbf{p}$, $i \in \mathbb{N}$. Then taking the difference of Eqs. (27) and (26) yields

$$\mathbf{L}_{\mathrm{pp}}\left(\mathbf{e}_{\mathbf{p}}^{i}-\mathbf{e}_{\mathbf{p}}^{i-1}\right)+\left(\mathbf{b}(\mathbf{p}^{i-1})-\mathbf{b}(\mathbf{p})\right)+\tau\mathbf{D}\mathbf{K}(\mathbf{p}^{i-1})\mathbf{D}^{\top}\mathbf{e}_{\mathbf{p}}^{i}+\tau\mathbf{D}\left(\mathbf{K}(\mathbf{p}^{i-1})-\mathbf{K}(\mathbf{p})\right)\left(\mathbf{f}_{\mathrm{q}}+\mathbf{D}^{\top}\mathbf{p}\right)=0.$$

Multiplying with $\mathbf{e}_\mathbf{p}^i$ and applying elementary algebraic manipulations, yields

$$\frac{L}{2}\left\|\mathbf{e}_\mathbf{p}^i\right\|_{\mathbf{M}_\mathrm{pp}}^2 + \frac{L}{2}\left\|\mathbf{e}_\mathbf{p}^i - \mathbf{e}_\mathbf{p}^{i-1}\right\|_{\mathbf{M}_\mathrm{pp}}^2 - \frac{L}{2}\left\|\mathbf{e}_\mathbf{p}^{i-1}\right\|_{\mathbf{M}_\mathrm{pp}}^2 \tag{A.1}$$

$$+\left\langle \mathbf{b}(\mathbf{p}^{i-1}) - \mathbf{b}(\mathbf{p}), \mathbf{e}_\mathbf{p}^{i-1}\right\rangle \tag{A.2}$$

$$+\left\langle \mathbf{b}(\mathbf{p}^{i-1}) - \mathbf{b}(\mathbf{p}), \mathbf{e}_\mathbf{p}^{i} - \mathbf{e}_\mathbf{p}^{i-1}\right\rangle \tag{A.3}$$

$$+\,\tau\left\langle \mathbf{K}(\mathbf{p}^{i-1})\mathbf{D}^\top\mathbf{e}_\mathbf{p}^i, \mathbf{D}^\top\mathbf{e}_\mathbf{p}^i\right\rangle \tag{A.4}$$

$$+\,\tau\left\langle \left(\mathbf{K}(\mathbf{p}^{i-1}) - \mathbf{K}(\mathbf{p})\right)\left(\mathbf{f}_\mathrm{q} + \mathbf{D}^\top\mathbf{p}\right), \mathbf{D}^\top\mathbf{e}_\mathbf{p}^i\right\rangle = 0. \tag{A.5}$$

By employing (L1), we obtain for the term (A.2)

$$\left\langle \mathbf{b}(\mathbf{p}^{i-1}) - \mathbf{b}(\mathbf{p}), \mathbf{e}_\mathbf{p}^{i-1}\right\rangle \geq \frac{1}{L_\mathrm{b}}\left\|\mathbf{b}(\mathbf{p}^{i-1}) - \mathbf{b}(\mathbf{p})\right\|_{\mathbf{M}_\mathrm{pp}^{-1}}^2. \tag{A.6}$$

By employing the Cauchy–Schwarz inequality and Young's inequality, we obtain for the term (A.3)

$$\left\langle \mathbf{b}(\mathbf{p}^{i-1}) - \mathbf{b}(\mathbf{p}), \mathbf{e}_\mathbf{p}^{i} - \mathbf{e}_\mathbf{p}^{i-1}\right\rangle \geq -\frac{1}{2L}\left\|\mathbf{b}(\mathbf{p}^{i-1}) - \mathbf{b}(\mathbf{p})\right\|_{\mathbf{M}_\mathrm{pp}^{-1}}^2 - \frac{L}{2}\left\|\mathbf{e}_\mathbf{p}^i - \mathbf{e}_\mathbf{p}^{i-1}\right\|_{\mathbf{M}_\mathrm{pp}}^2. \tag{A.7}$$

By employing Assumption (L2), we obtain for the term (A.4)

$$\left\langle \mathbf{K}(\mathbf{p}^{i-1})\mathbf{D}^\top\mathbf{e}_\mathbf{p}^i, \mathbf{D}^\top\mathbf{e}_\mathbf{p}^i\right\rangle \geq k_\mathrm{m}\left\|\mathbf{D}^\top\mathbf{e}_\mathbf{p}^i\right\|_{\mathbf{M}_\mathrm{qq}^{-1}}^2. \tag{A.8}$$

By employing Cauchy–Schwarz, Young's inequality, Assumptions (L2)–(L3), we obtain for the term (A.5)

$$\begin{aligned}\left\langle \left(\mathbf{K}(\mathbf{p}^{i-1}) - \mathbf{K}(\mathbf{p})\right)\left(\mathbf{f}_\mathrm{q} + \mathbf{D}^\top\mathbf{p}\right), \mathbf{D}^\top\mathbf{e}_\mathbf{p}^i\right\rangle &\geq -\frac{1}{2k_\mathrm{m}}\left\|\mathbf{M}_\mathrm{qq}^{-1}(\mathbf{f}_\mathrm{q} + \mathbf{D}^\top\mathbf{p})\right\|_\infty^2 \left\|(\mathbf{K}(\mathbf{p}^{i-1}) - \mathbf{K}(\mathbf{p}))\mathbf{M}_\mathrm{qq}\right\|_{\mathbf{M}_\mathrm{qq},\infty}^2 - \frac{k_\mathrm{m}}{2}\left\|\mathbf{D}^\top\mathbf{e}_\mathbf{p}^i\right\|_{\mathbf{M}_\mathrm{qq}^{-1}}^2\\ &\geq -\frac{1}{2k_\mathrm{m}}q_\infty^2 L_\mathrm{K}^2\left\|\mathbf{b}(\mathbf{p}^{i-1}) - \mathbf{b}(\mathbf{p})\right\|_{\mathbf{M}_\mathrm{qq}^{-1}}^2 - \frac{k_\mathrm{m}}{2}\left\|\mathbf{D}^\top\mathbf{e}_\mathbf{p}^i\right\|_{\mathbf{M}_\mathrm{qq}^{-1}}^2.\end{aligned} \tag{A.9}$$

Inserting Eqs. (A.6)–(A.9) into Eqs. (A.1)–(A.5), yields

$$\left(\frac{1}{L_\mathrm{b}} - \frac{1}{2L} - \tau\frac{q_\infty^2 L_\mathrm{K}^2}{2k_\mathrm{m}}\right)\left\|\mathbf{b}(\mathbf{p}^{i-1}) - \mathbf{b}(\mathbf{p})\right\|_{\mathbf{M}_\mathrm{pp}^{-1}}^2 + \frac{L}{2}\left\|\mathbf{e}_\mathbf{p}^i\right\|_{\mathbf{M}_\mathrm{pp}}^2 + \tau\frac{k_\mathrm{m}}{2}\left\|\mathbf{D}^\top\mathbf{e}_\mathbf{p}^i\right\|_{\mathbf{M}_\mathrm{qq}^{-1}}^2 \leq \frac{L}{2}\left\|\mathbf{e}_\mathbf{p}^{i-1}\right\|_{\mathbf{M}_\mathrm{pp}}^2. \tag{A.10}$$

Assuming $\frac{1}{L_\mathrm{b}} - \frac{1}{2L} - \tau\frac{q_\infty^2 L_\mathrm{K}^2}{2k_\mathrm{m}} \geq 0$ and applying an algebraic Poincaré inequality, introducing a Poincaré constant C_Ω, yields the final result. □

Appendix B. Auxiliary lemmas for convergence of Fixed-Stress-L-scheme

Before proving Lemma 2, we show that $\mathbf{b}$ is in some sense bi-Lipschitz continuous.

Lemma 7. *Let Assumptions (F1)–(F2) be satisfied. Furthermore, assume $W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$ yields an inf–sup stable discretization. Then for $\mathbf{b}$ as defined in Eq. (25), there exist mesh-independent constants $l_\mathrm{b}, L_\mathrm{b} \in \mathbb{R}_+$ satisfying for all $\mathbf{p}, \tilde{\mathbf{p}} \in \mathbf{P}_{\phi\geq 0}$*

$$l_\mathrm{b}\|\mathbf{p} - \tilde{\mathbf{p}}\|_{\mathbf{M}_\mathrm{pp}}^2 \leq \left\langle \mathbf{b}(\mathbf{p}) - \mathbf{b}(\tilde{\mathbf{p}}), \mathbf{p} - \tilde{\mathbf{p}}\right\rangle \leq L_\mathrm{b}\|\mathbf{p} - \tilde{\mathbf{p}}\|_{\mathbf{M}_\mathrm{pp}}^2.$$

Proof. As $\mathbf{b} \in C^1(\mathbb{R}^{n_\mathrm{p}}; \mathbb{R}^{n_\mathrm{p}})$, with Jacobian $\mathbf{D}_\mathbf{b}(\mathbf{p}) \in \mathbb{R}^{n_\mathrm{p}\times n_\mathrm{p}}$, $\mathbf{p} \in \mathbb{R}^{n_\mathrm{p}}$, and $\mathbf{M}_\mathrm{pp}$ is a diagonal matrix, it holds

$$\sup_{\substack{\mathbf{p},\tilde{\mathbf{p}}\in\mathbf{P}_{\phi\geq 0}\\ \mathbf{p}\neq\tilde{\mathbf{p}}}} \frac{\left\langle \mathbf{b}(\mathbf{p}) - \mathbf{b}(\tilde{\mathbf{p}}), \mathbf{p} - \tilde{\mathbf{p}}\right\rangle}{\|\mathbf{p} - \tilde{\mathbf{p}}\|_{\mathbf{M}_\mathrm{pp}}^2} = \sup_{\substack{\mathbf{p}\in\mathbf{P}_{\phi\geq 0},\mathbf{h}\in\mathbb{R}^{n_\mathrm{p}}\setminus\{\mathbf{0}\}\\ \mathbf{p}+\mathbf{h}\in\mathbf{P}_{\phi\geq 0}}} \frac{\langle \mathbf{D}_\mathbf{b}(\mathbf{p})\mathbf{h}, \mathbf{h}\rangle}{\|\mathbf{h}\|_{\mathbf{M}_\mathrm{pp}}^2} = \sup_{\substack{\mathbf{p}\in\mathbf{P}_{\phi\geq 0},\mathbf{h}\in\mathbb{R}^{n_\mathrm{p}}\setminus\{\mathbf{0}\}\\ \mathbf{p}+\mathbf{h}\in\mathbf{P}_{\phi\geq 0}}} \frac{\left\langle \mathbf{M}_\mathrm{pp}^{-1/2}\mathbf{D}_\mathbf{b}(\mathbf{p})\mathbf{M}_\mathrm{pp}^{-1/2}\mathbf{h}, \mathbf{h}\right\rangle}{\|\mathbf{h}\|^2}. \tag{B.1}$$

Employing the properties of $\mathbf{b}$, and making use of the specific choice of the equivalent pore pressure (5), the Jacobian of $\mathbf{b}$ is given by

$$\mathbf{D}_\mathbf{b}(\mathbf{p}) = \begin{bmatrix} s'(\mathbf{p}_1)\boldsymbol{\phi}_1(\mathbf{p}) & & \\ & \ddots & \\ & & s'(\mathbf{p}_{n_\mathrm{p}})\boldsymbol{\phi}_{n_\mathrm{p}}(\mathbf{p}) \end{bmatrix} + \alpha^2\mathbf{S}_\mathrm{pp}(\mathbf{p})\mathbf{D}_\mathrm{pu}\mathbf{A}_\mathrm{uu}^{-1}\mathbf{D}_\mathrm{pu}^\top\mathbf{S}_\mathrm{pp}(\mathbf{p})^\top + \frac{1}{N}\mathbf{S}_\mathrm{pp}(\mathbf{p})\mathbf{M}_\mathrm{pp}\mathbf{S}_\mathrm{pp}(\mathbf{p})^\top. \tag{B.2}$$

Hence, $\mathbf{D}_\mathbf{b}(\mathbf{p}) = \mathbf{D}_\mathbf{b}(\mathbf{p})^\top$ for all $\mathbf{p} \in \mathbf{P}_{\phi\geq 0}$ with eigenvalues greater than or equal to zero. After all, the largest value for the Rayleigh quotient (B.1) is given by the largest eigenvalue of $\mathbf{M}_\mathrm{pp}^{-1/2}\mathbf{D}_\mathbf{b}(\mathbf{p})\mathbf{M}_\mathrm{pp}^{-1/2}$ maximized over $\mathbf{p} \in \mathbf{P}_{\phi\geq 0}$. The components of the porosity vector $\boldsymbol{\phi}$ are by assumption positive. Additionally, due to stability of $W_h \times \boldsymbol{Z}_h \times \boldsymbol{V}_h$, $\mathbf{D}_\mathrm{pu}\mathbf{A}_\mathrm{uu}^{-1}\mathbf{D}_\mathrm{pu}^\top$ is norm equivalent with $\mathbf{M}_{\mathrm{pp}}$. Hence, also $\mathbf{D_b}(\mathbf{p})$ is norm equivalent with the standard mass matrix with mesh-independent bounds. Together with employing the assumptions, we see there exists a largest eigenvalue $L_b \in \mathbb{R}_+$ of $\mathbf{M}_{\mathrm{pp}}^{-1/2}\mathbf{D_b}\mathbf{M}_{\mathrm{pp}}^{-1/2}$ independent of the mesh. Analogously, it holds

$$\inf_{\substack{\mathbf{p},\tilde{\mathbf{p}}\in\mathbf{P}_{\phi\geq0}\\ \mathbf{p}\neq\tilde{\mathbf{p}}}} \frac{\left\langle \mathbf{b}(\mathbf{p})-\mathbf{b}(\tilde{\mathbf{p}}),\mathbf{p}-\tilde{\mathbf{p}}\right\rangle}{\|\mathbf{p}-\tilde{\mathbf{p}}\|^2_{\mathbf{M}_{\mathrm{pp}}}} = \inf_{\substack{\mathbf{p}\in\mathbf{P}_{\phi\geq0},\mathbf{h}\in\mathbb{R}^{n_\mathrm{p}}\setminus\{\mathbf{0}\}\\ \mathbf{p}+\mathbf{h}\in\mathbf{P}_{\phi\geq0}}} \frac{\left\langle \mathbf{M}_{\mathrm{pp}}^{-1/2}\mathbf{D_b}(\mathbf{p})\mathbf{M}_{\mathrm{pp}}^{-1/2}\mathbf{h},\mathbf{h}\right\rangle}{\|\mathbf{h}\|^2}$$

with the value given by the smallest eigenvalue l_b of $\mathbf{M}_{\mathrm{pp}}^{-1/2}\mathbf{D_b}(\mathbf{p})\mathbf{M}_{\mathrm{pp}}^{-1/2}$ minimized over $\mathbf{p}\in\mathbf{P}_{\phi\geq0}$. From above discussion it follows that $l_b\in\mathbb{R}_+$ is mesh-independent. All in all, the proposed thesis follows. □

