Development of robust and efficient solution strategies for coupled problems

Erlend Storvik

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2022

UNIVERSITY OF BERGEN

Development of robust and efficient solution strategies for coupled problems

Erlend Storvik

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

Date of defense: 10.10.2022

Year: 2022

Title: Development of robust and efficient solution strategies for coupled problems

Name: Erlend Storvik

Print: Skipnes Kommunikasjon / University of Bergen

Preface

This dissertation is submitted as a partial fulfillment of the requirements to achieve the degree of Philosophiae Doctor (PhD) at the University of Bergen. The advisory committee consisted of Florin Adrian Radu (University of Bergen), Jakub Wiktor Both (University of Bergen), Jan Martin Nordbotten (University of Bergen), and Rainer Helmig (University of Stuttgart and University of Bergen).

Acknowledgements

Doing a PhD has been a joyous experience. The amount of cool mathematics, physics and programming that I have learned over the last coupled of years have far surpassed what I could imagine, and I feel truly lucky and thankful to have been given the opportunity.

I have been even luckier with my supervisors. The team of Florin, Jan and Jakub have been incredibly inspiring, helpful, positive and constructive. In addition to our weekly meetings, your doors and zoom rooms have always been open, and I am grateful for all the good faith and knowledge that you have put in me. In particular, I want to thank Florin for including me in different projects, travels, and social activities, as well as significantly extending my academic network and making me a better mathematician, Jan for being ever insightful and inspiring, and Jakub for being an amazing role model and mentor. Moreover, I feel like you are my friends, and I will forever be grateful for the dinners, travels, beers, fishing trips, and padel matches that we have had together.

Since starting my position as a PhD candidate I have had the honor of being a part of the porous media group. This has, due to all the friendly and good colleagues, made both the social and scientific part of work very inspiring. The weekly seminars have been a great arena for learning new concepts as well as getting feedback on research, and through the connections of the group, I have had the pleasure to get to know many international colleagues. I would like to thank the entire porous media group and my other colleagues for the good times we have had together. In particular, my thanks go to Davide for the great friendship we have developed over the last couple of years, Jhabriel for being the most friendly office mate and colleague I could wish for, Anita for the friendship that we have built over countless coffee breaks and for being the main social engine in the department of mathematics, and to Ivar for early taking me into the porous media group and for being a polite jogging partner during his days of restitution.

I would also like to thank my wife Laila for the great times we always have together and my family and friends for always being great and inspiring supporters. As far as I am concerned, you all make me the luckiest person I know!

Abstract

There are many applications where the study of coupled physical processes is of great importance. These range from the life sciences with flow in deformable human tissue to structural engineering with fracture propagation in elastic solids. In this doctoral dissertation, there is a twofold focus on coupled problems. Firstly, robust and efficient solution strategies, with a focus on iterative decoupling methods, have been applied to several coupled systems of equations. Secondly, a new thermodynamically consistent coupled system of equations is proposed. Solution strategies are developed for three different coupled problems; the quasi-static linearized Biot equations that couples flow through porous materials and elastic deformation of the solid medium, variational phase-field models for brittle fracture that couple a phase-field equation for fracture evolution with linearized elasticity, and the Cahn-Larché equations that model elastic effects in a two-phase elastic material and couples an extended Cahn-Hilliard phase-field equation and linearized elasticity. Finally, the new system of equations that is proposed models flow through a two-phase deformable porous material where the solid phase evolution is governed by interfacial forces as well as effects from both the fluid and elastic properties of the material.

In the work that concerns the quasi-static linearized Biot equations, the focus is on the fixed-stress splitting scheme, which is a popular method for sequentially solving the flow and elasticity subsystems of the full model. Using such a method is beneficial as it allows for the use of readily available solvers for the subproblems; however, a stabilizing term is required for the scheme to converge. It is well known that the convergence properties of the method strongly depend on how this term is chosen, and here, the optimal choice of it is addressed both theoretically and practically. An interval where the optimal stabilization parameter lies is provided, depending on the material parameters. In addition, two different ways of optimizing the parameter are proposed. The first is a brute-force method that relies on the mesh independence of the scheme's optimal stabilization parameter, and the second is valid for low-permeable media and utilizes an equivalence between the fixed-stress splitting scheme and the modified Richardson iteration.

Regarding the variational phase-field model for brittle fracture propagation, the focus is on improving the convergence properties of the most commonly used solution strategy with an acceleration method. This solution strategy relies on a staggered scheme that alternates between solving the elasticity and phase-field subproblems in an iterative way. This is known to be a robust method compared to the monolithic Newton method. However, the staggered scheme often requires many iterations to converge to satisfactory precision. The contribution of this work is to accelerate the solver through a new acceleration method that combines Anderson acceleration and over-relaxation, dynamically switching back and forth between them depending on a criterion that takes the residual evolution into account. The acceleration scheme takes advantage of the strengths of both Anderson acceleration and over-relaxation, and the fact that they are complementary when applied to this problem, resulting in a significant speed-up of the convergence. Moreover, the method is applied as a post-processing technique to the increments of the solver, and can thus be implemented with minor modifications to readily available software.

The final contribution toward solution strategies for coupled problems focuses on the Cahn-Larché equations. This is a model for linearized elasticity in a medium with two elastic phases that evolve with respect to interfacial forces and elastic effects. The system couples linearized elasticity and an extended Cahn-Hilliard phase-field equation. There are several challenging features with regards to solution strategies for this system including nonlinear coupling terms, and the fourth-order term that comes from the Cahn-Hilliard subsystem. Moreover, the system is nonlinear and non-convex with respect to both the phase-field and the displacement. In this work, a new semi-implicit time discretization that extends the standard convex-concave splitting method applied to the double-well potential from the Cahn-Hilliard subsystem is proposed. The extension includes special treatment for the elastic energy, and it is shown that the resulting discrete system is equivalent to a convex minimization problem. Furthermore, an alternating minimization solver is proposed for the fully discrete system, together with a convergence proof that includes convergence rates. Through numerical experiments, it becomes evident that the newly proposed discretization method leads to a system that is far better conditioned for linearization methods than standard time discretizations.

Finally, a new model for flow through a two-phase deformable porous material is proposed. The two poroelastic phases have distinct material properties, and their interface evolves according to a generalized Ginzburg–Landau energy functional. As a result, a model that extends the Cahn-Larché equations to poroelasticity is proposed, and essential coupling terms for several applications are highlighted. These include solid tumor growth, biogrout, and wood growth. Moreover, the coupled set of equations is shown to be a generalized gradient flow. This implies that the system is thermodynamically

consistent and makes a toolbox of analysis and solvers available for further study of the model.

Sammendrag

Det er mange modeller i moderne vitenskap hvor sammenkoblingen mellom forskjellige fysiske prosesser er svært viktig. Disse finner man for eksempel i forbindelse med CO_2-lagring i undervannsreservoarer, flyt i kroppsvev, kreftsvulstvekst og geotermisk energiutvinning. Denne avhandlingen har to fokusområder som er knyttet til sammenkoblede modeller. Det første er å utvikle pålitelige og effektive tilnærmingsmetoder, og det andre er utviklingen av en ny modell som tar for seg flyt i et porøst medium som består av to forskjellige materialer.

For tilnærmingsmetodene har det vært et spesielt fokus på splittemetoder. Dette er metoder hvor hver av de sammenkoblede modellene håndteres separat, og så itererer man mellom dem. Dette gjøres i hovedsak fordi man kan utnytte tilgjengelig teori og programvare for å løse hver undermodell svært effektivt. Ulempen er at man kan ende opp med løsningsalgoritmer for den sammenkoblede modellen som er trege, eller ikke kommer frem til noen løsning i det hele tatt. I denne avhandlingen har tre forskjellige metoder for å forbedre splittemetoder blitt utviklet for tre forskjellige sammenkoblede modeller.

Den første modellen beskriver flyt gjennom deformerbart porøst medium og er kjent som Biot ligningene. For å anvende en splittemetode på denne modellen har et stabiliseringsledd blitt tilført. Dette sikrer at metoden konvergerer (kommer frem til en løsning), men dersom man ikke skalerer stabiliseringsleddet riktig kan det ta veldig lang tid. Derfor har et intervall hvor den optimale skaleringen av stabiliseringsleddet befinner seg blitt identifisert, og utfra dette presenteres det en måte å praktisk velge den riktige skaleringen på.

Den andre modellen er en fasefeltmodell for sprekkpropagering. Denne modellen løses vanligvis med en splittemetode som er veldig treg, men konvergent. For å forbedre dette har en ny akselerasjonsmetode har blitt utviklet. Denne anvendes som et postprosesseringssteg til den klassiske splittemetoden, og utnytter både overrelaksering og Anderson akselerasjon. Disse to forskjellige akselerasjonsmetodene har kompatible styrker i at overrelaksering akselererer når man er langt fra løsningen (som er tilfellet når sprekken

propagerer), og Anderson akselerasjon fungerer bra når man er nærme løsningen. For å veksle mellom de to metodene har et kriterium basert på residualfeilen blitt brukt. Resultatet er en pålitelig akselerasjonsmetode som alltid akselererer og ofte er svært effektiv.

Det siste modellen kalles Cahn-Larché ligningene og er også en fasefeltmodell, men denne beskriver elastisitet i et medium bestående av to elastiske materialer som kan bevege seg basert på overflatespenningen mellom dem. Dette problemet er spesielt utfordrende å løse da det verken er lineært eller konvekst. For å håndtere dette har en ny måte å behandle tidsavhengigheten til det underliggende koblede problemet på blitt utviklet. Dette leder til et diskret system som er ekvivalent med et konvekst minimeringsproblem, som derfor er velegnet til å løses med de fleste numeriske optimeringsmetoder, også splittemetoder.

Den nye modellen som har blitt utviklet er en utvidelse av Cahn-Larché ligningene og har fått navnet Cahn-Hilliard-Biot. Dette er fordi ligningene utgjør en fasefelt modell som beskriver flyt i et deformerbart porøst medium med to poroelastiske materialer. Disse kan forflytte seg basert på overflatespenning, elastisk spenning, og poretrykk, og det er tenkt at modellen kan anvendes i forbindelse med kreftsvulstmodellering.

Outline

This doctoral dissertation is divided into two main parts. Part I governs the theoretical foundations of the research and Part II is a collection of the papers that constitute the scientific results of the research.

The first part of the thesis is organized into four chapters. In Chapter 1, a brief introduction to the mathematical coupled models that are analyzed in this dissertation is provided, as well as a summary of the main contributions of the research. Then, in Chapter 2, the theoretical background for the mathematical models is discussed, before numerical solution strategies to coupled problems are presented in Chapter 3. Finally, a summary of the articles in Part II and an outlook are given in Chapter 4.

The second part of the thesis consists of five papers that make up the scientific results of the dissertation:

Paper A Storvik, E., Both, J.W., Kumar, K., Nordbotten, J.M., and Radu, F.A. On the optimization of the fixed-stress splitting for Biot's equations. *International Journal for Numerical Methods in Engineering*, 120, 179–194 (2019)

Paper B Storvik, E., Both, J.W., Nordbotten, J.M., and Radu, F.A. The Fixed-Stress Splitting Scheme for Biot's Equations as a Modified Richardson Iteration: Implications for Optimal Convergence. *Numerical Mathematics and Advanced Applications ENUMATH 2019*, Lecture Notes in Computational Science and Engineering, 139, 909–917 (2021)

Paper C Storvik, E., Both, J.W., Sargado, J.M., Nordbotten, J.M., and Radu, F.A. An accelerated staggered scheme for variational phase-field models of brittle fracture. *Computational Methods in Applied Mechanics and Engineering*, 381, 113822 (2021)

Paper D Storvik, E., Both, J.W., Nordbotten, J.M., and Radu, F.A. A Cahn-Hilliard-Biot system and its generalized gradient flow structure. *Applied*

Mathematics Letters, 381, 107799 (2021)

Paper E Storvik, E., Both, J.W., Nordbotten, J.M., and Radu, F.A. A robust solution strategy for the Cahn-Larché equations. *In review.* arXiv:2206.01541 [math.NA].

Contents

Part II: Scientific results

Chapter 1

Introduction

Coupled problems arise in many societally relevant applications, ranging from flow through deformable porous media in relation to CO_2 sequestration, groundwater extraction and flow in brain tissue to fracture mechanics in relation to structural engineering and geothermal energy extraction. This dissertation is concerned with three different coupled mathematical models:

- The quasi-static linearized Biot consolidation model of flow through deformable porous media, often known as poroelasticity or poromechanics.
- The variational phase-field approach to fracture propagation, where a phase-field evolution equation is coupled with linearized elasticity.
- The Cahn-Larché system that models the evolution of a composition of two disjoint different elastic materials with interface forces and swelling effects. Here, a phase-field evolution equation of Cahn-Hilliard type is coupled with linearized elasticity.

In particular, the focus of the research is on the development of robust and efficient solution strategies for these coupled problems. This is important because it enables the possibility of approximating solutions to problems for a wide variety of material parameters with high accuracy.

1.1 Poroelasticity

Poroelasticity governs flow through deformable porous materials. The first for flow and deformation was due to Terzaghi, who developed a one-dimensional consolidation model

and introduced the concept of effective stress [151]. Later, Biot extended that theory to three spatial dimensions [19]. There are many societally relevant applications related to poroelasticity ranging from the life sciences with modelling of the heart-circulatory system [30, 132] or flow in brain tissue [112, 153] to environmental science related to CO_2 sequestration in depleted hydrocarbon reservoirs [21, 93, 125] and modelling of ground subsidence due to extraction of groundwater [76, 157]. Probably the most widely used mathematical model for poroelasticity today is the quasi-static linearized Biot equations that account for the balance of mass and linear momentum, as well as incorporating pore-pressure to the effective stress by the Biot-Willis coupling coefficient [20], and Darcy's law.

There are two popular choices for solving the quasi-static linearized Biot equations: Either to use a monolithic method, i.e., solve for all unknowns (e.g., pore pressure and displacement) simultaneously or to apply a splitting method, alternating between solving the flow and elasticity subproblem in an iterative way. The monolithic method has the benefit that it is stable in the sense that one can use a linear solver and get a solution. It is, however, difficult to construct efficient and robust linear solvers for such coupled problems, although a lot of work has been made toward preconditioners for the Biot equations [1, 2, 24, 49, 92, 96, 111]. Splitting methods, on the other hand, have the benefit that one can use readily available solvers for flow and mechanics and iterate between them. Moreover, it can easily be extended to more complicated, and possibly nonlinear problems, such as unsaturated deformable porous media [33], finite-strain poroelasticity [27], the inclusion of thermal effects [41], Biot-Allard [15, 16], nonlinear poromechanics [26], multiple-network poroelasticity [95], and fluid-structure interactions [155], and be combined with time discretizations to make a partially-parallel-in-time solver [25].

The drawback of the splitting methods is that they typically require some stabilization to converge. The original stabilization is due to Settari [138], who, based on physical interpretations, proposed to fix the volumetric stress over the iterations, resulting in a simple pressure stabilization and leading to the notion of the fixed-stress splitting scheme. Later, in the work of Mikelić and Wheeler [121], it was mathematically proven that as long as the stabilization constant is larger than half of the one proposed in [138], the method would converge. The same was proved, using a different technique in [31]. In the Papers A and B, the optimal choice of this stabilization term is addressed and a finite interval where it always resides is found. Moreover, a way to compute it for general cases is provided in Paper A, and a more efficient approach for the case of low-permeable media can be found in Paper B. There is also a counterpart to the fixed-stress splitting scheme, that stabilizes the elasticity equation instead of the flow equation. This is often denoted as the undrained split [104]. Moreover, both splitting methods can be obtained as alternating minimization applied to a primal or a dual formulation of

the minimization problem associated with the thermodynamically motivated generalized gradient flow structure of the quasi-static linearized Biot equations [34].

1.2 The variational phase-field approach to fracture

Mathematical modeling of fracture propagation in elastic solid is an important topic in structural engineering. The thermodynamical principles of fracture propagation originated in the work of Griffith [89], who proposed that an existing fracture will propagate if the energy release rate associated with crack extension exceeds a critical value. Much later, Francfort and Marigo [71] introduced the variational approach to fracture mechanics in an effort to overcome some of the shortcomings of Griffith's framework, namely fracture nucleation or branching. At this stage, the central issue with the variational approach to fracture was the discontinuities in displacement across fractures.

The remedy, proposed by Bourdin, Francfort and Marigo [36, 37], was to introduce a smooth indicator function, called a phase-field, to track the crack location. Since then, numerous developments of the model have been made, eventually branching in different directions. Of the most important contributions, one finds the splitting of the elastic energy into tensile and compressive parts to account for fracture propagation only due to tensile forces. Several ways have been proposed to do this, but the most common choices are due to Miehe [119], and Amor [9].

One of the key challenges related to approximating solutions to the phase-field models for fracture propagation is in the linearization strategies. Typically, the Newton method struggles with convergence, which is to be expected, as it is only known for its local convergence properties, and in loading steps with crack propagation, the states between consecutive solutions might vary greatly. The most common choice for a robust method is to use a decoupling method that usually either is called a staggered scheme or alternating minimization. This method alternates between solving the Euler-Lagrange equations corresponding to the phase-field evolution and the elastic deformation. Although the common experience is that this solution strategy is robust, it is in certain scenarios very slow and can during loading steps where cracks are propagating demand thousands of iterations to converge to satisfactory precision. Known remedies in the literature that aim to stabilize the robustness issues of the monolithic Newton method include the modified Newton method from [159], the BFGS-type methods proposed in [106, 161], a line-search-based Newton method [83], and a truncated non-smooth Newton method [87]. Moreover, several attempts to accelerate the staggered scheme have been proposed, including the stabilized staggered scheme in [42], and a combined over-relaxed staggered

scheme and Newton method was proposed in [67]. In Paper C an acceleration method that can be implemented as a post-processing technique to the staggered solution strategy is developed. This method is, in addition to being easy to implement on top of readily available software, highly accelerating and robust in the sense that it never decelerates the convergence of the scheme.

1.3 The Cahn-Larché equations

The early work of Cahn and Hilliard [43, 45, 46], related to the free energy of nonuniform systems has had a lasting impact in several fields of applied mathematics, ranging from modelling of spinodal decomposition [44, 47, 108] to phase-field models of two-phase flow [48, 59, 68] and fingering effects in porous materials [57]. In this dissertation, the interest has primarily been in a model for elastic deformation in a composition of two solids. This model is often known as the Cahn-Larché system, due to the work of Cahn and Larché in [109, 110]. The equation, with problem-specific extensions, has been employed for predictive tumor growth modelling [74, 79, 80], lithium-ion intercalation into silicon [117], diffusional coarsening in binary alloys [61, 88] and was recently experimentally verified as a model for the connection between chemical and mechanical processes in alloys [139].

One of the main difficulties with approximating solutions to the Cahn-Larché equations lies with the non-convex nonlinearities in the system. These are present in both the double-well potential that is inherent in all Cahn-Hilliard-type equations, and in the nonlinear coupling between the phase-field and elasticity. One could try to use an explicit time-discretization and remove the need for linearization, however, that would lead to a non-gradient stable discretization, i.e., the free energy of the system might increase due to poor time-stepping choices. Therefore, it is more common to apply a semi-implicit time-discretization, which was proposed for the double well potential in the Cahn-Hilliard equation by Eyre in [65]. Here, the idea is to split the non-convex nonlinearity into two parts, one convex and one concave, and evaluate the convex part implicitly in time and the concave one explicitly. Doing so for the Cahn-Hilliard equation leads to an unconditionally gradient-stable time discretization that is well suited for linearization methods. For the Cahn-Larché equations the same splitting has been applied several times [77, 78, 81, 88]. In Paper E, the inherent minimization structure related to the generalized gradient flow structure of the Cahn-Larché equations is exploited. Some of the terms from the elastic free energy of the system are evaluated explicitly, resulting in a system that is well suited for linearization methods and can be proven to be gradient stable under certain assumptions on the material parameters.

1.4 The Cahn-Hilliard-Biot equations

During the last decade there has been an increasing interest in predictive tumor growth modeling utilizing phase-field methods of Cahn-Hilliard type to account for interfacial forces between cancerous and healthy cells (with surrounding tissue); see, e.g., [72, 73, 74, 75, 79, 80, 114, 115, 126, 141, 160]. The extension to account for elastic effects has been considered in [74, 79]. Paper D is concerned with the extension to a thermodynamically consistent model that also accounts for flow through the (poro)elastic material, with permeability and poroelastic coupling parameters that depend on the material phase. Moreover, the system is written as a generalized gradient flow, which could prove useful for further analysis to come.

1.5 Main contributions

This dissertation is concerned with the development and analysis of robust solution strategies for coupled problems, as well as one contribution to extend the Cahn-Larché equations to include flow through the poroelastic material in a thermodynamically consistent manner. The following summarizes the main contributions of the research herein:

1. **Optimal stabilization for decoupling of the Biot consolidation model.**

   In Paper A and B the optimal stabilization parameter for the fixed-stress splitting method applied to Biot's equations is addressed. In Paper A, a theoretical proof of convergence, in the form of a contraction result, including the contraction rate, is provided. That contraction rate is utilized to determine an interval that the optimal stabilization parameter resides in, provided that the system is discretized with an inf-sup stable spatial discretization. Moreover, a practical approach to compute the optimal stabilization parameter by exploiting the mesh independence of the fixed-stress splitting scheme's performance is proposed.

   In Paper B, the fixed-stress splitting scheme is identified with a modified Richardson iteration. Using the theory for the optimal convergence of that method, the optimal choice of stabilization parameter in the fixed-stress splitting scheme is determined. Some computational aspects with regards to finding the optimal stabilization parameter are discussed involving the computation of eigenvalues.

2. **Acceleration of the staggered scheme for variational phase-field models for brittle fracture propagation.**

In Paper C, an acceleration method for the staggered solution scheme applied to the phase-field models for fracture propagation is proposed. The acceleration method is a combination of two standard techniques; Anderson acceleration and over-relaxation, where either the Anderson acceleration or over-relaxation always is applied. The benefit of switching between these two methods compared to switching between the staggered scheme and a monolithic Newton method is that both of them are applied as post-processes to the standard staggered scheme, i.e., each update made by the staggered scheme is modified according to some rule. Therefore, the method is straightforward to implement on top of readily available software, and the cost of applying it is negligible compared to the cost of an iterative step. Moreover, a criterion for switching between the two acceleration methods, based on monitoring the algebraic residuals, is proposed. Several standard benchmark problems from the literature are considered and it is shown that the method always accelerates and in some examples more than 80% reduction in the number of iterations is experienced.

3. **Extension of the Cahn-Larché system to an equation that accounts for flow through a poroelastic material.**

   An extension of the Cahn-Larché equation including flow through the elastic material is proposed in Paper D. This model can be seen as a combination of the Cahn-Hilliard equation and the Biot equations and is, therefore, called the Cahn-Hilliard-Biot system. Moreover, the proposed model is shown to have a generalized gradient flow structure, and thereby be thermodynamically consistent. A numerical experiment that shows the impact of fluid flow on the phase-field evolution is provided.

4. **A robust solution strategy for the Cahn-Larché equations.**

   In Paper E, a new semi-implicit time-discretization for the Cahn-Larché equations is provided. This is shown theoretically to be gradient stable while the stiffness tensors are the same for both elastic materials. Moreover, the semi-implicit time discretization is identified with a discrete minimization problem, and a theoretical proof that alternating minimization converges for this minimization problem is provided. Additionally, several numerical examples show that linearization methods applied to this newly proposed time-discretization behave very well compared to standard implicit time-discretizations for which the linearization methods fail to converge at all, even when the classical convex-concave splitting method is applied to the double-well potential.

Chapter 2

Mathematical models

In this chapter, an introduction to the mathematical models that have been studied in Part II of the dissertation is presented. First, in Section 2.1, the subproblems that contribute to the coupled problems are discussed, and in Section 2.2 the fully coupled problems are introduced. Finally, in Section 2.3, a brief introduction to gradient flows is provided as it is a core concept that is utilized in Paper D and E in Part II of the dissertation.

2.1 Non-coupled mathematical models

In Section 2.1.1, a brief introduction to linearized elasticity is provided, before single-phase flow in porous materials is discussed in Section 2.1.2. Finally, Section 2.1.3 gives an introduction to general phase-field modeling, with the Cahn-Hilliard and Allen-Cahn equations as examples.

2.1.1 Linearized elasticity

Linearized elasticity constitutes a subproblem of all the coupled models that are discussed in this dissertation. The theory of elasticity governs physical effects and motions related to elastic bodies subject to deformation due to external forces, loading, or other effects such as those related to pressure of internal fluids or swelling. There are several comprehensive text books that cover the basic theory of linearized elasticity, and in this thesis the presentation is based on the textbooks by Ciarlet [53] and Coussy [56].

The core problem in the theory of elasticity is to find the relation between the position

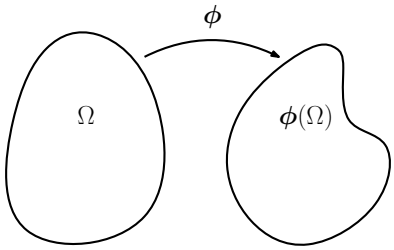


Figure 2.1: Two-dimensional domain in reference Ω and deformed $\boldsymbol{\phi}(\Omega)$ configuration.

of an elastic body after forces have been applied to it and its reference configuration, i.e., the state of the body in the absence of external forces. To this end, define the reference configuration of the elastic body as the domain Ω, and assume that the deformation can be defined by the differentiable vector-field $\boldsymbol{\phi} : \Omega \to \mathbb{R}^d$, where $\boldsymbol{\phi}(\Omega)$ is the deformed elastic body, and d is the spatial dimension, see Figure 2.1. It is common to work with the displacement of particles between the reference configuration and the deformed configuration, $\boldsymbol{u}(\boldsymbol{x}) := \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{x}$, where $\boldsymbol{x}$ is a position in the reference configuration, instead of taking the deformation function into account.

A central construction in the theory of elasticity is the *strain* tensor which gives a measure on the relative displacement of points in a body due to deformation. Consider a curve $\boldsymbol{\gamma}(t)$, parametrized by t, in the reference configuration and the deformed curve $\boldsymbol{\phi}(\boldsymbol{\gamma}(t))$. The lengths of the curves are given as

$$\text{length}(\boldsymbol{\gamma}) = \int_{t_0}^{t_1} |\boldsymbol{\gamma}'(t)| dt = \int_{t_0}^{t_1} (\boldsymbol{\gamma}'(t)_i \boldsymbol{\gamma}'(t)_i)^{\frac{1}{2}}\, dt,$$

and

$$\text{length}(\boldsymbol{\phi}(\boldsymbol{\gamma})) = \int_{t_0}^{t_1} \left| \frac{d}{dt} \boldsymbol{\phi}(\boldsymbol{\gamma}(t)) \right| dt = \int_{t_0}^{t_1} |\nabla \boldsymbol{\phi}(\boldsymbol{\gamma}(t)) \boldsymbol{\gamma}'(t)|\, dt = \int_{t_0}^{t_1} (\mathcal{C}_{ij} \boldsymbol{\gamma}'(t)_i \boldsymbol{\gamma}'(t)_j)^{\frac{1}{2}}\, dt,$$

where $\boldsymbol{\gamma}(t_0)$ and $\boldsymbol{\gamma}(t_1)$ are two points on the curve, the Einstein rule for summation is applied and the tensor $\mathcal{C} := (\nabla \boldsymbol{\phi})^\top \nabla \boldsymbol{\phi}$ is the right Cauchy-Green strain tensor. From this, one can define the infinitesimal length element $d\boldsymbol{l}$ in the reference configuration and in the deformed configuration $d\boldsymbol{l}^{\boldsymbol{\phi}}$ as

$$d\boldsymbol{l} = \left(d\boldsymbol{x}^\top d\boldsymbol{x}\right)^{\frac{1}{2}}, \quad \text{and} \quad d\boldsymbol{l}^{\boldsymbol{\phi}} = \left(d\boldsymbol{x}^\top \mathcal{C} d\boldsymbol{x}\right)^{\frac{1}{2}}.$$

Observe here, that the tensor $\mathcal{C}$ gives information about the strain of the material. Particularly, if $\mathcal{C} = \boldsymbol{I}$ the material has not been strained. Such motions are characterized (by Theorem 1.8-1 in [53]) as rigid-body motions, which are defined by the deformation

function

$$\boldsymbol{\phi}_{\text{rigid}}(\boldsymbol{x}) := \mathbf{a} + \mathbf{Q}\boldsymbol{x}$$

for some translation $\mathbf{a} \in \mathbb{R}^d$ and rotation $\mathbf{Q} \in \mathbb{R}^{d\times d}$, where $\mathbf{Q}$ is an orthogonal matrix.

With this in mind, the Green-St. Venant strain tensor is defined as $E := \dfrac{\mathcal{C} - \boldsymbol{I}}{2}$, and gives a measure on a deformations deviation from a rigid body motion. Moreover, it is common to work with the Green-St. Venant strain tensor given in terms of displacement

$$E(\boldsymbol{u}) = \frac{\mathcal{C}(\boldsymbol{u}) - \boldsymbol{I}}{2} = \frac{(\nabla \boldsymbol{u} + \boldsymbol{I})^\top (\nabla \boldsymbol{u} + \boldsymbol{I}) - \boldsymbol{I}}{2} = \frac{\nabla \boldsymbol{u}^\top \nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top + \nabla \boldsymbol{u}}{2}.$$

For the theory of linearized elasticity, one now assumes that the deformation gradient is small $|\nabla \boldsymbol{u}| \ll 1$, and thereby define the linearized strain tensor as

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) := \frac{\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top}{2} \approx E(\boldsymbol{u}), \quad \text{for} \quad |\nabla \boldsymbol{u}| \ll 1. \tag{2.1}$$

One of the fundamental concepts of continuum mechanics is the linear momentum balance. It states that the momentum of an object is balanced by the forces F acting on it. To express this in mathematical terms let $\omega \subseteq \Omega$ be an arbitrary control volume. The momentum related to the control volume ω is given by

$$\int_\omega (\rho \partial_t \boldsymbol{u}) \; d\boldsymbol{x} = \rho \int_\omega \partial_t \boldsymbol{u} \; d\boldsymbol{x},$$

where the mass density ρ is assumed to be constant in (time and) space, and $\partial_t \boldsymbol{u}$ represents the derivative of $\boldsymbol{u}$ with respect to time. The forces that act on the control volume consist of external body forces

$$\int_\omega \boldsymbol{f} \; d\boldsymbol{x},$$

and traction forces on the boundary of the control volume due to deformation of the remaining volume $\Omega \backslash \omega$

$$\int_{\partial\omega} \boldsymbol{\sigma} \boldsymbol{n} \; ds,$$

where $\boldsymbol{\sigma}$ is known as the symmetric Cauchy stress tensor, and $\boldsymbol{n}$ is the outward pointing normal vector on $\partial\omega$. Hence, the balance of momentum states that

$$\rho \partial_t \int_\omega \partial_t \boldsymbol{u} \; d\boldsymbol{x} = \int_\omega \boldsymbol{f} \; d\boldsymbol{x} + \int_{\partial\omega} \boldsymbol{\sigma} \boldsymbol{n} \; ds$$

for all control volumes $\omega \subseteq \Omega$. By applying the divergence theorem, the momentum balance equation becomes

$$\rho \partial_t^2 \boldsymbol{u} - \nabla \cdot \boldsymbol{\sigma} = \boldsymbol{f},$$

in differential form. If one, furthermore, ignores inertial effects ($\partial_t^2 \boldsymbol{u} = 0$), as will be done for all models considered later in the dissertation under the assumption that the deformation of the elastic body is so slow that the system is in elastic equilibrium at all times, the momentum balance equation is given by

$$-\nabla \cdot \boldsymbol{\sigma} = \boldsymbol{f}. \tag{2.2}$$

To close the system, a constitutive relation between the Cauchy stress tensor and the material strain is needed. The most common one is known as Hooke's law, which assumes a component-wise linear relation between stress and strain

$$\boldsymbol{\sigma} = \mathbb{C}\boldsymbol{\varepsilon}. \tag{2.3}$$

The tensor $\mathbb{C}$ is of fourth order, and is usually called the elasticity tensor or the stiffness tensor. In Einstein's summation notation, (2.3) reads

$$\sigma_{ij} = \mathbb{C}_{ijkl}\varepsilon_{kl},$$

and one can count that in three spatial dimensions ($d = 3$) the elasticity tensor contains the information of 81 relations. However, due to the symmetry of $\boldsymbol{\sigma}$, as a result of the balance of angular momentum, and $\boldsymbol{\varepsilon}$, there are only 21 independent relations in $\mathbb{C}$. This number can be further reduced depending on the material.

For the special case of isotropic materials, one can show that the elasticity tensor is determined by only two independent coefficients, and the stress-strain relation is often given in terms of the Lamé parameters

$$\boldsymbol{\sigma} = 2G\boldsymbol{\varepsilon} + \lambda \mathrm{tr}(\boldsymbol{\varepsilon})\boldsymbol{I}$$

where λ is the first Lamé parameter, and G is either referred to as the second Lamé parameter or the shear modulus.

Finally, one can compute the work that is required to deform a material and find that the potential elastic energy related to a deformed configuration is given as

$$\mathcal{E}_{\mathrm{e}} = \frac{1}{2}\int_{\Omega} \boldsymbol{\varepsilon} : \mathbb{C}\boldsymbol{\varepsilon} \, d\boldsymbol{x}. \tag{2.4}$$

2.1.2 Single-phase flow in porous materials

Since the geometry of porous materials such as tissue, rock, and sand is highly complex and difficult to determine precisely, it is common to describe flow processes through it as an up-scaled continuum model. This is commonly done by considering representative elementary volumes (REV) of a certain size that contains both void spaces (potentially filled with fluids) and parts of the solid material. Here, a very brief introduction to the theory of single-phase flow through porous media is provided. Text-books on the subject include [56, 58, 113, 125].

One of the most basic notions of a porous material is its porosity. This is defined as the ratio between the volume of all of the void spaces in a REV and the total volume of the REV. However, in the case of a fully saturated porous medium, the porosity is interchangeable with the volumetric fluid content, which will be denoted by θ in this dissertation. The standard equation of continuity for flow in porous media is the mass balance equation, which in terms of the volumetric fluid content, is given by the equation

$$\partial_t \int_\omega \rho_w \theta \, d\boldsymbol{x} = -\int_{\partial\omega} \rho_w \boldsymbol{q} \cdot \boldsymbol{n} \, ds + \int_\omega s_f \, d\boldsymbol{x},$$

where ω is a control volume, ρ_w is the fluid mass density, $\boldsymbol{q}$ is the volumetric fluid flux, $\boldsymbol{n}$ is the outward pointing normal vector to ω and s_f represents any sources or sinks. Utilizing now the divergence theorem and that the mass balance equation holds for all control volumes $\omega \subseteq \Omega$, as well as assuming that the density is constant, the equations in differential form reduce to

$$\partial_t \theta + \nabla \cdot \boldsymbol{q} = S_f, \tag{2.5}$$

where $S_f = \dfrac{s_f}{\rho_w}$.

The basic constitutive law for flow in porous materials is the Darcy law which gives a relation between the fluid flux $\boldsymbol{q}$ and the pore pressure p:

$$\boldsymbol{q} = -\kappa(\nabla p - \boldsymbol{g}), \tag{2.6}$$

where $\boldsymbol{g}$ is the gravitational vector, and the constant κ is proportional to the permeability of the porous medium and inversely proportional to the viscosity of the fluid.