Proof of Lemma 2. First, we show (28). From Lemma 7, it follows, $\mathbf{b}$ is invertible and $\mathbf{D_b}$ is symmetric. Using the Inverse Function theorem, it holds

$$\begin{aligned}
\sup_{\substack{\mathbf{p},\tilde{\mathbf{p}}\in\mathbf{P}_{\phi\geq0}\\ \mathbf{p}\neq\tilde{\mathbf{p}}}} \frac{\|\mathbf{b}(\mathbf{p})-\mathbf{b}(\tilde{\mathbf{p}})\|^2_{\mathbf{M}_{\mathrm{pp}}^{-1}}}{\left\langle \mathbf{b}(\mathbf{p})-\mathbf{b}(\tilde{\mathbf{p}}),\mathbf{p}-\tilde{\mathbf{p}}\right\rangle} &= \sup_{\substack{\mathbf{b}^{-1}(\mathbf{p}),\mathbf{b}^{-1}(\tilde{\mathbf{p}})\in\mathbf{P}_{\phi\geq0}\\ \mathbf{p}\neq\tilde{\mathbf{p}}}} \frac{\|\mathbf{p}-\tilde{\mathbf{p}}\|^2_{\mathbf{M}_{\mathrm{pp}}^{-1}}}{\left\langle \mathbf{b}^{-1}(\mathbf{p})-\mathbf{b}^{-1}(\tilde{\mathbf{p}}),\mathbf{p}-\tilde{\mathbf{p}}\right\rangle}\\
&= \left[\inf_{\substack{\mathbf{b}^{-1}(\mathbf{p}),\mathbf{b}^{-1}(\tilde{\mathbf{p}})\in\mathbf{P}_{\phi\geq0}\\ \mathbf{p}\neq\tilde{\mathbf{p}}}} \frac{\left\langle \mathbf{b}^{-1}(\mathbf{p})-\mathbf{b}^{-1}(\tilde{\mathbf{p}}),\mathbf{p}-\tilde{\mathbf{p}}\right\rangle}{\|\mathbf{p}-\tilde{\mathbf{p}}\|^2_{\mathbf{M}_{\mathrm{pp}}^{-1}}}\right]^{-1}\\
&= \sup_{\substack{\mathbf{p},\tilde{\mathbf{p}}\in\mathbf{P}_{\phi\geq0}\\ \mathbf{p}\neq\tilde{\mathbf{p}}}} \frac{\left\langle \mathbf{b}(\mathbf{p})-\mathbf{b}(\tilde{\mathbf{p}}),\mathbf{p}-\tilde{\mathbf{p}}\right\rangle}{\|\mathbf{p}-\tilde{\mathbf{p}}\|^2_{\mathbf{M}_{\mathrm{pp}}}}.
\end{aligned}$$

Analogously for the infimum. By Lemma 7, (L1) holds. Indeed, it follows for all $\mathbf{p},\tilde{\mathbf{p}}\in\mathbf{P}_{\phi\geq0}$

$$l_\mathrm{b}\left\langle \mathbf{b}(\mathbf{p})-\mathbf{b}(\tilde{\mathbf{p}}),\mathbf{p}-\tilde{\mathbf{p}}\right\rangle \leq \|\mathbf{b}(\mathbf{p})-\mathbf{b}(\tilde{\mathbf{p}})\|^2_{\mathbf{M}_{\mathrm{pp}}^{-1}} \leq L_\mathrm{b}\left\langle \mathbf{b}(\mathbf{p})-\mathbf{b}(\tilde{\mathbf{p}}),\mathbf{p}-\tilde{\mathbf{p}}\right\rangle. \tag{B.3}$$

Next, we show (29). As the underlying permeability $k_\mathrm{w}=k_\mathrm{w}(s_\mathrm{w})$ is Lipschitz continuous, together with a scaling argument, it follows, there exists a constant $\tilde{L}_\mathrm{K}\in\mathbb{R}_+$ satisfying

$$\|(\mathbf{K}(\mathbf{p})-\mathbf{K}(\tilde{\mathbf{p}}))\mathbf{M}_{\mathrm{qq}}\|_{\mathbf{M}_{\mathrm{qq}},\infty} \leq \tilde{L}_\mathrm{K}\|\mathbf{S}_{\mathrm{pp}}(\mathbf{p})-\mathbf{S}_{\mathrm{pp}}(\tilde{\mathbf{p}})\|_{\mathbf{M}_{\mathrm{pp}},\infty}.$$

Furthermore, as $s_\mathrm{w}=s_\mathrm{w}(p)$ is Lipschitz continuous, and $\mathbf{S}_{\mathrm{pp}}$ is a diagonal matrix, there exists a constant $L_\mathrm{s}\in\mathbb{R}_+$ satisfying

$$\|\mathbf{S}_{\mathrm{pp}}(\mathbf{p})-\mathbf{S}_{\mathrm{pp}}(\tilde{\mathbf{p}})\|_{\mathbf{M}_{\mathrm{pp}},\infty} \leq L_\mathrm{s}\|\mathbf{p}-\tilde{\mathbf{p}}\|_{\mathbf{M}_{\mathrm{pp}}}.$$

All in all, with Lemma 7 and inequality (B.3), Eq. (29) follows with $L_\mathrm{K}=\tilde{L}_\mathrm{K}L_\mathrm{s}l_\mathrm{b}^{-2}$. Eq. (30) follows directly from Assumption (F3) together with a scaling argument. □

Appendix C. Proof for contraction of AA*(1)

Proof of Lemma 4. First, an iteration-dependent error propagation matrix is derived, and second, an upper bound for its spectral radius is computed. For this purpose, we ignore Assumption (C4) for a moment.

Iterative error propagation. As we intend to relate $\mathbf{e}^{i+4}$ with $\mathbf{e}^i$, we explicitly write out the first four iterates and the corresponding errors. Given $\mathbf{x}^i$, by using $\mathbf{b}=\mathbf{x}^\star-\mathbf{A}\mathbf{x}^\star$ and $\mathbf{x}^i-\mathbf{x}^{i+1}=\mathbf{e}^i-\mathbf{e}^{i+1}$, we obtain

$$\mathbf{x}^{i+1}=\mathbf{A}\mathbf{x}^i+\mathbf{b}, \qquad \mathbf{e}^{i+1}=\mathbf{A}\mathbf{e}^i, \tag{C.1}$$

$$\mathbf{x}^{i+2}=\mathbf{A}\mathbf{x}^{i+1}+\mathbf{b}+\alpha^{(i+1)}\mathbf{A}(\mathbf{x}^i-\mathbf{x}^{i+1}), \qquad \mathbf{e}^{i+2}=\mathbf{A}\mathbf{e}^{i+1}+\alpha^{(i+1)}\mathbf{A}(\mathbf{e}^i-\mathbf{e}^{i+1}), \tag{C.2}$$

$$\mathbf{x}^{i+3}=\mathbf{A}\mathbf{x}^{i+2}+\mathbf{b}, \qquad \mathbf{e}^{i+3}=\mathbf{A}\mathbf{e}^{i+2}, \tag{C.3}$$

$$\mathbf{x}^{i+4}=\mathbf{A}\mathbf{x}^{i+3}+\mathbf{b}+\alpha^{(i+3)}\mathbf{A}(\mathbf{x}^{i+3}-\mathbf{x}^{i+2}), \qquad \mathbf{e}^{i+4}=\mathbf{A}\mathbf{e}^{i+3}+\alpha^{(i+3)}\mathbf{A}(\mathbf{e}^{i+2}-\mathbf{e}^{i+3}). \tag{C.4}$$

By plugging all together, we obtain

$$\mathbf{e}^{i+4}=\mathbf{A}(\mathbf{A}+\alpha^{(i+3)}(\mathbf{I}-\mathbf{A}))\mathbf{A}(\mathbf{A}+\alpha^{(i+1)}(\mathbf{I}-\mathbf{A}))\mathbf{e}^i.$$

It suffices to bound the largest eigenvalue of the error propagation matrix $\mathbf{A}(\mathbf{A}+\alpha^{(i+3)}(\mathbf{I}-\mathbf{A}))\mathbf{A}(\mathbf{A}+\alpha^{(i+1)}(\mathbf{I}-\mathbf{A}))$. From Assumption (C2) it follows that $\{\mathbf{v}_j\}_j$ defines an orthogonal basis of eigenvectors for the error propagation matrix with corresponding eigenvalues $\{\tilde{\lambda}_j\}_j$ defined by

$$\tilde{\lambda}_j=\lambda_j^2(\lambda_j+\alpha^{(i+1)}(1-\lambda_j))(\lambda_j+\alpha^{(i+3)}(1-\lambda_j)). \tag{C.5}$$

Explicit definition of $\alpha^{(i+1)}$ and $\alpha^{(i+3)}$. The minimization problem in Algorithm 2 can be solved explicitly, by solving adequate normal equations. It follows, that

$$\alpha^{(i+1)} = \frac{(\Delta\mathcal{FP}(\mathbf{x}^{i+1}) - \Delta\mathcal{FP}(\mathbf{x}^{i}))\cdot \Delta\mathcal{FP}(\mathbf{x}^{i+1})}{(\Delta\mathcal{FP}(\mathbf{x}^{i+1}) - \Delta\mathcal{FP}(\mathbf{x}^{i}))\cdot(\Delta\mathcal{FP}(\mathbf{x}^{i+1}) - \Delta\mathcal{FP}(\mathbf{x}^{i}))}.$$

After employing simple arithmetics and using Eq. (C.1), we obtain

$$\begin{aligned}\Delta\mathcal{FP}(\mathbf{x}^{i+1}) &= (\mathbf{A}-\mathbf{I})\mathbf{x}^{i+1} + \mathbf{b} = (\mathbf{A}-\mathbf{I})\mathbf{e}^{i+1} = (\mathbf{A}-\mathbf{I})\mathbf{A}\mathbf{e}^{i} = \mathbf{A}(\mathbf{A}-\mathbf{I})\mathbf{e}^{i},\\ \Delta\mathcal{FP}(\mathbf{x}^{i+1}) - \Delta\mathcal{FP}(\mathbf{x}^{i}) &= (\mathbf{A}-\mathbf{I})(\mathbf{x}^{i+1}-\mathbf{x}^{i}) = (\mathbf{A}-\mathbf{I})(\mathbf{e}^{i+1}-\mathbf{e}^{i}) = (\mathbf{A}-\mathbf{I})^2\mathbf{e}^{i}.\end{aligned}$$

Consequently, it holds

$$\alpha^{(i+1)} = \frac{((\mathbf{A}-\mathbf{I})^2\mathbf{e}^i)\cdot(\mathbf{A}(\mathbf{A}-\mathbf{I})\mathbf{e}^i)}{\|(\mathbf{A}-\mathbf{I})^2\mathbf{e}^i\|^2} = \hat{\mathbf{e}}^i\cdot\mathbf{A}(\mathbf{A}-\mathbf{I})^{-1}\hat{\mathbf{e}}^i, \tag{C.6}$$

where we define $\hat{\mathbf{e}}^i = (\mathbf{A}-\mathbf{I})^2\mathbf{e}^i/\|(\mathbf{A}-\mathbf{I})^2\mathbf{e}^i\|$, satisfying $\|\hat{\mathbf{e}}^i\| = 1$. Analogously, using Eqs. (C.1)–(C.4), we obtain

$$\alpha^{(i+3)} = \frac{((\mathbf{A}-\mathbf{I})^2\mathbf{e}^{i+2})\cdot(\mathbf{A}(\mathbf{A}-\mathbf{I})\mathbf{e}^{i+2})}{\|(\mathbf{A}-\mathbf{I})^2\mathbf{e}^{i+2}\|^2} = \frac{\hat{\mathbf{e}}^i\cdot\mathbf{A}^3(\mathbf{A}-\mathbf{I})^{-1}(\mathbf{A}+\alpha^{(i+1)}(\mathbf{I}-\mathbf{A}))^2\hat{\mathbf{e}}^i}{\|\mathbf{A}(\mathbf{A}+\alpha^{(i+1)}(\mathbf{I}-\mathbf{A}))\hat{\mathbf{e}}^i\|^2}. \tag{C.7}$$

Decomposition of $\hat{\mathbf{e}}^i$ and useful computations. Employing the orthogonal eigenvector basis $\{\mathbf{v}_j\}_j$, we can decompose $\hat{\mathbf{e}}^i = \sum_j \beta_j\mathbf{v}_j$. As $\|\hat{\mathbf{e}}^i\| = 1$ it holds $\sum_j\beta_j^2 = 1$. By inserting the decomposition into Eq. (C.6), we obtain

$$\alpha^{(i+1)} = \sum_j \beta_j^2\frac{\lambda_j}{\lambda_j - 1}.$$

Hence, for the eigenvalues of $\mathbf{A}+\alpha^{(i+1)}(\mathbf{I}-\mathbf{A})$ and also the second factor of Eq. (C.5), it follows

$$\eta_j(\boldsymbol{\beta}) := \lambda_j + \alpha^{(i+1)}(1-\lambda_j) = \sum_{k\neq j}\beta_k^2\frac{\lambda_k-\lambda_j}{\lambda_k - 1}, \tag{C.8}$$

where $\boldsymbol{\beta} = \left[\beta_1,\ldots,\beta_n\right]^\top \in \mathbb{R}^n$. Hence, for the contribution in the denominator of Eq. (C.7), we obtain

$$\mathbf{A}(\mathbf{A}+\alpha^{(i+1)}(\mathbf{I}-\mathbf{A}))\hat{\mathbf{e}}^i = \sum_j \beta_j\lambda_j\eta_j(\boldsymbol{\beta})\mathbf{v}_j.$$

By plugging into Eq. (C.7) and using the orthogonality of $\{\mathbf{v}_j\}_j$, we obtain for $\alpha^{(i+3)}$

$$\alpha^{(i+3)} = \left[\sum_j \beta_j^2\lambda_j^2\eta_j(\boldsymbol{\beta})^2\right]^{-1}\left[\sum_j\beta_j^2\frac{\lambda_j^3}{\lambda_j-1}\eta_j(\boldsymbol{\beta})^2\right].$$

By employing some arithmetics, for the third factor of Eq. (C.5), it follows

$$\lambda_j + \alpha^{(i+3)}(1-\lambda_j) = \left[\sum_k\beta_k^2\lambda_k^2\eta_k(\boldsymbol{\beta})^2\right]^{-1}\left[\sum_{k\neq j}\beta_k^2\lambda_k^2\frac{\lambda_k-\lambda_j}{\lambda_k-1}\eta_k(\boldsymbol{\beta})^2\right]. \tag{C.9}$$

Resulting eigenvalues. By inserting Eqs. (C.8)–(C.9) into Eq. (C.5), we obtain for the eigenvalues of the iteration-dependent error propagation matrix $\mathbf{A}(\mathbf{A}+\alpha^{(i+3)}(\mathbf{A}-\mathbf{I}))\mathbf{A}(\mathbf{A}+\alpha^{(i+1)}(\mathbf{A}-\mathbf{I}))$

$$\tilde{\lambda}_j = \left[\sum_k\beta_k^2\lambda_k^2\eta_k(\boldsymbol{\beta})^2\right]^{-1}\left[\lambda_j^2\eta_j(\boldsymbol{\beta})\sum_{k\neq j}\beta_k^2\lambda_k^2\eta_k(\boldsymbol{\beta})^2\frac{\lambda_k-\lambda_j}{\lambda_k-1}\right]. \tag{C.10}$$

Analysis for special decomposition. By Assumption (C4), the initial error is spanned by two orthogonal eigenvectors. Without loss of generality let $\mathbf{e}^0 \in \mathrm{span}\{\mathbf{v}_1,\mathbf{v}_2\}$. Then also $\hat{\mathbf{e}}^i \in \mathrm{span}\{\mathbf{v}_1,\mathbf{v}_2\}$ and there exist $\beta_1,\beta_2\in\mathbb{R}$ satisfying $\hat{\mathbf{e}}^i = \beta_1\mathbf{v}_1 + \beta_2\mathbf{v}_2$ and $\beta_1^2+\beta_2^2 = 1$. Consequently, Eq. (C.10) for $j = 1$ reduces to

$$\tilde{\lambda}_1 = \lambda_1^2\lambda_2^2(\lambda_2-\lambda_1)^2\frac{(1-\gamma)\gamma}{(1-\gamma)\lambda_1^2(\lambda_1-1)^2+\gamma\lambda_2^2(\lambda_2-1)^2},$$

where $\gamma = \beta_1^2 \in [0, 1]$. Maximizing the second factor with respect to $\gamma \in [0, 1]$, results in the upper bound

$$|\tilde{\lambda}_1| \le \frac{\lambda_1^2\lambda_2^2(\lambda_2 - \lambda_1)^2}{(|\lambda_1(\lambda_1 - 1)| + |\lambda_2(\lambda_2 - 1)|)^2} =: r(\lambda_1, \lambda_2).$$

Due to symmetry it holds $|\tilde{\lambda}_j| \le r(\lambda_1, \lambda_2), j = 1{,}2$. Consequently, we obtain the result. □

References

[1] M. Biot, General theory of three-dimensional consolidation, J. Appl. Phys. 12 (2) (1941) 155–164.
[2] O. Coussy, Poromechanics, Wiley, 2004.
[3] J. Kim, H.A. Tchelepi, R. Juanes, Rigorous coupling of geomechanics and multiphase flow with strong capillarity, Soc. Pet. Eng. (2013).
[4] van Genuchten, A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, Soil Sci. Am. J. 44 (5) (1980) 892–898.
[5] A. Settari, F. Mourits, A coupled reservoir and geomechanical simulation system, Soc. Pet. Eng. 3 (1998) 219–226.
[6] J.A. White, R.I. Borja, Block-preconditioned Newton–Krylov solvers for fully coupled flow and geomechanics, Comput. Geosci. 15 (4) (2011) 647.
[7] J. Kim, H.A. Tchelepi, R. Juanes, Stability, accuracy, and efficiency of sequential methods for coupled flow and geomechanics, Soc. Pet. Eng. (2011).
[8] A. Mikelić, M.F. Wheeler, Convergence of iterative coupling for coupled flow and geomechanics, Comput. Geosci. 17 (3) (2013) 455–461.
[9] J.W. Both, M. Borregales, J.M. Nordbotten, K. Kumar, F.A. Radu, Robust fixed stress splitting for Biot's equations in heterogeneous media, Appl. Math. Lett. 68 (2017) 101–108.
[10] M. Bause, F.A. Radu, U. Köcher, Space–time finite element approximation of the Biot poroelasticity system with iterative coupling, Comput. Methods Appl. Mech. Engrg. 320 (2017) 745–768.
[11] M. Borregales, K. Kumar, F.A. Radu, C. Rodrigo, F. José Gaspar, A parallel-in-time fixed-stress splitting method for Biot's consolidation model, 2018. arXiv:1802.00949 [math.NA].
[12] S. Dana, B. Ganis, M.F. Wheeler, A multiscale fixed stress split iterative scheme for coupled flow and poromechanics in deep subsurface reservoirs, J. Comput. Phys. 352 (2018) 1–22.
[13] N. Castelletto, J.A. White, H.A. Tchelepi, Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics, Int. J. Numer. Anal. Methods Geomech. 39 (14) (2015) 1593–1618.
[14] N. Castelletto, J.A. White, M. Ferronato, Scalable algorithms for three-field mixed finite element coupled poromechanics, J. Comput. Phys. 327 (2016) 894–918.
[15] J.A. White, N. Castelletto, H.A. Tchelepi, Block-partitioned solvers for coupled poromechanics: A unified framework, Comput. Methods Appl. Mech. Engrg. 303 (2016) 55–74.
[16] J.H. Adler, F.J. Gaspar, X. Hu, C. Rodrigo, L.T. Zikatanov, Robust block preconditioners for Biot's model, 2017. arXiv:1705.08842 [math.NA].
[17] F.J. Gaspar, C. Rodrigo, On the fixed-stress split scheme as smoother in multigrid methods for coupling flow and geomechanics, Comput. Methods Appl. Mech. Engrg. 326 (2017) 526–540.
[18] M. Slodicka, A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media, SIAM J. Sci. Comput. 23 (5) (2002) 1593–1614 https://doi.org/10.1137/S1064827500381860.
[19] I.S. Pop, F.A. Radu, P. Knabner, Mixed finite elements for the Richards' equation: Linearization procedure, J. Comput. Appl. Math. 168 (1) (2004) 365–373.
[20] F.A. Radu, J.M. Nordbotten, I.S. Pop, K. Kumar, A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media, J. Comput. Appl. Math. 289 (2015) 134–141.
[21] F.A. Radu, K. Kumar, J.M. Nordbotten, I.S. Pop, A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities, IMA J. Numer. Anal. 38 (2) (2018) 884–920.
[22] F. List, F.A. Radu, A study on iterative methods for solving Richards' equation, Comput. Geosci. 20 (2) (2016) 341–353.
[23] J.W. Both, K. Kumar, J.M. Nordbotten, I. Sorin Pop, F.A. Radu, Linear iterative schemes for doubly degenerate parabolic equations, 2018. arXiv:1801.00846 [math.NA].
[24] D. Seus, K. Mitra, I.S. Pop, F.A. Radu, C. Rohde, A linear domain decomposition method for partially saturated flow in porous media, Comput. Methods Appl. Mech. Engrg. 333 (2018) 331–355.
[25] M. Borregales, F.A. Radu, K. Kumar, J.M. Nordbotten, Robust iterative schemes for non-linear poromechanics, Comput. Geosci. (2018).
[26] M.A. Celia, E.T. Bouloutas, R.L. Zarba, A general mass-conservative numerical solution for the unsaturated flow equation, Water Resour. Res. 26 (7) (1990) 1483–1496.
[27] D.G. Anderson, Iterative procedures for nonlinear integral equations, J. Assoc. Comput. Mach. 12 (4) (1965) 547–560.
[28] P. Lott, H. Walker, C. Woodward, U. Yang, An accelerated Picard method for nonlinear systems related to variably saturated flow, Adv. Water Resour. 38 (2012) 92–101.
[29] H. Fang, Y. Saad, Two classes of multisecant methods for nonlinear acceleration, Numer. Linear Algebra Appl. 16 (3) (2009) 197–221.
[30] H.F. Walker, P. Ni, Anderson acceleration for fixed-point iterations, SIAM J. Numer. Anal. 49 (4) (2011) 1715–1735.
[31] A. Toth, C.T. Kelley, Convergence analysis for Anderson acceleration, SIAM J. Numer. Anal. 53 (2) (2015) 805–819.
[32] C. Paniconi, M. Putti, A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems, Water Resour. Res. 30 (12) (1994) 3357–3374.
[33] C. Rodrigo, X. Hu, P. Ohm, J.H. Adler, F.J. Gaspar, L. Zikatanov, New stabilized discretizations for poroelasticity and the Stokes' equations, 2017. arXiv:1706.05169 [math.NA].
[34] J.B. Haga, H. Osnes, H.P. Langtangen, On the causes of pressure oscillations in low-permeable and low-compressible porous media, Int. J. Numer. Anal. Methods Geomech. 36 (12) (2012) 1507–1522.
[35] J. Both, U. Köcher, Numerical investigation on the fixed-stress splitting scheme for Biot's equations: Optimality of the tuning parameter, 2018. arXiv:1801.08352 [math.NA].
[36] J.B. Haga, H. Osnes, H.P. Langtangen, A parallel block preconditioner for large-scale poroelasticity with highly heterogeneous material parameters, Comput. Geosci. 16 (3) (2012) 723–734.
[37] T. Arbogast, M. Obeyesekere, M.F. Wheeler, Numerical methods for the simulation of flow in root-soil systems, SIAM J. Numer. Anal. 30 (6) (1993) 1677–1702.
[38] T. Arbogast, M.F. Wheeler, N.-Y. Zhang, A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media, SIAM J. Numer. Anal. (33) (1996) 1669–1687.
[39] F. Radu, I.S. Pop, P. Knabner, Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation, SIAM J. Numer. Anal. 42 (4) (2004) 1452–1478.
[40] F.A. Radu, W. Wang, Convergence analysis for a mixed finite element scheme for flow in strictly unsaturated porous media, Nonlinear Anal. RWA 15 (2014) 266–275.

[41] R. Showalter, N. Su, Partially saturated flow in a poroelastic medium, Discrete Contin. Dyn. Syst. Ser. B 1 (4) (2001) 403–420.

[42] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, O. Sander, A generic grid interface for parallel and adaptive scientific computing. Part I: Abstract framework, Computing 82 (2) (2008) 103–119.

[43] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, O. Sander, A generic grid interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE, Computing 82 (2) (2008) 121–138.

[44] M. Blatt, A. Burchardt, A. Dedner, C. Engwer, J. Fahlke, B. Flemisch, C. Gersbacher, C. Gräser, F. Gruber, C. Grüninger, D. Kempf, R. Klöfkorn, T. Malkmus, S. Müthing, M. Nolte, M. Piatkowski, O. Sander, The distributed and unified numerics environment, version 2.4, Arch. Numer. Softw. 4 (100) (2016) 13–29.

Paper G

On the optimization of the fixed-stress splitting for Biot's equations

Storvik, E., Both, J.W., Kumar, K., Nordbotten, J.M., and Radu, F.A.

International Journal for Numerical Methods in Engineering **120** (2019), pg. 179–194.

doi:10.1002/nme.6130

Received: 6 November 2018 | Revised: 6 April 2019 | Accepted: 23 May 2019
DOI: 10.1002/nme.6130

RESEARCH ARTICLE

On the optimization of the fixed-stress splitting for Biot's equations

Erlend Storvik[1] | **Jakub W. Both**[1] | **Kundan Kumar**[1,2] | **Jan M. Nordbotten**[1,3] | **Florin A. Radu**[1]

[1]Department of Mathematics, University of Bergen, Bergen, Norway

[2]Department of Mathematics and Computer Science, Karlstad University, Karlstad, Sweden

[3]Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

Correspondence
Florin A. Radu, Department of Mathematics, University of Bergen, Allégaten 41, 5007 Bergen, Norway.
Email: florin.radu@uib.no

Funding information
Norges Forskningsrád, Grant/Award Number: 250223

Summary

In this work, we are interested in efficiently solving the quasi-static, linear Biot model for poroelasticity. We consider the fixed-stress splitting scheme, which is a popular method for iteratively solving Biot's equations. It is well known that the convergence properties of the method strongly depend on the applied stabilization/tuning parameter. We show theoretically that, in addition to depending on the mechanical properties of the porous medium and the coupling coefficient, they also depend on the fluid flow and spatial discretization properties. The type of analysis presented in this paper is not restricted to a particular spatial discretization, although it is required to be inf-sup stable with respect to the displacement-pressure formulation. Furthermore, we propose a way to optimize this parameter that relies on the mesh independence of the scheme's optimal stabilization parameter. Illustrative numerical examples show that using the optimized stabilization parameter can significantly reduce the number of iterations.

KEYWORDS

Biot model, convergence analysis, fixed-stress splitting, geomechanics, poroelasticity

1 | INTRODUCTION

There is currently a strong interest in the numerical simulation of poroelasticity, ie, fully coupled porous media flow and mechanics. This is due to its high number of societal relevant applications, such as geothermal energy extraction, life sciences, or CO_2 storage, to name a few. The most commonly used mathematical model for poroelasticity is the quasi-static, linear Biot model. It is the coupled problem arising when considering the balance of linear momentum for the porous medium allowing for only small deformations (1) and mass conservation and Darcy's law for the fluid flow (2) (see, eg, the work of Coussy[1]): find $(\mathbf{u}, p)$ such that

$$-\nabla \cdot (2\mu\epsilon(\boldsymbol{u}) + \lambda\nabla\cdot\boldsymbol{u}\boldsymbol{I}) + \alpha\nabla p = \boldsymbol{f}, \tag{1}$$

$$\frac{\partial}{\partial t}\left(\frac{p}{M} + \alpha\nabla\cdot\boldsymbol{u}\right) - \nabla\cdot(\kappa(\nabla p - \mathbf{g}\rho)) = S_f, \tag{2}$$

where $\boldsymbol{u}$ is the displacement; $\varepsilon(\boldsymbol{u}) := \frac{1}{2}\left(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^{\top}\right)$ is the (linear) strain tensor; μ and λ are the Lamé parameters; α is the Biot-Willis constant; p and ρ are the fluid's pressure and density, respectively; $1/M$ is the compressibility constant; $\boldsymbol{g}$ is the gravitational vector; and κ is the permeability. The source terms $\boldsymbol{f}$ and S_f represent the density of applied body forces and a forced fluid extraction or injection process, respectively.