Remark 2.1.1. *It might at first glance seem strange to include the time-derivative of the volumetric fluid content for the mass balance equation* (2.5) *when the porous material is fully saturated. However, in the context of deformable porous media, also known as poromechanics, the volumetric fluid content will change as the material deforms. To account for this, a constitutive relation between the volumetric fluid content, the pore*

pressure, and the material displacement will be given in Section 2.2.1.

2.1.3 Phase-field modelling

A phase-field is a regularization of an indicator function that represents certain physical configurations of a system. In this dissertation, two different settings are considered. The first is the variational approach to fracture in which the phase-field captures fractured and unbroken parts of an elastic material, and the second is the Cahn-Larché equations where the phase-field tracks to two distinct elastic solid materials that move due to interface forces or external forces, or change phase due to reactions. The former situation is discussed in Section 2.2.2, and the latter in Section 2.2.3. Here, some general features of phase-field models will be defined together with the classical Allen-Cahn and Cahn-Hilliard equations that arose in connection to the works [3, 4, 43, 45, 46] motivated by the modelling of spinodal decomposition in binary alloys.

To discuss phase-field modeling, some preliminary knowledge of variational modeling and minimization in Hilbert spaces is required. To that end, let H be a Hilbert space and consider the following minimization problem: Let $\mathcal{F} : H \to \mathbb{R}$ and find $y \in H$ such that $\mathcal{F}(y) \leq \mathcal{F}(\tilde{y})$ for all $\tilde{y} \in H$. If $\mathcal{F}$ is convex that corresponds to solving the variational problem: Find $y \in H$ such that

$$\lim_{\delta \to 0} \frac{\mathcal{F}(y + \delta \bar{y}) - \mathcal{F}(y)}{\delta} = 0, \quad \forall \bar{y} \in H. \tag{2.7}$$

The limit above is known as the Gateaux derivative, or variational derivative, and here, the notation

$$\langle \mathcal{D}\mathcal{F}(y), \bar{y} \rangle := \lim_{\delta \to 0} \frac{\mathcal{F}(y + \delta \bar{y}) - \mathcal{F}(y)}{\delta} = \left[\frac{\partial}{\partial \delta} \mathcal{F}(y + \delta \bar{y}) \right]_{|_{\delta = 0}},$$

is used. Moreover, the minimization problem can be written using "arg min"-notation, i.e., "the argument that minimizes", as

$$y = \operatorname*{arg\,min}_{s \in H} \mathcal{F}(s).$$

The equation (2.7) is known as the optimality condition to the minimization problem.

Remark 2.1.2. *It is also common to consider constrained minimization problems. When that is the case, the optimality condition* (2.7) *needs to be altered slightly to take into account that the test space and the Hilbert space containing the minimizer might not be the same.*

Example 2.1.1. Let $H = H_0^1(\Omega)$, where $H_0^1(\Omega)$ is the space of weakly differentiable functions on Ω with vanishing trace on the boundary $\partial\Omega$, and define the function $\mathcal{F} : H_0^1(\Omega) \to \mathbb{R}^+$ such that $\mathcal{F}(z) = \int_\Omega z^2 + |\nabla z|^2 \, dx$. Then, the variational derivative, with $z, v \in H_0^1(\Omega)$, is given as

$$\begin{aligned}
\langle \mathcal{DF}(z), v \rangle &= \left[\frac{\partial}{\partial \delta} \int_\Omega (z + \delta v)^2 |\nabla (z + \delta v)|^2 \, dx \right]_{|\delta=0} \\
&= \left[\frac{\partial}{\partial \delta} \int_\Omega z^2 + 2\delta v z + \delta^2 v^2 + |\nabla z|^2 + 2\delta \nabla v \cdot \nabla z + \delta^2 |\nabla v|^2 \, dx \right]_{|\delta=0} \\
&= \int_\Omega 2zv + 2\nabla z \cdot \nabla v \, dx = \int_\Omega 2(z - \Delta z) v \, dx,
\end{aligned}$$

and the notation $\mathcal{DF}(z) = 2(z - \Delta z)$ is used.

Consider now a domain Ω that is composed of two subdomains Ω_a and Ω_b, $\Omega_a \cup \Omega_b = \Omega$, that only intersect on a common lower-dimensional surface Γ, $\Omega_a \cap \Omega_b = \Gamma$. One can then define a phase-field as a regularized indicator function that describes the location of each of the two subdomains. In other words, the phase-field will be a function $\varphi \in H^1(\Omega)$ that in some sense is an approximation to the indicator function

$$\chi(x) = \begin{cases} 1, & x \in \Omega_a \backslash \Gamma \\ -1, & x \in \Omega_b \end{cases}.$$

Remark 2.1.3. *Note that the choice of value on the interface of the indicator function χ is insignificant as it will be regularized.*

As an example, consider the domain $\Omega = [-1, 1]$, with $\Omega_a = [0, 1]$ and $\Omega_b = [-1, 0]$, and let $\chi : \Omega \to [-1, 1]$ be the indicator function

$$\chi(x) = \begin{cases} 1, & x > 0 \\ -1, & x \leq 0. \end{cases}$$

The function, $\chi(x)$ is clearly in $L^2(\Omega)$, but it does not belong to $H^1(\Omega)$ as it does not have a weak derivative. To see this, recall the definition of a weak derivative:

Definition 2.1.1 (Weak derivative)**.** Let $g \in L^1_{\text{loc}}(\Omega)$ (the space of locally integrable functions) and $\boldsymbol{\beta}$ be a multi-index. The function g has a $\boldsymbol{\beta}$-th weak derivative if there exists a $w \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_\Omega g \partial_{\boldsymbol{\beta}} v \, dx = (-1)^{|\boldsymbol{\beta}|} \int_\Omega w v \, dx,$$

for all $v \in C_c^\infty(\Omega)$ (compactly supported smooth functions on Ω).

Assuming that χ has a weak derivative, there exists a $w \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_{-1}^{1} \chi v' \, dx = -\int_{-1}^{1} wv \, dx.$$

However, $\int_{-1}^{1} \chi v' \, dx = -2v(0)$ which gives

$$2v(0) = \int_{-1}^{1} wv \, dx. \tag{2.8}$$

If $w \in L^1_{\text{loc}}(\Omega)$ then $\lim_{r\to 0} \int_{-r}^{r} |w| \, dx = 0$, hence there exists a $\delta > 0$ such that $\int_{-\delta}^{\delta} |w| \, dx < 1$. Choosing a function $v \in C_c^\infty(\Omega)$ such that $\text{supp}(v) = [-\delta, \delta]$ and $\max(|v|) = v(0) = 1$ equation (2.8) gives

$$2 = 2v(0) = \int_{-1}^{1} wv \, dx = \int_{-\delta}^{\delta} wv \, dx \leq \max(v) \int_{-\delta}^{\delta} |w| \, dx < 1,$$

which is a contradiction. Hence, χ cannot be in $H^1(\Omega)$.

Remark 2.1.4. *Equation* (2.8) *is sufficient to realize that w is a scaled Dirac delta function.*

The objective is now to search for a phase-field $\varphi \in H^1(\Omega)$ that is close to χ in the $L^2(\Omega)$-norm, while still being a function in $H^1(\Omega)$. This is done by solving a minimization problem of the form

$$\varphi = \underset{s \in H^1(\Omega)}{\arg\min} \left(a_1 \|s - \chi\|^2_{L^2(\Omega)} + a_2 \|\nabla s\|^2_{L^2(\Omega)} \right), \tag{2.9}$$

where a_1, and a_2 are weights that determine the importance of φ being close to χ in the $L^2(\Omega)$-norm compared to how small the gradient of it should be. The optimality conditions of the problem read as follows: Find $\varphi \in H^1(\Omega)$ such that

$$a_1 (\varphi, q) + a_2 (\nabla\varphi, \nabla q) = a_1 (\chi, q), \quad \forall q \in H^1(\Omega),$$

where $(\cdot, \cdot)$ denotes the $L^2(\Omega)$ inner-product. Notice that by choosing $q = 1$ (or any other constant function) one obtains that the mean of the phase-field is equal to the mean of the function it approximates $(\varphi, 1) = (\chi, 1)$. An approximation to the solution of this problem for $a_1 = a_2 = 1$ is presented in Figure 2.2a, and the need to properly choose the weights is clear.

To get an idea of how to do this, consider the function

$$\tilde{\varphi}_\ell(x) = \begin{cases} -1, & x \leq -\frac{\ell}{2}, \\ \frac{2}{\ell}x, & x \in [-\frac{\ell}{2}, \frac{\ell}{2}], \\ 1, & x > \frac{\ell}{2}, \end{cases}$$

with $\ell \in (0, 2)$. This function transitions between the two phases in a region of width ℓ and the value of the potential in (2.9) for it is easily computed. More importantly, by comparing the values of the two terms in the minimization problem, it becomes clear how the weights can be chosen as functions of ℓ. The first term in (2.9) evaluated at $\tilde{\varphi}_\ell$ takes the value

$$a_1 \|\tilde{\varphi}_\ell - \chi\|^2_{L^2(\Omega)} = a_1 \int_{-\frac{\ell}{2}}^{0} \left(\frac{2}{\ell}x + 1\right)^2 \, dx + \int_{0}^{\frac{\ell}{2}} \left(\frac{2}{\ell}x - 1\right)^2 \, dx = a_1 \frac{\ell}{3},$$

and the second term

$$a_2 \|\nabla \tilde{\varphi}_\ell\|^2_{L^2(I)} = a_2 \int_{-\frac{\ell}{2}}^{\frac{\ell}{2}} \frac{4}{\ell^2} \, dx = a_2 \frac{4}{\ell}.$$

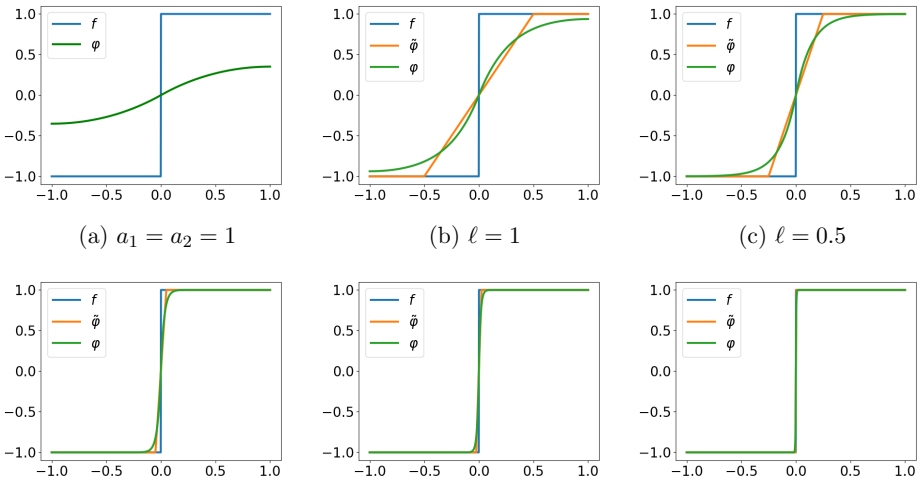
Hence, to make the two terms comparable, the choice of weights $a_1 = \frac{3}{\ell}$ and $a_2 = \frac{\ell}{4}$ allows control of the regularization width, ℓ. See Figure 2.2 for solutions to the minimization problem

$$\varphi = \underset{s \in H^1(\Omega)}{\arg\min} \left(\frac{3}{\ell} \|s - \chi\|^2_{L^2(\Omega)} + \frac{\ell}{4} \|\nabla s\|^2_{L^2(\Omega)} \right), \tag{2.10}$$

for different values of ℓ.

Remark 2.1.5. *The values* 3 *and* 4 *in* (2.10) *are in principle insignificant as the cases* $\ell \ll 1$ *or "the sharp interface limit"* $\ell \to 0$ *are of interest.*

Example 2.1.2 (The Allen-Cahn and Cahn-Hilliard equations)**.** As examples of classical phase-field models, the Allen-Cahn and Cahn-Hilliard equations are presented here. They can both be seen as models for interfaces that move due to surface tension, and were originally developed in the context of binary alloys [4, 44]. In recent years, however, it has been more common to consider the equations as a part of coupled diffuse interface models such as the Cahn-Larché equations [88, 146], tumor growth models [126, 160], two-phase flow [48, 57, 135] and Navier-Stokes-Cahn-Hilliard [156]. Suppose that two materials in a domain $\Omega \subseteq \mathbb{R}^d$ are separated by an interface Γ, and that surface tension or other interface forces are acting such that the preferred state is one where the lower-dimensional interface area $\mathcal{I}^{d-1}(\Gamma)$ is minimized. By introducing a phase-field that takes the value 1 for one of the materials and -1 for the other material, and transitions



(a) $a_1 = a_2 = 1$ (b) $\ell = 1$ (c) $\ell = 0.5$

(d) $\ell = 0.1$ (e) $\ell = 0.05$ (f) $\ell = 0.01$

Figure 2.2: (a): Solution to (2.9) for $a_1 = a_2 = 1$. (b) – (f): Solutions to the minimization problem (2.10) for different values of ℓ.

smoothly between, the interface area can be approximated, as in [4], by

$$\mathcal{I}^{d-1}(\Gamma) \approx \int_\Omega \frac{1}{\ell}\Psi(\varphi) + \frac{\ell}{2}|\nabla\varphi|^2 \, dx =: \mathcal{I}^{d-1}(\varphi), \tag{2.11}$$

where ℓ is a positive constant that is related to the width of the transition region for φ, and $\Psi(\varphi)$ is known as a double-well function that is evaluated to 0 in the two pure phases and is otherwise positive. A classical example is the function

$$\Psi(\varphi) := (1 - \varphi^2)^2. \tag{2.12}$$

Assume now that the system has an initial configuration that is described by the phase-field function $\varphi_0 \in H^1(\Omega)$, and then evolves in the direction of steepest decent of the interface area, with mobility m and surface tension γ, i.e.,

$$(\partial_t\varphi, q) = -m\gamma \left\langle \mathcal{D}\mathcal{I}^{d-1}(\varphi), q \right\rangle \quad \forall q \in H^1(\Omega), \tag{2.13}$$

where $(\cdot,\cdot)$ is the $L^2(\Omega)$ inner product, and $\left\langle \mathcal{D}\mathcal{I}^{d-1}(\varphi), q \right\rangle$ is the variational derivative of $\mathcal{I}^{d-1}(\varphi)$ with test function q. The equation (2.13) is the variational form of the Allen-Cahn equation, which corresponds to the partial differential equation

$$\partial_t\varphi + \frac{\gamma m}{\ell}\Psi'(\varphi) - \gamma m\Delta\varphi = 0,$$

with initial condition $\varphi(0) = \varphi_0$ and homogeneous Neumann conditions $\nabla\varphi \cdot \boldsymbol{n} = 0$ on $\partial\Omega$ ($\boldsymbol{n}$ representing the outward-pointing normal vector) in strong form.

Note that there is nothing in the Allen-Cahn equation, as it is given here, that prevents phases form disappearing. Indeed, this will eventually happen if the initial configuration contains more of one phase than the other, since the smallest interface area is obtained by having no interface at all. Remedies could be to add a source term to the equation, or a Neumann boundary condition that enforces influx of a phase.

An alternative is to prescribe a phase-conservation law

$$\partial_t\varphi + \nabla \cdot \boldsymbol{J} = 0, \tag{2.14}$$

with a phase-field flux $\boldsymbol{J}$ that is assumed to follow Fick's law of chemical diffusion [69, 129]

$$\boldsymbol{J} = -m\nabla\mu, \tag{2.15}$$

and have no-flux boundary conditions. Here, m is again the mobility, and μ is the potential that is defined as the rate of change of free energy with respect to changes in phase. In this setting, the free energy is given as the interface tension parameter γ multiplied by the interface measure $\mathcal{I}^{d-1}(\varphi)$, i.e., the potential is defined through the equation

$$(\mu, q) = \gamma\langle\mathcal{D}\mathcal{I}^{d-1}(\varphi), q\rangle, \quad \forall q \in H^1(\Omega). \tag{2.16}$$

By substituting equation (2.15) into equation (2.14) and multiplying with test functions from $H^1(\Omega)$, the variational equations

$$\begin{aligned} (\partial_t\varphi, q^\varphi) + (m\nabla\mu, \nabla q^\varphi) &= 0, \quad \forall q^\varphi \in H^1(\Omega) && (2.17)\\ (\mu, q^\mu) - \left(\frac{\gamma}{\ell}\Psi'(\varphi), q^\mu\right) - (\gamma\ell\nabla\varphi, \nabla q^\mu) &= 0, \quad \forall q^\mu \in H^1(\Omega), && (2.18) \end{aligned}$$

are obtained. This is known as the Cahn-Hilliard model and the corresponding strong partial differential equation is given as

$$\partial_t\varphi - \nabla \cdot m\nabla\left(\frac{\gamma}{\ell}\Psi'(\varphi) - \gamma\ell\Delta\varphi\right) = 0,$$

with initial condition $\varphi_0 \in H^1(\Omega)$ and boundary conditions $\nabla\mu \cdot \boldsymbol{n} = \nabla\varphi \cdot \boldsymbol{n} = 0$ on $\partial\Omega$.

2.2 Coupled problems

Here, several models that treat the coupling between the models from Section 2.1 are discussed. First, poroelasticity, and specifically the quasi-static Biot equations, are described. Then, in Section 2.2.2, the variational approach to fracture propagation is discussed, and finally the Cahn-Larché equation which couples the Cahn-Hilliard equation (2.17)–(2.18) to linearized elasticity are defined.

2.2.1 Poroelasticity and the quasi-static Biot equations

Poroelasticity governs flow in a deformable porous material. In its most general setting, poroelasticity governs both finite strain elasticity [27, 154], and partially saturated [33, 35] and multi-phase flow [100, 103]. Here, only the restriction to linearized elasticity, from Section 2.1.1, and fully saturated single-phase flow, from Section 2.1.2, is discussed. The resulting model is often known as the quasi-static linearized Biot equations, and in addition to the equations that have been discussed in Sections 2.1.1 and 2.1.2 two constitutive laws will be introduced that describe the relationship between material displacement and fluid flow. The standard textbook on poroelasticity is [56].

The change in volumetric fluid content is assumed to be proportional to both the change in pore pressure and the divergence of material displacement, i.e.,

$$\partial_t \theta = \partial_t \left(\frac{p}{M} + \alpha \nabla \cdot \boldsymbol{u} \right), \tag{2.19}$$

where α is the Biot-Willis coupling coefficient and M is a compressibility coefficient. Moreover, the solid matrix is assumed to be isotropically stressed by the pore pressure, which is accounted for using the principle of effective stress

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_e - \alpha p \boldsymbol{I}. \tag{2.20}$$

Here, $\boldsymbol{\sigma}_e$ is the effective stress tensor that is given by Hooke's law

$$\boldsymbol{\sigma}_e = \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}).$$

Notably, the coefficient α appears in both the volumetric fluid balance equation and the stress-strain relationship, which eventually leads to a saddle-point structure of the total problem [39]. Combining the linear momentum balance (2.2) with the modified Hooke's law (2.20), and the volumetric fluid balance equation (2.5) with the relation (2.19) and Darcy's law (2.6), the Biot consolidation model (see e.g., [56]) reads: Find $\boldsymbol{u}$ and p such

that

$$-\nabla \cdot \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) + \alpha \nabla p = \boldsymbol{f}, \tag{2.21}$$

$$\partial_t \left(\frac{p}{M} + \alpha \nabla \cdot \boldsymbol{u} \right) - \nabla \cdot (\kappa (\nabla p - \boldsymbol{g})) = S_f, \tag{2.22}$$

with initial conditions for pore pressure and displacement $p_0, \boldsymbol{u}_0$, Dirichlet boundary conditions $p = p_D$ on $\Gamma_D^p \subseteq \partial\Omega$ and $\boldsymbol{u} = \boldsymbol{u}_D$ on $\Gamma_D^{\boldsymbol{u}} \subseteq \partial\Omega$ and Neumann boundary conditions $\kappa (\nabla p - \boldsymbol{g}) \cdot \boldsymbol{n} = \boldsymbol{q}_N$ on $\partial\Omega \backslash \Gamma_D^p$ and $(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) - \alpha p \boldsymbol{I}) \cdot \boldsymbol{n} = \boldsymbol{\sigma}_N$ on $\partial\Omega \backslash \Gamma_D^{\boldsymbol{u}}$, $\boldsymbol{n}$ being the outward pointing normal vector.

2.2.2 The variational approach to brittle fracture propagation

In this section, a variational phase-field framework is presented to model fracture propagation in a brittle elastic material. The framework follows the theory of Griffith for fracture propagation [89], and the variational model was proposed by Bourdin, Francfort and Marigo in [36, 37, 71]. The fracture is tracked by a phase-field function $\varphi \in H^1(\Omega)$ that takes the value $\varphi = 1$ in the fractured part of the domain, $\varphi = 0$ in "healthy/intact" part of the domain, and transitions smoothly between the values in a region that will be controlled by the model parameter ℓ. This allows for modeling fractures without the need for conforming meshes or special path-tracking algorithms (as e.g. XFEM [70, 85, 86, 123]) when considering numerical discretization.

Let Ω be an elastic domain, and $\mathcal{S} \subseteq \Omega$ a lower-dimensional manifold in Ω that represents the fractured region. The variational model is derived as an energy minimization problem following Griffith's theory [89]. Let $\mathcal{S}_0$ be the initial crack and suppose that the medium is subject to loading through traction forces $\boldsymbol{\tau}$. The free energy of the system is the additive combination of the elastic (bulk) energy (which is extended to account for body forces and traction compared to (2.4))

$$\mathcal{E}_{\text{bulk}}(\boldsymbol{u}, \mathcal{S}) = \int_{\Omega \backslash \mathcal{S}} \boldsymbol{\varepsilon}(\boldsymbol{u}) : \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) \, dx - \int_{\Omega} \boldsymbol{f} \cdot \boldsymbol{u} \, dx - \int_{\Gamma} \boldsymbol{\tau} \cdot \boldsymbol{u} \, ds, \tag{2.23}$$

and the potential energy related to the fracture

$$\mathcal{E}_{\text{crack}}(\mathcal{S}) = G_c \mathcal{H}^{d-1}(\mathcal{S}), \tag{2.24}$$

where $\boldsymbol{u}$ is displacement, $\boldsymbol{\varepsilon}$ is the material strain, $\mathbb{C}$ is the elasticity tensor, $\boldsymbol{f}$ accounts for external body forces, τ represents traction to a part of the boundary $\Gamma \subseteq \partial\Omega$, G_c is the critical energy release rate, and $\mathcal{H}^{d-1}(\mathcal{S})$ is the lower-dimensional Hausdorff measure of the crack surface. In the variational approach to fracture [71], one then searches for

the pair of displacement $\boldsymbol{u}$ and fracture $\mathcal{S}$ such that the total energy is minimized

$$(\boldsymbol{u}, \mathcal{S}) = \arg\min_{\boldsymbol{w}, \mathcal{V}} \left(\mathcal{E}_{\text{bulk}}(\boldsymbol{w}, \mathcal{V}) + \mathcal{E}_{\text{crack}}(\mathcal{V}) \right), \tag{2.25}$$

subject to the non-healing constraint $\mathcal{S}_0 \subseteq \mathcal{S}$.

By regularizing the fracture with a phase-field φ, there are three components of the minimization problem (2.25) that need to be altered:

- First is the restricted integral in the bulk energy

$$\int_{\Omega \setminus \mathcal{C}} \boldsymbol{\varepsilon}(\boldsymbol{u}) : \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) \, dx \approx \int_{\Omega} g(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) : \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) \right) \, dx, \tag{2.26}$$

  where $g(\varphi)$ is called a degradation function that should satisfy $g(1) = 0$ and $g(0) = 1$. Typically, it is chosen as $g(\varphi) = (1 - \varphi)^2$, but other choices have also been discussed in the literature [137]. Moreover, a small artificial parameter ζ is often introduced to the degradation function $g_\zeta(\varphi) = (1 - \zeta)(1 - \varphi)^2 + \zeta$ to make sure that the elasticity subproblem does not degenerate.

- The second is the lower-dimensional Hausdorff-measure in (2.24), $\mathcal{H}^{n-1}(\mathcal{S})$, which will be approximated by

$$\mathcal{H}^{d-1}(\mathcal{S}) \approx \int_{\Omega} \frac{1}{2\ell} \varphi^2 + \frac{\ell}{2} |\nabla \varphi|^2 \, dx. \tag{2.27}$$

  This choice of functional is motivated by the work of Ambrosio and Tortorelli [7, 8] on a similar type of problem in image segmentation. An alternative is to exchange the quadratic term φ^2, with a linear term $\varphi = 1$ (see, for example, [149]).

- Finally, the non-healing constraint on the fracture, $\mathcal{S}_0 \subseteq \mathcal{S}$, is typically approximated by the inequality $\varphi \geq \varphi^0$ a.e. in Ω, where φ^0 is the approximation of the initial fracture. There have been suggestions that this might be too restrictive, as the physical meaning of $\varphi \in (0, 1)$ is unclear. A remedy is to only enforce the restriction to phase-field values that exceed some user-defined threshold [137].

Remark 2.2.1. *To initialize the fracture, one can either include the initial fracture into the domain or create an indicator function that represents it and regularize it by using the theory from Section 2.1.3.*

By choosing the solution space for displacement as

$$H^1_{\tilde{\boldsymbol{u}}}(\Omega)^d = \left\{ \boldsymbol{w} \in \left(H^1(\Omega) \right)^d : \operatorname{tr}(\boldsymbol{w}) = \tilde{\boldsymbol{u}} \right\},$$

where the $\tilde{\boldsymbol{u}}$ is the Dirichlet boundary conditions on $\boldsymbol{u}$ and the trace is taken on the appropriate part of the boundary, the regularized minimization problem becomes

$$(\boldsymbol{u},\varphi) = \underset{\boldsymbol{w}\in H^1_{\tilde{\boldsymbol{u}}}(\Omega)^d, s\in H^1(\Omega)}{\arg\min} \left(\int_\Omega g(s)\left(\boldsymbol{\varepsilon}(\boldsymbol{w}) : \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{w}) - \boldsymbol{f}\cdot\boldsymbol{w}\right)\, dx - \int_\Gamma \boldsymbol{\tau}\cdot\boldsymbol{w}\, ds \right.$$
$$\left. + G_c \int_\Omega \frac{1}{2\ell}s^2 + \frac{\ell}{2}|\nabla s|^2\, dx \right),$$

subject to $\varphi \geq \varphi^0$ a.e. in Ω. If several loading steps are performed, the loading step indicator n is introduced, and the evolution problem reads: Given $\varphi^{n-1} \in H^1(\Omega)$ solve

$$(\boldsymbol{u}^n,\varphi^n) = \underset{\boldsymbol{w}\in H^1_{\tilde{\boldsymbol{u}}}(\Omega)^d, s\in H^1(\Omega)}{\arg\min} \left(\int_\Omega g(s)\left(\boldsymbol{\varepsilon}(\boldsymbol{w}) : \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{w}) - \boldsymbol{f}^n\cdot\boldsymbol{w}\right)\, dx - \int_\Gamma \boldsymbol{\tau}^n\cdot\boldsymbol{w}\, ds \right.$$
$$\left. + G_c \int_\Omega \frac{1}{2\ell}s^2 + \frac{\ell}{2}|\nabla s|^2\, dx \right), \tag{2.28}$$

subject to $\varphi^n \geq \varphi^{n-1}$ a.e. in Ω. The most classical solution strategy for this constrained minimization problem is covered in Section 3.2.5.

Another standard modification of the minimization problem (2.28), is to split the elastic bulk energy $\psi(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}(\boldsymbol{u}) : \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u})$ by an additive decomposition $\psi(\boldsymbol{\varepsilon}) = \psi^+(\boldsymbol{\varepsilon}) + \psi^-(\boldsymbol{\varepsilon})$, and only let the degradation function $g(\varphi)$ act on the $\psi^+(\boldsymbol{\varepsilon})$ part, i.e., the term $g(s)\psi(\boldsymbol{\varepsilon}(\boldsymbol{w}))$ is replaced by $g(s)\psi^+(\boldsymbol{\varepsilon}(\boldsymbol{w})) + \psi^-(\boldsymbol{\varepsilon}(\boldsymbol{w}))$ in (2.28). There are several proposed splits in the literature, but the most popular are the spectral split into tensile ψ^+ and compressive ψ^- parts from [119], and the split into volumetric ψ^+ and deviatoric ψ^+ parts from [9].

2.2.3 The Cahn-Larché equations

The Cahn-Larché system can be seen as a combination of a Cahn-Hilliard equation and linearized elasticity with infinitesimal strains and displacements [81, 109]. Assume that the domain Ω is occupied by two elastic materials that cover the subdomains Ω_a and Ω_b, and that the phase-field φ acts as a regularization of the indicator function

$$\chi(x) = \begin{cases} 1, & \text{for} \quad x \in \Omega_a, \\ -1, & \text{for} \quad x \in \Omega\backslash\Omega_a. \end{cases} \tag{2.29}$$

Similar to the Cahn-Hilliard equation, local phase-balance is assumed (2.14);

$$\partial_t \varphi + \nabla \cdot \boldsymbol{J} = R,$$

where $\boldsymbol{J}$ is the phase-field flux and R accounts for reactions. Moreover, the stress follows quasi-static linear momentum balance (ignoring inertial effects) (2.2)

$$-\nabla \cdot \boldsymbol{\sigma} = \boldsymbol{f},$$

where $\boldsymbol{\sigma}$ is the stress tensor and $\boldsymbol{f}$ corresponds to external forces. The free energy $\mathcal{E}(\varphi, \boldsymbol{u})$ of the system is assumed to be an additive combination of the regularized interface energy

$$\mathcal{E}_{\mathrm{I}}(\varphi) = \gamma \mathcal{I}^{d-1}(\varphi), \tag{2.30}$$

and a modified potential elastic energy (2.4)

$$\mathcal{E}_{\mathrm{e}}(\varphi, \boldsymbol{u}) = \frac{1}{2} \int_\Omega \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\left(\varphi - \tilde{\varphi}\right) \boldsymbol{I}\right) : \mathbb{C}(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\left(\varphi - \tilde{\varphi}\right) \boldsymbol{I}\right) \, dx,$$

such that the total free energy becomes

$$\mathcal{E}(\varphi, \boldsymbol{u}) := \mathcal{E}_{\mathrm{I}}(\varphi) + \mathcal{E}_{\mathrm{e}}(\varphi, \boldsymbol{u}). \tag{2.31}$$

Here, the parameter γ is related to the interfacial tension between the two phases and can be considered to account for adhesive/cohesive forces between the phases, and the interface energy $\mathcal{I}^{d-1}(\varphi)$ from (2.11) is applied. Moreover, $\boldsymbol{\varepsilon}(\boldsymbol{u})$ is the linearized symmetric strain tensor (2.1), $\mathbb{C}(\varphi)$ is the elasticity tensor that now depends on the material phase, the term $\xi\left(\varphi - \tilde{\varphi}\right) \boldsymbol{I}$ accounts for swelling effects where $\tilde{\varphi}$ is a reference phase-field, and $\boldsymbol{I}$ is the identity tensor.

As for the Cahn-Hilliard equation, the phase-field flux is assumed to be governed by Fick's law

$$\boldsymbol{J} = -m(\varphi) \nabla \mu,$$

where $m(\varphi)$ is the mobility, and μ is the potential. Finally, the potential μ and the stress tensor $\boldsymbol{\sigma}$ are defined as the rates of change, variational derivatives, of the free energy (2.31) with respect to the phase-field φ and linearized strain $\boldsymbol{\varepsilon}$. Standard computation, as those demonstrated in Example 2.1.1, yield

$$\begin{aligned} \mu := \mathcal{D}_\varphi \mathcal{E}(\varphi, \boldsymbol{u}) \quad &= \quad \gamma \left(\frac{1}{\ell} \Psi'(\varphi) - \ell \Delta \varphi \right) - \xi \boldsymbol{I} : \mathbb{C}(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I}\right) \\ &\quad + \frac{1}{2} \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I}\right) : \mathbb{C}'(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I}\right), \end{aligned}$$

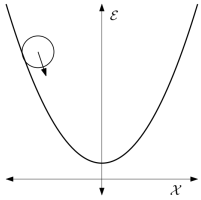


Figure 2.3: Representation of a gradient flow. The ball (representing the energy at some state) moves gradually towards its minimal value, in the opposite direction of its gradient.

and

$$\boldsymbol{\sigma} := \mathcal{D}_{\boldsymbol{\varepsilon}}\mathcal{E}(\varphi, \boldsymbol{\varepsilon}(\boldsymbol{u})) = \mathbb{C}(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\varphi\boldsymbol{I}\right),$$

where vanishing Neumann boundary conditions for the phase-field ($\nabla\varphi \cdot \boldsymbol{n} = 0$ on $\partial\Omega$) are assumed.

In total, the Cahn-Larché equations in strong form are given as

$$\begin{aligned}
\partial_t\varphi - \nabla\cdot(m\nabla\mu) &= R && \text{in} \quad \Omega\times[0,T],\\
\mu + \gamma\left(\ell\Delta\varphi - \frac{1}{\ell}\Psi'(\varphi)\right) - \delta_\varphi\mathcal{E}_{\mathrm{e}}(\varphi,\boldsymbol{u}) &= 0 && \text{in} \quad \Omega\times[0,T],\\
-\nabla\cdot\left(\mathbb{C}(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\varphi\boldsymbol{I}\right)\right) &= \boldsymbol{f} && \text{in} \quad \Omega\times[0,T],
\end{aligned}$$

with initial condition $\varphi = \varphi_0$ in $\Omega\times\{0\}$, and boundary conditions $\nabla\varphi\cdot\boldsymbol{n} = \nabla\mu\cdot\boldsymbol{n} = 0$ and $\boldsymbol{u} = \boldsymbol{u}_D$ on $\Gamma_D \subseteq \partial\Omega$, and $\left(\mathbb{C}(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\varphi\boldsymbol{I}\right)\right)\cdot\boldsymbol{n} = \boldsymbol{u}_N$ on $\partial\Omega\backslash\Gamma_D$. Here,

$$\mathcal{D}_\varphi\mathcal{E}_{\mathrm{e}}(\varphi,\boldsymbol{u}) = \frac{1}{2}\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\varphi\boldsymbol{I}\right):\mathbb{C}'(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\varphi\boldsymbol{I}\right) - \xi\boldsymbol{I}:\mathbb{C}(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\varphi\boldsymbol{I}\right), \tag{2.32}$$

and the elasticity tensor $\mathbb{C}(\varphi)$ is typically depending on the phase-field through some given interpolation function $\pi(\varphi)$; $\mathbb{C}(\varphi) = \mathbb{C}_{-1} + \pi(\varphi)(\mathbb{C}_1 - \mathbb{C}_{-1})$.

Remark 2.2.2. *A further extension of the Cahn-Larché equations are the Cahn-Hilliard-Biot system which either can be seen as a combination of the Cahn-Hilliard phase-field equation and the Biot consolidation model, or as an extension of the Cahn-Larché equations that includes flow though the poroelastic material. This system is developed in Paper D [145].*

2.3 Gradient flows

All of the physical processes that have been covered in this chapter, except for the variational phase-field model for brittle fracture, have an underlying minimization structure

and can be written as generalized gradient flows [34, 127]. A (generalized) gradient flow is an evolution equation where the state of the system evolves in the steepest decent direction of some energy related to the system state subject to a dissipation mechanism, see illustration in Figure 2.3. Identifying this structure in processes can be useful for both abstract analysis of the system [12, 34, 55, 124], and numerical solution strategies [34, 101, 146].

Let H be a Hilbert space. A gradient flow is an equation

$$\langle \partial_t y - \mathcal{P}_{\text{ext}}, \tilde{y} \rangle_D = -\langle \mathcal{D}\mathcal{E}(y), \tilde{y} \rangle, \quad \forall \tilde{y} \in H, \tag{2.33}$$

where $\langle \cdot, \cdot \rangle_D$ is an inner-product that includes information about the dissipation mechanism, $\mathcal{P}_{\text{ext}}$ accounts for external forces, and $\mathcal{E}$ is the energy of the system. The variable y is referred to as the state variable and H as the state space. A general feature of gradient flows is that they enforce dissipation of energy in the absence of external contributions $\mathcal{P}_{\text{ext}} = 0$. Formally, it holds that

$$\partial_t \mathcal{E}(y) = \langle \mathcal{D}\mathcal{E}(y), \partial_t y \rangle = -\langle \partial_t y, \partial_t y \rangle_D \leq 0, \tag{2.34}$$

where the second equality is due to (2.33).

Example 2.3.1 (The Allen-Cahn and the Cahn-Hilliard equations as gradient flows)**.** Both the Allen-Cahn and the Cahn-Hilliard equations, as described in Example 2.1.2, are gradient flows. This can be seen by choosing the phase-field φ as the state, $H^1(\Omega)$ as the state space, and (2.11) as the energy. By utilizing the weighted $L^2(\Omega)$-inner-product $\langle x, y \rangle_{L^2_m(\Omega)} := \int_\Omega \frac{xy}{m}\, dx$ as the dissipation mechanism, one obtains the Allen-Cahn equation (2.13). If the weighted H^{-1}-like inner-product

$$\langle x, y \rangle_{H^{-1}_m(\Omega)} = \int x\tilde{y}\, dx, \tag{2.35}$$

where $\tilde{y}$ is defined through $\int_\Omega y\bar{y}\, dx = \int_\Omega m\nabla\tilde{y}\cdot\nabla\bar{y}\, dx$ for all $\bar{y} \in H^1(\Omega)$, is used instead, then the Cahn-Hilliard equations (2.17)–(2.18) are obtained in a one-field formulation.

To extend the notion of a gradient flow to a generalized gradient flow, it is more convenient to consider it in terms of the dissipation potential $\mathcal{R}(\partial_t y)$. The generalized gradient flow is then defined as the evolution equation where the variational derivative of the dissipation potential with respect to the change of state is equal to the variational derivative of the free energy with respect to the state:

$$\langle \mathcal{D}_{\partial_t y}\mathcal{R}(\partial_t y), \tilde{y} \rangle = -\langle \mathcal{D}_y \mathcal{E}(y), \tilde{y} \rangle, \quad \forall \tilde{y} \in H.$$

Notice that for the special case of a quadratic dissipation potential, $\mathcal{R}(\partial_t y) = \frac{\|\partial_t y\|_D^2}{2}$, the classical gradient flow (2.33) is recovered. An important feature of generalized gradient flows, is that they can be described as minimization problems, see [127],

$$\partial_t y = \underset{s \in H}{\arg\min} \left\{ \mathcal{R}(s) + \langle \mathcal{D}_y \mathcal{E}(y), s \rangle \right\}. \tag{2.36}$$

This is exploited in Section 3.1.1 to define dissipation-preserving time-discretizations.

Chapter 3

Numerical solution strategies for coupled problems

In this chapter, numerical solution strategies for coupled problems are considered. First, in Section 3.1, both temporal and spatial discretization techniques are discussed. Then, solution strategies for discrete systems of equations are presented in Section 3. This section is heavily focused on iterative solution strategies such as decoupling and linearization methods but also covers basics of linear solvers and acceleration methods.

3.1 Discretization techniques

Here, discretization of variational problems is discussed. In Section 3.1.1, temporal discretization techniques are presented. In particular, the θ-methods, including the explicit and implicit Euler methods are discussed, as well as a special structure-preserving scheme that is applied to generalized gradient flows (2.36). Then, in Section 3.1.2, the finite element method for spatial discretization is presented.

3.1.1 Time discretization

Let $\Omega \subseteq \mathbb{R}$ be a domain, $[0,T]$ be a time interval, V be a Sobolev space on Ω (e.g., $V = H_0^1(\Omega)$), and B be the Bochner space $B = H^1([0,T],V)$. Consider now the general variational evolution equation: Given $u(0,x) = u_0$, find $u \in B$ such that

$$(\partial_t u, v) + b(u,v) = K(v), \tag{3.1}$$

for all $v \in V$ and almost all $t \in [0,T]$, where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner-product, $b(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a continuous function that is linear in the second component, and $K(\cdot) : V \to \mathbb{R}$ is a linear function.

Let τ be the time-step size and n denote the time step. By approximating the derivative in time with a backward difference, and evaluating the other time-dependent terms implicitly, the implicit (backward) Euler approximation to the system (3.1) is obtained: Given $u^{n-1} \in V$ (where $u^0 := u_0$), find $u^n \in V$ such that

$$\left(\frac{u^n - u^{n-1}}{\tau}, v\right) + b\left(u^n, v\right) = K(v) \tag{3.2}$$

for all $v \in V$. This method is a special case of the θ-methods that take the form: Given $u^{n-1} \in V$, find $u^n \in V$ such that

$$\left(\frac{u^n - u^{n-1}}{\tau}, v\right) + (1-\theta)\, b\left(u^n, v\right) + \theta b\left(u^{n-1}, v\right) = K(v) \tag{3.3}$$

for all $v \in V$ and some given $\theta \in [0,1]$. By choosing different values for θ different methods are obtained, and the most important choices include the implicit (backward) Euler method $\theta = 0$, the explicit (forward) Euler method $\theta = 1$, and the quadratic Crank-Nicholson (trapezoidal) method $\theta = 0.5$.

When generalized gradient flows are considered, it might be natural to discretize the system in time such that the minimization problem (2.36) naturally is satisfied. This is done by the energy-driven time discretization method (see e.g., [34, 146]): Given $u^{n-1} \in V$ solve

$$u^n = \underset{s \in Y}{\arg\min} \left\{ \tau \mathcal{R}\left(\frac{s - u^{n-1}}{\tau}\right) + \mathcal{E}(s) \right\}. \tag{3.4}$$

This time discretization naturally satisfies the energy dissipation that is inherent in gradient flows (2.34), which can be seen by subtracting the potential in (3.4) evaluated in u^{n-1} from the same potential evaluated in u^n:

$$\begin{aligned} \tau \mathcal{R}\left(\frac{u^n - u^{n-1}}{\tau}\right) + \mathcal{E}(u^n) - \tau \mathcal{R}\left(\frac{u^{n-1} - u^{n-1}}{\tau}\right) - \mathcal{E}(u^{n-1}) &\leq 0 \\ \tau \mathcal{R}\left(\frac{u^n - u^{n-1}}{\tau}\right) + \mathcal{E}(u^n) - \mathcal{E}(u^{n-1}) &\leq 0. \end{aligned}$$

A time-discretization with the property the energy is dissipated over the time steps, i.e., the energy evaluated at consecutive time-steps forms a decreasing sequence, is called a *gradient stable* (or *energy stable*) time discretization. If this holds true without any restrictions on the time-step size, the method is said to be *unconditionally gradient stable*. Moreover, the time-discretization (3.4) corresponds to the implicit Euler method

whenever $\mathcal{R}(s) = \dfrac{\|s\|^2_{L^2(\Omega)}}{2}$ and $\mathcal{E}(u)$ is such that $\langle \delta\mathcal{E}(u^n), q\rangle = b(u^n, q) - K(q)$.

3.1.2 Spatial discretization

For all of the numerical experiments that have been performed in relation to the scientific results of this dissertation, the equations have been discretized in space by finite elements. However, all of the techniques that have been developed to solve the corresponding discrete systems of equations could just as well have been applied to equations that were discretized by other means, such as finite volume, or finite difference methods. Here, a brief introduction to the finite element method is provided.

Consider the variational equation: Find $u \in V$ such that

$$a\left(u, v\right) = L(v), \quad \forall v \in V, \tag{3.5}$$

where V is a Sobolev space that, in this setting, is called the solution (and test) space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a continuous function that is linear in the second component, and $L(\cdot) : V \to \mathbb{R}$ is a linear function. Notice that the variational equation (3.5) corresponds to the time-discrete system (3.2), for $a(u, v) = \left(\frac{u}{\tau}, v\right) + b(u, v)$ and $L(v) = \left(\frac{u^{n-1}}{\tau}, v\right) + K(v)$. The existence and uniqueness of the solution to equation (3.5) is provided by the Lax-Milgram lemma (see, e.g., [38]) when $a(u, v)$ is bilinear and coercive ($a(u, u) \geq c\|u\|_V$ for some $c \in \mathbb{R}^+$) and $L(v)$ is linear.

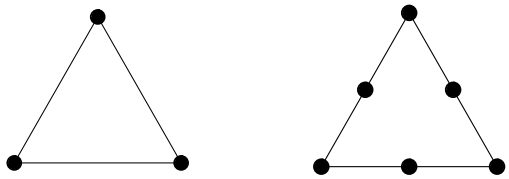
Remark 3.1.1. *The solution and test space might differ depending on the boundary conditions of the equations, however, for the simplicity of the exposition, that situation is disregarded here.*

To approximate solutions to systems of the form (3.5) with the *conforming* finite element method, let $\mathcal{T}_h = \{T_k\}_k$ be a subdivision of the domain Ω, where the *elements* T_k are non-overlapping and nonempty, and $h = \max_k \operatorname{diam}(T_k)$. Then, define the finite-dimensional subspace $V_h \subseteq V$ as

$$V_h := \left\{ v_h \in V : v_h|_{T_k} \in \mathcal{P}^l(T_k)\ \forall k, v_h \in C(\Omega) \right\}, \tag{3.6}$$

where $\mathcal{P}^l(T_k)$ is the space of polynomials of degree l on T_k. Let now $\{\eta_i\}_{i=1}^N$ be a basis for the finite-dimensional space V_h (with dimension N), and define the discrete approximation of the solution u to (3.5) as $u_h := \sum_{i=1}^N \alpha_i \eta_i$ where $\alpha_i \in \mathbb{R}$ for $i = 1, \ldots, N$. Then, the discrete counterpart to (3.5) is given as: Find $u_h \in V_h$ such that

$$a(u_h, \eta_j) = L(\eta_j), \quad \text{for} \quad j = 1, \ldots, N. \tag{3.7}$$



(a) $\mathcal{P}^1$ (b) $\mathcal{P}^2$

Figure 3.1: Linear $\mathcal{P}^1$ and quadratic $\mathcal{P}^2$ triangular Lagrange finite elements of degree 1 and 2. The black dots correspond to degrees of freedom for the polynomials.

This is now a (possibly nonlinear) system of N equations with N unknowns (α_i, for $i = 1 \ldots, N$).

It is clear, that different choices of finite dimensional subspaces, V_h, and different subdivisions $\mathcal{T}_h$ corresponds to different approximations of the continuous variational equation (3.5). To this end, the concise definition of a finite element, due to Ciarlet [54], is the triplet $(\mathcal{K}, \mathcal{P}, \mathcal{N})$, where $\mathcal{K}$ determines the geometry of the elements (triangular, quadrilateral, etc.), $\mathcal{P}$ is a finite dimensional space of functions on $\mathcal{K}$ (corresponds to the polynomial space) and $\mathcal{N}$ is a basis for the dual of $\mathcal{P}$, and relates to the choice of basis for the subspace V_h. Most notable are the Lagrange finite elements, where the degrees of freedom on each element, $\mathcal{N}$, are evaluations of the polynomials (in $\mathcal{P}$) on points on the element (as opposed to, for example, the Hermite conforming finite elements where also evaluations of the derivatives of the polynomials are considered), see Figure 3.1.

Remark 3.1.2 (Implementation of numerical solution strategies in DUNE)**.** *All the numerical tests that have been performed in relation to the scientific results in this dissertation have been implemented in the modular C++ library, DUNE (Distributed and Unified Numerics Environment) [13, 14, 22, 136]. Here, general implementations of many standard conforming and mixed finite elements are provided, as well as linear algebra and mesh generation tools. In particular, the dune-functions module for managing discrete global functions [63], and the dune-biot module that was developed for the doctoral dissertation of Both [28], have been used.*

Example 3.1.1 (Discretization of the Biot equations)**.** As an example, consider the Biot equations (2.21)–(2.22), which for simplicity are equipped with homogeneous Dirichlet boundary conditions. The corresponding variational equations are derived by multiplying the elasticity equation (2.21) by a test function $\boldsymbol{v} \in (H_0^1(\Omega))^d$ and the flow equation by $q \in H_0^1(\Omega)$ and applying the divergence theorem. Here, d is the spatial dimension and $H_0^1(\Omega)$ is the Sobolev space of $H^1(\Omega)$ functions with vanishing trace. The variational

problem then reads: Find $(\boldsymbol{u}, p) \in H^1\left([0,T], \left(H_0^1(\Omega)\right)^d\right) \times H^1\left([0,T], H_0^1(\Omega)\right)$ such that

$$\begin{aligned} (\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v})) - \alpha\,(p, \nabla\cdot\boldsymbol{v}) &= (\boldsymbol{f}, \boldsymbol{v})\,, && (3.8)\\ \left(\partial_t\left(\frac{p}{M} + \alpha\nabla\cdot\boldsymbol{u}\right), q\right) + \left(\frac{K}{\eta}\nabla p, \nabla q\right) &= (S_f, q)\,, && (3.9)\end{aligned}$$

for all $(\boldsymbol{v}, q) \in \left(H_0^1(\Omega)\right)^d \times H_0^1(\Omega)$ and almost all $t \in [0,T]$.

By applying the implicit Euler discretization in time and conforming finite elements in space with the test space $\boldsymbol{V}_h$ for displacement and Q_h for pressures, the discrete equations become: Given $\boldsymbol{u}_h^{n-1}, p_h^{n-1} \in \boldsymbol{V}_h \times Q_h$ find $(\boldsymbol{u}_h^n, \boldsymbol{p}_h^n) \in \boldsymbol{V}_h \times Q_h$ such that

$$\begin{aligned} (\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)) - \alpha\,(p_h^n, \nabla\cdot\boldsymbol{v}_h) &= (\boldsymbol{f}, \boldsymbol{v}_h)\,, && (3.10)\\ \left(\frac{p_h^n - p_h^{n-1}}{\tau M} + \alpha\frac{\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}}{\tau}, q_h\right) + \left(\frac{K}{\eta}\nabla p_h^n, \nabla q_h\right) &= (S_f, q_h)\,, && (3.11)\end{aligned}$$

for all $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h \times Q_h$ and almost all $t \in [0,T]$. To obtain an inf-sup stable discretization, typical choices for the pair of finite element spaces $\boldsymbol{V}_h \times Q_h$ include the Taylor-Hood elements [150, 23] (piece-wise quadratic polynomials, $\mathcal{P}^2$, for displacements and piece-wise linear polynomials, $\mathcal{P}^1$, for pressures) or $\mathcal{P}^1$ elements enriched with bubble functions for displacement and $\mathcal{P}^1$ elements for pressures [11]. Moreover, in the formulation (3.10)–(3.11), local mass conservation is lost, and a remedy (that uses the finite element method) is to apply a mixed formulation and use, for example, the Raviart-Thomas [133] element with piece-wise linear polynomials for the flow subproblem, see e.g., [31, 97].

Example 3.1.2 (Gradient stable time discretization and the convex-concave splitting method applied to the Cahn-Hilliard equation)**.** To see an example of a stable gradient time discretization, consider the Cahn-Hilliard equation (2.17)–(2.18). By using linear Lagrange finite elements, where the function space is denoted by $Q_h \subseteq H^1(\Omega)$, for both phase-field and potential, and the implicit Euler method, the discrete nonlinear system of equations reads: Given $\varphi_h^{n-1} \in Q_h$, find $(\varphi_h^n, \mu_h^n) \in Q_h \times Q_h$ such that

$$\begin{aligned} \left(\frac{\varphi_h^n - \varphi_h^{n-1}}{\tau}, q_h^\varphi\right) + (m\nabla\mu_h^n, \nabla q_h^\varphi) &= 0 && (3.12)\\ (\mu_h^n, q_h^\mu) - \left(\frac{\gamma}{\ell}\Psi'(\varphi_h^n), q_h^\mu\right) - (\gamma\ell\nabla\varphi_h^n, \nabla q_h^\mu) &= 0, && (3.13)\end{aligned}$$

for all $(q_h^\varphi, q_h^\mu) \in Q_h \times Q_h$. This is equivalent to the optimality conditions of the energy-based time discretization (3.4): Given $\varphi_h^{n-1} \in Q_h$ solve

$$\varphi_h^n = \operatorname*{arg\,min}_{s_h \in Q_h}\left\{\frac{1}{2\tau}\left\|s_h - \varphi^{n-1}\right\|^2_{Q^*_{m,h,0}} + \mathcal{E}_{\mathrm{I}}(s_h)\right\}, \tag{3.14}$$

subject to

$$\int_\Omega s_h - \varphi_h^{n-1}\, dx = 0,$$

where $\mathcal{E}_{\mathrm{I}}(s)$ is the energy defined in (2.30). Here, the space $Q_{h,m,0}^*$ is the discrete $H^{-1}(\Omega)$ space that is defined as the dual of

$$Q_{h,m,0} = \left\{ q_h \in Q_h \ : \ \int_\Omega q_h \, dx = 0 \right\},$$

with norm $\|m\nabla q_h\|_{L^2(\Omega)}$. Details of the equivalence between (3.12)–(3.13) and (3.14) can be found in Paper E [146].

Notice now that as $\varphi_h^{n-1} \in Q_h$ and $\int_\Omega \varphi_h^{n-1} - \varphi_h^{n-1}\, dx = 0$, the following inequality holds;

$$\frac{1}{2\tau}\left\|\varphi_h^n - \varphi_h^{n-1}\right\|_{Q_{m,h,0}^*}^2 + \mathcal{E}_{\mathrm{I}}(\varphi_h^n) - \frac{1}{2\tau}\left\|\varphi_h^{n-1} - \varphi_h^{n-1}\right\|_{Q_{m,h,0}^*}^2 - \mathcal{E}_{\mathrm{I}}(\varphi_h^{n-1}) \leq 0.$$

This implies that

$$\mathcal{E}_{\mathrm{I}}(\varphi_h^n) \leq \mathcal{E}_{\mathrm{I}}(\varphi_h^{n-1})$$

regardless of n and τ. Hence, the implicit Euler method is unconditionally gradient stable.

There is, however, another challenge that occurs when the implicit Euler method is applied to the Cahn-Hilliard equation; the minimization problem (3.14) is non-convex, due to the presence of the double-well potential $\Psi(\varphi)$ in the energy. Therefore, solving the nonlinear discrete problem (3.12)–(3.13) can be challenging, and conventional methods, such as the Newton scheme (see Section 3.2.2) often do not converge. A popular remedy proposed in [62, 65], and applied in different variations, e.g., in [90, 91, 134, 161], is to split the double-well potential $\Psi(\varphi)$ into the difference of two convex functions, e.g., (although many other variants exist)

$$\Psi(\varphi) = (1-\varphi^2)^2 = \left(1+\varphi^4\right) - 2\varphi^2 = \Psi_c(\varphi) - \Psi_e(\varphi)$$

and evaluate the convex part of its derivative implicitly $\Psi_c'(\varphi_h^n)$, and the concave (expansive) part explicitly $\Psi_e'(\varphi_h^{n-1})$. That corresponds to the discrete system of equations: Given $\varphi_h^{n-1} \in Q_h$ find $(\varphi_h^n, \mu_h^n) \in Q_h$ such that

$$\left(\frac{\varphi_h^n - \varphi_h^{n-1}}{\tau}, q_h^\varphi\right) + (m\nabla\mu_h^n, \nabla q_h^\varphi) = 0 \tag{3.15}$$

$$(\mu_h^n, q_h^\mu) - \left(\frac{\gamma}{\ell}\left(\Psi_c'(\varphi_h^n) - \Psi_e'(\varphi_h^{n-1})\right), q_h^\mu\right) - (\gamma\ell\nabla\varphi_h^n, \nabla q_h^\mu) = 0, \tag{3.16}$$

for all $(q_h^\varphi, q_h^\mu) \in Q_h \times Q_h$. The related minimization problem now reads: Given $\varphi_h^{n-1} \in Q_h$ solve

$$\varphi_h^n = \arg\min_{s_h \in Q_h} \mathcal{F}^n(s_h), \tag{3.17}$$

subject to

$$\int_\Omega s_h - \varphi_h^{n-1}\, dx = 0,$$

where

$$\mathcal{F}^n(s_h) := \frac{1}{2\tau} \left\| s_h - \varphi^{n-1} \right\|^2_{Q^*_{m,h,0}} + \gamma \int_\Omega \frac{1}{\ell} \left(\Psi_c(s_h) - \Psi'_e\left(\varphi_h^{n-1}\right) s_h \right) + \frac{\ell}{2} |\nabla s_h|^2 \, dx.$$

Notice that the minimization problem (3.17) is convex and thus is far better suited for linearization methods than (3.14). Moreover, by a similar computation as above, one can see that the discretization is unconditionally gradient stable;

$$\mathcal{F}^n\left(\varphi_h^n\right) - \mathcal{F}^n\left(\varphi_h^{n-1}\right) \leq 0, \tag{3.18}$$

which by applying the convexity of $\Psi_e(s_h)$

$$\int_\Omega \Psi_e\left(\varphi_h^n\right) - \Psi_e\left(\varphi_h^{n-1}\right)\, dx \geq \int_\Omega \Psi'\left(\varphi_h^{n-1}\right)\left(\varphi_h^{n-1} - \varphi_h^{n-1}\right)\, dx,$$

yields

$$\mathcal{E}_{\mathrm{I}}\left(\varphi_h^n\right) \leq \mathcal{E}_{\mathrm{I}}\left(\varphi_h^{n-1}\right),$$

for all n and τ.

3.2 Approximating solutions to the discrete system of equations

In the previous section, continuous systems were approximated by discrete systems of equations to be solved in each time step (3.7). Here, the focus is on finding good approximations of the solutions to these discrete systems of equations, in a robust and efficient manner. A robust method is, in this context, a method that will find an approximation without requiring special restrictions on material and discretization parameters, and an efficient method is one that finds the approximation fast, typically in a few iterative steps. Both linearization procedures, i.e., techniques for approximating the solution of a nonlinear problem by solving a sequence of linear problems iteratively, and decoupling techniques, i.e., techniques for approximating solutions to coupled problems by

sequentially solving the subproblems, are considered. Let

$$\mathbf{F} : \mathbb{R}^N \to \mathbb{R}^N,$$

be the function such that $[\mathbf{F}(\boldsymbol{\alpha}_h)]_j = a(u_h, \eta_j) - L(\eta_j)$ for $\boldsymbol{\alpha}_h = [\alpha_1, \ldots, \alpha_N]^\top$ and $u_h = \sum_{i=1}^N \alpha_i \eta_i$. Then, equation (3.7) can be reformulated as follows: Find $\boldsymbol{\alpha}_h \in \mathbb{R}^N$ such that

$$\mathbf{F}(\boldsymbol{\alpha}_h) = 0. \tag{3.19}$$

3.2.1 Solving linear problems

Regardless of whether $\mathbf{F}$ in itself is linear, i.e., $\mathbf{F}(\boldsymbol{\alpha}) = \mathbf{A}\boldsymbol{\alpha}_h - \mathbf{b}$ for $\mathbf{A} \in \mathbb{R}^{N \times N}$ and $\mathbf{b} \in \mathbb{R}^N$, or nonlinear (in which case the solution will be approximated by solving a sequence of linear problems), there is a need for solving linear systems of equations. Doing this efficiently is important, especially for time-dependent nonlinear problems, where linear systems are required to be solved several times in each time step. There are many techniques for doing this, and the most popular ones can either be classified as direct solvers or iterative solvers. Direct solvers have the benefit of not producing iteration errors themselves, but there is a significant memory consumption in storing large matrices, and matrix factorization can be time-consuming. Iterative solvers, on the other hand, reduce the memory requirement from the direct solvers and can in certain situations converge to an approximation of satisfactory precision rather quickly. It is, however, not easy to design fast methods that guarantee convergence, and often the choice of initial guess can be quite important. None of the methodologies will be covered here, but it is worth it to mention that for problems related to PDEs, multigrid and domain decomposition methods are very popular, and are often used as preconditioners to the iterative Krylov subspace methods like the generalized minimal residual method (GMRES) and the conjugate gradient method (CG).

3.2.2 Iterative linearization techniques

To approximate solutions to nonlinear problems, iterative methods are usually applied. Here, as well as for iterative methods for linear systems of equations, it is desirable to design methods that converge fast, and are robust.

The most classical iterative method for solving nonlinear problems is the Newton method (often called the Newton-Raphson) [60], and it is known for its quadratic convergence rate, and that it only converges locally (the initial guess has to be chosen sufficiently

close to the real solution), see the Newton-Kantorovich theorem [60, 102]. Consider the discrete problem (3.19). The Newton method approximates the solution by using the first-order Taylor expansion of the nonlinearity. Let i be the iteration index, and let $\boldsymbol{\alpha}_h^{i-1} \in \mathbb{R}^N$ be given. The Newton method then reads: Find $\boldsymbol{\alpha}_h^i \in \mathbb{R}^N$ such that

$$\mathbf{F}\left(\boldsymbol{\alpha}_h^{i-1}\right) + \nabla\mathbf{F}\left(\boldsymbol{\alpha}_h^{i-1}\right)\left(\boldsymbol{\alpha}_h^i - \boldsymbol{\alpha}_h^{i-1}\right) = 0, \tag{3.20}$$

where $\boldsymbol{\alpha}_h^0$ is known as the initial guess and is often chosen as the solution to the discrete problem at the previous time step in time-dependent problems. Equivalently, one could work directly on discrete variational equations like (3.7) by using the Gateaux derivative of the first argument of $a(\cdot,\cdot)$ instead of the gradient and get the method: Given $u_h^{i-1} \in V_h$, find $u_h^i \in V_h$ such that

$$a\left(u_h^{i-1}, v_h\right) + \lim_{\delta\to 0} \frac{a\left(u_h^{i-1} + \delta\left(u_h^i - u_h^{i-1}\right), v_h\right) - a\left(u_h^{i-1}, v_h\right)}{\delta} = L\left(v_h\right), \tag{3.21}$$

for all $v_h \in V_h$. Note that, if $a(\cdot,\cdot)$ is linear in the first entry as well, then equation (3.21) reduces to equation (3.7).

The main drawback of the Newton method is that it is only locally convergent. There are many ways to modify the method such that it becomes more robust, but this usually results in a linearly convergent scheme. Most of these methods fall under the classification of quasi-Newton methods, where the Jacobian term $\nabla\mathbf{F}$ in (3.20) is replaced by some approximation of it. Two common approximations are explained here:

- The modified Picard method [51] that computes only parts of the Jacobian and evaluates the rest of the nonlinearity in the previous iteration. Suppose that $\mathbf{F}(\mathbf{x}) = \mathbf{F}_1(\mathbf{x}) + \mathbf{F}_2(\mathbf{x})$, and that it is more desirable to compute the Jacobian of $\mathbf{F}_1(\mathbf{x})$ than of $\mathbf{F}_2(\mathbf{x})$, e.g., if $\mathbf{F}_2(\mathbf{x})$ is non-smooth. Then the modified Picard method reads: Given $\boldsymbol{\alpha}_h^{i-1} \in \mathbb{R}^N$, find $\boldsymbol{\alpha}_h^i \in \mathbb{R}^N$ such that

$$\mathbf{F}\left(\boldsymbol{\alpha}_h^{i-1}\right) + \nabla\mathbf{F}_1\left(\boldsymbol{\alpha}_h^{i-1}\right)\left(\boldsymbol{\alpha}_h^i - \boldsymbol{\alpha}_h^{i-1}\right) = 0. \tag{3.22}$$

- The L-scheme [99, 116, 131], which replaces the Jacobian with a scaled identity matrix. The method reads: Given $\boldsymbol{\alpha}_h^{i-1} \in \mathbb{R}^N$ and $L \in \mathbb{R}$, find $\boldsymbol{\alpha}_h^i \in \mathbb{R}^N$ such that

$$\mathbf{F}\left(\boldsymbol{\alpha}_h^{i-1}\right) + L\boldsymbol{I}\left(\boldsymbol{\alpha}_h^i - \boldsymbol{\alpha}_h^{i-1}\right) = 0. \tag{3.23}$$

  The L-scheme can also be seen as a stabilized Picard method, with stabilization parameter L.

All iterative solution strategies require a stopping criterion that determines when the iterative procedures should end. This is usually based on the increments $u_h^i - u_h^{i-1}$ (or $\boldsymbol{\alpha}_h^i - \boldsymbol{\alpha}_h^{i-1}$), and the residual $\mathbf{F}(\boldsymbol{\alpha}_h^i)$, and can be computed both *absolute*, and *relative*. Absolute and relative increments are defined as

$$\mathrm{Inc}_{\mathrm{abs}}^i := \|u_h^i - u_h^{i-1}\|, \quad \text{and} \quad \mathrm{Inc}_{\mathrm{rel}}^i := \frac{\|u_h^i - u_h^{i-1}\|}{\|u_h^1 - u_h^0\|},$$

where $\|\cdot\|$ typically is the $L^2(\Omega)$-norm, or any other norm on the function space V_h. The absolute and relative residual values are computed as

$$\mathrm{Res}_{\mathrm{abs}}^i := \|\mathbf{F}\left(\boldsymbol{\alpha}_h^i\right)\|_2, \quad \text{and} \quad \mathrm{Res}_{\mathrm{rel}}^i := \frac{\|\mathbf{F}\left(\boldsymbol{\alpha}_h^i\right)\|_2}{\|\mathbf{F}\left(\boldsymbol{\alpha}_h^0\right)\|_2},$$

where $\|\cdot\|_2$ is the Euclidean 2-norm. The iterative procedure is terminated when

$$\mathrm{Inc}_{\mathrm{abs}}^i \leq \mathrm{Tol}_{\mathrm{inc,abs}}, \ \mathrm{Inc}_{\mathrm{rel}}^i \leq \mathrm{Tol}_{\mathrm{inc,rel}}, \ \mathrm{Res}_{\mathrm{abs}}^i \leq \mathrm{Tol}_{\mathrm{res,abs}}, \ \text{and}, \ \mathrm{Res}_{\mathrm{rel}}^i \leq \mathrm{Tol}_{\mathrm{res,rel}}, \quad (3.24)$$

for some predetermined tolerance values $\mathrm{Tol}_{\mathrm{inc,abs}}$, $\mathrm{Tol}_{\mathrm{inc,rel}}$, $\mathrm{Tol}_{\mathrm{res,abs}}$, and $\mathrm{Tol}_{\mathrm{res,rel}}$.

3.2.3 Iterative decoupling of coupled systems

When considering discretized coupled systems of equations, there are two popular classes of solution strategies: monolithic approaches where the entire system is treated as one, or iterative decoupling approaches where the subproblems are solved sequentially, iterating back and forth. Monolithic approaches have the advantage of reducing one step of complexity, i.e., there is no requirement to wrap another iterative method around the solution strategy. They do, however, often lack optimized solvers, both in terms of robust linearization methods, and good preconditioners for the linear problems. The decoupling methods, on the other hand, has the advantage that one can apply already optimized solvers for each of the subproblems, and iterate back and forth between them. The drawback is that one often needs some stabilization to ensure convergence, and the convergence properties of the iterative decoupling method are often highly dependent on the choice of this stabilization.

Consider the coupled variational problem: Find $(u_h, p_h) \in V_h \times Q_h$ such that

$$c(u_h, p_h, v_h) = 0, \quad \forall v_h \in V_h, \quad (3.25)$$

$$d(u_h, p_h, q_h) = 0, \quad \forall q_h \in Q_h. \quad (3.26)$$

Here, V_h and Q_h are solution and test spaces for each of the subproblems, and $c(\cdot,\cdot,\cdot)$:

$V_h \times Q_h \times V_h \to \mathbb{R}$ and $d(\cdot,\cdot,\cdot) : V_h \times Q_h \times Q_h \to \mathbb{R}$ are generic functions that are linear in the third entry. To solve this problem with a general iterative decoupling method, let i denote the iteration index, make an initial guess p_h^0, and define the update procedure: Given $u_h^{i-1} \in V_h$ find $(u_h^i, p_h^i) \in V_h \times Q_h$ such that

$$c(u_h^i, p_h^i, v_h) = 0, \quad \forall v_h \in V_h, \tag{3.27}$$

$$d(u_h^{i-1}, p_h^i, q_h) = 0, \quad \forall q_h \in Q_h. \tag{3.28}$$

where equation (3.28) is solved first to get an updated p_h^i which then is used in (3.27) to get an updated u_h^i. The iterations proceed until some prescribed stopping criterion (3.24) is satisfied.

Remark 3.2.1. *Notice that the subproblems* (3.27) *and* (3.28) *may still be nonlinear and require some linearization procedure. One then has the possibility to fully resolve the linearization procedure in each decoupling iteration, or to combine the linearization and the decoupling iterations and do just one linearization iteration within each decoupling iteration. Although such a procedure might result in an increase in the number of decoupling iterations, it might decrease the total number of iterations within the solution step. The behavior of this type of combined iterative method is problem dependent and has been analyzed for the Richards equation coupled with transport in [99], and for poroelasticity with large deformations in [27].*

3.2.4 Stabilization of iterative decoupling methods

In general, the iterative decoupling method (3.27)–(3.28) is not guaranteed to converge and certainly not to converge fast (in relatively few iterations). A remedy is to add a stabilization/tuning term to the system which might help in providing convergence, or to accelerate the convergence speed. One would then typically modify the method as follows: Given $\left(u_h^{i-1}, p^{i-1}\right) \in Q_h$, find $(u_h^i, p_h^i) \in V_h \times Q_h$ such that

$$c(u_h^i, p_h^i, v_h) = 0, \quad \forall v_h \in V_h \tag{3.29}$$

$$d(u_h^{i-1}, p_h^i, q_h) + l(p_h^i - p_h^{i-1}, q_h) = 0, \quad \forall q_h \in Q_h. \tag{3.30}$$

where $l(\cdot,\cdot) : Q_h \times Q_h \to \mathbb{R}$ is a bilinear function that is known as a stabilization or tuning term.