A lot of work has been done concerning the discretization of Biot's equations (1) and (2). Various spatial discretizations, combined with the backward Euler method as temporal discretization, have been proposed and analyzed. We mention cell-centered finite volumes,[2] continuous Galerkin for the mechanics and mixed finite elements for the flow,[3-6] mixed finite elements for flow and mechanics,[4,7] nonconforming finite elements,[8] the MINI element,[9] continuous or discontinuous Galerkin,[10-12] or multiscale methods.[13-15] Continuous and discontinuous higher-order Galerkin space-time finite elements were proposed in the work of Bause et al.[16] Adaptive computations were considered, for example, in the work of Ern and Meunier.[17] A Monte Carlo approach was proposed in the work of Rahrah and Vermolen.[18] For a discussion on the stability of different spatial discretizations, we refer to the recent papers.[19,20]

Independently of the chosen discretization, there are two popular alternatives for solving Biot's equations: monolithically or by using an iterative splitting algorithm. The former has the advantage of being unconditionally stable, whereas a splitting method is much easier to implement, typically building on already available, tailored, separate numerical codes for porous media flow and for mechanics. However, a naive splitting of Biot's equations will lead to an unstable scheme.[21] To overcome this, one adds a stabilization term in either the mechanics equation (the so-called *undrained splitting scheme*[22]) or the flow equation (the *fixed-stress splitting scheme*).[23] The splitting methods have very good convergence properties, making them a valuable alternative to monolithic solvers for simulation of the linear Biot model (see, eg, the works of Both et al,[5] Kim et al,[21] Settari and Mourits,[23] and Mikelić and Wheeler[24]). In the present work, we will discuss the fixed-stress splitting scheme. For other splitting schemes, see, for example, the works of Turska and Schrefler[25] and Turska et al.[26]

After applying the backward Euler method in time to (1) and (2) and discretizing in space (using finite elements or finite volumes), one has to solve a fully coupled, discrete system at each time step. The fixed-stress splitting scheme is an iterative splitting scheme to solve this system. Let i denote the iteration index, and look for a pair $(\mathbf{u}^{\mathbf{i}}, p^i)$ to converge to the solution $(\mathbf{u}, p)$, when $i \to +\infty$. Algorithmically, one first solves the flow equation (2) using the displacement from the previous iteration, and then, one solves the mechanics equation (1) with the updated pressure and iterates until convergence is achieved. To ensure convergence,[5,21,24] one needs to add a stabilizing term $L(p^i - p^{i-1})$ to the flow equation (2). The free-to-be-chosen parameter $L \geq 0$ is called the stabilization or tuning parameter. Choosing the value of this parameter is of major importance to the performance of the algorithm, because the number of iterations strongly depends on its value (see the works of Both et al,[5] Bause et al,[16] Both and Köcher,[27] Mikelić et al,[28] and Dana et al[29]). Moreover, a too small or too big L will lead to slow or no convergence.

The initial derivation of the fixed-stress splitting scheme had a physical motivation[21,23]: one "fixes the (volumetric) stress," ie, imposes $K_{\mathrm{dr}}\nabla\cdot\mathbf{u}^i - \alpha p^i = K_{\mathrm{dr}}\nabla\cdot\mathbf{u}^{i-1} - \alpha p^{i-1}$ and uses this to replace $\alpha\nabla\cdot\mathbf{u}^i$ in the flow equation. Here, K_{dr} is the physical drained bulk modulus. The resulting stabilization parameter L, called from now on the *physical* stabilization parameter, is $L_{\mathrm{phys}} = \frac{\alpha^2}{K_{\mathrm{dr}}}$ (depending on the mechanical properties and the Biot coefficient). In 2013, a rigorous mathematical analysis of the fixed-stress splitting scheme was performed for the first time in the work of Mikelić and Wheeler.[24] The authors show that the scheme is a contraction for any stabilization parameter $L \geq \frac{L_{\mathrm{phys}}}{2}$. This analysis was confirmed in the work of Both et al[5] for heterogeneous media using a simpler technique, and the same result was obtained for both continuous and discontinuous Galerkin higher-order space-time finite elements in the works of Bause et al[16] and Bause,[30] implying that the value of the stabilization parameter does not depend on the order of the spatial discretization. The question of which stabilization parameter is the optimal one (in the sense that it requires the least number of iterations to converge) arises, and the aim of this paper is to answer this open question.

In a recent study,[27] the authors studied the convergence of the fixed-stress splitting scheme for different test cases with varying material parameters. They determined numerically the optimal stabilization parameter for each considered case. This study, together with the previous results presented in the works of Mikelić et al[28] and Both et al,[5] suggests that the optimal parameter actually is a value in the interval $[\frac{L_{\mathrm{phys}}}{2}, L_{\mathrm{phys}}]$, depending on the data. In particular, the optimal parameter depends on the problem's boundary conditions and flow parameters, and not only on its mechanical properties and coupling coefficient. Nevertheless, to the best of our knowledge, there exists no theoretical evidence for this in the literature so far.

In this paper, we propose for the first time that the optimal stabilization parameter for the fixed-stress splitting scheme lies in the interval $[\frac{\alpha^2}{4\mu+2\lambda}, \frac{\alpha^2}{K_{\mathrm{dr}}}) \supseteq [\frac{L_{\mathrm{phys}}}{2}, L_{\mathrm{phys}})$ and depends also on the fluid flow properties and stability properties of the spatial discretization. This is achieved through refining the proof techniques in the work of Both et al[5] to obtain an improved linear rate of convergence; minimizing this rate with respect to the stabilization parameter gives the "theoretical" optimal choice. Although the trends for the practical and the proposed theoretically optimal stabilization parameter are sound for varying material parameters, the theoretically calculated one does not show great practical promise in terms of being optimal (see the work of Storvik et al[31] for a supplementary numerical study). This is due to harsh bounds that have been used in the proof. Therefore, we propose a brute-force approach for optimizing the stabilization parameter, utilizing the newly found interval $[\frac{\alpha^2}{4\mu+2\lambda}, \frac{\alpha^2}{K_{\mathrm{dr}}})$.

In contrast to previous works, the spatial discretization is required to be inf-sup stable, which essentially allows for the control of errors in the pressure by those in the stress. A novel consequence of our theoretical result is that under the use of an inf-sup–stable discretization, the fixed-stress splitting scheme also converges robustly in the limit case of incompressible fluids and impermeable porous media.

In Section 4, numerical experiments are performed, which show the soundness and efficiency of the proposed optimization technique. In particular, we show that the optimized stabilization parameter can be far superior to a naive choice among the classical stabilization parameters, L_{phys} or $\frac{L_{\mathrm{phys}}}{2}$.

To summarize, the main contributions of this work are as follows:

- an improved, theoretical convergence result for the fixed-stress splitting scheme under the assumption of an inf-sup–stable discretization;
- the derivation of an explicit interval for the optimal stabilization parameter, depending solely on the material parameters;
- a brute-force approach for optimizing the stabilization parameter, relying on a nearly mesh-independent performance of the fixed-stress splitting.

We mention that the fixed-stress splitting scheme also can be applied to more involved extensions of Biot's equations, for example, including nonlinear water compressibility,[32] unsaturated poroelasticity,[33,34] the multiple-network poroelasticity theory,[35,36] finite-strain poroplasticity,[37] fractured porous media,[38] and fracture propagation.[39,40] For nonlinear problems, one combines a linearization technique, eg, the L-scheme,[41,42] with the splitting algorithm; the convergence of the resulting scheme can be proved rigorously.[32,33] Finally, we would like to mention some valuable variants of the fixed-stress splitting scheme: the multirate fixed-stress method,[43] the multiscale fixed-stress method,[29] and the parallel-in-time fixed-stress method.[44]

This paper is structured as follows. The notation, the discretization, and the fixed-stress splitting scheme are presented in Section 2. The theoretical analysis of the convergence and the optimization technique are the subject of Section 3. In Section 4, numerical experiments that test the optimization technique are presented. Finally, conclusions are given in Section 5.

2 | THE NUMERICAL SCHEME FOR SOLVING BIOT'S MODEL

In this paper, we use common notations in functional analysis. Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain where d is the spatial dimension. The space $L^2(\Omega)$ is the Hilbert space of Lebesgue-measurable, square-integrable functions on Ω, and $H^1(\Omega)$ is the Hilbert space of functions in $L^2(\Omega)$ with derivatives (in the weak sense) in $L^2(\Omega)$. The inner product and its associated norm in $L^2(\Omega)$ are denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively, and $\|\cdot\|_{H^1(\Omega)}$ is the standard $H^1(\Omega)$-norm. Vectors and tensors are written bold, and, sometimes, the scalar product and the norm will be taken for vectors and tensors. Vectorial functions are written bold-italic. T will denote the final time.

Biot equations (1) and (2) are solved in the domain $\Omega \times (0, T)$ together with (for simplicity) homogeneous Dirichlet boundary conditions and a given initial condition. In time, the backward Euler method is applied with a constant time-step size $\tau := \frac{T}{N}, N \in \mathbb{N}$. Throughout this work, the index n will refer to the time level. For the spatial discretization, a two-field Galerkin finite element formulation is considered, and two generic discrete spaces $\mathbf{V}_h$ and Q_h, associated with displacements and pressures, are introduced. Later, we require $\mathbf{V}_h \times Q_h$ to be inf-sup stable with respect to the divergence operator; the most prominent inf-sup–stable example is the Taylor-Hood element, ie, P2-P1 for displacement and pressure.[45] Nevertheless, the analysis below can be extended without difficulties to a three-field formulation as, for example, in the works of Phillips and Wheeler,[3] Both et al,[5] and Berger et al.[6]

In this way, the fully discrete, weak problem reads: let $n \geq 1$ and assume $(\boldsymbol{u}_h^{n-1}, p_h^{n-1}) \in \mathbf{V}_h \times Q_h$ are given. Find $(\boldsymbol{u}_h^n, p_h^n) \in \mathbf{V}_h \times Q_h$ such that

$$2\mu \left\langle \varepsilon\left(\boldsymbol{u}_h^n\right), \varepsilon\left(\boldsymbol{v}_h\right)\right\rangle + \lambda \left\langle \nabla \cdot \boldsymbol{u}_h^n, \nabla \cdot \boldsymbol{v}_h \right\rangle - \alpha \left\langle p_h^n, \nabla \cdot \boldsymbol{v}_h \right\rangle = \left\langle \boldsymbol{f}^n, \boldsymbol{v}_h \right\rangle, \tag{3}$$

$$\frac{1}{M}\left\langle p_h^n - p_h^{n-1}, q_h \right\rangle + \alpha \left\langle \nabla \cdot \left(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right), q_h \right\rangle + \tau \left\langle \kappa \nabla p_h^n, \nabla q_h \right\rangle - \tau \left\langle \kappa \mathbf{g}\rho, \nabla q_h \right\rangle = \tau \left\langle S_f^n, q_h \right\rangle \tag{4}$$

for all $\boldsymbol{v}_h \in \mathbf{V}_h, q_h \in Q_h$. For $n = 1$, the functions $(\boldsymbol{u}_h^{n-1}, p_h^{n-1})$ are obtained by using the initial condition.

The fixed-stress splitting scheme[5,21,23,28] is now introduced. Denote by i the iteration index. Iterate until convergence.

For $i \geq 1$, given a stabilization parameter $L \geq 0$ and $(\boldsymbol{u}_h^{n-1}, p_h^{n-1}), (\boldsymbol{u}_h^{n,i-1}, p_h^{n,i-1}) \in \mathbf{V}_h \times Q_h$, find $(\boldsymbol{u}_h^{n,i}, p_h^{n,i}) \in \mathbf{V}_h \times Q_h$ such that

$$2\mu \left\langle \varepsilon\left(\boldsymbol{u}_h^{n,i}\right), \varepsilon\left(\boldsymbol{v}_h\right)\right\rangle + \lambda \left\langle \nabla \cdot \boldsymbol{u}_h^{n,i}, \nabla \cdot \boldsymbol{v}_h \right\rangle - \alpha \left\langle p_h^{n,i}, \nabla \cdot \boldsymbol{v}_h \right\rangle = \left\langle \boldsymbol{f}^n, \boldsymbol{v}_h \right\rangle, \tag{5}$$

$$\begin{aligned} \frac{1}{M}\left\langle p_h^{n,i} - p_h^{n-1}, q_h \right\rangle + \alpha \left\langle \nabla \cdot \left(\boldsymbol{u}_h^{n,i-1} - \boldsymbol{u}_h^{n-1}\right), q_h \right\rangle + L \left\langle p_h^{n,i} - p_h^{n,i-1}, q_h \right\rangle \\ + \tau \left\langle \kappa \nabla p_h^{n,i}, \nabla q_h \right\rangle - \tau \left\langle \kappa \mathbf{g}\rho, \nabla q_h \right\rangle = \tau \left\langle S_f^n, q_h \right\rangle \end{aligned} \tag{6}$$

for all $\boldsymbol{v}_h \in \mathbf{V}_h, q_h \in Q_h$. The initial guess for the iterations is chosen to be the solution at the last time step, ie, $(\boldsymbol{u}_h^{n,0}, p_h^{n,0}) := (\boldsymbol{u}_h^{n-1}, p_h^{n-1})$. Notice that the mechanics and flow problems decouple, allowing for the use of separate simulators for both subproblems.

3 | CONVERGENCE ANALYSIS AND OPTIMIZATION

In this section, the convergence of the scheme (5)-(6) is analyzed. We are particularly interested in finding an *optimal* stabilization parameter L, in the sense that the scheme requires the least amount of iterations, ie, has the smallest possible convergence rate. Before we proceed with the main result, we need some preliminaries.