The decoupling method (3.29)–(3.30), can for linear equations be identified with block partitioned, Schur-complement-based iterative solvers. To see this, suppose that $c(\cdot,\cdot,\cdot)$

and $d(\cdot,\cdot,\cdot)$ are trilinear and that (3.25)–(3.26) forms the block-linear system

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_h^u \\ \boldsymbol{\alpha}_h^p \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{h} \end{pmatrix}, \tag{3.31}$$

where $u_h = \sum_l [\boldsymbol{\alpha}_h^u]_l \eta_l^u$, and $p_h = \sum_k [\boldsymbol{\alpha}_h^p]_k \eta_k^p$, with $\{\eta_l^u\}_l$ and $\{\eta_k^p\}_k$ being bases for V_h and Q_h, respectively, and $\mathbf{g}$ and $\mathbf{h}$ are source terms. One can quite easily see that (3.31) is equivalent with the block system

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} - \mathbf{S} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_h^u \\ \boldsymbol{\alpha}_h^p \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{h} - \mathbf{C}\mathbf{A}^{-1}\mathbf{g} \end{pmatrix}, \tag{3.32}$$

where $\mathbf{S} = \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ is the Schur-complement. This formulation is essentially not better suited for solving than (3.31) as it involves the computation of $\mathbf{A}^{-1}$, but it motivates the choice of tuning terms in (3.29)–(3.30). Let again i be the iteration index, and suppose that $\boldsymbol{\alpha}_h^{u,i-1}$ and $\boldsymbol{\alpha}_h^{p,i-1}$ are given (through the iterative procedure). The top row of the block linear system (3.32) then gives

$$\mathbf{A}\boldsymbol{\alpha}_h^{u,i-1} + \mathbf{B}\boldsymbol{\alpha}_h^{p,i-1} = \mathbf{g},$$

which is equivalent with

$$\mathbf{C}\boldsymbol{\alpha}_h^{u,i-1} + \mathbf{S}\boldsymbol{\alpha}_h^{p,i-1} = \mathbf{C}\mathbf{A}^{-1}\mathbf{g}.$$

This can be substituted into (3.32) to get the block system at iteration i

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} - \mathbf{S} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_h^{u,i} \\ \boldsymbol{\alpha}_h^{p,i} \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \mathbf{h} - \mathbf{C}\boldsymbol{\alpha}_h^{u,i-1} - \mathbf{S}\boldsymbol{\alpha}_h^{p,i-1} \end{pmatrix}, \tag{3.33}$$

which corresponds to choosing l_d as the linear function related to the Schur complement $\mathbf{S}$ in (3.29)–(3.30). As the computation of $\mathbf{S}$ is unfeasible, a good and cheap approximation to $\mathbf{S}$ is regarded as a good tuning term.

Remark 3.2.2. *Notice that the Schur-complement* $\mathbf{S}$ *in the block system* (3.33) *is subtracted from the original block diagonal. This leads to the discussion of whether to call it a stabilization term or a tuning term. It is generally seen as stabilizing for a linear system to add something to the diagonal (block diagonal in this case), and this would be the case for Schur-complement for saddle-point problems like the Biot equations, where* $\mathbf{C} = -\mathbf{B}^\top$. *For "block-symmetric" problems, like the Cahn-Larché equations, on the other hand, the Schur-complement approach actually suggests to "destabilize" the block-system, and the added term should only be regarded as a tuning term. This can be seen in practice and theory for both the Biot equations and the Cahn-Larché equation, where the Biot equation actually requires stabilization to converge when decoupled (both for theo-*

retical and practical convergence) [31, 121, 143, 144], while the Cahn-Larch 'e equations can be decoupled (with theoretical and practical convergence) without the addition of a stabilizing term [146].

Example 3.2.1 (The fixed-stress splitting scheme applied to Biot's equations)**.** As an example of a decoupling method, the widely used fixed-stress splitting scheme applied to Biot's equations will be presented. The scheme alternates between solving a porous media flow equation (2.22) and an elasticity equation (3.10). One of the main motivations for using such a decoupling scheme is that it enables the use of readily available and optimized solvers for both flow and elasticity instead of constructing new ones for the full monolithic problem. Moreover, splitting methods, such as fixed stress, can also be used as preconditioners for the monolithic problem [82].

Consider first a generic stabilized decoupling method of the type (3.29)–(3.30) applied to the discretized Biot equations (3.10)–(2.22): Given $\left(\boldsymbol{u}_h^{n,i-1}, p_h^{n,i-1}, \boldsymbol{u}_h^{n-1}, p_h^{n-1}\right) \in \boldsymbol{V}_h \times Q_h \times \boldsymbol{V}_h \times Q_h$, find $\left(\boldsymbol{u}_h^{n,i}, p_h^{n,i}\right) \in Q_h \times \boldsymbol{V}_h$ such that

$$\begin{aligned} \left(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_h^{n,i}), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\right) - \alpha\left(p_h^{n,i}, \nabla\cdot\boldsymbol{v}_h\right) &= \left(\boldsymbol{f}, \boldsymbol{v}_h\right), \\ \left(\frac{p_h^{n,i}-p_h^{n-1}}{\tau M} + \alpha\frac{\nabla\cdot\boldsymbol{u}_h^{n,i-1}-\nabla\cdot\boldsymbol{u}_h^{n-1}}{\tau}, q_h\right) + \left(\frac{K}{\eta}\nabla p_h^{n,i}, \nabla q_h\right) & \\ + L\left(p_h^{n,i}-p_h^{n,i-1}, q_h\right) &= \left(S_f, q_h\right), \end{aligned}$$

for all $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h \times Q_h$, where L is a real number, which is often known as stabilization or tuning parameter. This method is known as the fixed-stress splitting scheme [5, 31, 32, 50, 96, 105, 121, 138] and dates back to the work in [138], where the (volumetric) stress is fixed over the iterations, i.e., the condition

$$K_{\mathrm{dr}}\nabla\cdot\boldsymbol{u}_h^{n,i} - \alpha p_h^{n,i} = K_{\mathrm{dr}}\nabla\cdot\boldsymbol{u}_h^{n,i-1} - \alpha p_h^{n,i-1},$$

is imposed and used to eliminate the term $\alpha\nabla\cdot\boldsymbol{u}_h^n$ in equation (3.11) in each decoupling iteration. Here, K_{dr} is the drained bulk modulus which is defined as $K_{\mathrm{dr}} := \frac{2\mu}{d} + \lambda$, where μ and λ are the Lamé parameters and d is the spatial dimension. The resulting stabilization term is then $L = \frac{\alpha^2}{K_{\mathrm{dr}}}$. Later in [121] it was shown that choosing the stabilization parameter $L \geq \frac{\alpha^2}{2K_{\mathrm{dr}}}$, that is, greater than half of the physically motivated parameter, results in a convergent method. The same was shown using different techniques in [31]. In Paper A [143], the optimal choice of stabilization parameter is discussed, and an interval where the optimal parameters resides is provided, and given as $\left[\frac{\alpha^2}{4\mu+2\lambda}, \frac{\alpha^2}{\frac{2\mu}{d}+\lambda}\right)$. Moreover, a method for determining that parameter, utilizing its mesh independency, is proposed. In Paper B [144], the optimal stabilization parameter in the special case of low-permeable medium is discussed, which results in a formula for computing it by

approximating the eigenvalues of the Schur complement.

Another possibility is to stabilize the elasticity equation with a term of the form $L\left(\nabla\cdot\boldsymbol{u}_h^{n,i}-\nabla\cdot\boldsymbol{u}_h^{n,i-1},\nabla\cdot\boldsymbol{v}_h\right)$. This is known as the undrained splitting method, where it is assumed, for $L=\alpha^2 M$, that the volumetric fluid content,

$$\frac{p_h^{n,i}}{M}+\alpha\nabla\cdot\boldsymbol{u}_h^{n,i}=\frac{p_h^{n,i-1}}{M}+\alpha\nabla\cdot\boldsymbol{u}_h^{n,i-1},$$

remains constant over the iterations. The undrained splitting method is analyzed in [6, 104].

3.2.5 Decoupling methods as alternating minimization

For coupled variational problems that correspond to the optimality conditions of minimization problems, one can also regard the decoupling methods as alternating minimization. Consider the minimization problem

$$(u_h,p_h)=\underset{v_h\in V_h,q_h\in Q_h}{\arg\min}\ \mathcal{F}(v_h,q_h), \tag{3.34}$$

where $\mathcal{F}(\cdot,\cdot):V_h\times Q_h\to\mathbb{R}$. The optimality conditions for this minimization problem read: Find $(u_h,p_h)\in V_h\times Q_h$ such that

$$\begin{aligned}\langle\mathcal{D}_1\mathcal{F}(u_h,p_h),v_h\rangle &= 0,\quad \forall v_h\in V_h,\\ \langle\mathcal{D}_2\mathcal{F}(u_h,p_h),q_h\rangle &= 0,\quad \forall q_h\in Q_h,\end{aligned}$$

where $\mathcal{D}_1\mathcal{F}$ and $\mathcal{D}_2\mathcal{F}$ represent the Gateaux derivative of $\mathcal{F}$ with respect to the first and second argument, respectively. Both the energy-based time-discretization for gradient flows (3.4) and the variational phase-field models for fracture propagation, see Chapter 2.2.2, are examples of this type of discrete minimization problem.

Alternating minimization applied to (3.34) is the algorithm: Given u_h^{i-1} solve first

$$p_h^i=\underset{q_h\in Q_h}{\arg\min}\,\mathcal{F}\left(u_h^{i-1},q_h\right),$$

then, using the newly computed p_h^i, solve

$$u_h^i=\underset{v_h\in V_h}{\arg\min}\,\mathcal{F}\left(v_h,p_h^i\right).$$

Considering decoupling methods in this way has the benefit that one can apply the

theory of (convex) optimization to analyze the convergence properties of the decoupling methods [18].

Remark 3.2.3. *If* $\langle \mathcal{D}_1\mathcal{F}(u_h, p_h), v_h\rangle = c(u_h, p_h, v_h)$ *and* $\langle \mathcal{D}_2\mathcal{F}(u_h, p_h), v_h\rangle = d(u_h, p_h, q_h)$ *then the solution to the minimization problem* (3.34) *is the same as the solution to* (3.25)–(3.26)*.*

Example 3.2.2 (Alternating minimization to solve phase-field for fracture equations)**.** As an example of alternating minimization, consider the variational approach to fracture propagation, see Chapter 2.2.2. First, the continuous minimization problem (2.28) is discretized with conforming finite elements (the typical choice is linear Lagrange finite elements $\mathcal{P}^1$ for both displacement and phase-field)

$$(\boldsymbol{u}_h^n, \varphi_h^n) = \underset{\boldsymbol{w}_h \in \boldsymbol{V}_h, s_h \in Q_h}{\arg\min} \mathcal{E}_{\text{frac}}(\boldsymbol{w}_h, s_h), \tag{3.35}$$

subject to

$$\varphi_h^n(x) \geq \varphi_h^{n-1}(x), \quad \forall x \in \Omega,$$

where

$$\begin{aligned}\mathcal{E}_{\text{frac}}(\boldsymbol{w}_h, s_h) \;=\; & \int_\Omega g(s_h)\left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h) : \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{w}_h) - \boldsymbol{f}^n \cdot \boldsymbol{w}_h\right) dx - \int_\Gamma \boldsymbol{\tau}^n \cdot \boldsymbol{w}_h \, ds \\ & + G_c \int_\Omega \frac{1}{2\ell} s_h^2 + \frac{\ell}{2} |\nabla s_h|^2 \, dx.\end{aligned}$$

Here, $\boldsymbol{V}_h$ is the discrete test and solution space for the displacement variable $\boldsymbol{u}_h^n$, and Q_h is the discrete test and solution space for the phase-field variable φ_h^n.

Now, alternating minimization is applied to the minimization problem (3.35). First, the potential $\mathcal{E}_{\text{frac}}$ is minimized with respect to the displacement variable using the previous iterate for the phase-field (the solution at the previous loading step in the first iteration), then the potential is minimized with respect to the phase-field variable using the newly computed displacement function. The solution strategy becomes as follows: Given $\varphi_h^{n,i-1} \in Q_h$ find $\boldsymbol{u}_h^i \in \boldsymbol{V}_h$ such that

$$\left\langle \mathcal{D}_1\mathcal{E}(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i-1}), \boldsymbol{v}_h \right\rangle = 0, \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h,$$

which is equivalent to

$$\left(g(\varphi_h^{n,i-1})\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_h^{n,i}); \boldsymbol{v}_h\right) - (\boldsymbol{\tau}, \boldsymbol{v}_h)_\Gamma = (\boldsymbol{f}^n, \boldsymbol{v}_h), \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h.$$

Then, using the newly computed $\boldsymbol{u}_h^{n,i}$, find $\varphi_h^{n,i} \in Q_h$ such that

$$\varphi_h^{n,i} = \underset{s_h \in Q_h}{\arg\min}\, \mathcal{E}(\boldsymbol{u}_h^{n,i}, s_h) \tag{3.36}$$

and $\varphi_h^{n,i}(x) \geq \varphi_h^{n-1}(x)$ for all $x \in \Omega$.

To solve the constrained minimization problem (3.36) there are several options that exist in the literature, including penalization methods [83, 119], augmented Lagrangian penalization [42, 158], primal-dual active set methods [94], fixing values by using Dirichlet nodes [36, 107], and using a history field [118].

In Paper C [147], the history field approach is utilized, and therefore it is the only one that will be discussed here. The method replaces the elasticity contribution $\psi(\boldsymbol{u}_h^{n,i}) = \boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right) \mathbb{C}\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right)$ in (3.36) by a history field $\mathcal{H}_h^{n,i}$ that is defined recursively by the pointwise equation

$$\mathcal{H}_h^{n,i}(x) := \max\left\{\mathcal{H}_h^{n-1}(x), \psi(\boldsymbol{u}_h^{n,i}(x))\right\},$$

and $\mathcal{H}_h^n := \mathcal{H}_h^{n,i}$ where i is the iteration number of the accepted approximation.

The resulting iterative solution method is then: Given $\varphi_h^{n,i-1} \in Q_h$ find $\left(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i}\right) \in \boldsymbol{V}_h \times Q_h$ such that

$$\begin{aligned}
\left(g\left(\varphi_h^{n,i-1}\right) \mathbb{C}\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right); \boldsymbol{v}_h\right) - (\boldsymbol{\tau}, \boldsymbol{v}_h)_\Gamma - (\boldsymbol{f}^n, \boldsymbol{v}_h) &= 0, \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h, \quad (3.37)\\
\left(g'\left(\varphi_h^{n,i}\right) \mathcal{H}_h^{n,i}, q_h\right) + \frac{G_c}{\ell}\left(\left(\varphi_h^{n,i}, q_h\right) + \ell^2\left(\nabla\varphi_h^{n,i}, \nabla q_h\right)\right) &= 0, \quad \forall q_h \in Q_h. \quad (3.38)
\end{aligned}$$

Although decoupling methods of this type are known in the literature to be convergent [42], the convergence rates are at time very slow. In Paper C [147], this is handled by the use of a new acceleration method, see Remark 3.3.1.

3.3 Acceleration of fixed-point methods

All iterative methods in Section 3.2 can be written as fixed-point iterations, i.e., methods that after some initial guess $x^0 \in \mathbb{R}^N$ are updated by the procedure

$$\mathbf{x}^i = \mathcal{G}\left(\mathbf{x}^{i-1}\right), \tag{3.39}$$

where $\mathcal{G} : \mathbb{R}^N \to \mathbb{R}^N$. Typically, fixed-point iterations converge if the operator $\mathcal{G}$ is contractive, due to the Banach fixed-point theorem [52], which, additionally, provides certain estimates for the convergence rates. However, not all fixed-point iterations are

contractive, and several might be rather slow. Here, two acceleration methods, that act as postprocessing techniques for fixed-point iterations (that is applied after every iterative step) are introduced. The general purpose of an acceleration method is to reduce the number of iterations required for the fixed-point iteration to convergence. Moreover, in some scenarios, acceleration techniques can also be used as methods for making non-convergent fixed-point iterations convergent [33].

3.3.1 Relaxation

One of the simplest acceleration techniques is the relaxation method. Rewrite first the fixed-point iteration (3.39) to the "increment problem"

$$\mathbf{x}^i = \mathcal{G}\big(\mathbf{x}^{i-1}\big) = \mathbf{x}^{i-1} + \Delta\mathcal{G}\big(\mathbf{x}^{i-1}\big)\,, \tag{3.40}$$

where

$$\Delta\mathcal{G}\big(\mathbf{x}^{i-1}\big) := \mathbf{x}^i - \mathbf{x}^{i-1},$$

is the increment of the iterative method. With a relaxation method, one first computes the increment $\Delta\mathcal{G}(\mathbf{x}^{i-1})$, then updates iterate with a scaled increment:

$$\mathbf{x}^i = \mathbf{x}^{i-1} + \omega\Delta\mathcal{G}\big(\mathbf{x}^{i-1}\big) \tag{3.41}$$

where $\omega \in (0, 2)$. It is often known as under-relaxation when $\omega < 1$, and over-relaxation when $\omega > 1$. Choosing the correct value for the relaxation parameter ω is essential to achieve optimal acceleration, and the most common way of doing this in a systematic manner is through the use of line-search algorithms. One can then either do a residual-based line-search:

$$\omega = \underset{w\in(0,2)}{\arg\min}\, \mathrm{Res}\big(\mathbf{x}^{i-1} + w\Delta\mathcal{G}\big(\mathbf{x}^{i-1}\big)\big)\,,$$

where the residual of the fixed-point problem is given as $\mathrm{Res}(\mathbf{x}) = \mathcal{G}(\mathbf{x})-\mathbf{x}$. Alternatively, if the fixed-point problem is minimizing some energy-potential $\mathcal{E}$ (e.g., with alternating minimization), the line-search can be based on minimizing the energy in the search direction $\Delta\mathcal{G}(\mathbf{x}^{i-1})$:

$$\omega = \underset{w\in(0,2)}{\arg\min}\, \mathcal{E}\big(\mathbf{x}^{i-1} + w\Delta\mathcal{G}\big(\mathbf{x}^{i-1}\big)\big)\,.$$

3.3.2 Anderson acceleration

Anderson acceleration was proposed for integral equations in [10], and has during the last decade been studied extensively and been applied to solve problems related to the Navier-

Stokes equations [130], electronic structure computations [66], geometry optimization [128], machine-learning [17, 140], flow and transport in unsaturated porous media [98], and flow in unsaturated deformable porous media [33].

Consider again the fixed-point problem (3.40). Anderson acceleration is a post-processing technique, that updates the current iterate based on the m previous iterations, where m is called the acceleration depth. The iterate $\mathbf{x}^i$ is updated as

$$\mathbf{x}^i = \sum_{k=0}^{m} \alpha_i^k \mathcal{G}\left(\mathbf{x}^{i-k-1}\right), \tag{3.42}$$

where $\boldsymbol{\alpha}_i = [\alpha_i^0, \ldots, \alpha_i^m]^\top \in \mathbb{R}^{m+1}$ solves the constrained minimization problem

$$\boldsymbol{\alpha}_i = \underset{\substack{\boldsymbol{c}_i \in \mathbb{R}^{m+1} \\ \sum_{k=0}^{m} c_i^k = 1}}{\arg\min} \left\| \sum_{k=0}^{m} c_i^k \Delta\mathcal{G}\left(\mathbf{x}^{k-1}\right) \right\|_2, \tag{3.43}$$

where the norm $\|\cdot\|_2$ is the Euclidean 2-norm and $\boldsymbol{c}_i = [c_i^0, \ldots, c_i^m]^\top$.

Remark 3.3.1. *In [64], the authors prove that Anderson acceleration is accelerating for linearly convergent fixed-point iterations, but not for quadratically convergent ones. However, the convergence of Anderson accelerated methods is local in nature [152]. This is exploited in Paper C [147], where the decoupling method for the variational phase-field for fracture problem* (3.37)–(3.38)*, is accelerated using a combination of over-relaxation and Anderson acceleration. Here, Anderson acceleration with relatively low depth ($m \in [1, 5]$, although tuning the depth results in a very limited gain in computational efficiency) is used as the default acceleration method. However, due to its local convergence properties, a safe-guard [148, 162], based on the residual evolution, changes the acceleration technique to over-relaxation when cracks are propagating. This results in a very robust and highly accelerating technique.*

Chapter 4

Summary and outlook

In this chapter, a summary and an outlook of the scientific results that constitute Part II of the dissertation are provided. First, all the papers are presented in the order that they appear in Part II, then, Part I of the dissertation is concluded with an outlook.

4.1 Summary of the included papers

The papers that are summarized below, are related in several ways. The first two papers Paper A and Paper B are concerned with the optimal stabilization parameter for the quasi-static Biot equations, and are as such closely connected, and can be seen as two approaches to solve the same problem. Then, Paper D and Paper E are both concerned with extensions of the Cahn-Hilliard equation, however, where Paper D is a modelling paper, Paper E focuses on solution strategies. Furthermore, Paper A, Paper B, Paper C and Paper E are concerned with iterative decoupling solvers for different coupled problems.

Paper A [143]

Title:	*On the optimization of the fixed-stress splitting for Biot's equations*
Authors:	Erlend Storvik, Jakub Wiktor Both, Kundan Kumar, Jan Martin Nordbotten and Florin Adrian Radu
Journal:	International Journal for Numerical Methods in Engineering 120, 179–194, (2019).
DOI:	10.1002/nme.6130

This paper is concerned with finding the optimal stabilization parameter for the fixed-stress splitting scheme applied to Biot's equations (2.21)–(2.22). The fixed-stress splitting scheme iteratively decouples the flow and elasticity subsystem of the quasi-static linearized Biot equations and stabilizes the flow subsystem, as described in Section 3.2.4. The method was first developed in [138] by fixing the volumetric stress over the iterations, resulting in a stabilizing term that depends on the coupling coefficient α and the elasticity coefficients, through the drained bulk modulus K_{dr}, $L = \frac{\alpha^2}{K_{\mathrm{dr}}}$. Later, in [121, 31] theoretical results show that stabilization parameters that are larger than half of the original $L \geq \frac{\alpha^2}{2K_{\mathrm{dr}}}$ result in a convergent method. However, numerical experiments show that although convergence was achieved, different stabilization parameters often result in better performance of the numerical solution strategies [32, 120]. In this paper, a new theoretical convergence proof shows that the optimal stabilization parameter also depends on the flow parameters. A formula for computing the stabilization parameter is provided, but some of the parameters that are required to compute it are quite demanding to precisely determine, e.g., the inf-sup constant and the Poincaré constant appear in the expression. Moreover, some of the bounds that were used to find the optimal parameter might have been too coarse, i.e., there might for specific cases exist better bounds. On the other hand, an interval where the optimal stabilization parameter is always found is provided, and numerical experiments show that the optimal stabilization parameter is independent of the mesh size. Therefore, a simple and effective *brute force* method is proposed. Here, several stabilization parameters in the provided optimal region are tested on a coarse mesh for one time step, and the optimal one is chosen for the full simulation. Several numerical experiments, including the Mandel benchmark problem, and a 3D footing problem, show that the proposed brute force method is highly effective. However, it does require that coarser mesh sizes are available, although this might not be the case for industrial applications.

Paper B [144]

Title: *The fixed-stress splitting scheme for Biot's equations as a modified Richardson iteration: Implications for optimal convergence*

Authors: Erlend Storvik, Jakub Wiktor Both, Jan Martin Nordbotten and Florin Adrian Radu

Book: Numerical Mathematics and Advanced Applications ENUMATH 2019, 909–917, (2021).

DOI: 10.1007/978-3-030-55874-1_90

This paper is a continuation of the work in Paper A. The fixed-stress splitting method

is first considered in the special case of impermeable media, and for that case, it is shown to be equivalent to a modified Richardson iteration. The relation between the constant in the Richardson iteration and the stabilization parameter in the fixed stress splitting method is identified. Then, using the theory for the optimal constant in the modified Richardson iteration, the fixed-stress splitting scheme is analyzed. The optimal fixed-stress stabilization parameter is shown to be $\frac{\alpha^2}{2}\left(\frac{1}{K_{\mathrm{dr}}^*}+\frac{1}{\beta}\right)$, where K_{dr}^* is called the mathematical bulk modulus and is related to the drained bulk modulus in that they both satisfy a similar inequality, but K_{dr}^* is a sharper bound and β is related to the inf-sup constant of the discretized pressure-displacement coupling. This stabilization parameter is, of course, also in the interval that is proposed in Paper A. The two parameters K_{dr} and β are, however, difficult to determine exactly, and to rectify this, an inexact eigenvalue solver, e.g., the power iteration, is proposed to find them.

It is important to emphasize that the undrained splitting scheme, for particular discretizations, is exact for the special case of impermeable media. However, for only slightly permeable media, the fixed-stress method typically outperforms it, especially with the stabilization parameter that is proposed in this paper. Numerical experiments show that the method for computing the optimal stabilization parameter is effective and that the inexact eigenvalue solver helps in finding a stabilization parameter that is truly optimal for both impermeable and low-permeable porous materials.

Paper C [147]

Title: *An accelerated staggered scheme for variational phase-field models of brittle fracture*

Authors: Erlend Storvik, Jakub Wiktor Both, Juan Michael Sargado, Jan Martin Nordbotten and Florin Adrian Radu

Journal: Computational Methods in Applied Mechanics and Engineering 381, 113822, (2021).

DOI: 10.1016/j.cma.2021.113822

In this paper, the main focus is to improve the performance of solvers for the variational phase-field models for brittle fracture propagation, see Section 2.2.2. The standard solution strategy in the literature for this problem is to use an alternating minimization type solver that often is called a staggered solution scheme. This is due to its robustness, in the sense that the method converges for most problem setups. However, the staggered solution scheme is known to be notoriously slow, requiring thousands of iterations to converge in single loading steps when fractures are propagating. Using a plain Newton

method, on the other hand, is not considered to be a good remedy as it often does not converge at all. In this paper, a novel acceleration method for the staggered solution scheme is provided. The method is a combination of a safe-guarded Anderson acceleration and over-relaxation, and the main principle is that the staggered scheme should be over-relaxed when fractures are propagating, and thus are far from the final configuration, and Anderson accelerated in all other scenarios. A criterion based on the norm of the residual is used to switch between the two acceleration methods, which makes the total acceleration cost very low. The motivation for applying over-relaxation when the fractures are propagating is based on the observation that fractures gradually propagate over consecutive iterations of the staggered scheme, and therefore moving further in each iteration is beneficial.

The method is tested in several numerical experiments that are widely used in the community. For all the examples, the staggered solution scheme is significantly accelerated, sometimes with more than 80% reduction in the number of iterations. Moreover, many combinations of the over-relaxation parameter and Anderson acceleration depth are tested, and in all of the situations, Anderson acceleration depth of at least one and over-relaxation parameter of approximately 1.6 seem to be the optimal choice. However, a very beneficial trait of the acceleration method is that the potential gain in computational efficiency by tuning the acceleration parameters is very small. Therefore, any combination of over-relaxation parameter and Anderson acceleration depth can be used.

Paper D [145]

Title:	*A Cahn–Hilliard–Biot system and its generalized gradient flow structure*
Authors:	Erlend Storvik, Jakub Wiktor Both, Jan Martin Nordbotten and Florin Adrian Radu
Journal:	Applied Mathematics Letters 126, 107799, (2022).
DOI:	10.1016/j.aml.2021.107799

During the last decade, there has been an increasing interest in using phase-fields in predictive tumor growth modelling. Due to the assumptions of local phase-field balance with a diffusive flux law, and the influence of interface tension, due to cohesion and adhesion, the models take the form of extended Cahn-Hilliard equations; see, e.g., [72, 73, 74, 75, 79, 80, 114, 115, 126, 141, 160]. Furthermore, these models have been extended to account for elasticity effects [74, 79]. The phase-field is used to represent different stages of cancerous and healthy tissue. This inspired the work in this paper to develop a thermodynamically consistent model for flow through a two-phase poroelastic material

where the two phases move depending on the interfacial tension as well as the fluid flow and elasticity properties. Moreover, this general model can be considered in relation to wood growth simulation, where sapwood transforms into heartwood as the tree grows.

The system is called the Cahn-Hilliard-Biot model as it can be seen as a combination of the quasi-static Biot equations and the Cahn-Hilliard equation. In the paper, the essential coupling terms are highlighted. Moreover, the full model is shown to be a generalized gradient flow, which in itself is not obvious, although all of the subsystems have this structure. This is an important feature that provides thermodynamical consistency of the model, in the sense that the free energy is dissipated in the absence of external forces. Additionally, the generalized gradient flow structure can be utilized to develop solution strategies for the model, as well as for theoretical analysis. Finally, a numerical experiment is performed in which the effects of flow and elasticity on the phase-field evolution are analyzed.

Paper E [146]

Title: *A robust solution strategy for the Cahn-Larché equations*

Authors: Erlend Storvik, Jakub Wiktor Both, Jan Martin Nordbotten and Florin Adrian Radu

Preprint: arXiv:2206.01541 [math.NA].

This paper is concerned with developing robust solution strategies for the Cahn-Larché equations, see Section 2.2.3. The equations are coupled, have non-convex nonlinearities, and are of fourth order in space. The fourth order term is handled as normally done with the Cahn-Hilliard equations with a mixed formulation of the Cahn-Hilliard subsystem that solves for phase-field and potential.

One could try to apply a fully explicit time discretization to avoid the difficult nonlinearities, but for the Cahn-Hilliard equation, this is known to not be energy stable. A fully implicit formulation, on the other hand, can be seen to be equivalent to a non-convex minimization problem, and therefore standard linearization methods such as the Newton method fail to converge for all but very small time steps and restrictive material properties. In this paper, a novel convex-concave time discretization is proposed that takes inspiration from the classical convex-concave splitting methods for the Cahn-Hilliard equation [62, 65] and extends it to also handle the non-convex elasticity term.

In the paper, the Cahn-Larché equation for phase-field independent elasticity tensor is first analyzed, and the equivalence between the time discretized system of equations and

a convex discrete minimization problem is established. Moreover, the discretization is shown to be unconditionally energy stable, and an alternating minimization method to solve the discrete system of equations is proposed. A convergence proof that includes convergence rates for the alternating minimization method is provided, using the theory of [29].

Then, the full system with phase-field dependent elasticity tensor is considered, and the same analysis for the alternating minimization is shown to work for the newly proposed convex-concave time discretization for this problem as well. Finally, numerical experiments show that in several situations where classical time-discretizations, including the convex-concave treatment of the Cahn-Hilliard equation, lead to a system that is unfeasible to solve with standard linearization methods, the newly proposed discretization method leads to a system that is very well-suited for linearization methods.

4.2 Outlook

This dissertation focuses on developing robust and efficient solution strategies for several coupled problems. In Paper A and B the quasi-static linearized Biot equations for poroelasticity are considered, Paper C is concerned with the variational phase-field models for brittle fracture propagation, and in Paper E the focus is on the Cahn-Larché system that couples the Cahn-Hilliard equation with linearized elasticity. Moreover, in Paper D a new thermodynamically consistent model is developed that couples poroelasticity with a Cahn-Hilliard phase-field equation. There are of course many further enhancements and developments that can be made to all of these contributions. Here, some of them are highlighted.

Regarding the work on the optimal stabilization parameter for the quasi-static Biot equations, there are several aspects that can be improved. There is still no general *a priori* way to compute the optimal stabilization parameter based on material parameters and boundary conditions. This could perhaps be achieved with sharper bounds and different proof techniques. Moreover, getting a good theory regarding optimal stabilization parameters for a broader class of problems, also with more than two coupled subproblems would be very beneficial. Although there are several results already available in the literature for many different extensions of the quasi-static Biot equations, e.g., extension to large deformations [27] and thermo-poroelasticity [40], and other coupled problems such as coupled flow and transport [99] and the variational phase-field approach to brittle fracture propagation [42], the theory is still missing even for the coupling of some of the processes that are discussed in this dissertation, such as poroelasticity with frac-

tures [84, 122, 142] and the Cahn-Hilliard-Biot system [145]. Additionally, for certain problems one could consider to "destabilize" the decoupling method to enhance the convergence speed, specifically for block-symmetric problems, as was experienced in [30] and not saddle-point problems like the quasi-static Biot equations, this is a promising possibility.

The acceleration method from Paper C should be more carefully tested, especially with simulations including several simultaneously propagating fractures, other methods of enforcing the non-healing constraint, and different splits of the elastic energy. Furthermore, the extension to fluid-filled fractures and pressurized fracture propagation could be considered. A possible enhancement of the acceleration method could be to use a line-search algorithm to choose the over-relaxation parameter and find an adaptive and robust way to choose the acceleration depth.

Regarding the Cahn-Hilliard-Biot model, several important aspects are not addressed. Firstly, a well-posedness analysis should be performed. Then, an investigation of robust and efficient solution algorithms for the problem should be developed. One could for example try to use an extension of the time-discretization that was proposed in Paper E, together with alternating minimization or a monolithic Newton method. So far only standard alternating minimization has been applied and, only implicit evaluations in time have been done in the Biot-subsystem. Finally, the main motivation for developing the model was to utilize it to simulate tumor growth, and to do so it is reasonable to couple the system with transport of chemo-taxis and nutrients.

Regarding the solution strategies that were proposed in Paper E, there are several possible improvements to be made. A proper analysis of the time-discretization including error estimates and sensitivity on material parameters for the unconditional energy stability is missing. Moreover, proper treatment of the degenerate mobility term is not provided, and getting a good theory to cover this might be essential. It is also possible to destabilize the alternating minimization method that was proposed to improve the convergence speed of the method. This has, so far, only been tested in practice for very few problem setups, and although it looks promising, there is no theory on how to tune the destabilization. Finally, there are no available linearization methods that are robust and efficient for the fully implicit discretization in time.

Bibliography

[1] Adler, J., Gaspar, F., Hu, X., Ohm, P., Rodrigo, C., and Zikatanov, L. Robust preconditioners for a new stabilized discretization of the poroelastic equations. *SIAM Journal on Scientific Computing 42*, 3 (2020), B761–B791.

[2] Adler, J., Gaspar, F., Hu, X., Rodrigo, C., and Zikatanov, L. Robust block preconditioners for Biot's model. *International Conference on Domain Decomposition Methods* (2017), 3–16.

[3] Allen, S., and Cahn, J. Ground state structures in ordered binary alloys with second neighbor interactions. *Acta Metallurgica 20*, 3 (1972), 423–433.

[4] Allen, S., and Cahn, J. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metallurgica 27*, 6 (1979), 1085–1095.

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

[6] Almani, T., Manea, A., Kumar, K., and Dogru, A. Convergence of the undrained split iterative scheme for coupling flow with geomechanics in heterogeneous poroelastic media. *Computational Geosciences 24*, 2 (2020), 551–569.

[7] Ambrosio, L., and Tortorelli, V. Approximation of functional depending on jumps by elliptic functional via γ-convergence. *Communications on Pure and Applied Mathematics 43*, 8 (1990), 999–1036.

[8] Ambrosio, L., and Tortorelli, V. On the approximation of functionals depending on jumps by quadratic, elliptic functionals. *Bollettino dell'Unione Matematica Italiana 6* (1992), 105–123.

[9] Amor, H., Marigo, J., and Maurini, C. Regularized formulation of the variational brittle fracture with unilateral contact: Numerical experiments. *Journal of the Mechanics and Physics of Solids 57*, 8 (2009), 1209–1229.

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

[11] Arnold, D., Brezzi, F., and Fortin, M. A stable finite element for the Stokes equations. *Calcolo 21*, 4 (1984), 337–344.

[12] Bartels, S., Nochetto, R., and Salgado, A. Discrete total variation flows without regularization. *IAM Journal on Numerical Analysis 52*, 1 (2014), 363–385.

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

[14] Bastian, P., Blatt, M., Dedner, A., Engwer, C., Klöfkorn, R., Ohlberger, M., and Sander, O. A generic grid interface for parallel and adaptive scientific computing. Part I: abstract framework. *Computing 82*, 2 (2008), 103–119.

[15] Bause, M., Both, J., and Radu, F. Iterative coupling for fully dynamic poroelasticity. *Numerical Mathematics and Advanced Applications ENUMATH 2019* (2021), 115–123.

[16] Bause, M., Köcher, U., and Radu, F. Convergence of a continuous galerkin method for mixed hyperbolic-parabolic systems. *arXiv:2201.12014* (2022).

[17] Bertrand, Q., and Massias, M. Anderson acceleration of coordinate descent. *International Conference on Artificial Intelligence and Statistics* (2021), 1288–1296.

[18] Bertsekas, D. *Nonlinear programming*, vol. 48. Taylor & Francis, 1997.

[19] Biot, M. General theory of three-dimensional consolidation. *Journal of applied physics 12*, 2 (1941), 155–164.

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

[21] Bjørnarå, T., Nordbotten, J., and Park, J. Vertically integrated models for coupled two-phase flow and geomechanics in porous media. *Water Resources Research 52*, 2 (2016), 1398–1417.

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

[23] Boffi, D., Brezzi, F., and Fortin, M. *Mixed finite element methods and applications*, vol. 44. Springer, 2013.

[24] Boon, W., Kuchta, M., Mardal, K., and Ruiz-Baier, R. Robust preconditioners for perturbed saddle-point problems and conservative discretizations of Biot's equations utilizing total pressure. *SIAM Journal on Scientific Computing 43*, 4 (2021), B961–B983.

[25] Borregales, M., Kumar, K., Radu, F., Rodrigo, C., and Gaspar, F. A partially parallel-in-time fixed-stress splitting method for biot's consolidation model. *Computers & Mathematics with Applications 77*, 6 (2019), 1466–1478.

[26] Borregales, M., Radu, F., Kumar, K., and Nordbotten, J. Robust iterative schemes for non-linear poromechanics. *Computational Geosciences 22*, 4 (2018), 1021–1038.

[27] Borregales R, M., Kumar, K., Nordbotten, J., and Radu, F. Iterative solvers for Biot model under small and large deformations. *Computational Geosciences 25*, 2 (2021), 687–699.

[28] Both, J. Mathematical and Numerical Analysis of Flow in Deformable Porous Media.

[29] Both, J. On the rate of convergence of alternating minimization for non-smooth non-strongly convex optimization in Banach spaces. *Optimization Letters* (2021), 1–15.

[30] Both, J., Barnafi, N., Radu, F., Zunino, P., and Quarteroni, A. Iterative splitting schemes for a soft material poromechanics model. *Computer Methods in Applied Mechanics and Engineering 388* (2022), 114183.

[31] Both, J., Borregales, M., Nordbotten, J., Kumar, K., and Radu, F. Robust fixed stress splitting for Biot's equations in heterogeneous media. *Applied Mathematics Letters 68* (2017), 101–108.

[32] Both, J., and Köcher, U. Numerical investigation on the fixed-stress splitting scheme for Biot's equations: Optimality of the tuning parameter. *European Conference on Numerical Mathematics and Advanced Applications* (2017), 789–797.

[33] Both, J., Kumar, K., Nordbotten, J., and Radu, F. Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media. *Computers & Mathematics with Applications 77*, 6 (2019), 1479–1502.

[34] Both, J., Kumar, K., Nordbotten, J., and Radu, F. The gradient flow structures of thermo-poro-visco-elastic processes in porous media. *arXiv:1907.03134* (2019).

[35] Both, J., Pop, I., and Yotov, I. Global existence of weak solutions to unsaturated poroelasticity. *ESAIM: Mathematical Modelling and Numerical Analysis 55*, 6 (2021), 2849–2897.

[36] Bourdin, B., Francfort, G., and Marigo, J. Numerical experiments in revisited brittle fracture. *Journal of the Mechanics and Physics of Solids 48*, 4 (2000), 797–826.

[37] Bourdin, B., Francfort, G., and Marigo, J. The variational approach to fracture. *Journal of elasticity 91*, 1 (2008), 5–148.

[38] Brenner, S., and Scott, L. *The mathematical theory of finite element methods*, vol. 3. Springer, 2008.

[39] Brezzi, F. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Publications mathématiques et informatique de Rennes S4* (1974), 1–26.

[40] Brun, M., Ahmed, E., Berre, I., Nordbotten, J., and Radu, F. Monolithic and splitting solution schemes for fully coupled quasi-static thermo-poroelasticity with nonlinear convective transport. *Computers & Mathematics with Applications 80*, 8 (2020), 1964–1984.

[41] Brun, M., Berre, I., Nordbotten, J., and Radu, F. Upscaling of the coupling of hydromechanical and thermal processes in a quasi-static poroelastic medium. *Transport in Porous Media 124*, 1 (2018), 137–158.

[42] Brun, M., Wick, T., Berre, I., Nordbotten, J., and Radu, F. An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters. *Computer Methods in Applied Mechanics and Engineering 361* (2020), 112752.

[43] Cahn, J. Free energy of a nonuniform system. II. thermodynamic basis. *Journal of Chemical Physics 30*, 5 (1959), 1121–1124.

[44] Cahn, J. On spinodal decomposition. *Acta Metallurgica 9*, 9 (1961), 795–801.

[45] Cahn, J., and Hilliard, J. Free energy of a nonuniform system. I. Interfacial free energy. *Journal of Chemical Physics 28*, 2 (1958), 258–267.

[46] Cahn, J., and Hilliard, J. Free energy of a nonuniform system. III. nucleation in a two-component incompressible fluid. *Journal of Chemical Physics 31*, 3 (1959), 688–699.

[47] Cahn, J., and Hilliard, J. Spinodal decomposition: A reprise. *Acta Metallurgica 19*, 2 (1971), 151–161.

[48] Cancès, C., Matthes, D., and Nabet, F. A two-phase two-fluxes degenerate Cahn–Hilliard model as constrained Wasserstein gradient flow. *Archive for Rational Mechanics and Analysis 233*, 2 (2019), 837–866.

[49] Castelletto, N., Klevtsov, S., Hajibeygi, H., and Tchelepi, H. Multiscale two-stage solver for Biot's poroelasticity equations in subsurface media. *Computational geosciences 23*, 2 (2019), 207–224.

[50] Castelletto, N., White, J., and Tchelepi, H. 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 (2015), 1593–1618.

[51] Celia, M., and Binning, P. A mass conservative numerical solution for two-phase flow in porous media with application to unsaturated flow. *Water Resources Research 28*, 10 (1992), 2819–2828.

[52] Cheney, W. *Analysis for applied mathematics*, vol. 1. Springer, 2001.

[53] Ciarlet, P. *Three-dimensional elasticity*. Elsevier, 1988.

[54] Ciarlet, P. *The finite element method for elliptic problems*. SIAM, 2002.

[55] Colli, P. On some doubly nonlinear evolution equations in Banach spaces. *Japan Journal of Industrial and Applied Mathematics 9*, 2 (1992), 181–203.

[56] Coussy, O. *Poromechanics*. John Wiley & Sons, 2004.

[57] Cueto-Felgueroso, L., and Juanes, R. A phase field model of unsaturated flow. *Water Resources Research 45*, 10 (2009).

[58] De Boer, R. *Theory of porous media: highlights in historical development and current state*. Springer Science & Business Media, 2012.

[59] Dede, L., Garcke, H., and Lam, K. A Hele–Shaw–Cahn–Hilliard model for incompressible two-phase flows with different densities. *Journal of Mathematical Fluid Mechanics 20*, 2 (2018), 531–567.

[60] Deuflhard, P. *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*, vol. 35. Springer Science & Business Media, 2005.

[61] Dreyer, W., and Müller, W. Modeling diffusional coarsening in eutectic tin/lead solders: a quantitative approach. *International Journal of Solids and Structures 38*, 8 (2001), 1433–1458.

[62] Elliott, C., and Stuart, A. The global dynamics of discrete semilinear parabolic equations. *SIAM Journal on Numerical Analysis 30*, 6 (1993), 1622–1663.

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

[64] Evans, C., Pollock, S., Rebholz, L., and Xiao, M. A proof that anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically). *SIAM Journal on Numerical Analysis 58*, 1 (2020), 788–810.

[65] Eyre, D. Unconditionally gradient stable time marching the Cahn-Hilliard equation. *MRS Online Proceedings Library (OPL) 529* (1998).

[66] Fang, H., and Saad, Y. Two classes of multisecant methods for nonlinear acceleration. *Numerical linear algebra with applications 16*, 3 (2009), 197–221.

[67] Farrell, P., and Maurini, C. Linear and nonlinear solvers for variational phase-field models of brittle fracture. *International Journal for Numerical Methods in Engineering 109*, 5 (2017), 648–667.

[68] Feng, X. Fully Discrete Finite Element Approximations of the Navier–Stokes–Cahn-Hilliard Diffuse Interface Model for Two-Phase Fluid Flows. *SIAM Journal on Numerical Analysis 44*, 3 (2006), 1049–1072.

[69] Fick, A. V. On liquid diffusion. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 10*, 63 (1855), 30–39.

[70] Flemisch, B., Fumagalli, A., and Scotti, A. A review of the XFEM-based approximation of flow in fractured porous media. *Advances in discretization methods* (2016), 47–76.

[71] Francfort, G., and Marigo, J. Revisiting brittle fracture as an energy minimization problem. *Journal of the Mechanics and Physics of Solids 46*, 8 (1998), 1319–1342.

[72] Frieboes, H., Jin, F., Chuang, Y., Wise, S., Lowengrub, J., and Cristini, V. Three-dimensional multispecies nonlinear tumor growth—II: tumor invasion and angiogenesis. *Journal of Theoretical Biology 264*, 4 (2010), 1254–1278.

[73] Fritz, M., Jha, P., Köppl, T., Oden, J., and Wohlmuth, B. Analysis of a new multispecies tumor growth model coupling 3D phase-fields with a 1D vascular network. *Nonlinear Analysis: Real World Applications 61* (2021), 103331.

[74] Fritz, M., Kuttler, C., Rajendran, M., Scarabosio, L., and Wohlmuth, B. On a subdiffusive tumour growth model with fractional time derivative. *IMA Journal of Applied Mathematics 86* (2021), 688 – 729.

[75] Fritz, M., Lima, E., Nikolić, V., Oden, J., and Wohlmuth, B. Local and nonlocal phase-field models of tumor growth and invasion due to ECM degradation. *Mathematical Models and Methods in Applied Sciences 29*, 13 (2019), 2433–2468.

[76] Galloway, D., and Burbey, T. Regional land subsidence accompanying groundwater extraction. *Hydrogeology Journal 19*, 8 (2011), 1459–1486.

[77] Garcke, H. On Cahn–Hilliard systems with elasticity. *Proceedings of the Royal Society of Edinburgh Section A 133*, 2 (2003), 307.

[78] Garcke, H. Mechanical effects in the Cahn-Hilliard model: A review on mathematical results. *Mathematical methods and models in phase transitions* (2005), 43–77.

[79] Garcke, H., Lam, K., and Signori, A. On a phase field model of Cahn–Hilliard type for tumour growth with mechanical effects. *Nonlinear Analysis: Real World Applications 57* (2021), 103192.

[80] Garcke, H., Lam, K., and Signori, A. Sparse optimal control of a phase field tumor model with mechanical effects. *SIAM Journal on Control and Optimization 59*, 2 (2021), 1555–1580.

[81] Garcke, H., and Weikard, U. Numerical approximation of the Cahn-Larché equation. *Numerische Mathematik 100*, 4 (2005), 639–662.

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

[83] Gerasimov, T., and De Lorenzis, L. A line search assisted monolithic approach for phase-field computing of brittle fracture. *Computer Methods in Applied Mechanics and Engineering 312* (2016), 276–303.

[84] Giovanardi, B., Formaggia, L., Scotti, A., and Zunino, P. Unfitted FEM for modelling the interaction of multiple fractures in a poroelastic medium. *Geometrically Unfitted Finite Element Methods and Applications* (2017), 331–352.

[85] Giovanardi, B., Scotti, A., and Formaggia, L. A hybrid XFEM–phase field (Xfield) method for crack propagation in brittle elastic materials. *Computer Methods in Applied Mechanics and Engineering 320* (2017), 396–420.

[86] Gordeliy, E., and Peirce, A. Coupling schemes for modeling hydraulic fracture propagation using the XFEM. *Computer Methods in Applied Mechanics and Engineering 253* (2013), 305–322.

[87] Gräser, C., Kienle, D., and Sander, O. Truncated Nonsmooth Newton Multigrid for phase-field brittle-fracture problems. *arXiv:2007.12290* (2020).

[88] Gräser, C., Kornhuber, R., and Sack, U. Numerical simulation of coarsening in binary solder alloys. *Computational Materials Science 93* (2014), 221–233.

[89] Griffith, A. VI. The phenomena of rupture and flow in solids. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character 221*, 582-593 (1921), 163–198.

[90] Guillén-González, F., and Tierra, G. Second order schemes and time-step adaptivity for Allen–Cahn and Cahn–Hilliard models. *Computers & Mathematics with Applications 68*, 8 (2014), 821–846.

[91] Guo, J., Wang, C., Wise, S., and Yue, X. An H^2 convergence of a second-order convex-splitting, finite difference scheme for the three-dimensional Cahn–Hilliard equation. *Communications in Mathematical Sciences 14*, 2 (2016), 489–515.

[92] Haga, J., Osnes, H., and Langtangen, H. Efficient block preconditioners for the coupled equations of pressure and deformation in highly discontinuous media. *International Journal for Numerical and Analytical Methods in Geomechanics 35*, 13 (2011), 1466–1482.

[93] Hawkes, C., McLellan, P., Zimmer, U., and Bachu, S. Geomechanical factors affecting geological storage of CO2 in depleted oil and gas reservoirs. *Canadian International Petroleum Conference* (2004).

[94] Heister, T., Wheeler, M., and Wick, T. A primal-dual active set method and predictor-corrector mesh adaptivity for computing fracture propagation using a phase-field approach. *Computer Methods in Applied Mechanics and Engineering 290* (2015), 466–495.

[95] Hong, Q., Kraus, J., Lymbery, M., and Philo, F. Parameter-robust Uzawa-type iterative methods for double saddle point problems arising in Biot's consolidation and multiple-network poroelasticity models. *Mathematical Models and Methods in Applied Sciences 30*, 13 (2020), 2523–2555.

[96] Hong, Q., Kraus, J., Lymbery, M., and Wheeler, M. Parameter-robust convergence analysis of fixed-stress split iterative method for multiple-permeability poroelasticity systems. *Multiscale Modeling & Simulation 18*, 2 (2020), 916–941.

[97] Hu, X., Rodrigo, C., Gaspar, F., and Zikatanov, L. A nonconforming finite element method for the Biot's consolidation model in poroelasticity. *Journal of Computational and Applied Mathematics 310* (2017), 143–154.

[98] Illiano, D., Both, J., Pop, I., and Radu, F. Efficient solvers for nonstandard models for flow and transport in unsaturated porous media. *arXiv:2012.14773* (2020).

[99] Illiano, D., Pop, I., and Radu, F. Iterative schemes for surfactant transport in porous media. *Computational Geosciences 25*, 2 (2021), 805–822.

[100] Jha, B., and Juanes, R. Coupled multiphase flow and poromechanics: A computational model of pore pressure effects on fault slip and earthquake triggering. *Water Resources Research 50*, 5 (2014), 3776–3808.

[101] Jüngel, A., Stefanelli, U., and Trussardi, L. Two structure-preserving time discretizations for gradient flows. *Applied Mathematics & Optimization 80*, 3 (2019), 733–764.

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

[103] Kim, J. *Sequential methods for coupled geomechanics and multiphase flow.* Stanford University, 2010.

[104] Kim, J., Tchelepi, H., and Juanes, R. Stability and convergence of sequential methods for coupled flow and geomechanics: Drained and undrained splits. *Computer Methods in Applied Mechanics and Engineering 200*, 23-24 (2011), 2094–2116.

[105] Kim, J., Tchelepi, H., and Juanes, R. 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*, 13-16 (2011), 1591–1606.

[106] Kristensen, P., and Martínez-Pañeda, E. Phase field fracture modelling using quasi-Newton methods and a new adaptive step scheme. *Theoretical and Applied Fracture Mechanics 107* (2020), 102446.

[107] Kuhn, C., and Müller, R. A continuum phase field model for fracture. *Engineering Fracture Mechanics 77*, 18 (2010), 3625–3634.

[108] Langer, J. Theory of spinodal decomposition in alloys. *Annals of Physics 65*, 1 (1971), 53–86.

[109] Larché, F., and Cahn, J. A linear theory of thermochemical equilibrium of solids under stress. *Acta Metallurgica 21*, 8 (1973), 1051–1063.

[110] Larché, F., and Cahn, J. The effect of self-stress on diffusion in solids. *Acta Metallurgica 30*, 10 (1982), 1835–1845.

[111] Lee, J., Mardal, K., and Winther, R. Parameter-robust discretization and preconditioning of Biot's consolidation model. *SIAM Journal on Scientific Computing 39*, 1 (2017), A1–A24.

[112] Lee, J., Piersanti, E., Mardal, K., and Rognes, M. A mixed finite element method for nearly incompressible multiple-network poroelasticity. *SIAM Journal on Scientific Computing 41*, 2 (2019), A722–A747.

[113] Lewis, R., and Schrefler, B. *The finite element method in the static and dynamic deformation and consolidation of porous media.* John Wiley & Sons, 1998.

[114] Lima, E., Almeida, R., and Oden, J. Analysis and numerical solution of stochastic phase-field models of tumor growth. *Numerical methods for partial differential equations 31*, 2 (2015), 552–574.

[115] Lima, E., Oden, J., Hormuth, D., Yankeelov, T., and Almeida, R. Selection, calibration, and validation of models of tumor growth. *Mathematical Models and Methods in Applied Sciences 26*, 12 (2016), 2341–2368.

[116] List, F., and Radu, F. A study on iterative methods for solving Richards' equation. *Computational Geosciences 20*, 2 (2016), 341–353.

[117] Meca, E., Münch, A., and Wagner, B. Sharp-interface formation during lithium intercalation into silicon. *European Journal of Applied Mathematics 29*, 1 (2018), 118–145.

[118] Miehe, C., Hofacker, M., and Welschinger, F. A phase field model for rate-independent crack propagation: Robust algorithmic implementation based on operator splits. *Computer Methods in Applied Mechanics and Engineering 199*, 45-48 (2010), 2765–2778.

[119] Miehe, C., Welschinger, F., and Hofacker, M. Thermodynamically consistent phase-field models of fracture: Variational principles and multi-field FE implementations. *International Journal for Numerical Methods in Engineering 83*, 10 (2010), 1273–1311.

[120] Mikelić, A., Wang, B., and Wheeler, M. Numerical convergence study of iterative coupling for coupled flow and geomechanics. *Computational Geosciences 18*, 3 (2014), 325–341.

[121] Mikelić, A., and Wheeler, M. Convergence of iterative coupling for coupled flow and geomechanics. *Computational Geosciences 17*, 3 (2013), 455–461.

[122] Mikelic, A., Wheeler, M., and Wick, T. A phase-field method for propagating fluid-filled fractures coupled to a surrounding porous medium. *Multiscale Modeling & Simulation 13*, 1 (2015), 367–398.

[123] Moës, N., Dolbow, J., and Belytschko, T. A finite element method for crack growth without remeshing. *International Journal for Numerical Methods in Engineering 46*, 1 (1999), 131–150.

[124] Nochetto, R., Savaré, G., and Verdi, C. A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations. *Communications on Pure and Applied Analysis 53*, 5 (2000), 525–589.

[125] Nordbotten, J., and Celia, M. *Geological Storage of CO2: Modeling Approaches for Large-Scale Simulation.* John Wiley and Sons, 2011.

[126] Oden, J., Hawkins, A., and Prudhomme, S. General diffuse-interface theories and an approach to predictive tumor growth modeling. *Mathematical Models and Methods in Applied Sciences 20*, 03 (2010), 477–517.

[127] Peletier, M. Variational modelling: Energies, gradient flows, and large deviations. *arXiv:1402.1990* (2014).

[128] Peng, Y., Deng, B., Zhang, J., Geng, F., Qin, W., and Liu, L. Anderson acceleration for geometry optimization and physics simulation. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 1–14.

[129] Philibert, J. One and a Half Century of Diffusion: Fick, Einstein. *Diffusion Fundamentals: Leipzig 2005 1* (2005), 8.

[130] Pollock, S., Rebholz, L., and Xiao, M. Anderson-accelerated convergence of Picard iterations for incompressible Navier–Stokes equations. *SIAM Journal on Numerical Analysis 57*, 2 (2019), 615–637.

[131] Pop, I., Radu, F., and Knabner, P. Mixed finite elements for the Richards' equation: linearization procedure. *Journal of computational and applied mathematics 168*, 1-2 (2004), 365–373.

[132] Quarteroni, A. *Modeling the heart and the circulatory system*, vol. 14. Springer, 2015.

[133] Raviart, P., and Thomas, J. A mixed finite element method for 2-nd order elliptic problems. *Mathematical aspects of finite element methods* (1977), 292–315.

[134] Roudbari, M., Şimşek, G., van Brummelen, E., and van der Zee, K. Diffuse-interface two-phase flow models with different densities: A new quasi-incompressible form and a linear energy-stable method. *Mathematical Models and Methods in Applied Sciences 28*, 04 (2018), 733–770.

[135] Roudbari, M., van Brummelen, E. H., and Verhoosel, C. V. A multiscale diffuse-interface model for two-phase flow in porous media. *Computers & Fluids 141* (2016), 212–222.

[136] Sander, O. *DUNE—The Distributed and Unified Numerics Environment*, vol. 140. Springer Nature, 2020.

[137] Sargado, J., Keilegavlen, E., Berre, I., and Nordbotten, J. High-accuracy phase-field models for brittle fracture based on a new family of degradation functions. *Journal of the Mechanics and Physics of Solids 111* (2018), 458–489.

[138] Settari, A., and Mourits, F. A coupled reservoir and geomechanical simulation system. *Spe Journal 3*, 03 (1998), 219–226.

[139] Shi, S., Markmann, J., and Weissmüller, J. Verifying Larché–Cahn elasticity, a milestone of 20th-century thermodynamics. *Proceedings of the National Academy of Sciences 115*, 43 (2018), 10914–10919.

[140] Shi, W., Song, S., Wu, H., Hsu, Y., Wu, C., and Huang, G. Regularized Anderson acceleration for off-policy deep reinforcement learning. *Advances in Neural Information Processing Systems 32* (2019).

[141] Signori, A. Optimal treatment for a phase field system of Cahn-Hilliard type modeling tumor growth by asymptotic scheme. *Mathematical Control and Related Fields 10* (2020), 305–331.

[142] Singh, N., Verhoosel, C., and van Brummelen, E. Finite element simulation of pressure-loaded phase-field fractures. *Meccanica 53*, 6 (2018), 1513–1545.

[143] Storvik, E., Both, J., Kumar, K., Nordbotten, J., and Radu, F. On the optimization of the fixed-stress splitting for biot's equations. *International Journal for Numerical Methods in Engineering 120*, 2 (2019), 179–194.

[144] Storvik, E., Both, J., Nordbotten, J., and Radu, F. The fixed-stress splitting scheme for Biot's equations as a modified Richardson iteration: Implications for optimal convergence. *Numerical Mathematics and Advanced Applications ENUMATH 2019* (2021), 909–917.

[145] Storvik, E., Both, J., Nordbotten, J., and Radu, F. A Cahn–Hilliard–Biot system and its generalized gradient flow structure. *Applied Mathematics Letters 126* (2022), 107799.

[146] Storvik, E., Both, J., Nordbotten, J., and Radu, F. A robust solution strategy for the Cahn-Larché equations. *arXiv:2206.01541* (2022).

[147] Storvik, E., Both, J., Sargado, J., Nordbotten, J., and Radu, F. An accelerated staggered scheme for variational phase-field models of brittle fracture. *Computer Methods in Applied Mechanics and Engineering 381* (2021), 113822.

[148] Suryanarayana, P., Pratapa, P., and Pask, J. Alternating Anderson–Richardson method: An efficient alternative to preconditioned Krylov methods for large, sparse linear systems. *Computer Physics Communications 234* (2019), 278–285.

[149] Tanné, E., Li, T., Bourdin, B., Marigo, J., and Maurini, C. Crack nucleation in variational phase-field models of brittle fracture. *Journal of the Mechanics and Physics of Solids 110* (2018), 80–99.

[150] Taylor, C., and Hood, P. A numerical solution of the Navier-Stokes equations using the finite element technique. *Computers & Fluids 1*, 1 (1973), 73–100.

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

[152] Toth, A., and Kelley, C. Convergence analysis for Anderson acceleration. *SIAM Journal on Numerical Analysis 53*, 2 (2015), 805–819.

[153] Tully, B., and Ventikos, Y. Cerebral water transport using multiple-network poroelastic theory: application to normal pressure hydrocephalus. *Journal of Fluid Mechanics 667* (2011), 188–215.

[154] Uzuoka, R., and Borja, R. I. Dynamics of unsaturated poroelastic solids at finite strain. *International Journal for Numerical and Analytical Methods in Geomechanics 36*, 13 (2012), 1535–1573.

[155] van Brummelen, E. Partitioned iterative solution methods for fluid–structure interaction. *International Journal for Numerical Methods in Fluids 65*, 1-3 (2011), 3–27.

[156] van Brummelen, E., Roudbari, M., Simsek, G., and van der Zee, K. Binary-fluid–solid interaction based on the Navier–Stokes–Cahn–Hilliard equations. *Fluid Structure Interaction 20* (2017), 283–328.

[157] Wang, S., Lee, C., and Hsu, K. A technique for quantifying groundwater pumping and land subsidence using a nonlinear stochastic poroelastic model. *Environmental Earth Sciences 73*, 12 (2015), 8111–8124.

[158] Wheeler, M., Wick, T., and Wollner, W. An augmented-Lagrangian method for the phase-field approach for pressurized fractures. *Computer Methods in Applied Mechanics and Engineering 271* (2014), 69–85.

[159] Wick, T. Modified Newton methods for solving fully monolithic phase-field quasi-static brittle fracture propagation. *Computer Methods in Applied Mechanics and Engineering 325* (2017), 577–611.

[160] Wise, S., Lowengrub, J., Frieboes, H., and Cristini, V. Three-dimensional multispecies nonlinear tumor growth—I: model and numerical method. *Journal of Theoretical Biology 253*, 3 (2008), 524–543.

[161] Wu, J., Huang, Y., and Nguyen, V. P. On the BFGS monolithic algorithm for the unified phase field damage theory. *Computer Methods in Applied Mechanics and Engineering 360* (2020), 112704.

[162] Zhang, J., O'Donoghue, B., and Boyd, S. Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations. *SIAM Journal on Optimization 30*, 4 (2020), 3170–3197.

Part II
Scientific results

Paper A

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**, 179–194 (2019)

Received: 6 November 2018 | Revised: 6 April 2019 | Accepted: 23 May 2019

DOI: 10.1002/nme.6130

RESEARCH ARTICLE

WILEY

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 Forskningsråd, 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

$$-\nabla \cdot (2\mu\boldsymbol{\epsilon}(\boldsymbol{u}) + \lambda\nabla \cdot \boldsymbol{u}\boldsymbol{I}) + \alpha\nabla p = \boldsymbol{f}, \tag{1}$$

$$\frac{\partial}{\partial t}\left(\frac{p}{M} + \alpha\nabla \cdot \boldsymbol{u}\right) - \nabla \cdot (\kappa(\nabla p - \boldsymbol{g}\rho)) = S_f, \tag{2}$$

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

$$2\mu \left\langle \boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^n\right), \boldsymbol{\varepsilon}\left(\boldsymbol{v}_h\right)\right\rangle + \lambda \left\langle \nabla \cdot \boldsymbol{u}_h^n, \nabla \cdot \boldsymbol{v}_h\right\rangle - \alpha \left\langle p_h^n, \nabla \cdot \boldsymbol{v}_h\right\rangle = \left\langle \boldsymbol{f}^n, \boldsymbol{v}_h\right\rangle, \tag{3}$$

$$\frac{1}{M}\left\langle p_h^n - p_h^{n-1}, q_h\right\rangle + \alpha \left\langle \nabla \cdot \left(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\right), q_h\right\rangle + \tau \left\langle \kappa \nabla p_h^n, \nabla q_h\right\rangle - \tau \left\langle \kappa \mathbf{g}\rho, \nabla q_h\right\rangle = \tau \left\langle S_f^n, q_h\right\rangle \tag{4}$$

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

$$2\mu \left\langle \boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right), \boldsymbol{\varepsilon}\left(\boldsymbol{v}_h\right)\right\rangle + \lambda \left\langle \nabla \cdot \boldsymbol{u}_h^{n,i}, \nabla \cdot \boldsymbol{v}_h\right\rangle - \alpha \left\langle p_h^{n,i}, \nabla \cdot \boldsymbol{v}_h\right\rangle = \left\langle \boldsymbol{f}^n, \boldsymbol{v}_h\right\rangle, \tag{5}$$

$$\begin{aligned}\frac{1}{M}\left\langle p_h^{n,i} - p_h^{n-1}, q_h\right\rangle + \alpha \left\langle \nabla \cdot \left(\boldsymbol{u}_h^{n,i-1} - \boldsymbol{u}_h^{n-1}\right), q_h\right\rangle + L \left\langle p_h^{n,i} - p_h^{n,i-1}, q_h\right\rangle \\ + \tau \left\langle \kappa \nabla p_h^{n,i}, \nabla q_h\right\rangle - \tau \left\langle \kappa \mathbf{g}\rho, \nabla q_h\right\rangle = \tau \left\langle S_f^n, q_h\right\rangle\end{aligned} \tag{6}$$

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

$$2\mu \|\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h\right)\|^2 + \lambda \|\nabla \cdot \boldsymbol{u}_h\|^2 \geq K_{\mathrm{dr}}^\star \|\nabla \cdot \boldsymbol{u}_h\|^2 \qquad \text{for all } \boldsymbol{u}_h \in \mathbf{V}_h. \tag{7}$$

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:

$$2\mu \|\boldsymbol{\varepsilon}(\boldsymbol{u}_h)\|^2 + \lambda \|\nabla \cdot \boldsymbol{u}_h\|^2 \leq C \|\boldsymbol{u}_h\|_{H^1(\Omega)}^2 \leq C \|L_B\|^2 \|p_h\|^2,$$

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,

$$\|L_B\| = \sup_{0 \neq p_h \in Q_h} \frac{\|L_B\left(\langle p_h, \cdot \rangle\right)\|_{H^1(\Omega)}}{\|\langle p_h, \cdot \rangle\|_{L^2(\Omega)'}} = \sup_{\substack{0 \neq p_h \in Q_h \\ \boldsymbol{u}_h = L_B\left(\langle p_h, \cdot \rangle\right)}} \frac{\|\boldsymbol{u}_h\|_{H^1(\Omega)}}{\|p_h\|}.$$

We obtain our desired inequality, as follows:

$$2\mu \|\boldsymbol{\varepsilon}(\boldsymbol{u}_h)\|^2 + \lambda \|\nabla \cdot \boldsymbol{u}_h\|^2 \leq \frac{C}{\gamma^2} \|p_h\|^2 = \beta \|p_h\|^2.$$

□

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 $K_{\mathrm{dr}}^{\star}$, 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 $L \geq \frac{\alpha^2}{\delta K_{\mathrm{dr}}^{\star}}$, 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

$$\left\| e_p^{n,i} \right\|^2 \leq rate\,(L, \delta) \left\| e_p^{n,i-1} \right\|^2, \tag{10}$$

$$2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 \leq \frac{\alpha^2}{K_{\mathrm{dr}}^{\star}} \left\| e_p^{n,i} \right\|^2, \tag{11}$$

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

$$\begin{cases} (i)\ \ 2\mu \left\langle \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right), \boldsymbol{\varepsilon}\left(\boldsymbol{v}_h\right) \right\rangle + \lambda \left\langle \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i}, \nabla \cdot \boldsymbol{v}_h \right\rangle - \alpha \left\langle e_p^{n,i}, \nabla \cdot \boldsymbol{v}_h \right\rangle = 0, \\ (ii)\ \ \frac{1}{M} \left\langle e_p^{n,i}, q_h \right\rangle + \alpha \left\langle \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1}, q_h \right\rangle + L \left\langle e_p^{n,i} - e_p^{n,i-1}, q_h \right\rangle + \tau \left\langle \kappa \nabla e_p^{n,i}, \nabla q_h \right\rangle = 0, \end{cases} \tag{12}$$

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

$$2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i}\right) \right\|^2 + \lambda \left\| \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 \leq \frac{\alpha^2}{2K_{\mathrm{dr}}^{\star}} \left\| e_p^{n,i} \right\|^2 + \frac{K_{\mathrm{dr}}^{\star}}{2} \left\| \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2. \tag{13}$$

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

$$2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right) \right\|^2 + \lambda \left\| \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2 + \frac{1}{M} \left\| e_p^{n,i} \right\|^2 - \alpha \left\langle e_p^{n,i}, \nabla \cdot \left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} - \boldsymbol{e}_{\boldsymbol{u}}^{n,i-1} \right) \right\rangle + \tau\kappa \left\| \nabla e_p^{n,i} \right\|^2 + \frac{L}{2} \left\| e_p^{n,i} - e_p^{n,i-1} \right\|^2 + \frac{L}{2} \left\| e_p^{n,i} \right\|^2 = \frac{L}{2} \left\| e_p^{n,i-1} \right\|^2.$$

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

$$2\mu \|\boldsymbol{\varepsilon}(\boldsymbol{v}_h)\|^2 + \lambda \|\nabla \cdot \boldsymbol{v}_h\|^2 \leq \beta \left\| e_p^{n,i} \right\|^2. \tag{20}$$

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

$$\frac{\alpha^2}{\beta} \left\| e_p^{n,i} \right\|^2 \leq 2\mu \left\| \boldsymbol{\varepsilon}\left(\boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right) \right\|^2 + \lambda \left\| \nabla \cdot \boldsymbol{e}_{\boldsymbol{u}}^{n,i} \right\|^2, \tag{22}$$

which, together with (19), implies

$$\left(\frac{1}{M} + \frac{L}{2} + \frac{\tau\kappa}{C_\Omega^2} + \left(1 - \frac{\delta}{2}\right) \frac{\alpha^2}{\beta} \right) \left\| e_p^{n,i} \right\|^2 \leq \frac{L}{2} \left\| e_p^{n,i-1} \right\|^2.$$

This gives the following rate of convergence, for $\delta \in (0, 2]$ and $L \geq \frac{\alpha^2}{\delta K_{\mathrm{dr}}^\star}$:

$$\mathrm{rate}(L, \delta) = \frac{L}{L + \frac{2}{M} + \frac{2\tau\kappa}{C_\Omega^2} + (2 - \delta)\frac{\alpha^2}{\beta}}.$$

□

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 $\kappa_m \geq 0$. 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 $\mathrm{rate}(L, \delta)$ 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 $\mathrm{rate}(\delta)$ 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	$\frac{a^2}{K_{\text{dr}}}$	$\frac{a^2}{2K_{\text{dr}}}$	Section 3.2	$\frac{a^2}{4\mu+2\lambda}$

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

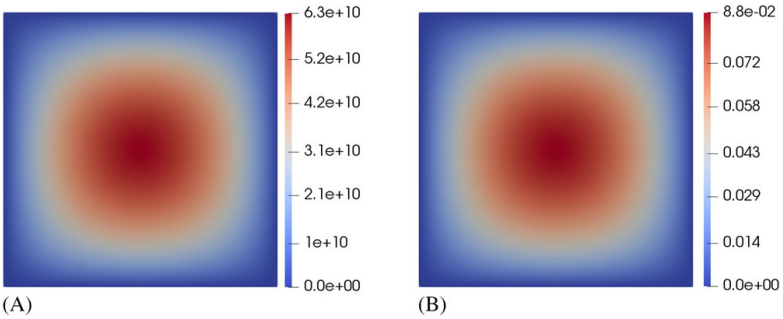


FIGURE 1 Unit square test case: solution—BC1. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$)