Definition 1. The mathematical bulk modulus, $K_{\mathrm{dr}}^{\star} > 0$, is defined as the largest constant such that

$$2\mu\|\varepsilon(\boldsymbol{u}_h)\|^2 + \lambda\|\nabla \cdot \boldsymbol{u}_h\|^2 \geq K_{\mathrm{dr}}^{\star}\|\nabla \cdot \boldsymbol{u}_h\|^2 \qquad \text{for all } \boldsymbol{u}_h \in \mathbf{V}_h. \tag{7}$$

By the Cauchy-Schwarz inequality, we get that the physical drained bulk modulus $K_{\mathrm{dr}} = \frac{2\mu}{d} + \lambda$ is a lower bound for $K_{\mathrm{dr}}^{\star}$. However, for effectively lower-dimensional situations, eg, a one-dimensional–like compression, d can be replaced by a value closer to 1. Lemma 1 below guarantees an upper bound for $K_{\mathrm{dr}}^{\star}$. Nevertheless, there is a strong indication (based on numerical experiments; see, eg, Section 4 and the work of Both and Köcher[27]) that $K_{\mathrm{dr}}^{\star} \in [K_{\mathrm{dr}} = \frac{2\mu}{d} + \lambda, 2\mu + \lambda]$. We remark that the exact value, depending on the physical situation, can be computed as a generalized eigenvalue.

Throughout this paper, we make use of the following two assumptions.

Assumption 1. The constants μ, λ, α, and ρ are strictly positive, the constants $1/M$ and κ are nonnegative, and the vector $\boldsymbol{g}$ is constant.

Assumption 2. The discretization $\boldsymbol{V}_h \times Q_h$ is inf-sup stable with respect to the bilinear form $b(\boldsymbol{v}_h, q_h) = \langle \nabla \cdot \boldsymbol{v}_h, q_h \rangle$.

From Assumption 2 follows Lemma 1 by applying corollary 4.1.1 in the work of Boffi et al,[45] which states as follows.

Corollary 1. *Let V and Q be Hilbert spaces, and let B be a linear continuous operator from V to Q′; here, Q′ denotes the dual space of Q. Denote by B^t the transposed operator of B. Then, the following two statements are equivalent.*

- *B^t is bounding: $\exists \gamma > 0$ such that $\|B^t q\|_{V'} \geq \gamma \|q\|_Q \; \forall q \in Q$.*
- *$\exists L_B \in \mathcal{L}(Q', V)$ such that $B(L_B(\xi)) = \xi \; \forall \xi \in Q'$ with $\|L_b\| = \frac{1}{\gamma}$.*

Lemma 1. *Assume Assumption 2. There exists $\beta > 0$ such that, for any $p_h \in Q_h$, there exists $\boldsymbol{u}_h \in \boldsymbol{V}_h$ satisfying $\langle \nabla \cdot \boldsymbol{u}_h, q_h \rangle = \langle p_h, q_h \rangle$ for all $q_h \in Q_h$ and*

$$2\mu\|\boldsymbol{\varepsilon}(\boldsymbol{u}_h)\|^2 + \lambda\|\nabla \cdot \boldsymbol{u}_h\|^2 \leq \beta\|p_h\|^2. \tag{8}$$

Proof. Consider Corollary 1. Let the continuous linear function $B : \mathbf{V}_h \to Q_h'$ be defined by $B(\boldsymbol{u}_h)(q_h) = \langle \nabla \cdot \boldsymbol{u}_h, q_h \rangle$. The first statement of Corollary 1 is a characterization of an inf-sup–stable discretization Assumption 2, with inf-sup constant γ. Hence, the second statement of Corollary 1 holds; there exists a linear function $L_B \in \mathcal{L}(Q_h', \boldsymbol{V}_h)$ such that $B(L_B(\langle p_h, \cdot \rangle)) = \langle p_h, \cdot \rangle$ for all $p_h \in Q_h$ with $\|L_B\| = 1/\gamma$. In particular, L_B is mapping $p_h \in Q_h$ to the corresponding $\boldsymbol{u}_h \in \mathbf{V}_h$ such that

$$\langle \nabla \cdot \boldsymbol{u}_h, q_h \rangle = B\left(L_B\left(\langle p_h, \cdot \rangle\right)\right)(q_h) = \langle p_h, q_h \rangle$$

for all $q_h \in Q_h$. Additionally, the following chain of inequalities holds true:

$$2\mu \|\varepsilon(\boldsymbol{u}_h)\|^2 + \lambda \|\nabla \cdot \boldsymbol{u}_h\|^2 \leq C\, \|\boldsymbol{u}_h\|_{H^1(\Omega)}^2 \leq C \|L_B\|^2 \|p_h\|^2,$$

where the first inequality follows from Young's inequality with C depending only on the Lamé parameters, and the second inequality results from the operator norm, ie,

$$\|L_B\| = \sup_{0 \neq p_h \in Q_h} \frac{\|L_B\left(\langle p_h, \cdot \rangle\right)\|_{H^1(\Omega)}}{\|\langle p_h, \cdot \rangle\|_{L^2(\Omega)'}} = \sup_{\substack{0 \neq p_h \in Q_h \\ \boldsymbol{u}_h = L_B(\langle p_h, \cdot \rangle)}} \frac{\|\boldsymbol{u}_h\|_{H^1(\Omega)}}{\|p_h\|}.$$

We obtain our desired inequality, as follows:

$$2\mu \|\varepsilon(\boldsymbol{u}_h)\|^2 + \lambda \|\nabla \cdot \boldsymbol{u}_h\|^2 \leq \frac{C}{\gamma^2} \|p_h\|^2 = \beta \|p_h\|^2.$$

□

Remark 1. The constant β above depends on μ, λ, and the domain Ω and on the choice of the finite-dimensional spaces $\mathbf{V}_h$ and Q_h. Similar to $K_{\text{dr}}^{\star}$, β can be computed as a generalized eigenvalue.

We can now give our main convergence result.

Theorem 1. *Assume that Assumptions 1 and 2 hold true, and let $\delta \in (0, 2]$. Define the iteration errors as $\boldsymbol{e}_{\boldsymbol{u}}^{n,i} := \boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^n$ and $e_p^{n,i} := p_h^{n,i} - p_h^n$, where $(\boldsymbol{u}_h^{n,i}, p_h^{n,i})$ is a solution to (5) and (6), and $(\boldsymbol{u}_h^n, p_h^n)$ is a solution to (3) and (4). The fixed-stress splitting scheme (5)-(6) converges linearly for any $L \geq \frac{\alpha^2}{\delta K_{\text{dr}}^{\star}}$, with a convergence rate given by*

$$rate\,(L, \delta) = \frac{L}{L + \frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2 - \delta)\, \frac{\alpha^2}{\beta}}, \tag{9}$$

through the error inequalities

$$\left\| e_p^{n,i} \right\|^2 \leq rate\,(L, \delta) \left\| e_p^{n,i-1} \right\|^2, \tag{10}$$

$$2\mu \left\| \varepsilon\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 \leq \frac{\alpha^2}{K_{\text{dr}}^{\star}} \left\| e_p^{n,i} \right\|^2, \tag{11}$$

where C_Ω is the Poincaré constant and β is the constant from (8).

Proof. Subtract (5) and (6) from (3) and (4), respectively, to obtain the error equations

$$\begin{cases} (i)\;\; 2\mu \left\langle \varepsilon\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right), \varepsilon\left(\boldsymbol{v}_h\right) \right\rangle + \lambda \left\langle \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \nabla \cdot \boldsymbol{v}_h \right\rangle - \alpha \left\langle e_p^{n,i}, \nabla \cdot \boldsymbol{v}_h \right\rangle = 0, \\ (ii)\;\; \frac{1}{M} \left\langle e_p^{n,i}, q_h \right\rangle + \alpha \left\langle \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}, q_h \right\rangle + L \left\langle e_p^{n,i} - e_p^{n,i-1}, q_h \right\rangle + \tau \left\langle \kappa \nabla e_p^{n,i}, \nabla q_h \right\rangle = 0, \end{cases} \tag{12}$$

holding for all $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h \times Q_h$. To prove (11), test (12)(i) with $\boldsymbol{v}_h = \boldsymbol{e}_{\boldsymbol{u}}^{n,i}$, and apply the Cauchy-Schwarz inequality and Young's inequality to the pressure term to obtain

$$2\mu \left\| \varepsilon\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 \leq \frac{\alpha^2}{2K_{\text{dr}}^{\star}} \left\| e_p^{n,i} \right\|^2 + \frac{K_{\text{dr}}^{\star}}{2} \left\| \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2. \tag{13}$$

We now get (11) by applying (7).

In order to prove (10), test (12) with $q_h = e_p^{n,i}$ and $\boldsymbol{v}_h = \boldsymbol{e}_{\boldsymbol{u}}^{n,i}$, add the resulting equations, and use the algebraic identity

$$\left\langle e_p^{n,i} - e_p^{n,i-1}, e_p^{n,i} \right\rangle = \frac{1}{2}\left(\left\| e_p^{n,i} - e_p^{n,i-1} \right\|^2 + \left\| e_p^{n,i} \right\|^2 - \left\| e_p^{n,i-1} \right\|^2 \right)$$

to get

$$\begin{aligned} &2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla\cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 + \frac{1}{M}\left\| e_p^{n,i} \right\|^2 - \alpha \left\langle e_p^{n,i}, \nabla\cdot\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\rangle + \tau\kappa \left\| \nabla e_p^{n,i} \right\|^2 + \frac{L}{2}\left\| e_p^{n,i} - e_p^{n,i-1} \right\|^2 + \frac{L}{2}\left\| e_p^{n,i} \right\|^2 \\ &\quad = \frac{L}{2}\left\| e_p^{n,i-1} \right\|^2. \end{aligned}$$

Using now Equation (12)(i), tested with $\boldsymbol{v}_h = \boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}$ in the above, yields

$$\begin{aligned} &2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla\cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 + \frac{1}{M}\left\| e_p^{n,i} \right\|^2 + \tau\kappa \left\| \nabla e_p^{n,i} \right\|^2 + \frac{L}{2}\left\| e_p^{n,i} \right\|^2 + \frac{L}{2}\left\| e_p^{n,i} - e_p^{n,i-1} \right\|^2 \\ &\quad = \frac{L}{2}\left\| e_p^{n,i-1} \right\|^2 + 2\mu \left\langle \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right), \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\rangle + \lambda \left\langle \nabla\cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \nabla\cdot\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\rangle. \end{aligned} \tag{14}$$

By applying Young's inequality in (14), we obtain that, for any $\delta > 0$, there holds

$$\begin{aligned} &2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla\cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 + \frac{1}{M}\left\| e_p^{n,i} \right\|^2 + \tau\kappa \left\| \nabla e_p^{n,i} \right\|^2 + \frac{L}{2}\left\| e_p^{n,i} \right\|^2 + \frac{L}{2}\left\| e_p^{n,i} - e_p^{n,i-1} \right\|^2 \\ &\quad = \frac{L}{2}\left\| e_p^{n,i-1} \right\|^2 + \frac{\delta}{2}\left(2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla\cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 \right) + \frac{1}{2\delta}\left(2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\|^2 + \lambda \left\| \nabla\cdot\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\|^2 \right). \end{aligned} \tag{15}$$

To take care of the last term in (15), consider Equation (12)(i), subtract iteration $i-1$ from iteration i, let $\boldsymbol{v}_h = \boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}$ in the result, and apply the Cauchy-Schwarz inequality to get

$$2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) - \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\|^2 + \lambda \left\| \nabla\cdot\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\|^2 \leq \alpha \left\| e_p^{n,i} - e_p^{n,i-1} \right\| \left\| \nabla\cdot\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\|. \tag{16}$$

By using (7), (16) implies

$$K_{\mathrm{dr}}^{\star} \left\| \nabla\cdot\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\| \leq \alpha \left\| e_p^{n,i} - e_p^{n,i-1} \right\|. \tag{17}$$

Inserting (17) into (16) yields

$$2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) - \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\|^2 + \lambda \left\| \nabla\cdot\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}\right) \right\|^2 \leq \frac{\alpha^2}{K_{\mathrm{dr}}^{\star}} \left\| e_p^{n,i} - e_p^{n,i-1} \right\|^2. \tag{18}$$

By rearranging terms and inserting (18) into (15), we immediately get

$$\begin{aligned} &\left(1 - \frac{\delta}{2}\right)\left(2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla\cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 \right) + \frac{1}{M}\left\| e_p^{n,i} \right\|^2 + \tau\kappa \left\| \nabla e_p^{n,i} \right\|^2 + \frac{L}{2}\left\| e_p^{n,i} \right\|^2 + \frac{L}{2}\left\| e_p^{n,i} - e_p^{n,i-1} \right\|^2 \\ &\quad \leq \frac{L}{2}\left\| e_p^{n,i-1} \right\|^2 + \frac{\alpha^2}{2\delta K_{\mathrm{dr}}^{\star}}\left\| e_p^{n,i} - e_p^{n,i-1} \right\|^2. \end{aligned}$$

Using that $L \geq \frac{\alpha^2}{\delta K_{\mathrm{dr}}^{\star}}$ and the Poincaré inequality, we obtain from the above

$$\left(1 - \frac{\delta}{2}\right)\left(2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla\cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 \right) + \left(\frac{1}{M} + \frac{L}{2} + \frac{\tau\kappa}{C_\Omega^2} \right)\left\| e_p^{n,i} \right\|^2 \leq \frac{L}{2}\left\| e_p^{n,i-1} \right\|^2. \tag{19}$$

The result, (19), already implies that we have convergence of the scheme. In previous works, particularly that of Both et al[5] (where the proof so far is very similar), the conclusion at this point is that $L = \frac{\alpha^2}{2K_{\mathrm{dr}}^\star}$ is the optimal parameter. However, this does not consider the influence of the first term in (19). By Lemma 1, we get that there exists $\boldsymbol{v}_h \in V_h$ such that $e_p^{n,i} = \nabla \cdot \boldsymbol{v}_h$ in a weak sense and

$$2\mu\|\boldsymbol{\varepsilon}(\boldsymbol{v}_h)\|^2 + \lambda\|\nabla \cdot \boldsymbol{v}_h\|^2 \le \beta\left\|e_p^{n,i}\right\|^2. \tag{20}$$

By testing now (12)(i) with this $\boldsymbol{v}_h$, we get

$$\alpha\left\|e_p^{n,i}\right\|^2 = 2\mu\left\langle \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\right\rangle + \lambda\left\langle \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \nabla \cdot \boldsymbol{v}_h\right\rangle. \tag{21}$$

From (20) and (21) and the Cauchy-Schwarz inequality, we immediately obtain

$$\frac{\alpha^2}{\beta}\left\|e_p^{n,i}\right\|^2 \le 2\mu\left\|\boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right)\right\|^2 + \lambda\left\|\nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right\|^2, \tag{22}$$

which, together with (19), implies

$$\left(\frac{1}{M} + \frac{L}{2} + \frac{\tau\kappa}{C_\Omega^2} + \left(1 - \frac{\delta}{2}\right)\frac{\alpha^2}{\beta}\right)\left\|e_p^{n,i}\right\|^2 \le \frac{L}{2}\left\|e_p^{n,i-1}\right\|^2.$$

This gives the following rate of convergence, for $\delta \in (0,2]$ and $L \ge \frac{\alpha^2}{\delta K_{\mathrm{dr}}^\star}$:

$$\mathrm{rate}(L,\delta) = \frac{L}{L + \frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2-\delta)\frac{\alpha^2}{\beta}}.$$

□

Remark 2. Assumptions 1 and 2 are valid in various relevant physical situations. Therefore, our analysis has a wide range of applications. One can easily extend the result to heterogeneous media, ie, $\kappa = \kappa(\mathbf{x})$ as long as κ is bounded from below by $\kappa_m \ge 0$. Moreover, any of the other parameters can be chosen spatially dependent as long as they are bounded from below by appropriate constants satisfying Assumption 1.