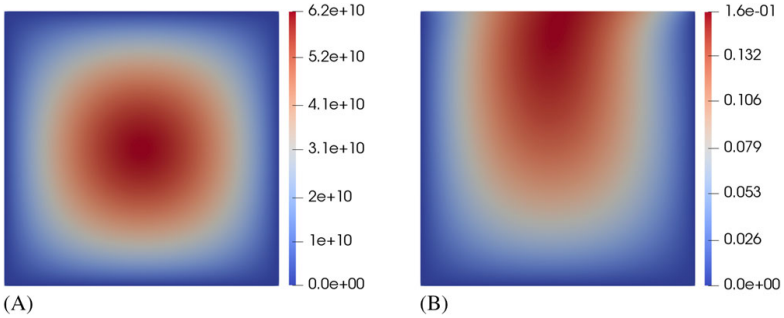


FIGURE 2 Unit square test case: solution—BC2. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$)

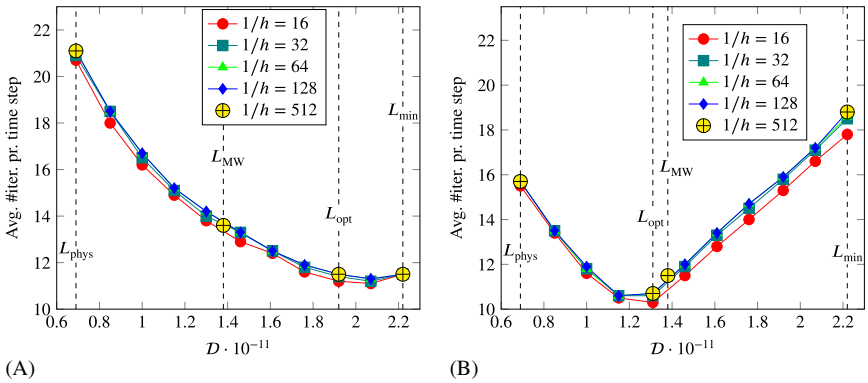


FIGURE 3 Unit square test case: average number of iterations per time step for different stabilization parameters, $L = \frac{\alpha^2}{\mathcal{D}}$, using parameters from Table 2. The largest value of $\mathcal{D}$ corresponds to L_{min}, whereas the smallest value of $\mathcal{D}$ corresponds to L_{phys}. 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

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

$$\mu = \frac{E}{2(1+\nu)} \quad \text{and} \quad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}. \tag{26}$$

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}

TABLE 3 Parameters used in Section 4.3

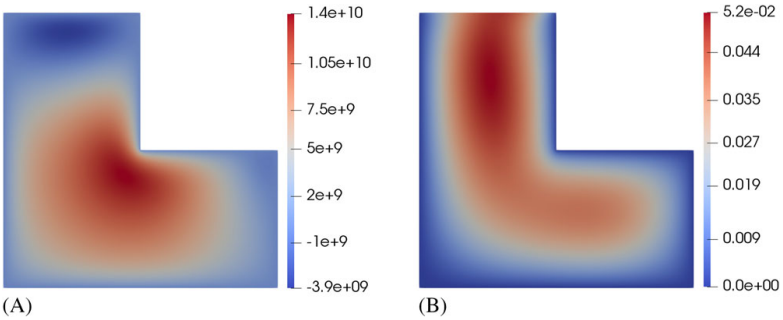


FIGURE 4 L-shaped domain test case: solution for $\nu = 0$. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$)

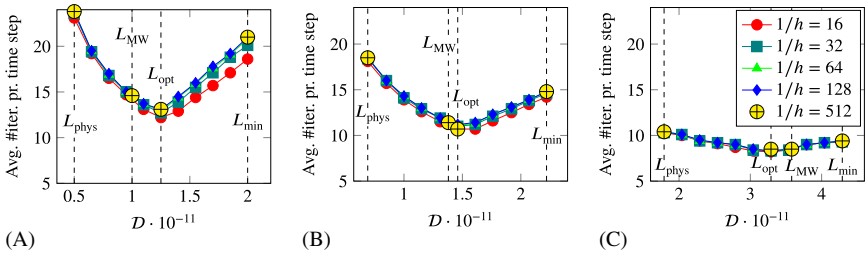


FIGURE 5 L-shaped domain test case: number of iterations for different stabilization parameters, $L = \frac{\alpha^2}{D}$, using parameters from Table 3. The largest value of D corresponds to $L_{\min}$, whereas the smallest value of D corresponds to L_{phys}. Notice that the axes are different. A, $\nu = 0$; B, $\nu = 0.2$; C, $\nu = 0.4$

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 $[L_{\min}, L_{\text{phys}})$ 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 $[L_{\min}, L_{\text{phys}}]$ 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

$$p = \frac{2FB(1+\nu_u)}{3a}\sum_{n=1}^{\infty}\frac{\sin(\alpha_n)}{\alpha_n - \sin(\alpha_n)\cos(\alpha_n)}\left(\cos\left(\frac{\alpha_n x}{a}\right) - \cos(\alpha_n)\right)e^{-\frac{\alpha_n^2 c_f t}{a^2}}, \tag{27}$$

$$u_x = \left[\frac{F\nu}{2\mu a} - \frac{F\nu_u}{\mu a}\sum_{n=1}^{\infty}\frac{\sin(\alpha_n)\cos(\alpha_n)}{\alpha_n - \sin(\alpha_n)\cos(\alpha_n)}e^{-\frac{\alpha_n^2 c_f t}{a^2}}\right]x + \frac{F}{\mu}\sum_{n=1}^{\infty}\frac{\cos(\alpha_n)}{\alpha_n - \sin(\alpha_n)\cos(\alpha_n)}\sin\left(\frac{\alpha_n x}{a}\right)e^{-\frac{\alpha_n^2 c_f t}{a^2}}, \tag{28}$$

$$u_y = \left[\frac{-F(1-\nu)}{2\mu a} + \frac{F(1-\nu_u)}{\mu a}\sum_{n=1}^{\infty}\frac{\sin(\alpha_n)\cos(\alpha_n)}{\alpha_n - \sin(\alpha_n)\cos(\alpha_n)}e^{-\frac{\alpha_n^2 c_f t}{a^2}}\right]y, \tag{29}$$

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}	–

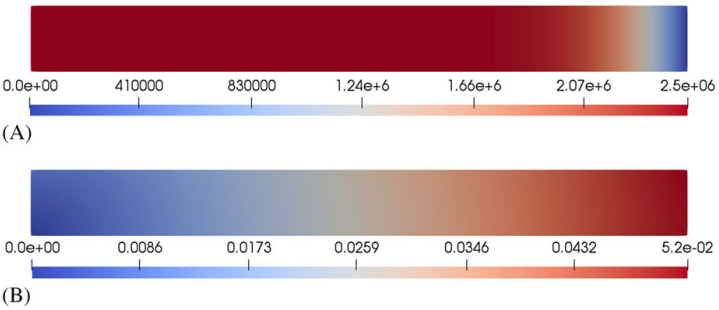


FIGURE 6 Mandel's problem: solution after 10 time steps with $N_x = N_y = 80$. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$)

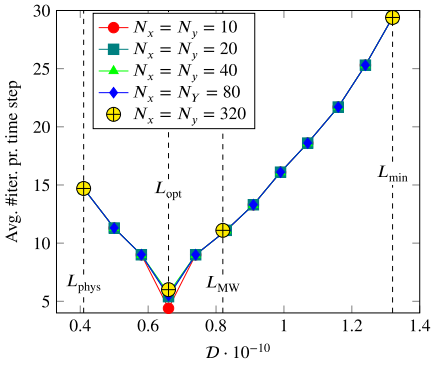


FIGURE 7 Mandel's problem: number of iterations for different stabilization parameters, $L = \frac{\alpha^2}{D}$, using parameters from Table 4. The largest value of D corresponds to L_{min}, whereas the smallest value of D corresponds to L_{phys}

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

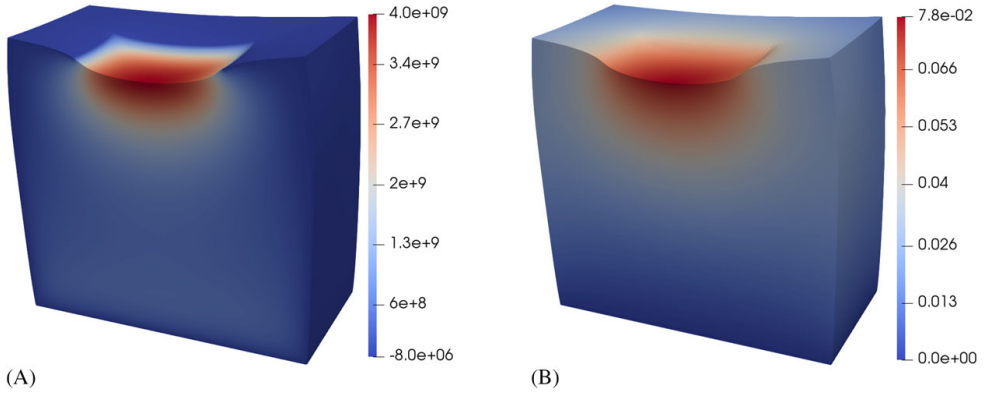


FIGURE 8 Three-dimensional footing problem: solution with a deformed configuration magnified by a factor of 2 at the final time $T = 1$. Notice that the figure only displays half of the domain but that the other half is symmetric. A, Pressure; B, Displacement($|\boldsymbol{u}_h|$)

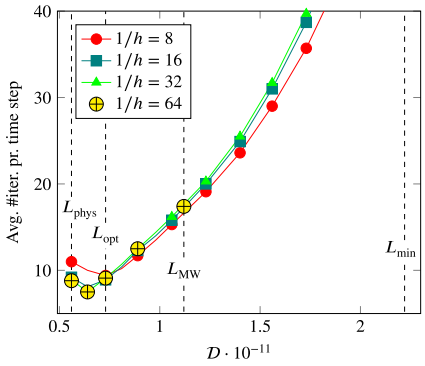


FIGURE 9 Three-dimensional footing problem: average number of iterations per time step for different stabilization parameters, $L = \frac{\alpha^2}{\mathcal{D}}$, using parameters from Table 2. The largest value of $\mathcal{D}$ corresponds to L_{min}, whereas the smallest value of $\mathcal{D}$ corresponds to L_{phys}

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;120:179–194. https://doi.org/10.1002/nme.6130

Paper C

An accelerated staggered scheme for variational phase-field models of brittle fracture

Storvik, E., Both, J.W., Sargado, J.M., Nordbotten, J.M., and Radu, F.A.
Computational Methods in Applied Mechanics and Engineering, **381**, 113822 (2021)

Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 381 (2021) 113822
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

An accelerated staggered scheme for variational phase-field models of brittle fracture

Erlend Storvik[a,*], Jakub Wiktor Both[a], Juan Michael Sargado[b], Jan Martin Nordbotten[a], Florin Adrian Radu[a]

[a] Department of Mathematics, University of Bergen, Allégaten 44, 5007 Bergen, Norway
[b] Danish Hydrocarbon Research and Technology Centre, Technological University of Denmark, Elektrovej Bygning 375, 2800 Kgs Lyngby, Denmark

Received 10 September 2020; received in revised form 17 March 2021; accepted 21 March 2021
Available online 19 April 2021

Abstract

There is currently an increasing interest in developing efficient solvers for variational phase-field models of brittle fracture. The governing equations for this problem originate from a constrained minimization of a non-convex energy functional, and the most commonly used solver is a staggered solution scheme. This is known to be robust compared to the monolithic Newton method, however, the staggered scheme often requires many iterations to converge when cracks are evolving. The focus of our work is to accelerate the solver through a scheme that sequentially applies Anderson acceleration and over-relaxation, switching back and forth depending on the residual evolution, and thereby ensuring a decreasing tendency. The resulting scheme takes advantage of the complementary strengths of Anderson acceleration and over-relaxation to make a robust and accelerating method for this problem. The new method is applied as a post-processing technique to the increments of the solver. Hence, the implementation merely requires minor modifications to already available software. Moreover, the cost of the acceleration scheme is negligible. The robustness and efficiency of the method are demonstrated through numerical examples.

Keywords: Variational brittle fracture; Phase-field modeling; Staggered scheme; Anderson acceleration; Relaxation; Nonlinear solver

1. Introduction

Mathematical modeling of brittle fracture propagation is an important and challenging topic in engineering sciences. The main difficulty arises in the transition between the distinct material properties in the fracture and the bulk domain. In this paper, we consider a variational phase-field model, as introduced by Bourdin, Francfort, and Marigo [1,2]. A smooth indicator that marks the broken and unbroken parts of the material regularizes the sharp crack topology. This enables modeling of fractures without conforming meshes or path-tracking algorithms (as in XFEM [3]). However, fine meshes are needed to resolve the regularized region between the fracture and the bulk domain.

The system is modeled by minimizing its energy as a function of material displacement and the indicator function under a non-healing constraint. This leads to a system of coupled, nonlinear equations which is challenging to

* Corresponding author.
E-mail address: erlend.storvik@uib.no (E. Storvik).

https://doi.org/10.1016/j.cma.2021.113822

solve. The most common technique, due to its robust nature, is a staggered scheme. This method decouples the system and sequentially updates the displacement and indicator variable by solving their respective subproblems. However, the convergence properties are at times very bad, and iterating to satisfactory precision can result in large numbers of iterations [4,5]. The monolithic Newton method, on the other hand, does not show the same numerical robustness. Therefore, several attempts have been made to find a method that is both fast and robust. A monolithic, modified Newton method was proposed in [6], a monolithic quasi-Newton method of Broyden–Fletcher–Goldfarb–Shanno type was applied in [7] and [8], a monolithic line-search Newton method (dependent on the system energy) was applied in [4], and the truncated nonsmooth Newton multigrid method was proposed in [9]. In [10], the L-scheme [11,12] was applied in the context of an augmented Lagrangian solver, and a combination of an over-relaxed staggered scheme and the monolithic Newton method was applied in [5].

In this paper, we propose a novel strategy to accelerate the classical staggered solution scheme solely utilizing two techniques for post-processing increments: Anderson acceleration and over-relaxation. In addition to accelerating the staggered scheme without sacrificing robustness, the new method allows the use of already available staggered scheme solvers with minor modifications to the implementation.

Anderson acceleration was first developed in [13] for integral equations. Since then, it has seen many applications, including electronic structure computations [14] and flow in deformable porous media [15]. It is a multi-secant, quasi-Newton method that has been related to a preconditioned GMRES [16]. Moreover, the method post-processes the increments of the solver by approximating the inverse of the Jacobian of the system by reusing previous iterations. It can, therefore, easily be applied in combination with splitting techniques such as the staggered scheme while maintaining the decoupled nature of the scheme.

In [17], the authors show theoretically that the Anderson acceleration improves the convergence rate of linearly convergent schemes, which is the case of the staggered scheme. However, as proved in [18], the convergence is only local. In the case of phase-field modeling of brittle fracture, this is a challenge; when fractures are initiated or propagating, the system state usually jumps drastically between consecutive loading steps. Therefore, a "naive" application of Anderson acceleration is not suitable for this application, as will be demonstrated in the numerical examples of this work. Recently, there has been an increasing interest in modified Anderson acceleration methods to overcome issues of local convergence. In [19] a safeguard, based on the residual norm of the problem, is applied to restart Anderson acceleration, and in [20] a periodically restarted Anderson acceleration is applied within a Richardson fixed-point iteration to accelerate the convergence of iterative solvers for large sparse linear systems.

Relaxation was applied to the staggered scheme on a phase-field model of brittle fracture in [5]. It is a post-processing method that updates each iterate by relaxing (scaling) its increment. For the purpose of this work, over-relaxation (a scaling larger than one) is of particular interest. This is because the staggered scheme steadily moves towards the final configuration of each loading step, and over-relaxation might move the iterates further during each iteration, potentially accelerating the convergence. For the particular loading steps in which fractures are propagating, the gain can be quite substantial. There is, however, a drawback with over-relaxation: Near the solution of each loading step one might end up over- and undershooting the solution sequentially leading to poor performance.

The most important observation of this paper is the complementary strengths of these two acceleration techniques; Anderson acceleration accelerates close to the solution, while over-relaxation accelerates during loading steps with large jumps in the solution (e.g., during crack propagation). We propose an acceleration algorithm that switches between Anderson acceleration and over-relaxation during each loading-step ensuring convergence at an accelerated rate. This scheme is related to the one in [5] where the authors switch between over-relaxation and monolithic Newton. However, for the new acceleration scheme, proposed in this paper, both of the combined acceleration methods function as post-processes to the increments of the standard staggered scheme. In other words, the new acceleration method can be implemented with minor modifications to already available software. Moreover, switching between the two acceleration techniques does not change the sparsity of the underlying linear systems. The switch criterion is based on the history of the residual norms of the staggered solution steps.

To summarize, the main contributions in this paper are:

- Presentation of the difficulties encountered with the application of plain Anderson acceleration and over-relaxation applied to the staggered solution scheme for variational phase-field modeling of brittle fracture.

- A new acceleration algorithm that exploits the complementary strengths of Anderson acceleration and over-relaxation, utilizing residual norm evolution as a rule for switching between the methods.
- The performance of the proposed acceleration scheme is demonstrated through thorough numerical examples including classical benchmark problems.

The paper is structured as follows: The mathematical model and numerical discretization are presented in Section 2. Here, we introduce the energy functional which is subject to minimization together with the discretization. In Section 3, the staggered scheme and the acceleration techniques are presented. Both Anderson acceleration and relaxation are described before the combined acceleration scheme is presented together with the inexact Newton modification. Section 4 contains the numerical study of the accelerations applied to the staggered scheme. We test the staggered scheme both with and without the combinations of Anderson acceleration and relaxation. Moreover, the optimal depth of Anderson acceleration and the choice of relaxation parameter is discussed. Finally, some concluding remarks are made in Section 5.

2. Mathematical problem

In this section, the mathematical problem that is considered throughout the paper is presented. An elastic medium, represented by the domain $\Omega \subset \mathbb{R}^d$ with $d = 2$ (or 3), is subject to loading through traction forces, $\boldsymbol{t}$, along Γ_N and displacement, $\boldsymbol{u}_D$, along Γ_D to the extent that it might break. Here, $\Gamma_N \cup \Gamma_D = \partial\Omega$ are subsets of the boundary of the domain, Ω, and Γ_D has nonzero measure. The state of the material is modeled by Griffith's criterion [21], with constant G_c, and a smooth indicator function (the phase-field variable) $\varphi : \Omega \to [0, 1]$ describes the state of the damage to the material. The phase-field is defined to take the value 0 whenever the material is unbroken, and 1 when the material is broken, and a model parameter ℓ determines the width of the regularized zone where the phase-field transitions from 0 to 1.

2.1. The energy of the system

Following the work of [1,22], we can express the total energy of the system as a sum of the medium's elastic energy, the surface energy dissipation associated with the broken parts of the material and external work related to traction. Now, let $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ denote the material displacement and define the total energy functional as

$$\mathcal{E}(\boldsymbol{u}, \varphi) \coloneqq \int_\Omega \mathcal{E}_c(\varphi) + \mathcal{E}_m(\boldsymbol{u}, \varphi)\, d\mathbf{x} - \int_{\Gamma_N} \boldsymbol{t} \cdot \boldsymbol{u}\, ds \tag{1}$$

where

$$\mathcal{E}_c(\varphi) \coloneqq \frac{G_c}{2}\left(\frac{\varphi^2}{\ell} + \ell \nabla\varphi \cdot \nabla\varphi\right), \tag{2}$$

and

$$\mathcal{E}_m(\boldsymbol{u}, \varphi) \coloneqq g(\varphi)\Psi^+(\boldsymbol{\varepsilon}) + \Psi^-(\boldsymbol{\varepsilon}) - \boldsymbol{b} \cdot \boldsymbol{u}. \tag{3}$$

Here, we have applied the degradation function

$$g(\varphi) \coloneqq (1 - \kappa)(1 - \varphi)^2 + \kappa,$$

where κ is a "small" constant. Other choices have been proposed in [23]. Moreover, the material is assumed to be homogeneous and isotropic, and the elastic strain energy functional

$$\Psi(\boldsymbol{\varepsilon}) \coloneqq \frac{1}{2}\boldsymbol{\varepsilon} : \mathbb{C} : \boldsymbol{\varepsilon} = \mu(\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon}) + \frac{\lambda \operatorname{tr}(\boldsymbol{\varepsilon})^2}{2}, \tag{4}$$

where $\boldsymbol{\varepsilon} = \frac{\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\top}{2}$ is the linearized elastic strain tensor and μ and λ are the Lamé parameters, has been decomposed into "tensile", Ψ^+, and "compressive", Ψ^-, parts. The additive spectral decomposition

$$\Psi^\pm(\boldsymbol{\varepsilon}) \coloneqq \mu(\boldsymbol{\varepsilon}_\pm : \boldsymbol{\varepsilon}_\pm) + \frac{\lambda \langle \operatorname{tr}(\boldsymbol{\varepsilon})\rangle_\pm^2}{2},$$

proposed in [24] has been employed. Here, $\langle a \rangle_\pm \coloneqq \frac{1}{2}(a \pm |a|)$ and $\boldsymbol{\varepsilon}_\pm \coloneqq \sum_i \langle \varepsilon_i \rangle_\pm \boldsymbol{n}_i \otimes \boldsymbol{n}_i$ where $\{\varepsilon_i\}_i$ and $\{\boldsymbol{n}_i\}_i$ are the principal strains and principal strain directions, respectively. Additionally, the material is unable to heal, and the constraint $\partial_t \varphi \geq 0$ is applied accordingly.

2.2. Time discretized, continuous-in-space equations

The loading procedure is discretized by the implicit Euler scheme, giving the non-healing constraint at loading step $n \geq 1$:

$$\varphi^n(x) - \varphi^{n-1}(x) \geq 0 \quad \forall x \in \Omega. \tag{5}$$

Now, we define the displacement solution space $\boldsymbol{V}^n = \left\{ \boldsymbol{v} \in \left(H^1(\Omega)\right)^d \mid \boldsymbol{v}|_{\Gamma_D} = \boldsymbol{u}_D^n \right\}$, the displacement test space $\boldsymbol{V}^0 = \left\{ \boldsymbol{v} \in \left(H^1(\Omega)\right)^d \mid \boldsymbol{v}|_{\Gamma_D} = 0 \right\}$, and the phase-field solution and test space $Q = H^1(\Omega)$. Then, the solution $(\boldsymbol{u}^n, \varphi^n) \in \boldsymbol{V}^n \times Q$ at loading step $n \geq 1$ is given by

$$(\boldsymbol{u}^n, \varphi^n) := \underset{\boldsymbol{u},\varphi}{\arg\min} \left\{ \mathcal{E}(\boldsymbol{u}, \varphi, \boldsymbol{t}^n) \mid \boldsymbol{u} \in \boldsymbol{V}^n, \varphi \in Q \right\}. \tag{6}$$

Letting $\langle \cdot, \cdot \rangle_X$ denote the usual L^2 inner product over the domain X and denoting

$$\boldsymbol{\sigma}^{\pm}(\boldsymbol{u}) := \frac{\partial \Psi^{\pm}(\boldsymbol{\varepsilon}(\boldsymbol{u}))}{\partial \boldsymbol{\varepsilon}(\boldsymbol{u})},$$

we find the variation of the energy (1) with respect to $\boldsymbol{u}$ and φ respectively:

$$\mathcal{E}_{\delta \boldsymbol{u}}(\boldsymbol{u}, \varphi, \boldsymbol{v}) = \left\langle \left(g(\varphi)\boldsymbol{\sigma}^+(\boldsymbol{u}) + \boldsymbol{\sigma}^-(\boldsymbol{u}) \right), \boldsymbol{\varepsilon}(\boldsymbol{v}) \right\rangle_\Omega - \langle \boldsymbol{b}, \boldsymbol{v} \rangle_\Omega - \langle \boldsymbol{t}, \boldsymbol{v} \rangle_{\Gamma_N} \tag{7}$$

$$\mathcal{E}_{\delta \varphi}(\boldsymbol{u}, \varphi, q) = \left\langle g'(\varphi)\Psi^+(\boldsymbol{\varepsilon}), q \right\rangle_\Omega + \frac{G_c}{\ell} \left(\langle \varphi, q \rangle_\Omega + \ell^2 \langle \nabla \varphi, \nabla q \rangle_\Omega \right). \tag{8}$$

It is now easy to see that the solution to (6), $(\boldsymbol{u}^n, \varphi^n)$, satisfies the system of equations

$$\mathcal{E}_{\delta \boldsymbol{u}}(\boldsymbol{u}^n, \varphi^n, \boldsymbol{v}) = 0 \tag{9}$$

$$\mathcal{E}_{\delta \varphi}(\boldsymbol{u}^n, \varphi^n, q) = 0 \tag{10}$$

for all $\boldsymbol{v} \in \boldsymbol{V}^0$ and $q \in Q$.

The inequality (5) still requires some special treatment, and in this paper, we follow the approach of [24], where non-healing for the phase-field is enforced by never allowing $\Psi^+(\boldsymbol{\varepsilon})$ to decrease in Eq. (10). To achieve this, we introduce a history variable;

$$\mathcal{H}^n := \max_{k \leq n} \Psi^+ \left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}^k\right) \right), \tag{11}$$

and define a modified version of the variation with respect to φ by

$$\widetilde{\mathcal{E}}_{\delta \varphi}(\boldsymbol{u}^n, \varphi^n, q) = \left\langle g'(\varphi^n)\mathcal{H}^n, q \right\rangle_\Omega + \frac{G_c}{\ell} \left(\left\langle \varphi^n, q \right\rangle_\Omega + \ell^2 \left\langle \nabla \varphi^n, \nabla q \right\rangle_\Omega \right). \tag{12}$$

The solution at loading step n will be defined as the pair $(\boldsymbol{u}^n, \varphi^n) \in \boldsymbol{V}^n \times Q$ that satisfies

$$\mathcal{E}_{\delta \boldsymbol{u}}(\boldsymbol{u}^n, \varphi^n, \boldsymbol{v}) = 0 \tag{13}$$

$$\widetilde{\mathcal{E}}_{\delta \varphi}(\boldsymbol{u}^n, \varphi^n, q) = 0 \tag{14}$$

for all $\boldsymbol{v} \in \boldsymbol{V}^0$ and $q \in Q$. Other methods such as penalization [25] and the augmented Lagrangian method [26] have also been applied in this context. It is worth noting that the use of a global constraint such as (5) to impose irreversibility of diffuse fractures has been called into question previously, on the grounds that it may not lead to the correct profile of the phase-field for a fully evolved crack [27]. Within the context of a history variable approach, this issue can be addressed by updating $\mathcal{H}$ only when the phase-field exceeds a certain threshold as done in [23,28] to enforce a localized constraint. In the current work, we have chosen to implement (11) as written in order to facilitate a clearer comparison with other solution schemes that utilize a history variable approach according to the definition given in [24].

2.3. Spatial discretization

To solve (13)–(14) we apply conforming linear finite elements [2,4,5,23], both for the phase-field variable and the displacement. Let $\mathcal{T}_\Omega^h = \{T_k\}_k$ be a decomposition of the domain Ω into simplices, T_k, and define, at loading

step $n \geq 1$, the spaces

$$\boldsymbol{V}_h^n = \left\{ \boldsymbol{v}_h \in (H^1(\Omega))^d \;\middle|\; \boldsymbol{v}_h|_{T_k} \in (\mathcal{P}^1(T_k))^d \;\forall\, T_k \in \mathcal{T}_\Omega,\; \boldsymbol{v}_h|_{\Gamma_D} = \boldsymbol{u}_D^n \right\},$$
$$Q_h = \left\{ q_h \in H^1(\Omega) \;\middle|\; q_h|_{T_k} \in \mathcal{P}^1(T_k) \;\forall\, T_k \in \mathcal{T}_\Omega \right\},$$

and $\boldsymbol{V}_h^0$ accordingly with zero trace. The system of equations to be solved is then: Find $(\boldsymbol{u}_h^n, \varphi_h^n) \in \boldsymbol{V}_h^n \times Q_h$ such that

$$\mathcal{E}_{\delta \boldsymbol{u}}(\boldsymbol{u}_h^n, \varphi_h^n, \boldsymbol{v}_h) = 0 \tag{15}$$
$$\widetilde{\mathcal{E}}_{\delta \varphi}(\boldsymbol{u}_h^n, \varphi_h^n, q_h) = 0 \tag{16}$$

for all $\boldsymbol{v}_h \in \boldsymbol{V}_h^0$ and $q_h \in Q_h$. We emphasize that this problem is challenging not just due to the coupling between the equation for mechanics and the evolution of the phase-field, but in particular due to the non-linearities associated with this coupling through the terms $g(\varphi^n)\mathcal{H}^n$, $g(\varphi^n)\boldsymbol{\sigma}^+(\boldsymbol{u}^n)$, and $\boldsymbol{\sigma}^-(\boldsymbol{u}^n)$.

Following standard procedures, the system (15)–(16) naturally translates to the algebraic residual equations

$$\mathrm{Res}_{\boldsymbol{u}}\left(\boldsymbol{u}_h^n, \varphi_h^n\right) = 0 \tag{17}$$
$$\mathrm{Res}_{\varphi}\left(\boldsymbol{u}_h^n, \varphi_h^n\right) = 0, \tag{18}$$

where $\mathrm{Res}_{\boldsymbol{u}}$ and Res_{φ} denote the algebraic residuals corresponding to (15) and (16), respectively.

3. Staggered scheme and acceleration

The discrete governing Eqs. (15)–(16) are strongly nonlinear and coupled. In this paper, we apply the staggered scheme [5,29,23] to solve them, decoupling the equations. We let $i \geq 1$ be the iteration index and define the staggered scheme as: Given $\varphi_h^{n,i-1} \in Q_h$, find $(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i}) \in \boldsymbol{V}_h^n \times Q_h$ such that

$$\mathcal{E}_{\delta \boldsymbol{u}}(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i-1}, \boldsymbol{v}_h) = 0 \tag{19}$$
$$\widetilde{\mathcal{E}}_{\delta \varphi}(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i}, q_h) = 0 \tag{20}$$

for all $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h^0 \times Q_h$ and $\varphi_h^{n,0} := \varphi_h^{n-1}$. The iterations are terminated when either the absolute (21)–(22) or the relative (23)–(24) stopping criteria are reached:

$$\left\| \mathrm{Res}_{\boldsymbol{u}}\left(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i}\right) \right\|_2 \leq \mathrm{Tol}_{\mathrm{Res,Abs}}, \tag{21}$$
$$\left\| \boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^{n,i-1} \right\|_{L^2(\Omega)} + \left\| \varphi_h^{n,i} - \varphi_h^{n,i-1} \right\|_{L^2(\Omega)} \leq \mathrm{Tol}_{\mathrm{Inc,Abs}}, \tag{22}$$
$$\frac{\left\| \mathrm{Res}_{\boldsymbol{u}}\left(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i}\right) \right\|_2}{\left\| \mathrm{Res}_{\boldsymbol{u}}\left(\boldsymbol{u}_h^{n,1}, \varphi_h^{n,0}\right) \right\|_2} \leq \mathrm{Tol}_{\mathrm{Res,Rel}}, \tag{23}$$
$$\frac{\left\| \boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^{n,i-1} \right\|_{L^2(\Omega)}}{\left\| \boldsymbol{u}_h^{n,1} \right\|_{L^2(\Omega)}} + \frac{\left\| \varphi_h^{n,i} - \varphi_h^{n,i-1} \right\|_{L^2(\Omega)}}{\left\| \varphi_h^{n,0} \right\|_{L^2(\Omega)}} \leq \mathrm{Tol}_{\mathrm{Inc,Rel}}, \tag{24}$$

for given tolerances $\mathrm{Tol}_{\mathrm{Res,Abs}}$, $\mathrm{Tol}_{\mathrm{Res,Rel}}$, $\mathrm{Tol}_{\mathrm{Inc,Abs}}$ and $\mathrm{Tol}_{\mathrm{Inc,Rel}}$. Notice that controlling the residuals corresponding to the phase-field equation (20) is redundant due to it being solved second in the staggered scheme by an exact linear solver.

To solve the nonlinear equation (19) we apply the Newton method with the relative stopping criterion

$$\frac{\left\| \mathrm{Res}_{\boldsymbol{u}}\left(\boldsymbol{u}_h^{n,i,j}, \varphi_h^{n,i-1}\right) \right\|_2}{\left\| \mathrm{Res}_{\boldsymbol{u}}\left(\boldsymbol{u}_h^{n,1}, \varphi_h^{n,0}\right) \right\|_2} \leq \mathrm{Tol}_{\mathrm{inner}}. \tag{25}$$

Here, $j \geq 1$ is the iteration index for the Newton method and the initial guess is chosen as the previous staggered iteration $\boldsymbol{u}_h^{n,i,0} := \boldsymbol{u}_h^{n,i-1}$.

The staggered scheme (19)–(20) is closely related to the alternate minimization method (it differs in the application of the history variable (11)) and is known to be a robust solution method [30]. However, it might require

a large number of iterations to reach satisfactory tolerances [4,8,5]. We aim to accelerate this slow convergence and propose a combination of Anderson acceleration and over-relaxation. We note that the staggered solution scheme can be written as the fixed-point iteration

$$\mathbf{x}_h^{n,i} := \mathcal{S}(\mathbf{x}_h^{n,i-1}) = \mathbf{x}_h^{n,i-1} + \Delta\mathcal{S}(\mathbf{x}_h^{n,i-1}) \tag{26}$$

where $\mathcal{S}$ is the staggered solution scheme operator, $\Delta\mathcal{S}$ is the increment of the staggered scheme and $\mathbf{x}_h^{n,i}$ is the vector $\begin{pmatrix} \boldsymbol{u}_h^{n,i} \\ \varphi_h^{n,i} \end{pmatrix}$.

Now, we present both the Anderson acceleration and the relaxed staggered scheme and describe their strengths and weaknesses. Then, taking advantage of the strengths of both schemes, a combined scheme is presented.

3.1. Anderson acceleration

Anderson acceleration is a multi-secant method that mimics the monolithic Newton method. The acceleration acts as a post-processing procedure that updates the current iterate by a linear combination of the m previous iterates, according to their respective increments. The value of m is free to be chosen and is known as the depth of the acceleration. At loading step n, the Anderson accelerated staggered scheme of depth m reads:

Algorithm 1: Anderson acceleration

Given $\mathbf{x}^0$;
for $i = 1, 2, \ldots$ *until convergence* **do**
  Set depth $m_i = \min\{m, i-1\}$;
  Define $\mathbf{F}^i := \left[\Delta\mathcal{S}\left(\mathbf{x}_h^{n,i-m_i-1}\right), \ldots, \Delta\mathcal{S}\left(\mathbf{x}_h^{n,i-1}\right)\right]$;
  Let $\boldsymbol{\alpha}^i = \left[\alpha_0^i, \ldots, \alpha_{m_i}^i\right]^\top \in \mathbb{R}^{m_i+1}$ be the minimizer of $\left\|\mathbf{F}^i\boldsymbol{\alpha}^i\right\|_2$ subject to $\sum_k \alpha_k^i = 1$;
  Define the accelerated iterate $\mathbf{x}_h^i := \sum_{k=0}^{m_i} \alpha_k \mathcal{S}\left(\mathbf{x}_h^{n,k+i-m_i-1}\right)$

Algorithm 1 is independent of the underlying fixed-point iteration, but is presented for the application to the staggered scheme here. An important feature of Anderson acceleration is that it preserves the decoupled nature of the staggered scheme, hence, the subproblem solvers are unaffected by it.

It is demonstrated in the numerical section that Anderson acceleration improves the convergence when close to the solution. However, the acceleration might deteriorate otherwise. In particular, we have observed that this happens for brutal crack propagation, where the method sometimes fails to converge at all. Therefore, we do not expect a staggered scheme that is merely enhanced by the Anderson acceleration to be a good choice. On the other hand, we provide a remedy to this in Section 3.3, where Anderson acceleration is combined with over-relaxation.

3.2. Over-relaxation

Relaxation applied to each subproblem of the staggered solution scheme was described and applied in [5]. The method first calculates the increment $\Delta\boldsymbol{u}_h^{n,i-1}$ obtained by solving Eq. (19), before defining the updated iterate as

$$\boldsymbol{u}_h^{n,i} := \boldsymbol{u}_h^{n,i-1} + \omega\Delta\boldsymbol{u}_h^{n,i-1},$$

where $\omega \in (0, 2)$ is a parameter. This new iterate $\boldsymbol{u}_h^{n,i}$ is now passed on to Eq. (20) and the same procedure is executed for the phase-field resulting in the updated iterate $\varphi_h^{n,i}$. Following standard literature on iterative methods, we refer to the choice $\omega \in (1, 2)$ as over-relaxation and $\omega \in (0, 1)$ as under-relaxation. At the nth loading step the relaxed staggered scheme reads:

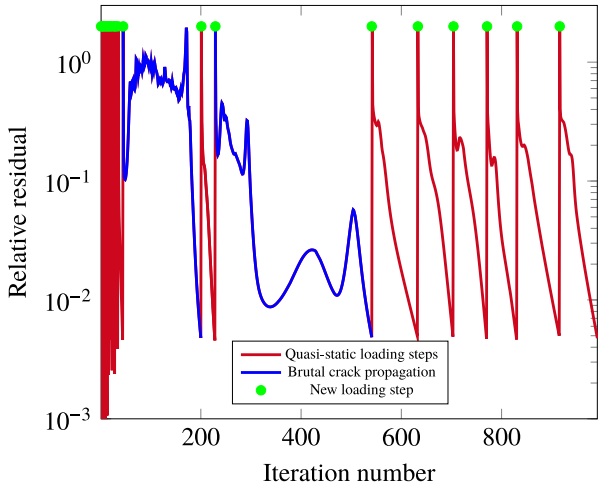


Fig. 1. Asymmetrical bending test: Relative residual evolution (see Eq. (23)) over the simulation. See Section 4.3 for an explanation of the test case. Similar behavior is experienced for all proposed test cases.

Algorithm 2: Relaxed staggered scheme

Given $\varphi_h^{n,0}$ and $\omega \in (0, 2)$;
for $i = 1, 2, \ldots$ *until convergence* **do**
  Find $\hat{\boldsymbol{u}}_h^{n,i} \in \boldsymbol{V}_h^n$ satisfying $\mathcal{E}_{\delta \boldsymbol{u}}(\hat{\boldsymbol{u}}_h^{n,i}, \varphi_h^{n,i-1}, \boldsymbol{v}_h) = 0, \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h^0$;
  Define $\Delta \boldsymbol{u}_h^{n,i-1} := \hat{\boldsymbol{u}}_h^{n,i} - \boldsymbol{u}_h^{n,i-1}$;
  Update the iterate $\boldsymbol{u}_h^{n,i} := \boldsymbol{u}_h^{n,i-1} + \omega \Delta \boldsymbol{u}_h^{n,i-1}$;
  Find $\hat{\varphi}_h^{n,i} \in Q_h$ satisfying $\widetilde{\mathcal{E}}_{\delta \varphi}(\boldsymbol{u}_h^{n,i}, \hat{\varphi}_h^{n,i}, q_h) = 0, \quad \forall q_h \in Q_h$;
  Define $\Delta \varphi_h^{n,i-1} := \hat{\varphi}_h^{n,i} - \varphi_h^{n,i-1}$;
  Update the iterate $\varphi_h^{n,i} := \varphi_h^{n,i-1} + \omega \Delta \varphi_h^{n,i-1}$;

Under-relaxation is robust when applied to the staggered scheme, however, it usually slows down the scheme. Over-relaxation, on the other hand, tends to accelerate the loading steps of the staggered solution scheme where cracks occur, while it might slow down the process for quasi-static loading steps.

3.3. Robust and efficient solution by combining Anderson acceleration and over-relaxation

As neither Anderson acceleration nor over-relaxation should be applied naively to the staggered scheme (19)–(20), due to their mentioned weaknesses, we propose a combined robust acceleration scheme. The key observations that motivate such a method are:

- Anderson acceleration is locally accelerating, while over-relaxation might struggle close to the solution.
- Anderson acceleration is applied as a post-processing algorithm to the increments of the staggered scheme. Hence, switching between relaxation and Anderson acceleration merely requires minor modifications to the implementation.
- During crack propagation the residuals for the staggered scheme show a stagnating, oscillatory behavior, and during quasi-static steps they are strictly decreasing, see [4] and Fig. 1. Therefore, it is possible to use residual evolution as a rule for switching between the acceleration techniques.

A new parameter $N_{\omega\to AA} \in \mathbb{N}$, related to the switch from relaxation to Anderson acceleration is defined, and at loading step n the combined accelerated staggered scheme reads:

1. Apply Anderson acceleration of given depth m.
2. While the norms of the residuals are strictly decreasing, continue with Anderson acceleration until convergence.
3. If the norms of the residuals are not strictly decreasing, switch to relaxation with given parameter ω.
4. When the norms of the $N_{\omega\to AA}$ previous residuals are strictly decreasing go back to 1, and restart[1] Anderson acceleration.

Below, we give a pseudo-code for the new combined acceleration method. Define the residual norm $\mathrm{Res}_i := \left\| \mathrm{Res}_{\boldsymbol{u}} \left(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i} \right) \right\|_2$ as in (21), and notice that the application of Anderson acceleration and relaxation in the pseudo-code denotes the ith step of the accelerations (see Algorithms 1 and 2).

Algorithm 3: Combined algorithm

Given depth m, relaxation ω, initial guess $\varphi_h^{n,0}$, and switch $N_{\omega\to AA}$;
relaxing := False;
for $i = 1, 2, \ldots$ *until convergence* **do**
  if *not(relaxing)* **then**
    if $i = 1$ *or* $\mathrm{Res}_i \leq \mathrm{Res}_{i-1}$ **then**
      apply Anderson accelerated staggered scheme, giving $(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i})$;
    else
      apply relaxed staggered scheme, giving $(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i})$;
      relaxing := True;
  else
    if *not(*$\mathrm{Res}_i \leq \mathrm{Res}_{i-1} \leq \cdots \leq \mathrm{Res}_{i-N_{\omega\to AA}-1}$*)* **then**
      apply relaxed staggered scheme, giving $(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i})$;
    else
      restart[1] *and apply Anderson accelerated staggered scheme, giving* $(\boldsymbol{u}_h^{n,i}, \varphi_h^{n,i})$;
      relaxing := False;

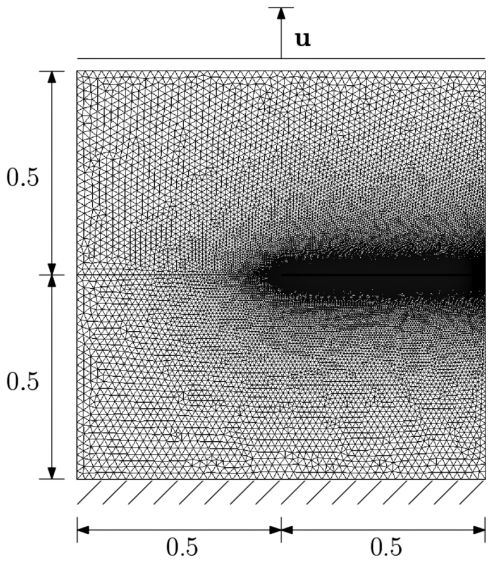
4. Numerical examples

This section explores the effects of the proposed acceleration methods from Section 3 applied to the staggered scheme (19)–(20). Both Anderson acceleration (see Fig. 16(b)) and over-relaxation (see Fig. 11(b)) alone are shown to be infeasible acceleration methods when plainly applied to the staggered scheme, while the combined scheme is superior to the unaccelerated scheme for all tests. We consider four different test cases which are widely used for numerical studies in the literature:

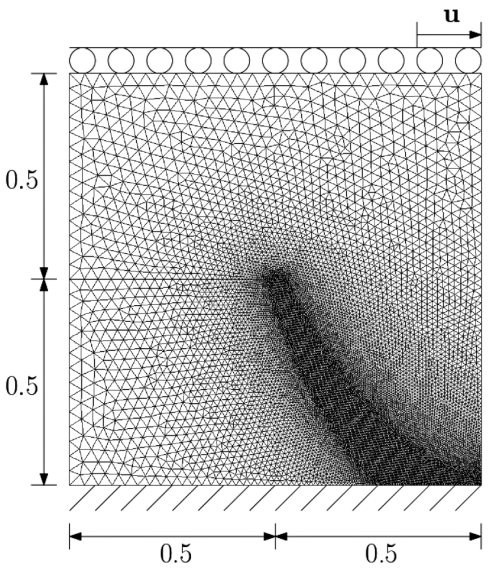
- A domain with a single notch subject to
  - tensile load;
  - shear load.
- An L-shaped domain subject to monotonic loading.
- Bending of an asymmetrically notched beam with holes.

All the numerical examples have been implemented using modules from the DUNE project [31], specifically dune-functions [32,33].

[1] Restart means that Anderson acceleration should be applied as it is in the first iteration, i.e., using no information of previous increments and iterates.



(a) **Single notch tensile test:** The bottom boundary of the domain is fixed ($\boldsymbol{u} = \boldsymbol{0}$), and the top boundary is uniformly displaced over the loading steps n in the vertical direction ($u_y = \bar{u}n$) while fixed in the horizontal direction ($u_x = 0$). The mesh is refined according to the expected crack path and contains a total of 36995 nodes.



(b) **Single notch shear test:** The bottom boundary is fixed ($\boldsymbol{u} = \boldsymbol{0}$), and the top boundary is uniformly displaced over the loading steps n in the horizontal direction ($u_x = \bar{u}n$). The left, right and top boundaries, and the lower lip of the prescribed crack are fixed in the vertical direction ($u_y = 0$). The mesh is refined according to the expected crack path and contains a total of 12660 nodes.

Fig. 2. Domain, boundary conditions, and mesh for the single-edge notch test cases.

The mesh for all numerical examples has been locally refined close to where the crack is expected to propagate. Choosing uniform fine meshes is naturally another option, and there are several algorithms for adaptive mesh refinement for these types of problems in the literature, see e.g., [34,35].

4.1. Single notch test

Two of the most commonly found test cases in the literature are both based on the same single notch geometry [36,37]. They consist of a square domain with a pre-existing crack that penetrates half the domain, see Fig. 2. The domain is held still at the bottom, and a displacement driven load is applied at the top boundary.

4.1.1. Single notch tensile test case

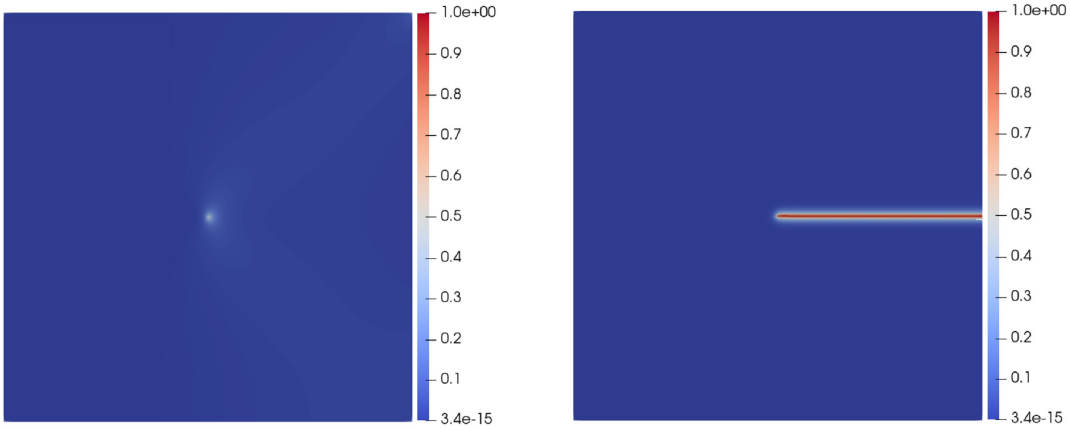
A tensile load is applied on the top boundary, and at loading step n we have

$$\boldsymbol{u}^n_{|_{\Gamma^{\text{Top}}}} = \begin{pmatrix} 0 \\ \bar{u}n \end{pmatrix},$$

where the load size $\bar{u}$ is given in Table 1 and Γ^{Top} is the top part of the boundary in Fig. 2(a). Due to the load being strictly tensile, there is no need to split the elastic strain energy functional (4) into tensile and compressive parts, which would effectively add nonlinearities to the system. Therefore, the first term in (7) is replaced by $\langle g(\varphi)\boldsymbol{\sigma}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v})\rangle$, for

$$\boldsymbol{\sigma}(\boldsymbol{u}) := \frac{\partial \psi(\boldsymbol{\varepsilon}(\boldsymbol{u}))}{\partial \boldsymbol{\varepsilon}(\boldsymbol{u})}. \tag{27}$$

Material parameter values are chosen as in e.g., [36], and can be found in Table 1. We employ a triangular mesh, which has been locally refined in the region where the crack is expected to propagate, see Fig. 2(a).



(a) Solution before crack growth at loading step 27.

(b) Solution after crack propagation at loading step 28.

Fig. 3. Solution for φ for the single notch tensile test case.

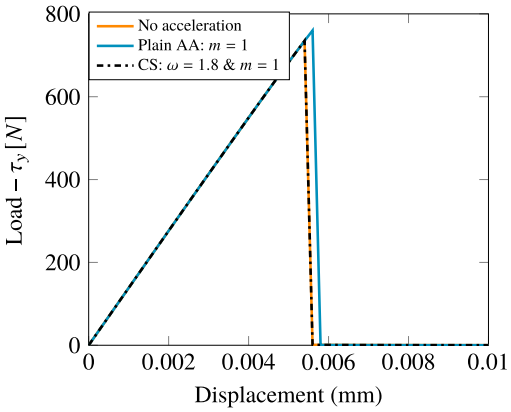
In this test case, the crack fully propagates in one single critical loading step, see Fig. 3, in which the crack gradually expands through the domain with increasing staggered iteration count (not displayed here). Fig. 4(b) shows that, as expected, the staggered scheme under Anderson acceleration alone struggles as a consequence of its local convergence. Aside of mitigating the issue and only applying Anderson acceleration in suitable situations, the combined scheme, in addition, takes advantage of over-relaxation and its ability to move further each iteration and accelerates this particular loading step significantly. For the remaining loading steps, the combined scheme accelerates by Anderson acceleration, as its local convergence is sufficient. The total number of iterations (Fig. 5) for the combined scheme is, therefore, smaller than those of the unaccelerated staggered scheme and the Anderson accelerated staggered scheme. The figure shows that the combined scheme accelerates by more than 50% for large relaxation parameters. We also observe that the depth of Anderson acceleration is not influential as long as it is larger than one. Moreover, there is a trend that more aggressive over-relaxation (higher ω) results in faster computations. Additionally, the combined acceleration scheme is robust with respect to the tuning parameters, m and ω, and exhibits convergence for all tested combinations.

The traction vector is defined by

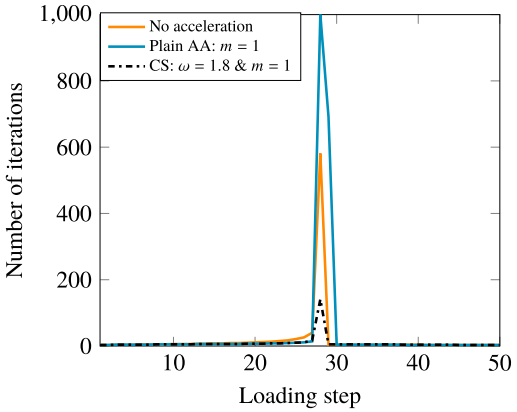
$$\boldsymbol{\tau} = (\tau_x, \tau_y) = \int_{\Gamma^{\text{Top}}} \boldsymbol{\sigma} \cdot \boldsymbol{\nu} \, dS, \tag{28}$$

where $\boldsymbol{\nu}$ is the outward pointing normal vector, and $\boldsymbol{\sigma}$ is defined in (27). For this problem, the load in the direction of interest is τ_y, and we observe in Fig. 4(a) that the load–displacement curves remain unchanged after the combined acceleration. This is an important observation that demonstrates that the acceleration method only affects the convergence properties of the solver, not the quality of the solution. The Anderson accelerated staggered scheme, however, does not converge in the maximal prescribed iterations for each loading step and we observe that its load–displacement curve is affected.

Remark 1. Notice that the plots of the number of iterations for depth $m = 0$ in Figs. 5, 8, 12 and 17 are not corresponding to plain relaxation. Here, relaxation is switched on and off depending on residual evolution, turning it into a safeguarded relaxation. The same goes for the plots of plain Anderson acceleration with over-relaxation parameter $\omega = 1$. These correspond to safeguarded Anderson accelerations, similar to those that are proposed in [19].



(a) Load-displacement curves.



(b) Number of iterations per loading step.

Fig. 4. Single notch tensile test: Load curves and number of iterations per loading step. "AA" is an abbreviation of Anderson acceleration, and "CS" is the combined acceleration scheme. "Plain AA" means that Anderson acceleration is applied without any form of safeguard or combination with relaxation.

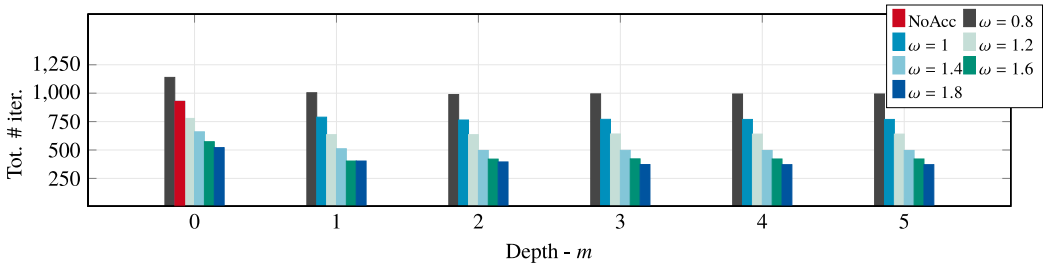
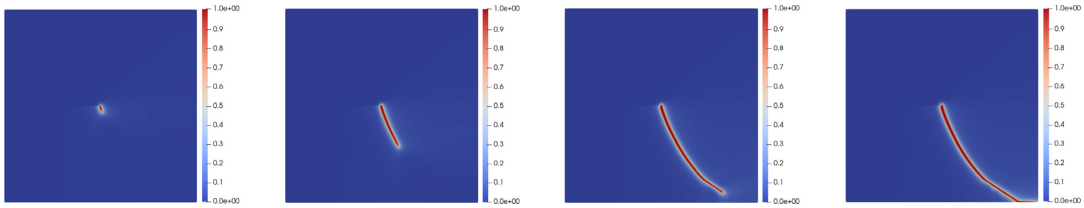


Fig. 5. Single notch tensile test: Total number of iterations for the combined scheme with different relaxation parameters and Anderson acceleration depths. "NoAcc" is the unaccelerated staggered scheme.

Table 1
Parameter values for the single notch test cases.

Parameter	Symbol	Value – Tensile	Value – Shear
Lamé's 1. parameter	λ	121.15 kN/mm^2	121.15 kN/mm^2
Lamé's 2. parameter	μ	80.77 kN/mm^2	80.77 kN/mm^2
Regularization width	ℓ	0.0075 mm	0.0075 mm
Griffith's constant	$\mathcal{G}_c$	2.7 N/mm	2.7 N/mm
Energy regularization	κ	10^{-10}	10^{-10}
Tot. # loading steps	N	50	150
Load size	$\bar{u}$	$2 \cdot 10^{-4}$ mm	10^{-4} mm
Fine mesh size	h	0.001 mm	0.00375 mm
Min. relax. steps	$N_{\omega \to \text{AA}}$	5	5
Abs. tol.	$\text{Tol}_{\text{Res/Inc,Abs}}$	10^{-8}	10^{-8}
Rel. residual tol.	$\text{Tol}_{\text{Res,Rel}}$	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
Rel. increment tol.	$\text{Tol}_{\text{Inc,Rel}}$	10^{-2}	10^{-2}
Max. iter. pr. load. step	Max_{iter}	1000	1000
Inner Newton tol.	$\text{Tol}_{\text{Inner}}$	10^{-4}	10^{-4}



(a) Solution at loading step 100. (b) Solution at loading step 117. (c) Solution at loading step 133. (d) Solution at loading step 150.

Fig. 6. Solution for φ for the single notch shear test case.

4.1.2. Single notch shear case

In this test case, a shear load is applied on the top boundary of a unit square domain with a prescribed crack that halfway penetrates the domain. The displacement boundary condition

$$\boldsymbol{u}^n_{|\Gamma^{\text{Top}}} = \begin{pmatrix} \bar{u}n \\ 0 \end{pmatrix}$$

is applied at loading step n. The load size $\bar{u}$ is presented in Table 1, and the top part of the boundary Γ^{Top} is displayed together with more details on the domain and boundary conditions in Fig. 2(b). The material properties are taken from [36] and displayed in Table 1. A triangular mesh, which is refined according to where the crack is expected to propagate, has been employed, see Fig. 2(b).

Contrary to the tensile test case, the crack propagation happens gradually over the course of many loading steps, see Fig. 6. Therefore, solutions at subsequent loading steps do not differ as significantly as for the brutal crack growth in the tensile test case. We expect that the Anderson acceleration is a more suitable choice for accelerating the staggered scheme. Indeed, Fig. 7(b) shows that even with the plain Anderson acceleration the staggered scheme is quite significantly accelerated. Moreover, the combined scheme is even better, and we see that it reaches convergence in every single loading step.

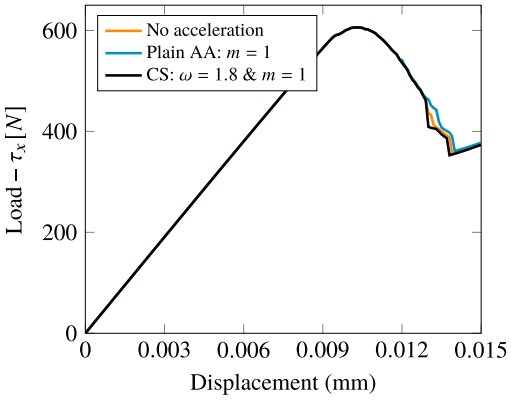
In the load–displacement curves, Fig. 7(a), the load τ_x as defined in (28) is displayed for each loading step. The plot shows minor differences towards the end of the displacement. This is due to the scheme not converging in its given maximal iterations per loading step (see Table 1) for both the unaccelerated staggered scheme and the plain Anderson accelerated scheme in all the loading steps. This is similar to the tensile case where the load–displacement curves, Fig. 4(a), also are affected.

In Fig. 8, we see the total iteration count for several acceleration depths in combination with over-relaxation. It can be observed that the staggered scheme is accelerated significantly for all combinations of Anderson acceleration and over-relaxation as long as the depth is greater than one. In fact, we have more than 80% reduction in the total number of iterations when choosing a high relaxation parameter. Moreover, for this test case the plain Anderson acceleration is in itself a suitable alternative to the unaccelerated staggered scheme. Notice the difference between plain Anderson acceleration and Anderson acceleration combined with over-relaxation of depth one described in Remark 1. Additionally, the total time in seconds to complete the simulations is plotted in Fig. 9, in a similar fashion as for the total number of iterations. We observe that the trend of what we save in computational time is equal to that of the total iteration count. This does, indeed, hold true for all of the test cases.

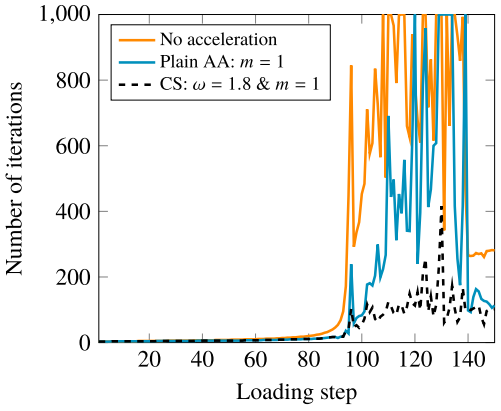
4.2. L-shaped domain subject to loading

An L-shaped domain with a displacement boundary condition applied on the right part of the boundary is considered, see Fig. 10(a) for details. The displacement is uniformly increased on the boundary segment over 800 loading steps. As a result, a crack occurs in the inner corner, propagating into the domain, see Fig. 10(b). A uniform quadrilateral mesh with a mesh diameter of $\frac{125}{32}$ mm is employed. See Table 2 for material and computational parameters.

Here, the crack propagation has a character somewhere between the single notch tensile test and the single notch shear test. Crack initiation shows similar behavior as brutal crack propagation, but not as extreme as for the single



(a) Load-displacement curves.



(b) Number of iterations per loading step. Maximum number of iterations per loading step is 1000.

Fig. 7. Single notch shear test: Load curves and number of iterations per loading step. "AA" is an abbreviation of Anderson acceleration, and "CS" is the combined acceleration scheme. "Plain AA" means that Anderson acceleration is applied without any form of safeguard or combination with relaxation.

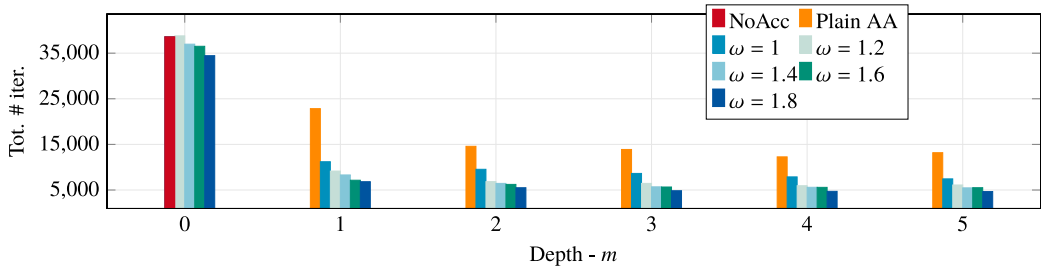


Fig. 8. Single notch shear test: Total number of iterations for different relaxation parameters and Anderson acceleration depths. "NoAcc" is the unaccelerated staggered scheme, and "Plain AA" is Anderson acceleration without the combination with relaxation.

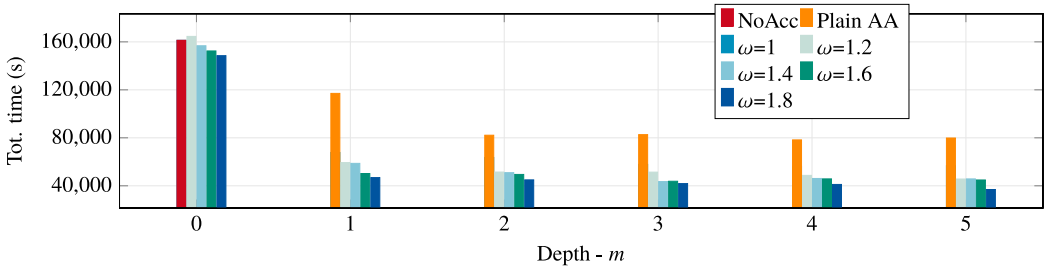
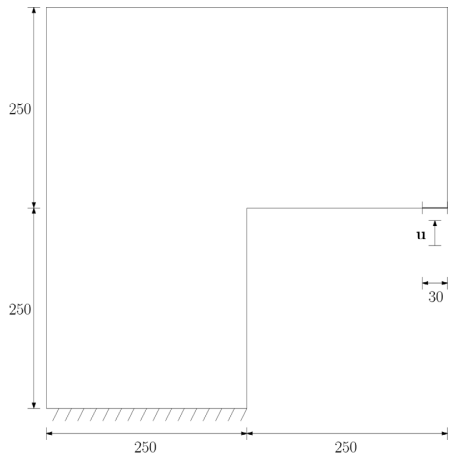
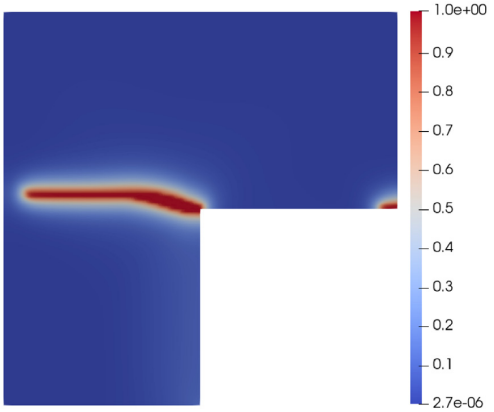


Fig. 9. Single notch shear test: Total time in seconds for different relaxation parameters and Anderson acceleration depths. "NoAcc" is the unaccelerated staggered scheme.

notch tensile test. A large peak in the number of iterations is experienced when the crack initiates, see Fig. 11(a). We observe that both the combined scheme and plain Anderson acceleration with depth $m = 1$ accelerate for all loading steps. Moreover, the only difference between the combined acceleration and Anderson acceleration is in the large peak where the combined scheme outperforms Anderson acceleration. For the rest of the simulation, the



(a) **L-shaped test:** Domain and boundary conditions. The bottom boundary of the domain is fixed ($\boldsymbol{u} = \boldsymbol{0}$), and the boundary-segment $\Gamma_D = \{(x, y) \in \Omega \mid 470 \leq x \leq 500,\ y = 250\}$ ($(0, 0)$ is in the lower left corner) is uniformly displaced over time in the vertical direction ($u_y = \bar{u}n$).



(b) **L-shaped test:** The solution for φ for the L-shape test after 800 loading steps.

Fig. 10. Domain with boundary conditions and solution for the L-shaped test.

staggered scheme converges in relatively few iterations, but the accelerated method converges faster in almost every loading step.

Fig. 11(b) displays the total number of iterations for plain over-relaxation with several relaxation parameters. A parabolic dependence on the relaxation parameter is observed, and choosing it to be too large results in more than three times the number of iterations that are required by the unaccelerated staggered scheme. This is due to over-relaxation struggling near the solution of the loading steps resulting in successive over- and undershooting of the solution. Therefore, a plain application of over-relaxation is not recommended. The total number of iterations required by the combined acceleration, however, is significantly smaller than those of the unaccelerated staggered scheme (and the optimally over-relaxed scheme), as observed in Fig. 12. Although the reduction in the number of iterations is not as good as for the single notch shear test case the combined scheme accelerates robustly with respect to the tuning parameters. It is clear that any combination of Anderson acceleration and over-relaxation is superior to the unaccelerated staggered scheme, accelerating by approximately 40%.

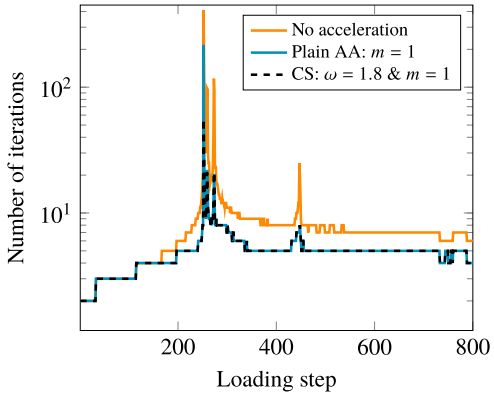
The load–displacement curve for this test case is displayed in Fig. 13(a). Here, the traction vector, see Eq. (28), is calculated on the bottom boundary and the vertical component τ_y is considered. We observe that, as all acceleration schemes converge within each loading step, the curves are completely overlapping.

4.3. Asymmetrical bending test

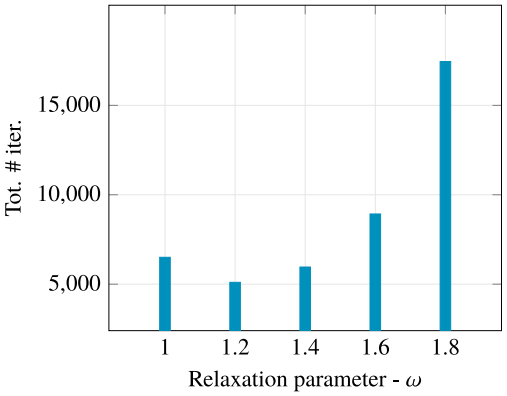
This test case considers a rectangular domain with three holes, slightly to the left, and a notch in the lower left part of the domain. It is subject to symmetrical displacement loading on the top boundary,

$$\boldsymbol{u}^n|_{\Gamma^{\text{Top}}} = \begin{pmatrix} 0 \\ \bar{u}n \end{pmatrix}. \tag{29}$$

The beam is simply supported as shown in Fig. 14(a). See Fig. 14(a) or [38] for details on boundary conditions and domain. Experimental results from [39] have shown that the crack path should hit the second hole, and we see from the numerical solution, Fig. 15, that this also happens here. The mesh has been refined in the region where the crack is expected to propagate, see Fig. 14(b). The problem parameters are chosen similarly to [10,36,37], and are presented in Table 2.



(a) Number of iterations per loading step. 'AA" is an abbreviation of Anderson acceleration, and "CS" is the combined acceleration scheme. "Plain AA" means that Anderson acceleration is applied without any form of safeguard or combination with relaxation. Notice that the plot has a log-scale on the y-axis.



(b) Total number of iterations for different over relaxations parameters applied without safeguard or combination.

Fig. 11. L-shaped test: number of iterations per loading step and total iterations for several over-relaxation parameters.

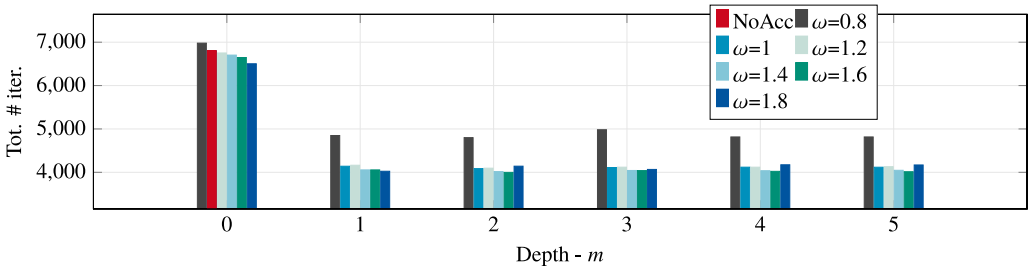
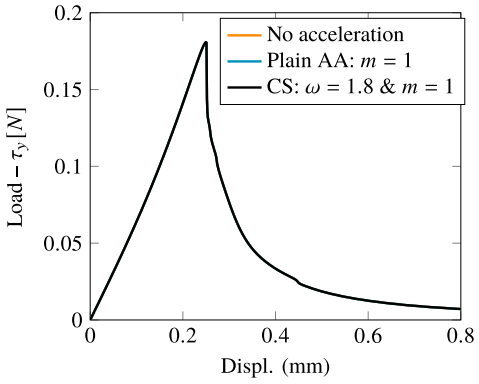
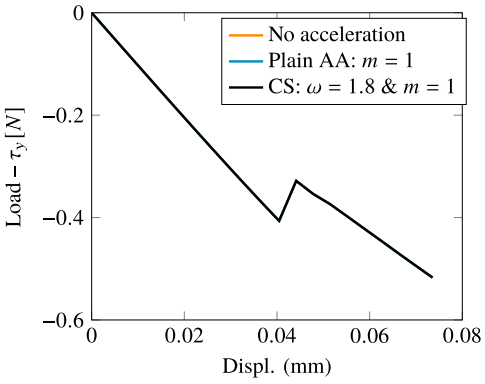


Fig. 12. L-shaped test: Total number of iterations for different relaxation parameters and Anderson acceleration depths. "NoAcc" is the unaccelerated staggered scheme.