3.1 | Optimality

Consider the rate obtained in (9). As rate(L,δ) is an increasing function of L, it follows that, for all $\delta \in (0,2]$, its minimum is obtained at $L = \frac{\alpha^2}{\delta K_{\mathrm{dr}}^\star}$, giving the rate

$$\mathrm{rate}(\delta) = \frac{\frac{\alpha^2}{K_{\mathrm{dr}}^\star}}{\frac{\alpha^2}{K_{\mathrm{dr}}^\star} + \delta\left(\frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2-\delta)\frac{\alpha^2}{\beta}\right)}. \tag{23}$$

Minimizing (23) with respect to δ corresponds to maximizing

$$\delta\left(\frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2-\delta)\frac{\alpha^2}{\beta}\right).$$

Let $A := \frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + 2\frac{\alpha^2}{\beta}$ and $B := \frac{\alpha^2}{\beta}$. It is easily seen that the maximum of $\delta(A - \delta B)$ is attained at $\delta = \frac{A}{2B}$. Therefore, the minimizer of rate(δ) is

$$\delta = \min\left\{\frac{A}{2B}, 2\right\} \in (1,2], \tag{24}$$

since $A \geq 2B$. This suggests that the theoretical optimal choice of L is

$$L = \frac{\alpha^2}{K_{\mathrm{dr}}^{\star} \min\left\{\frac{A}{2B}, 2\right\}} \in \left[\frac{\alpha^2}{2K_{\mathrm{dr}}^{\star}}, \frac{\alpha^2}{K_{\mathrm{dr}}^{\star}}\right) \subset \left[\frac{\alpha^2}{4\mu + 2\lambda}, \frac{\alpha^2}{\frac{2\mu}{d} + \lambda}\right). \tag{25}$$

Remark 3 (Consequence for low-compressible fluids and low-permeable porous media**).**
Previous convergence results in the literature for the fixed-stress splitting scheme have not predicted or guaranteed any robust convergence in the limit cases $M \to \infty$ and $\kappa \to 0$ (for a fixed time-step size τ). Now, by Theorem 1, for inf-sup–stable discretizations, robust convergence of the fixed-stress splitting scheme is guaranteed, even in the limit case. This was studied numerically in the work of Storvik et al.[31] Convergence was showed to be robust with respect to material parameters for P2-P1 elements and deteriorating for P1-P1.

3.2 | Brute-force optimization of the stabilization parameter

The rate obtained in Theorem 1 is not necessarily sharp, and it is rather viewed as theoretical evidence that the optimal stabilization parameter resides in the interval $[\frac{\alpha^2}{4\mu+2\lambda}, \frac{\alpha^2}{\frac{2\mu}{d}+\lambda})$. Additionally, convergence is predicted to be robust with respect to the mesh size. It can be, indeed, verified numerically that the performance of the fixed-stress splitting scheme is nearly mesh independent (see, for instance, the numerical examples in Section 4 or in the work of Adler et al[46]). Based on that, we propose the following brute-force search for optimizing the stabilization parameter for a fixed test case: test the fixed-stress splitting scheme using different stabilization parameters in the interval $[\frac{\alpha^2}{4\mu+2\lambda}, \frac{\alpha^2}{\frac{2\mu}{d}+\lambda})$ for a coarse mesh and a single time step. Choose the parameter that gives the fewest number of iterations, and employ it for any arbitrary mesh. Section 4 shows the effectiveness of the proposed method.

4 | NUMERICAL EXAMPLES

In this section, we demonstrate the effectiveness of the proposed brute-force method for optimizing the stabilization parameter for the fixed-stress splitting scheme. In particular, we show for several numerical test cases that the optimal stabilization parameter is close to being mesh independent and that the method for choosing it optimally, as described in Section 3.2, indeed yields a preferable alternative to the classical choices of $L = \frac{\alpha^2}{2K_{\mathrm{dr}}}$ and $L = \frac{\alpha^2}{K_{\mathrm{dr}}}$.

We consider four different test cases, as follows:

1. a unit square domain;
2. an L-shaped domain;
3. Mandel's problem;
4. three-dimensional (3D) footing problem on the unit cube.

For the implementation of the numerical examples, we use modules from the DUNE project,[47] particularly dune-functions.[48,49] If not mentioned otherwise, the inf-sup–stable Taylor-Hood pair P2-P1 is utilized as spatial discretization. As stopping criteria, we have applied relative L_2-norms for the pressure, ie, iterations stop when $\|p_h^i - p_h^{i-1}\| \leq \epsilon_r \|p_h^{i-1}\|$, consistent with Theorem 1. Constant material and fluid parameters are applied and given for each individual test case.

4.1 | Notations

During the numerical experiments, we apply some specific choices of stabilization parameters several times. Therefore, we give them names here. Recall the definition of the physical drained bulk modulus $K_{\mathrm{dr}} = \frac{2\mu}{d} + \lambda$. The original stabilization parameter will be called the physical one due to the fixed-stress splitting scheme's physical origin, ie, $L_{\mathrm{phys}} = \frac{\alpha^2}{K_{\mathrm{dr}}}$. The other classical choice of stabilization parameter will be named after Mikelić and Wheeler due to their paper,[24] ie, $L_{\mathrm{MW}} = \frac{L_{\mathrm{phys}}}{2} = \frac{\alpha^2}{2K_{\mathrm{dr}}}$. The stabilization parameter obtained by the brute-force method described in Section 3.2 will be called L_{opt}. The final parameter is the one that is proposed to be the smallest possible choice in Section 3.1, ie, $L_{\min} = \frac{\alpha^2}{4\mu+2\lambda}$ (see Table 1).

Name	L_{phys}	L_{MW}	L_{opt}	L_{min}
Value	$\frac{\alpha^2}{K_{\text{dr}}}$	$\frac{\alpha^2}{2K_{\text{dr}}}$	Section 3.2	$\frac{\alpha^2}{4\mu+2\lambda}$

TABLE 1 Names of specific stabilization parameters

Name	Symbol	Value	Unit
Shear modulus	μ	$41.667 \cdot 10^9$	Pa
First Lamé parameter	λ	$27.778 \cdot 10^9$	Pa
Permeability	κ	10^{-13}	m^2
Compressibility	$\frac{1}{M}$	10^{-11}	Pa^{-1}
Initial time	t_0	0	s
Time-step size	τ	0.1	s
Stop time	T	1	s
Biot-Willis coefficient	α	1	–
Relative error tolerance	ϵ_r	10^{-6}	–
Inverse of mesh size [a]	$1/h$	16, 32, 64, 128, 512	m^{-1}

[a] Mesh sizes are only used in Section 4.2.

TABLE 2 Parameters used in Sections 4.2 and 4.5

4.2 | Dependence on boundary conditions—the unit square

We consider two test cases differing solely in the applied boundary conditions. Common for both, the domain is the unit square discretized by structured triangles, and the constant material parameters from Table 2 are considered. Moreover, we employ source terms corresponding to the analytical solution

$$u_1(x,y,t) = u_2(x,y,t) = \frac{1}{p_{\text{ref}}} p(x,y,t) = txy(1-x)(1-y), \qquad (x,y) \in (0,1)^2, \quad t \in (0,1),$$

of the continuous problem (1)-(2). The pressure, p, is scaled by $p_{\text{ref}} = 10^{11}$ Pa in order to balance the magnitude of the mechanical and fluid stresses for the chosen physical parameters. Regarding the different sets of boundary conditions, we consider the following.

- BC1: homogeneous Dirichlet data on the entire boundary for displacement and pressure.
- BC2: homogeneous Dirichlet data for the pressure; homogeneous Neumann data on top in the mechanics equation and homogeneous Dirichlet data everywhere else for the displacement.

Solutions after 10 time steps using a mesh size of $h = 1/128$ are displayed in Figures 1 and 2.

To motivate the brute-force approach from Section 3.2, the performance of the fixed-stress splitting scheme has been measured for a variety of stabilization parameters and mesh sizes (see Figure 3). We observe that the numbers of iterations vary significantly for different stabilization parameters but that the optimal choice is within our proposed interval $[L_{\text{min}}, L_{\text{phys}})$. Additionally, for fixed stabilization parameters, we observe that the numbers of iterations are close to constant with respect to the mesh size.

Now, we test the brute-force approach of Section 3.2. In order to calculate L_{opt}, we start by applying the fixed-stress splitting scheme for 11 equidistant stabilization parameters in $[L_{\text{min}}, L_{\text{phys}}]$ while only computing one time step for a mesh

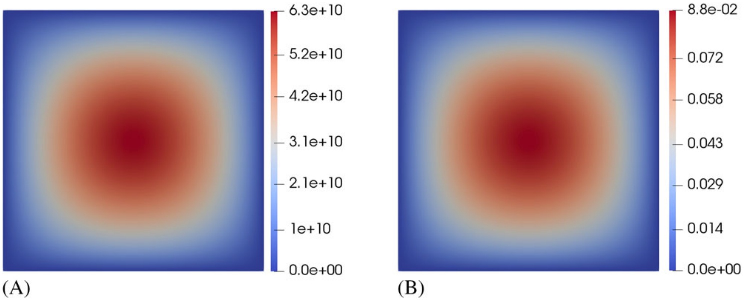


FIGURE 1 Unit square test case: solution—BC1. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$) [Colour figure can be viewed at wileyonlinelibrary.com]

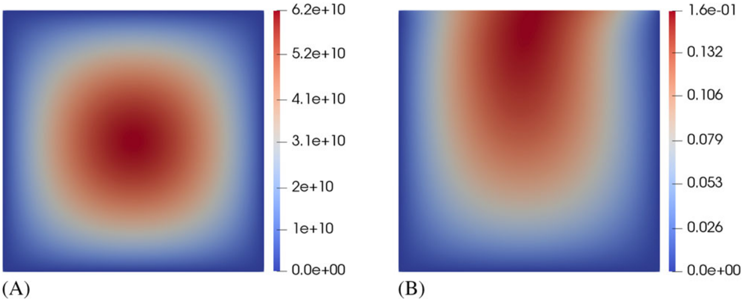


FIGURE 2 Unit square test case: solution—BC2. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$) [Colour figure can be viewed at wileyonlinelibrary.com]

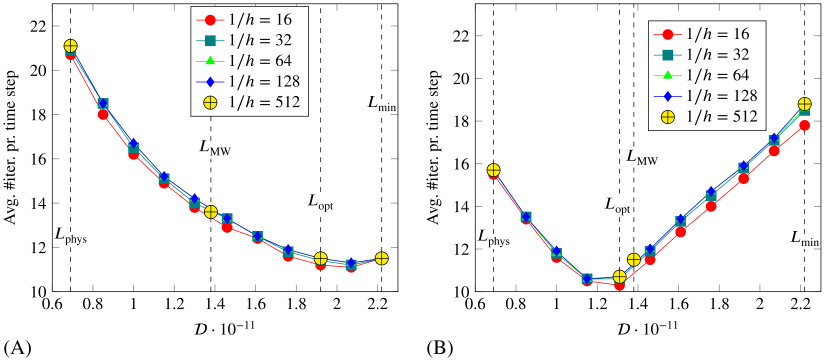


FIGURE 3 Unit square test case: average number of iterations per time step for different stabilization parameters, $L = \frac{\alpha^2}{D}$, using parameters from Table 2. The largest value of D corresponds to $L_{\min}$, whereas the smallest value of D corresponds to L_{phys}. Recall that L_{opt} is calculated using only one time step, and therefore, there is a slight deviation between L_{opt} and the actual optimal choice. A, BC1; B, BC2 [Colour figure can be viewed at wileyonlinelibrary.com]

size of $h = 1/16$. Then, using the stabilization parameter that needed the least amount of iterations to converge, we apply the fixed-stress splitting scheme for the full problem using a mesh size of $h = 1/512$. In Figure 3, the average numbers of iterations over 10 time steps are displayed for this “optimal” stabilization parameter, for the two classical choices L_{phys} and L_{MW}, and for the stabilization parameter that we consider to be the smallest possible choice, ie, $L_{\min}$. We see that the optimized stabilization parameter requires the least amount of iterations for both boundary conditions. It is also worth noticing that the optimal choice differs considerably for the two sets of boundary conditions.

4.3 | Dependence on Poisson's ratio—L-shaped domain

To further analyze the proposed brute-force optimization of the stabilization parameter for the fixed-stress splitting scheme, we test it on an L-shaped domain as well. The L-shaped domain is considered as a subdomain of the unit square domain where the top-right quarter square has been removed, ie, $L = [0,1]^2 \backslash (0.5,1]^2$. The material and implementation parameters from Table 3 are applied, whereas the right-hand side is the same as for the unit square test case. Zero Dirichlet boundary conditions are applied everywhere, but at the top boundary ($[0,0.5] \times \{1\}$) for the mechanics equation where zero Neumann conditions are considered. A solution to this problem after 10 time steps with $\nu = 0$ and mesh size $1/h = 128$ is given in Figure 4.

Given Young's modulus E and Poisson's ratio ν, the corresponding Lamé parameters have been determined by

$$\mu = \frac{E}{2(1+\nu)} \quad \text{and} \quad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}. \tag{26}$$

TABLE 3 Parameters used in Section 4.3

Name	Symbol	Value	Unit
Young's modulus	E	10^{11}	Pa
Poisson's ratio	ν	$0, 0.2, 0.4$	–
Permeability	κ	10^{-13}	m^2
Compressibility	$\frac{1}{M}$	10^{-11}	m^{-1}
Initial time	t_0	0	s
Time-step size	τ	0.1	s
Stop time	T	1	s
Biot-Willis coefficient	α	1	–
Relative error tolerance	ϵ_r	10^{-6}	–
Inverse of mesh size	$1/h$	16, 32, 64, 128, 512	m^{-1}

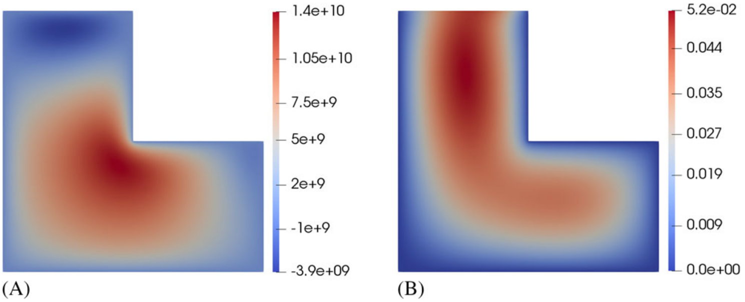


FIGURE 4 L-shaped domain test case: solution for $\nu = 0$. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$) [Colour figure can be viewed at wileyonlinelibrary.com]

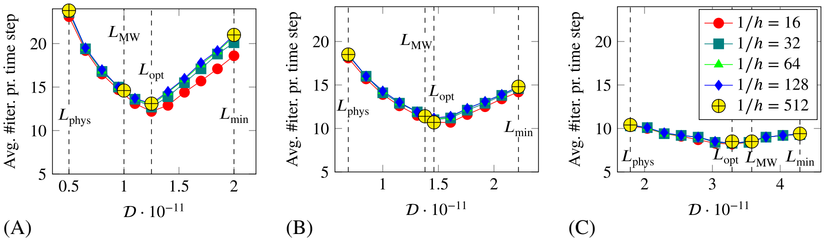


FIGURE 5 L-shaped domain test case: number of iterations for different stabilization parameters, $L = \frac{\alpha^2}{\mathcal{D}}$, using parameters from Table 3. The largest value of $\mathcal{D}$ corresponds to L_{min}, whereas the smallest value of $\mathcal{D}$ corresponds to L_{phys}. Notice that the axes are different. A, $\nu = 0$; B, $\nu = 0.2$; C, $\nu = 0.4$ [Colour figure can be viewed at wileyonlinelibrary.com]

Again, as for the unit square test case, we test the brute-force optimization technique that is described in Section 3.2, but now for three different Poisson's ratios. In Figure 5, the fixed-stress splitting scheme is applied to a variety of mesh sizes and with a variety of stabilization parameters to three problems with different Poisson's ratios. There are several key observations to make. First, the scheme is close to being mesh independent for all mesh sizes, stabilization parameters, and Poisson's ratios. Second, we see that the optimal stabilization parameter is in the proposed interval $[L_{\mathrm{min}}, L_{\mathrm{phys}})$ for all Poisson's ratios and all mesh sizes. The final observation is that when the Poisson's ratio increases, the choice of stabilization parameter becomes less important. This is due to the fact that an increase in the Poisson's ratio can be seen as an effective decrease in the coupling strength.

To calculate the optimal stabilization parameter, we follow the recipe of Section 3.2. We apply 11 equidistant stabilization parameters in the interval $[L_{\mathrm{min}}, L_{\mathrm{phys}}]$ for the fixed-stress splitting scheme on a coarse mesh $(1/h = 16)$ for only one time step. Counting the numbers of iterations it takes to reach convergence, we choose the parameter that corresponds to the smallest number and use this for the finer mesh $(1/h = 512)$ and more time steps (10). We see that the parameter that is the optimal choice for the coarse mesh is also the optimal one for the finer mesh for all Poisson's ratios.

4.4 | Mandel's problem

Here, we consider Mandel's problem, a relevant two-dimensional problem with a known analytical solution that is often used as a benchmark problem for discretizations. The analytical solution is derived in the works of Coussy[1] and Abousleiman et al,[50] and its expressions for pressure and displacement are given by

$$p = \frac{2FB(1+\nu_u)}{3a} \sum_{n=1}^{\infty} \frac{\sin(\alpha_n)}{\alpha_n - \sin(\alpha_n)\cos(\alpha_n)} \left(\cos\left(\frac{\alpha_n x}{a}\right) - \cos(\alpha_n) \right) e^{-\frac{\alpha_n^2 c_f t}{a^2}}, \tag{27}$$

$$\begin{aligned} u_x = {} & \left[\frac{F\nu}{2\mu a} - \frac{F\nu_u}{\mu a} \sum_{n=1}^{\infty} \frac{\sin(\alpha_n)\cos(\alpha_n)}{\alpha_n - \sin(\alpha_n)\cos(\alpha_n)} e^{-\frac{\alpha_n^2 c_f t}{a^2}} \right] x \\ & + \frac{F}{\mu} \sum_{n=1}^{\infty} \frac{\cos(\alpha_n)}{\alpha_n - \sin(\alpha_n)\cos(\alpha_n)} \sin\left(\frac{\alpha_n x}{a}\right) e^{-\frac{\alpha_n^2 c_f t}{a^2}}, \end{aligned} \tag{28}$$

$$u_y = \left[\frac{-F(1-\nu)}{2\mu a} + \frac{F(1-\nu_u)}{\mu a} \sum_{n=1}^{\infty} \frac{\sin(\alpha_n)\cos(\alpha_n)}{\alpha_n - \sin(\alpha_n)\cos(\alpha_n)} e^{-\frac{\alpha_n^2 c_f t}{a^2}} \right] y, \tag{29}$$

where α_n, $n \in \mathbb{N}$, correspond to the positive solutions of the equation

$$\tan(\alpha_n) = \frac{1-\nu}{\nu_u - \nu} \alpha_n,$$

and ν_u, F, B, c_f, and a are input parameters, partially depending on the physical problem parameters. Here, we apply the values listed in Table 4. For a thorough explanation of the problem and the coefficients in (27)-(29), we refer to the works of Coussy[1] and Phillips and Wheeler.[3]

We consider the domain, $\Omega = (0,100) \times (0,10)$, discretized by a regular triangular mesh. An equidistant partition of the time interval is applied with time-step size $\tau = 10$ from $t_0 = 0$ to $T = 100$. Initial conditions are inherited from the analytic solutions (27)-(29). As boundary conditions, we apply exact Dirichlet boundary conditions for the normal displacement on the top, left, and bottom boundaries. For pressure, we apply homogeneous boundary conditions on the right boundary. On the remaining boundaries, homogeneous natural boundary conditions are applied. The tolerance ϵ_r is set to 10^{-6}. The solution after 10 time steps with 80 vertical and horizontal nodes is displayed in Figure 6.

Similar to the unit square and L-shaped domain test cases, we test the mesh independence and the brute-force optimization technique for Mandel's problem. This time, the parameters from Table 4 are applied. In Figure 7, the mesh dependence of the fixed-stress splitting scheme is tested, and it is clear that the performance of the scheme is independent of this choice. At the same time, we confirm that the optimal stabilization parameters actually are in the proposed interval $[L_{\min}, L_{\text{phys}})$.

TABLE 4 Parameters for Mandel's problem

Name	Symbol	Value	Unit
Young's modulus	E	$5.94 \cdot 10^9$	Pa
Poisson's ratio	ν	0.2	–
Skempton coefficient	B	0.833	–
Undrained Poisson's ratio	ν_u	0.44	–
Applied force	F	$6 \cdot 10^8$	N
Biot-Willis constant	α	1	–
Compressibility coefficient	M	$1.650 \cdot 10^{10}$	Pa
Fluid diffusivity constant	c_f	0.47	m^2/s
Permeability	κ	10^{-10}	m^2
Width of domain	a	100	m
Height of domain	b	10	m
Horizontal number of nodes	N_x	10, 20, 40, 80, 320	–
Vertical number of nodes	N_y	10, 20, 40, 80, 320	–
Time-step size	τ	10	s
Initial time	t_0	0	s
Final time	T	100	s
Relative error tolerance	ϵ_r	10^{-9}	–

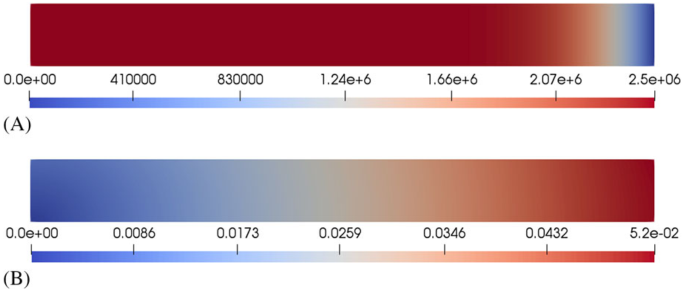


FIGURE 6 Mandel's problem: solution after 10 time steps with $N_x = N_y = 80$. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$) [Colour figure can be viewed at wileyonlinelibrary.com]

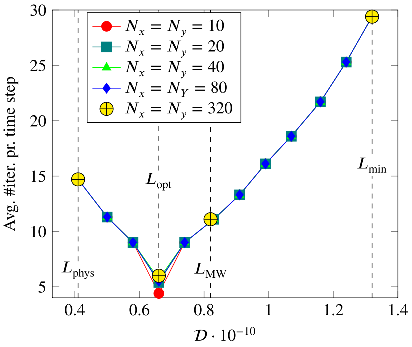


FIGURE 7 Mandel's problem: number of iterations for different stabilization parameters, $L = \frac{\alpha^2}{\mathcal{D}}$, using parameters from Table 4. The largest value of $\mathcal{D}$ corresponds to L_{min}, whereas the smallest value of $\mathcal{D}$ corresponds to L_{phys} [Colour figure can be viewed at wileyonlinelibrary.com]

To calculate the optimal stabilization parameter, we have applied the optimization technique of Section 3.2. First, the fixed-stress splitting scheme is applied for one time step using a coarse mesh with 10 horizontal and 10 vertical nodes for 11 different stabilization parameters in the interval $[L_{\text{min}}, L_{\text{phys}}]$. Choosing the parameter that yields the lowest number of iterations, we apply the scheme for finer meshes and count the number of iterations. As for the other test cases, we see that the optimal parameter indeed is optimal. Moreover, a poor choice of stabilization parameter can result in a huge number of iterations.

4.5 | 3D footing problem

The numerical section is concluded with a three-dimensional example, ie, a footing problem similar to a test case studied in the work of Adler et al.[46] We consider a unit cube subject to normal compression, ramped in time $\sigma_n(t) = t \cdot 10^{10}\,\text{N} \cdot \text{m}^2/\text{s}$, applied to a part of the top boundary $\Gamma_N := [0.25, 0.75] \times [0.25, 0.75] \times \{1\}$. The bottom is fixed in all directions, and the remaining boundary is considered to be stress free. A no-flow boundary condition is applied at the compression zone Γ_N, and zero pressure is enforced on the remaining boundary. Furthermore, zero body forces are applied. The medium is considered isotropic with the same material parameters as used in Section 4.2 (cf Table 2). For the numerical discretization, we consider a set of four meshes with mesh size $h \in \{1/8,\ 1/16,\ 1/32,\ 1/64\}$ and employ the inf-sup–stable MINI element.[51] The simulation result for the final time step is visualized in Figure 8.

Due to high computational cost, optimizing the stabilization parameter of the fixed-stress splitting becomes tedious for fine meshes in 3D. Motivated by the previous results, the optimal stabilization parameter is assumed to be nearly mesh independent. This allows for a brute-force search for the optimal, practical stabilization parameter utilizing the coarsest grid (cf Section 3.2). For validation of the optimization strategy, the performance of the splitting scheme is measured in the range $[L_{\text{min}}, L_{\text{phys}}]$ suggested by Theorem 1; for the finest mesh, we restrict the validation only to a neighborhood of the optimized stabilization parameter. The performance measured in terms of the number of iterations is presented in Figure 9. A large contrast in the performance can be observed for different stabilization parameters, emphasizing the need for a suitable stabilization parameter. Finally, as before, we observe that, indeed, the optimal, practical stabilization parameter is only slightly mesh dependent; it is close to the physical bulk modulus $K_{\text{dr}} = \frac{2\mu}{d} + \lambda$. All in all, the brute-force

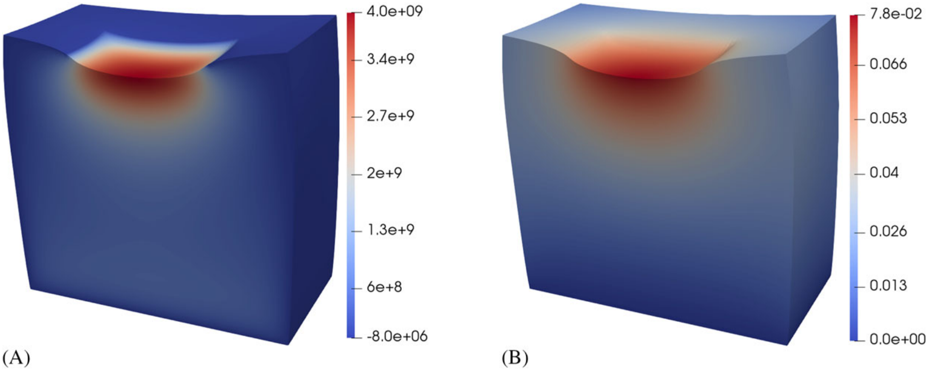


FIGURE 8 Three-dimensional footing problem: solution with a deformed configuration magnified by a factor of 2 at the final time $T = 1$. Notice that the figure only displays half of the domain but that the other half is symmetric. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$) [Colour figure can be viewed at wileyonlinelibrary.com]

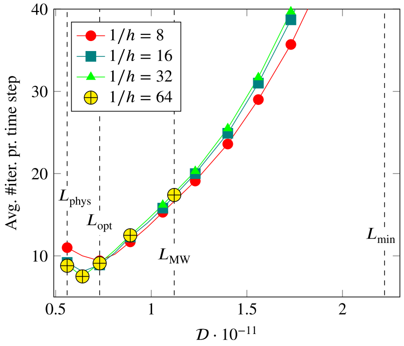


FIGURE 9 Three-dimensional footing problem: average number of iterations per time step for different stabilization parameters, $L = \frac{\alpha^2}{\mathcal{D}}$, using parameters from Table 2. The largest value of $\mathcal{D}$ corresponds to L_{min}, whereas the smallest value of $\mathcal{D}$ corresponds to L_{phys} [Colour figure can be viewed at wileyonlinelibrary.com]

search strategy from Section 3.2 has, again, been confirmed to be a suitable method to obtain a satisfactory stabilization parameter for finer meshes.

5 | CONCLUSIONS

In this work, we have considered the quasi-static, linear Biot model for poroelasticity and studied theoretically and numerically the convergence of the fixed-stress splitting scheme. An improved convergence result has been proved, indicating the nontrivial dependence of the optimal stabilization parameters on not only mechanical properties but also fluid flow properties and discretization properties. We observe numerically that the fixed-stress splitting scheme is close to being mesh independent and determine a novel domain in which the optimal stabilization/tuning parameter is found, ie, $[\frac{\alpha^2}{4\mu+2\lambda}, \frac{\alpha^2}{\frac{2\mu}{d}+\lambda})$. On the basis of these observations, we propose a brute-force method with low cost for choosing the optimal stabilization parameter, ie, the parameter that corresponds to the smallest amount of fixed-stress iterations. Through numerical experiments, we have showed that this optimization method results in a much faster fixed-stress splitting scheme than those obtained by choosing the classical stabilization parameters $L = \frac{\alpha^2}{K_{\text{dr}}}$ and $L = \frac{\alpha^2}{2K_{\text{dr}}}$.

ORCID

Florin A. Radu https://orcid.org/0000-0002-2577-5684

REFERENCES

1. Coussy O. *Poromechanics*. Hoboken, NJ: John Wiley & Sons; 2004.
2. Nordbotten JM. Stable cell-centered finite volume discretization for Biot equations. *SIAM J Numer Anal*. 2016;54(2):942-968. https://doi.org/10.1137/15M1014280
3. Phillips PJ, Wheeler MF. A coupling of mixed and continuous Galerkin finite element methods for poroelasticity I: the continuous in time case. *Computational Geosciences*. 2007;11(2):131. https://doi.org/10.1007/s10596-007-9045-y
4. Yi S-Y, Bean ML. Iteratively coupled solution strategies for a four-field mixed finite element method for poroelasticity. *Int J Numer Anal Methods Geomech*. 2016;41(2):159-179. https://onlinelibrary.wiley.com/doi/abs/10.1002/nag.2538
5. Both JW, Borregales M, Nordbotten JM, Kumar K, Radu FA. Robust fixed stress splitting for Biot's equations in heterogeneous media. *Appl Math Lett*. 2017;68:101-108. http://www.sciencedirect.com/science/article/pii/S0893965917300034
6. Berger L, Bordas R, Kay D, Tavener S. A stabilized finite element method for finite-strain three-field poroelasticity. *Computational Mechanics*. 2017;60(1):51-68. https://doi.org/10.1007/s00466-017-1381-8
7. Lee JJ. Robust error analysis of coupled mixed methods for Biot's consolidation model. *J Sci Comput*. 2016;69(2):610-632. https://doi.org/10.1007/s10915-016-0210-0
8. Hu X, Rodrigo C, Gaspar FJ, Zikatanov LT. A nonconforming finite element method for the Biot's consolidation model in poroelasticity. *J Comput Appl Math*. 2017;310:143-154. http://www.sciencedirect.com/science/article/pii/S0377042716302734
9. Rodrigo C, Gaspar FJ, Hu X, Zikatanov LT. Stability and monotonicity for some discretizations of the Biot's consolidation model. *Comput Methods Appl Mech Eng*. 2016;298:183-204. http://www.sciencedirect.com/science/article/pii/S0045782515003138
10. Chaabane N, Rivière B. A splitting-based finite element method for the Biot poroelasticity system. *Comput Math Appl*. 2018;75(7):2328-2337. http://www.sciencedirect.com/science/article/pii/S0898122117307721
11. Chaabane N, Rivière B. A sequential discontinuous Galerkin method for the coupling of flow and geomechanics. *J Sci Comput*. 2018;74(1):375-395. https://doi.org/10.1007/s10915-017-0443-6
12. Simoni L, Secchi S, Schrefler BA. Numerical difficulties and computational procedures for thermo-hydro-mechanical coupled problems of saturated porous media. *Computational Mechanics*. 2008;43(1):179-189. https://doi.org/10.1007/s00466-008-0302-2
13. Dana S, Wheeler MF. Convergence analysis of fixed stress split iterative scheme for anisotropic poroelasticity with tensor Biot parameter. *Computational Geosciences*. 2018;22(5):1219-1230. https://doi.org/10.1007/s10596-018-9748-2
14. Castelletto N, Klevtsov S, Hajibeygi H, Tchelepi HA. Multiscale two-stage solver for Biot's poroelasticity equations in subsurface media. *Computational Geosciences*. 2018;23(2):207-224. https://doi.org/10.1007/s10596-018-9791-z
15. Castelletto N, Hajibeygi H, Tchelepi HA. Multiscale finite-element method for linear elastic geomechanics. *J Comput Phys*. 2017;331:337-356. http://www.sciencedirect.com/science/article/pii/S0021999116306362
16. Bause M, Radu FA, Köcher U. Space-time finite element approximation of the Biot poroelasticity system with iterative coupling. *Comput Methods Appl Mech Eng*. 2017;320:745-768. http://www.sciencedirect.com/science/article/pii/S0045782516316164
17. Ern A, Meunier S. A posteriori error analysis of Euler-Galerkin approximations to coupled elliptic-parabolic problems. *ESAIM Math Model Numer Anal*. 2009;43(2):353-375.
18. Rahrah M, Vermolen F. Monte Carlo assessment of the impact of oscillatory and pulsating boundary conditions on the flow through porous media. *Transp Porous Media*. 2018;123(1):125-146. https://link.springer.com/article/10.1007/s11242-018-1028-z
19. Haga JB, Osnes H, Langtangen HP. On the causes of pressure oscillations in low-permeable and low-compressible porous media. *Int J Numer Anal Methods Geomech*. 2012;36(12):1507-1522. https://onlinelibrary.wiley.com/doi/abs/10.1002/nag.1062
20. Rodrigo C, Hu X, Ohm P, Adler JH, Gaspar FJ, Zikatanov LT. New stabilized discretizations for poroelasticity and the Stokes' equations. *Comput Methods Appl Mech Eng*. 2018;341:467-484. http://www.sciencedirect.com/science/article/pii/S0045782518303347
21. Kim J, Tchelepi HA, Juanes R. Stability and convergence of sequential methods for coupled flow and geomechanics: fixed-stress and fixed-strain splits. *Comput Methods Appl Mech Eng*. 2011;200(13-16):1591-1606. http://www.sciencedirect.com/science/article/pii/S0045782510003786
22. Kim J, Tchelepi HA, Juanes R. Stability and convergence of sequential methods for coupled flow and geomechanics: drained and undrained splits. *Comput Methods Appl Mech Eng*. 2011;200(23-24):2094-2116. http://www.sciencedirect.com/science/article/pii/S0045782511000466
23. Settari A, Mourits FM. A coupled reservoir and geomechanical simulation system. *Soc Petroleum Eng*. 1998;3:219-226.
24. Mikelić A, Wheeler MF. Convergence of iterative coupling for coupled flow and geomechanics. *Computational Geosciences*. 2013;17(3):455-461. https://doi.org/10.1007/s10596-012-9318-y
25. Turska E, Schrefler BA. On convergence conditions of partitioned solution procedures for consolidation problems. *Comput Methods Appl Mech Eng*. 1993;106(1-2):51-63. http://www.sciencedirect.com/science/article/pii/004578259390184Y
26. Turska E, Wisniewski K, Schrefler BA. Error propagation of staggered solution procedures for transient problems. *Comput Methods Appl Mech Eng*. 1994;114(1-2):177-188. http://www.sciencedirect.com/science/article/pii/0045782594901686
27. Both JW, Köcher U. Numerical investigation on the fixed-stress splitting scheme for Biot's equations: Optimality of the tuning parameter. In: *Numerical Mathematics and Advanced Applications ENUMATH 2017*. Cham, Switzerland: Springer; 2019:789-797.
28. Mikelić A, Wang B, Wheeler MF. Numerical convergence study of iterative coupling for coupled flow and geomechanics. *Computational Geosciences*. 2014;18(3-4):325-341. https://doi.org/10.1007/s10596-013-9393-8
29. Dana S, Ganis B, Wheeler MF. A multiscale fixed stress split iterative scheme for coupled flow and poromechanics in deep subsurface reservoirs. *J Comput Phys*. 2018;352:1-22. http://www.sciencedirect.com/science/article/pii/S002199911730709X

30. Bause M. Iterative coupling of mixed and discontinuous Galerkin methods for poroelasticity. In: *Numerical Mathematics and Advanced Applications ENUMATH 2017*. Cham, Switzerland: Springer; 2019:551-560.
31. Storvik E, Both JW, Kumar K, Nordbotten JM, Radu FA. On the optimization of the fixed-stress splitting for Biot's equations. arXiv preprint arXiv:1811.06242. 2018.
32. Borregales M, Radu FA, Kumar K, Nordbotten JM. Robust iterative schemes for non-linear poromechanics. *Computational Geosciences*. 2018;22(4):1021-1038. https://doi.org/10.1007/s10596-018-9736-6
33. Both JW, Kumar K, Nordbotten JM, Radu FA. Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media. *Comput Math Appl*. 2019;77(6):1479-1502. http://www.sciencedirect.com/science/article/pii/S0898122118304048
34. Both JW, Kumar K, Nordbotten JM, Radu FA. Iterative methods for coupled flow and geomechanics in unsaturated porous media. *Poromechanics VI*. 2017:411-418. https://ascelibrary.org/doi/abs/10.1061/9780784480779.050
35. Hong Q, Kraus J, Lymbery M, Philo F. Conservative discretizations and parameter-robust preconditioners for Biot and multiple-network flux-based poroelastic models. arXiv preprint arXiv: 1806.00353v2. 2018.
36. Lee JJ, Piersanti E, Mardal K-A, Rognes ME. A mixed finite element method for nearly incompressible multiple-network poroelasticity. *SIAM J Sci Comput*. 2019;41(2):A722-A747. https://doi.org/10.1137/18M1182395
37. Kim J. A new numerically stable sequential algorithm for coupled finite-strain elastoplastic geomechanics and flow. *Comput Methods Appl Mech Eng*. 2018;335:538-562.
38. Girault V, Kumar K, Wheeler MF. Convergence of iterative coupling of geomechanics with flow in a fractured poroelastic medium. *Computational Geosciences*. 2016;20(5):997-1011. https://doi.org/10.1007/s10596-016-9573-4
39. Giovanardi B, Formaggia L, Scotti A, Zunino P. Unfitted FEM for modelling the interaction of multiple fractures in a poroelastic medium. In: Bordas SPA, Burman E, Larson MG, Olshanskii MA, eds. *Geometrically Unfitted Finite Element Methods and Applications*. Cham, Switzerland: Springer International Publishing; 2017:331-352.
40. Lee S, Wheeler MF, Wick T. Iterative coupling of flow, geomechanics and adaptive phase-field fracture including level-set crack width approaches. *J Comput Appl Math*. 2017;314:40-60. http://www.sciencedirect.com/science/article/pii/S0377042716305118
41. List F, Radu FA. A study on iterative methods for solving richards' equation. *Computational Geosciences*. 2016;20(2):341-353. https://doi.org/10.1007/s10596-016-9566-3
42. Pop IS, Radu FA, Knabner P. Mixed finite elements for the richards' equation: linearization procedure. *J Comput Appl Math*. 2004;168(1):365-373. http://www.sciencedirect.com/science/article/pii/S037704270301001X
43. Almani T, Kumar K, Dogru A, Singh G, Wheeler MF. Convergence analysis of multirate fixed-stress split iterative schemes for coupling flow with geomechanics. *Comput Methods Appl Mech Eng*. 2016;311:180-207. http://www.sciencedirect.com/science/article/pii/S0045782516308180
44. Borregales M, Kumar K, Radu FA, Rodrigo C, Gaspar FJ. A partially parallel-in-time fixed-stress splitting method for Biot's consolidation model. *Comput Math Appl*. 2018;77(6):1466-1478. http://www.sciencedirect.com/science/article/pii/S0898122118305091
45. Boffi D, Brezzi F, Fortin M. *Mixed Finite Element Methods and Applications*. Berlin, Germany: Springer; 2013. *Springer Series in Computational Mathematics*; vol. 44.
46. Adler JH, Gaspar FJ, Hu X, Rodrigo C, Zikatanov LT. Robust block preconditioners for Biot's model. In: *Domain Decomposition Methods in Science and Engineering XXIV*. Cham, Switzerland: Springer; 2017:3-16.
47. Blatt M, Burchardt A, Dedner A, et al. The distributed and unified numerics environment, version 2.4. *Arch Numer Softw*. 2016;4(100):13-29.
48. Engwer C, Gräser C, Müthing S, Sander O. The interface for functions in the dune-functions module. arXiv preprint arXiv:1512.06136. 2015.
49. Engwer C, Gräser C, Müthing S, Sander O. Function space bases in the dune-functions module. arXiv preprint arXiv:1806.09545. 2018.
50. Abousleiman Y, Cheng AH-D, Cui L, Detournay E, Roegiers J-C. Mandel's problem revisited. *Géotechnique*. 1996;46(2):187-195. https://doi.org/10.1680/geot.1996.46.2.187
51. Arnold DN, Brezzi F, Fortin M. A stable finite element for the stokes equations. *CALCOLO*. 1984;21(4):337-344. https://doi.org/10.1007/BF02576171

How to cite this article: Storvik E, Both JW, Kumar K, Nordbotten JM, Radu FA. On the optimization of the fixed-stress splitting for Biot's equations. *Int J Numer Methods Eng*. 2019;1–16. https://doi.org/10.1002/nme.6130

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230856413 (print)
9788230858745 (PDF)