(a) **L-shaped test** Load-displacement curve.

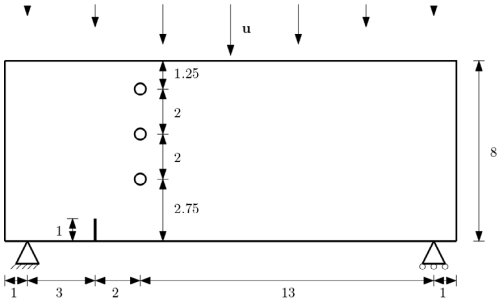


(b) **Asymmetrical bending test:** Load-displacement curve.

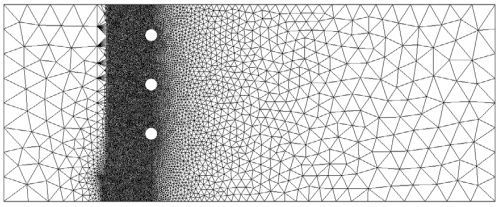
Fig. 13. "CS" is the combined acceleration scheme. "Plain AA" means that Anderson acceleration is applied without any form of safeguard or combination with relaxation.

Table 2
Parameter values for the L-shaped and asymmetrical bending tests.

Parameter	Symbol	L-shaped	Bend. test
Lamé's 1. parameter	λ	6.16 kN/mm^2	8 kN/mm^2
Lamé's 2. parameter	μ	10.95 kN/mm^2	12 kN/mm^2
Regularization width	ℓ	10 mm	0.1 mm
Griffith's constant	$\mathcal{G}_c$	$9.5 \cdot 10^{-5}$ kN/mm	10^{-3} kN/mm
Energy regularization	κ	10^{-10}	10^{-10}
Load size	$\bar{u}$	10^{-3} mm	$-10^{-2} e^{-\frac{(x-10)^2}{100}}$ mm
Fine mesh size	h	$\frac{125}{32}$ mm	0.05 mm
Min. relax. steps	$N_{\omega \to \mathrm{AA}}$	5	5
Abs. tol.	$\mathrm{Tol}_{\mathrm{Res/Inc,Abs}}$	10^{-8}	10^{-8}
Rel. residual tol.	$\mathrm{Tol}_{\mathrm{Res,Rel}}$	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
Rel. increment tol.	$\mathrm{Tol}_{\mathrm{Inc,Rel}}$	10^{-2}	10^{-2}
Max. iter. pr. load. step	$\mathrm{Max}_{\mathrm{iter}}$	1000	1000
Inner Newton tol.	$\mathrm{Tol}_{\mathrm{Inner}}$	10^{-4}	10^{-4}



(a) **Asymmetrical bending test:** Domain and boundary conditions. The three holes have 0.5 mm diameter. At the right and left boundaries, $\varphi = 0$ is enforced to prevent artificial crack initiation, see [39]. The beam is fixed ($\boldsymbol{u} = \boldsymbol{0}$) at the left foot and fixed in the vertical direction ($u_y = 0$) at the right foot. Displacement condition $u_y = \bar{u}n$ is applied at the top boundary Γ^{Top}.



(b) **Asymmetrical bending test:** The mesh is refined according to the expected crack path and contains a total of 9598 nodes.

Fig. 14. Domain with boundary conditions and mesh for the asymmetrical bending test case.

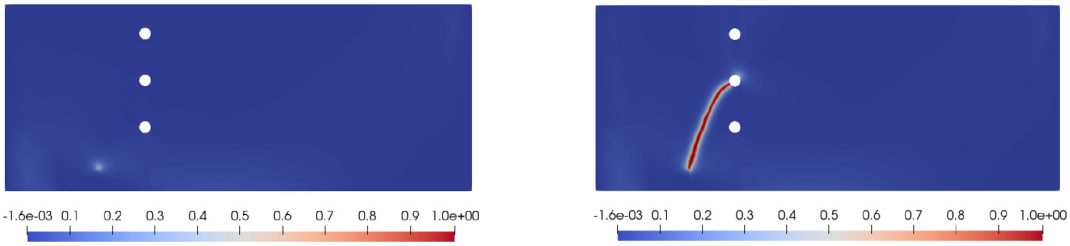
Here, we have two "critical" loading steps, in which the crack evolves and a large number of iterations is required, see Fig. 16(a). For these loading steps, we see that the plain Anderson acceleration does not accelerate, while the combined acceleration performs very well.

In Fig. 16(b) the total number of iterations for the plain Anderson acceleration is displayed for several depths. We clearly observe that the staggered scheme is significantly decelerated for depths larger than one. In other words, Anderson acceleration is not a robust method in itself for this problem. The combined scheme, on the other hand, reduces the total number of iterations for all combinations of over-relaxation and Anderson acceleration, see Fig. 17. There is, however, a tendency that larger relaxation parameters accelerate more, which is expected due to the brutal nature of the crack propagation in loading step 12, see Fig. 15.

The traction vector (28) is here calculated on the top boundary and the component of interest is τ_y. In Fig. 13(b), the load–displacement curves are displayed, and the displacement is calculated at the left corner of the top boundary. They are, as expected, overlapping as there are no loading steps for any configurations in which the convergence is not achieved in the given maximal amount of iterations.

5. Conclusion

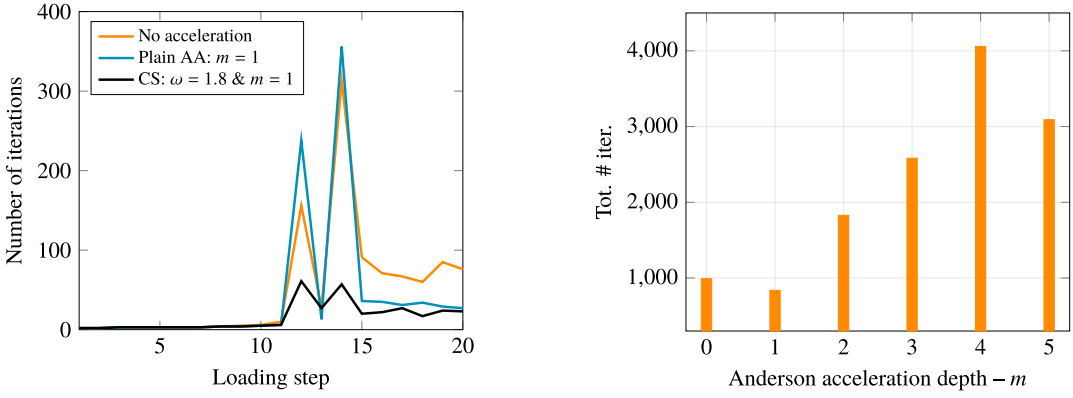
The staggered solution scheme is, due to its robustness, a popular method for solving variational phase-field models of brittle fracture. As it often requires a large number of iterations to converge we have proposed a



(a) **Asymmetrical bending test:** Solution before crack propagation at loading step 11.

(b) **Asymmetrical bending test:** Solution after crack propagation at loading step 12.

Fig. 15. Solution for φ for the asymmetrical bending test case.



(a) Number of iterations per loading step.

(b) Total number of iterations for several depths of plain Anderson acceleration.

Fig. 16. Asymmetrical bending test: Number of iterations per loading step, and total number of iterations for several depths of Anderson acceleration. "CS" is the combined acceleration scheme. "Plain AA" means that Anderson acceleration is applied without any form of safeguard or combination with relaxation.

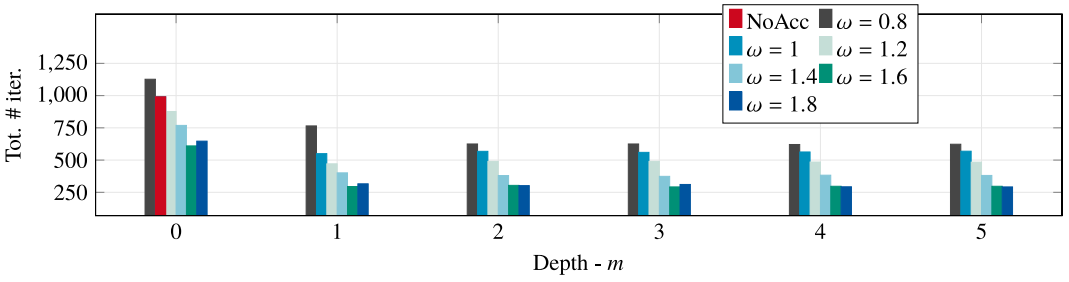


Fig. 17. Asymmetrical bending test: Total number of iterations for different relaxation parameters and Anderson acceleration depths. "NoAcc" is the unaccelerated staggered scheme.

method to accelerate it that exploits the complementary advantages of Anderson acceleration and over-relaxation. The acceleration method alternates between Anderson acceleration and over-relaxation according to a switch that depends on the norms of the previous residuals of the scheme. For problems without brutal crack growth, Anderson acceleration is quite efficient. It is, however, unstable for problems with brutal crack growth, and therefore, not a

technique that can be applied without modifications. Over-relaxation, on the other hand, works well within regimes of brutal crack propagation, but might struggle when the iterates get close to the solution within a single loading step. The scheme shows robustness with respect to the tuning parameters, Anderson acceleration depth and relaxation parameter, and converges for all combinations. Moreover, there is a tendency that choosing Anderson acceleration depth larger than one is insignificant, and that over-relaxation with parameters of at least 1.6 are the best choices. Therefore, we propose to apply the method with depth one and over-relaxation 1.6, although one might gain some speed in tuning these parameters to specific problems, or choose them adaptively. The success of the proposed combined scheme builds upon the following problem-specific phenomena:

- The observation that for fixed loading steps, fractures gradually propagate towards their final configuration by the staggered scheme. This suggests a suitable application of over-relaxation where fractures are forced to grow further each iteration.
- A characteristic residual history for the staggered scheme, motivated by experience from phase-field models for brittle fracture, suggests the residual-based strategy for the switch between Anderson acceleration and over-relaxation.

It is therefore expected that the success of the proposed scheme will be seen for other variational models for brittle fracture propagation as well.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partially supported by the Research Council of Norway Project 250223, the FracFlow project funded by Equinor, Norway through Akademiaavtalen, and by VISTA, Norway, a collaboration between the Norwegian Academy of Science and Letters and Equinor, project AdaSim #6367.

References

[1] G.A. Francfort, J.-J. Marigo, Revisiting brittle fracture as an energy minimization problem, J. Mech. Phys. Solids 46 (8) (1998) 1319–1342, http://dx.doi.org/10.1016/S0022-5096(98)00034-9.
[2] B. Bourdin, G.A. Francfort, J.-J. Marigo, Numerical experiments in revisited brittle fracture, J. Mech. Phys. Solids 48 (4) (2000) 797–826, http://dx.doi.org/10.1016/S0022-5096(99)00028-9.
[3] N. Möes, J. Dolbow, T. Belytschko, A finite element method for crack growth without remeshing, Internat. J. Numer. Methods Engrg. 46 (1) (1999) 131–150, http://dx.doi.org/10.1002/(SICI)1097-0207(19990910)46:1<131::AID-NME726>3.0.CO;2-J.
[4] T. Gerasimov, L. De Lorenzis, A line search assisted monolithic approach for phase-field computing of brittle fracture, Comput. Methods Appl. Mech. Engrg. 312 (2016) 276–303, http://dx.doi.org/10.1016/j.cma.2015.12.017, Phase Field Approaches to Fracture.
[5] P. Farrell, C. Maurini, Linear and nonlinear solvers for variational phase-field models of brittle fracture, Internat. J. Numer. Methods Engrg. 109 (5) (2017) 648–667, http://dx.doi.org/10.1002/nme.5300.
[6] T. Wick, Modified Newton methods for solving fully monolithic phase-field quasi-static brittle fracture propagation, Comput. Methods Appl. Mech. Engrg. 325 (2017) 577–611, http://dx.doi.org/10.1016/j.cma.2017.07.026.
[7] P.K. Kristensen, E. Martínez-Pañeda, Phase field fracture modelling using quasi-Newton methods and a new adaptive step scheme, Theor. Appl. Fract. Mech. 107 (2020) 102446, http://dx.doi.org/10.1016/j.tafmec.2019.102446.
[8] J. Wu, Y. Huang, V.P. Nguyen, On the BFGS monolithic algorithm for the unified phase field damage theory, Comput. Methods Appl. Mech. Engrg. 360 (2020) 112704, http://dx.doi.org/10.1016/j.cma.2019.112704.
[9] C. Gräser, D. Kienle, O. Sander, Truncated Nonsmooth Newton Multigrid for phase-field brittle-fracture problems, 2020, arXiv:2007.12290.
[10] M.K. Brun, T. Wick, I. Berre, J.M. Nordbotten, F.A. Radu, An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters, Comput. Methods Appl. Mech. Engrg. 361 (2020) 112752, http://dx.doi.org/10.1016/j.cma.2019.112752.
[11] 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, http://dx.doi.org/10.1016/j.cam.2003.04.008.
[12] F. List, F.A. Radu, A study on iterative methods for solving richards' equation, Comput. Geosci. 20 (2) (2016) 341–353, http://dx.doi.org/10.1007/s10596-016-9566-3.
[13] D.G. Anderson, Iterative procedures for nonlinear integral equations, J. Assoc. Comput. Mach. 12 (4) (1965) 547–560.
[14] H. Fang, Y. Saad, Two classes of multisecant methods for nonlinear acceleration, Numer. Linear Algebra Appl. 16 (3) (2009) 197–221, http://dx.doi.org/10.1002/nla.617.

[15] J.W. Both, K. Kumar, J.M. Nordbotten, F.A. Radu, Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media, Comput. Math. Appl. 77 (6) (2019) 1479–1502, http://dx.doi.org/10.1016/j.camwa.2018.07.033, 7th International Conference on Advanced Computational Methods in Engineering (ACOMEN 2017).
[16] H.F. Walker, P. Ni, Anderson acceleration for fixed-point iterations, SIAM J. Numer. Anal. 49 (4) (2011) 1715–1735, http://dx.doi.org/10.1137/10078356X.
[17] C. Evans, S. Pollock, L.G. Rebholz, M. Xiao, A proof that Anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically), SIAM J. Numer. Anal. 58 (1) (2020) 788–810, http://dx.doi.org/10.1137/19M1245384.
[18] A. Toth, C.T. Kelley, Convergence analysis for Anderson acceleration, SIAM J. Numer. Anal. 53 (2) (2015) 805–819, http://dx.doi.org/10.1137/130919398.
[19] J. Zhang, B. O'Donoghue, S. Boyd, Globally convergent type-I Anderson acceleration for non-smooth fixed-point iterations, 2018, arXiv:1808.03971.
[20] P. Suryanarayana, P.P. Pratapa, J.E. Pask, Alternating Anderson–Richardson method: An efficient alternative to preconditioned Krylov methods for large, sparse linear systems, Comput. Phys. Comm. 234 (2019) 278–285, http://dx.doi.org/10.1016/j.cpc.2018.07.007.
[21] A.A. Griffith, G.I. Taylor, VI. The phenomena of rupture and flow in solids, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 221 (582–593) (1921) 163–198, http://dx.doi.org/10.1098/rsta.1921.0006.
[22] B. Bourdin, G.A. Francfort, J.-J. Marigo, The variational approach to fracture, J. Elasticity 91 (2008) 5–148, http://dx.doi.org/10.1007/s10659-007-9107-3.
[23] J.M. Sargado, E. Keilegavlen, I. Berre, J.M. Nordbotten, High-accuracy phase-field models for brittle fracture based on a new family of degradation functions, J. Mech. Phys. Solids 111 (2018) 458–489, http://dx.doi.org/10.1016/j.jmps.2017.10.015.
[24] C. Miehe, M. Hofacker, F. Welschinger, A phase field model for rate-independent crack propagation: Robust algorithmic implementation based on operator splits, Comput. Methods Appl. Mech. Engrg. 199 (45) (2010) 2765–2778, http://dx.doi.org/10.1016/j.cma.2010.04.011.
[25] T. Gerasimov, L. De Lorenzis, On penalization in variational phase-field models of brittle fracture, Comput. Methods Appl. Mech. Engrg. 354 (2019) 990–1026, http://dx.doi.org/10.1016/j.cma.2019.05.038.
[26] M.F. Wheeler, T. Wick, W. Wollner, An augmented-Lagrangian method for the phase-field approach for pressurized fractures, Comput. Methods Appl. Mech. Engrg. 271 (2014) 69–85, http://dx.doi.org/10.1016/j.cma.2013.12.005.
[27] C. Miehe, F. Aldakheel, S. Teichtmeister, Phase-field modeling of ductile fracture at finite strains: A robust variational-based numerical implementation of a gradient-extended theory by micromorphic regularization, Internat. J. Numer. Methods Engrg. 111 (2017) 816–863, http://dx.doi.org/10.1002/nme.5484.
[28] J.M. Sargado, E. Keilegavlen, I. Berre, J.M. Nordbotten, A combined finite element–finite volume framework for phase-field fracture, Comput. Methods Appl. Mech. Engrg. 373 (2021) 113474, http://dx.doi.org/10.1016/j.cma.2020.113474.
[29] T. Gerasimov, N. Noii, O. Allix, L. De Lorenzis, A non-intrusive global/local approach applied to phase-field modeling of brittle fracture, Adv. Model. Simul. Eng. Sci. 5 (14) (2018) http://dx.doi.org/10.1186/s40323-018-0105-8.
[30] B. Bourdin, Numerical implementation of the variational formulation for quasi-static brittle fracture, Interfaces Free Bound. 9 (3) (2007) 411, http://dx.doi.org/10.4171/IFB/171.
[31] 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, http://dx.doi.org/10.11588/ans.2016.100.26526.
[32] C. Engwer, C. Gräser, S. Müthing, O. Sander, The interface for functions in the dune-functions module, 2015, arXiv preprint arXiv:1512.06136.
[33] C. Engwer, C. Gräser, S. Müthing, O. Sander, Function space bases in the dune-functions module, 2018, arXiv preprint arXiv:1806.09545.
[34] M.J. Borden, C.V. Verhoosel, M.A. Scott, T.J.R. Hughes, C.M. Landis, A phase-field description of dynamic brittle fracture, Comput. Methods Appl. Mech. Engrg. 217–220 (2012) 77–95, http://dx.doi.org/10.1016/j.cma.2012.01.008.
[35] T. Heister, M.F. Wheeler, T. Wick, A primal-dual active set method and predictor-corrector mesh adaptivity for computing fracture propagation using a phase-field approach, Comput. Methods Appl. Mech. Engrg. 290 (2015) 466–495, http://dx.doi.org/10.1016/j.cma.2015.03.009.
[36] C. Miehe, F. Welschinger, M. Hofacker, Thermodynamically consistent phase-field models of fracture: Variational principles and multi-field FE implementations, Internat. J. Numer. Methods Engrg. 83 (10) (2010) 1273–1311, http://dx.doi.org/10.1002/nme.2861.
[37] M. Ambati, T. Gerasimov, L. De Lorenzis, A review on phase-field models of brittle fracture and a new fast hybrid formulation, Comput. Mech. 55 (2015) 383–405, http://dx.doi.org/10.1007/s00466-014-1109-y.
[38] A. Mesgarnejad, B. Bourdin, M.M. Khonsari, Validation simulations for the variational approach to fracture, Comput. Methods Appl. Mech. Engrg. 290 (2015) 420–437, http://dx.doi.org/10.1016/j.cma.2014.10.052.
[39] T.N. Bittencourt, P.A. Wawrzynek, A.R. Ingraffea, J.L. Sousa, Quasi-automatic simulation of crack propagation for 2D LEFM problems, Eng. Fract. Mech. 55 (2) (1996) 321–334, http://dx.doi.org/10.1016/0013-7944(95)00247-2.

Paper D

A Cahn-Hilliard-Biot system and its generalized gradient flow structure

Storvik, E., Both, J.W., Nordbotten, J.M., and Radu, F.A.
Applied Mathematics Letters, **381**, 107799 (2021)

Applied Mathematics Letters 126 (2022) 107799

Contents lists available at ScienceDirect

Applied Mathematics Letters

www.elsevier.com/locate/aml

A Cahn–Hilliard–Biot system and its generalized gradient flow structure

Erlend Storvik *, Jakub Wiktor Both, Jan Martin Nordbotten, Florin Adrian Radu

Department of Mathematics, University of Bergen, Allégaten 41, 5007 Bergen, Norway

ARTICLE INFO

Article history:
Received 7 September 2021
Received in revised form 12 November 2021
Accepted 12 November 2021
Available online 22 November 2021

Keywords:
Cahn–Hilliard equation
Biot's equations
Generalized gradient flow
Mathematical modeling
Tumor growth modeling

ABSTRACT

In this work, we propose a new model for flow through deformable porous media, where the solid material has two phases with distinct material properties. The two phases of the porous material evolve according to a generalized Ginzburg–Landau energy functional, with additional impact from both elastic and fluid effects, and the coupling between flow and deformation is governed by Biot's theory. This results in a three-way coupled system which can be seen as an extension of the Cahn–Larché equations with the inclusion of a fluid flowing through the medium. The model covers essential coupling terms for several relevant applications, including solid tumor growth, biogrout, and wood growth simulation. Moreover, we show that this coupled set of equations follow a generalized gradient flow framework. This opens a toolbox of analysis and solvers which can be used for further study of the model. Additionally, we provide a numerical example showing the impact of the flow on the solid phase evolution in comparison to the Cahn–Larché system.

1. Introduction

In this letter, we develop a general model with the ability to capture situations with flow through a deformable porous medium, at Darcy scale, that changes character in terms of stiffness, permeability, compressibility, and poroelastic coupling strength due to phase changes in the solid matrix. The phase changes are governed by a generalized Ginzburg–Landau energy functional, and there are several applications where this type of behavior exists. One example being solid tumor evolution, where it is argued that stress effects resulting from tumor growth impact the tumor evolution itself [1] and that stress can inhibit tumor growth [2]. Moreover, the elastic properties of the surrounding matrix and the interstitial fluid pressure are elevated in most solid malignant tumors [3]. One can then consider the two-phase porous medium as cancerous and healthy cells with the surrounding extracellular matrix, and the fluid as the interstitial fluid.

* Corresponding author.
E-mail address: erlend.storvik@uib.no (E. Storvik).

https://doi.org/10.1016/j.aml.2021.107799

Additional applications of poroelastic media with solid phase changes range from biogrout to wood growth, where sapwood transforms to heartwood.

The proposed system is an extension of the Cahn–Hilliard model and the quasi-static linear Biot equations, where the Cahn–Hilliard contribution governs the solid phase changes in the system through a smooth phase-field variable, and the Biot equations govern flow and elasticity. The Cahn–Hilliard equation originates from the work of Cahn and Hilliard [4], where the interfacial free energy of a non-uniform composition was introduced to model phase separation. Coupling the Cahn–Hilliard model with elasticity, is often called the Cahn–Larché model due to its origination [5], and several applications have been considered with this model in mind, including li-ion batteries [6], and tumor evolution [7,8]. In this work, we assume small deformations and negligible inertial effects. Moreover, we include fluid to the system, which is assumed to flow through the poroelastic medium with Biot-type coupling between flow and elasticity [9].

We show that the resulting model has a generalized gradient flow structure, i.e., a dissipative system where the state of the system evolves with the negative gradient of its free energy. The extension to generalized gradient flows allows for non-quadratic, and partially degenerate, dissipation potentials, and there is currently an increasing interest in the mathematics of generalized gradient flows, both with respect to modeling [10,11], abstract analysis [12–15] and numerical solution strategies [15,16]. It is long known that the Cahn–Hilliard equation and single-phase flow through porous media can be written as standard gradient flows, and it was showed in [15] that the Biot equations have a generalized gradient flow structure. Here, we show that even though it is not obvious that the combination of two gradient flows retains the structure, the Cahn–Hilliard–Biot model does, indicating the thermodynamical consistency of the model. This will be a valuable toolbox for further study and development of mathematics for the model, both with respect to well-posedness analysis and numerical solution strategies.

The letter is structured as follows: In Section 2, the Cahn–Hilliard–Biot model is presented. Conservation laws for each of the three coupled processes; phase-field evolution, elasticity, and fluid flow are introduced, then the free energy of the system is proposed together with constitutive relations to close the system. In Section 3, the system is showed to be a generalized gradient flow, and in Section 4, a numerical example compares the newly proposed model with the Cahn–Larché system.

2. The derivation of the Cahn–Hilliard–Biot model

We consider a saturated porous medium with one fluid phase, and two solid phases with distinct material properties. The solid phases are modeled by a diffuse interface approach of Cahn–Hilliard type, where surface tension, deformation of the solid material, and pore pressure are acting as driving forces.

Let the medium $\Omega \subset \mathbb{R}^d$ be a bounded domain, d the spatial dimension, and $[0,T]$ be a time interval where T denotes the final time. In the matrix, the smooth phase-field, $\varphi : \Omega \times [0,T] \to [-1,1]$, tracks the two phases $\varphi = -1$ and $\varphi = 1$. We consider linearized elasticity with infinitesimal displacement $\boldsymbol{u}$, and $\|\nabla \boldsymbol{u}\| \ll 1$, the pore pressure is denoted by p, and $\boldsymbol{q}$ is the fluid flux.

2.1. Balance laws

Balance laws are imposed for each of the three coupled systems. For the phase-field equation, we assume that the phase change is balanced by a phase-field flux $\boldsymbol{J}$ and reactions R,

$$\partial_t \varphi + \nabla \cdot \boldsymbol{J} = R, \tag{1}$$

where the form of the reaction term differs depending on the application. In [7], a suitable reaction term is given in the context of tumor simulation with elastic effects. The elastic behavior of the material is governed by a quasi-static force balance equation where $\boldsymbol{\sigma}$ denotes the stress tensor and $\boldsymbol{f}$ external body forces

$$-\nabla \cdot \boldsymbol{\sigma} = \boldsymbol{f}. \tag{2}$$

Finally, the fluid is assumed to follow a volume balance law with negligible density gradients,

$$\partial_t \theta + \nabla \cdot \boldsymbol{q} = S_f, \tag{3}$$

where θ is the volumetric fluid content which changes due to the fluid flux $\boldsymbol{q}$ and source S_f. Notice that, as we are considering a saturated porous medium of a single fluid phase, the volumetric fluid content is proportional to the porosity of the medium which might change depending on the solid phase.

2.2. Free energy

The system is then closed through its free energy together with appropriate constitutive relations. We assume that the energy can be decomposed into three parts; the regularized interface energy, containing chemical energy and interfacial energy between the solid phases, the elastic energy, and the fluid energy

$$\mathcal{E}(\varphi, \boldsymbol{u}, \theta) = \mathcal{E}_{\text{ch}}(\varphi) + \mathcal{E}_{\text{e}}(\varphi, \boldsymbol{u}) + \mathcal{E}_{\text{f}}(\varphi, \boldsymbol{u}, \theta). \tag{4}$$

The regularized interface energy [4] is given as

$$\mathcal{E}_{\text{ch}}(\varphi) \coloneqq \int_\Omega \Psi(\varphi) + \frac{\gamma}{2}|\nabla\varphi|^2 \, dx, \tag{5}$$

where deviations from pure phases are penalized through the double-well potential $\Psi(\varphi)$, and transitions between phases are penalized by the second term which is related to the interfacial energy. Here, the parameter γ corresponds to interfacial tension between the phases and will account for adhesive and cohesive forces. The double-well potential takes minimal values in the two phases, $\varphi = -1$ and $\varphi = 1$, and is, in this work, given as

$$\Psi(\varphi) \coloneqq E_\Psi \left(1 - \varphi^2\right)^2, \tag{6}$$

where $E_\Psi > 0$ is a chemical energy density parameter.

We assume that the elastic energy takes the form that is typical to the Cahn–Larché equations,

$$\mathcal{E}_{\text{e}}(\varphi, \boldsymbol{u}) = \int_\Omega \frac{1}{2}\big(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{T}(\varphi)\big) : \mathbb{C}(\varphi)\big(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{T}(\varphi)\big) \, dx, \tag{7}$$

where $\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\left(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\top\right)$ is the linearized strain at displacement $\boldsymbol{u}$. The second term, $\mathcal{T}(\varphi)$, is the eigenstrain at φ (often called *stress-free strain*, or *intrinsic strain*) which corresponds to the state of the strain tensor if the material was uniform and unstressed [17]. Moreover, it can be considered to account for swelling effects [6] and takes different values depending on the solid phase φ. Here, we consider the form $\mathcal{T}(\varphi) = \xi\varphi\boldsymbol{I}$, where ξ is a swelling parameter. The elastic stiffness tensor $\mathbb{C}(\varphi)$, which can be anisotropic, depends on the phase-field.

Finally, we consider a natural extension of the classical fluid energy which is given as in [15] by

$$\mathcal{E}_{\text{f}}(\varphi, \boldsymbol{u}, \theta) = \int_\Omega \frac{M(\varphi)}{2} \left(\theta - \alpha(\varphi)\nabla \cdot \boldsymbol{u}\right)^2 \, dx \tag{8}$$

where both the compressibility parameter $M(\varphi)$ and the Biot–Willis coupling coefficient $\alpha(\varphi)$ depend on the phase-field φ.

2.3. Constitutive relations

Assuming that the phase-field follows Fick's law for non-ideal mixtures, the flux $\boldsymbol{J}$ is proportional to the negative gradient of the chemical potential

$$\boldsymbol{J} = -m(\varphi)\nabla\mu, \tag{9}$$

where $m(\varphi)$ is the chemical mobility. The chemical potential μ is defined to be the variational derivative of the free energy with respect to φ. Here, we denote the variational derivative of $\mathcal{E}$ with respect to y by $\delta_y\mathcal{E}$, and standard computations yield

$$\mu \coloneqq \delta_\varphi \mathcal{E} = \Psi'(\varphi) - \gamma\Delta\varphi + \delta_\varphi \mathcal{E}_{\mathrm{e}}(\varphi, \boldsymbol{u}) + \delta_\varphi \mathcal{E}_{\mathrm{f}}(\varphi, \boldsymbol{u}, \theta), \tag{10}$$

where zero Neumann or periodic boundary conditions have been applied to φ,

$$\delta_\varphi \mathcal{E}_{\mathrm{e}}(\varphi, \boldsymbol{u}) = \frac{1}{2}\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{T}(\varphi)\right) : \mathbb{C}'(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{T}(\varphi)\right) - \mathcal{T}'(\varphi) : \mathbb{C}(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{T}(\varphi)\right), \tag{11}$$

and

$$\delta_\varphi \mathcal{E}_{\mathrm{f}}(\varphi, \boldsymbol{u}, \theta) = \frac{M'(\varphi)}{2}(\theta - \alpha(\varphi)\nabla\cdot\boldsymbol{u})^2 - M(\varphi)(\theta - \alpha(\varphi)\nabla\cdot\boldsymbol{u})\alpha'(\varphi)\nabla\cdot\boldsymbol{u}. \tag{12}$$

According to thermodynamical principles [9], we define the stress tensor to be the rate of change of energy with respect to strain

$$\boldsymbol{\sigma} \coloneqq \delta_{\boldsymbol{\varepsilon}} \mathcal{E} = \mathbb{C}(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{T}(\varphi)\right) - M(\varphi)\alpha(\varphi)\left(\theta - \alpha\nabla\cdot\boldsymbol{u}\right)\boldsymbol{I}, \tag{13}$$

and the pore pressure p to be the rate of change of energy with respect to volumetric fluid content

$$p \coloneqq \delta_\theta \mathcal{E} = M(\varphi)\left(\theta - \alpha(\varphi)\nabla\cdot\boldsymbol{u}\right). \tag{14}$$

Finally, the flow through the porous medium is assumed to follow Darcy's law

$$\boldsymbol{q} = -\kappa(\varphi)\nabla p, \tag{15}$$

where the permeability $\kappa(\varphi)$ is assumed to depend on the solid phase.

Combining the balance laws with the constitutive relations, and making the identification (14) in (12) and (13), the Cahn–Hilliard–Biot model becomes

$$\partial_t\varphi - \nabla\cdot(m(\varphi)\nabla\mu) = R \tag{16}$$

$$\mu + \gamma\Delta\varphi - \Psi'(\varphi) - \delta_\varphi \mathcal{E}_{\mathrm{e}}(\varphi, \boldsymbol{u}) - \delta_\varphi \mathcal{E}_{\mathrm{f}}(\varphi, \boldsymbol{u}, p) = 0 \tag{17}$$

$$-\nabla\cdot\left(\mathbb{C}(\varphi)\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{T}(\varphi)\right)\right) + \nabla\left(\alpha(\varphi)p\right) = \boldsymbol{f} \tag{18}$$

$$\partial_t\left(\frac{p}{M(\varphi)} + \alpha(\varphi)\nabla\cdot\boldsymbol{u}\right) + \nabla\cdot\boldsymbol{q} = S_f \tag{19}$$

$$\boldsymbol{q} + \kappa(\varphi)\nabla p = 0, \tag{20}$$

equipped with suitable boundary and initial conditions.

3. The Cahn–Hilliard–Biot model as a generalized gradient flow

In this section, we identify the proposed Cahn–Hilliard–Biot model (16)–(20) as a generalized gradient flow, which in contrast to regular gradient flows allows for non-quadratic and even degenerate dissipation potentials. By making this identification for the newly proposed model, a wide toolbox of well-posedness analysis [12,15], numerical error analysis [13,14], and numerical solution algorithms [15,16] are made available, which will be a valuable asset for further study. Moreover, generalized gradient flows are inherently thermodynamically consistent in the sense that the free energy of the system decreases through dissipation, and can only increase through external forces. A generalized gradient flow takes the form

$$\mathcal{D}_{\partial_t \boldsymbol{z}}\mathcal{R}(\partial_t\boldsymbol{z}, \boldsymbol{z}) = -\mathcal{D}_{\boldsymbol{z}}\mathcal{E}(\boldsymbol{z}) + \mathcal{P}_{\mathrm{ext}}, \tag{21}$$

where $\boldsymbol{z}$ is a state variable, $\mathcal{R}$ is a dissipation potential, $\mathcal{E}$ is the energy at state $\boldsymbol{z}$, $\mathcal{D}_{\boldsymbol{x}}$ is the Gateaux derivative with respect to $\boldsymbol{x}$, and $\mathcal{P}_{\text{ext}}$ corresponds to external forces. Alternatively, one can reformulate the generalized gradient flow and split between states evolving with ($\boldsymbol{z}_{\text{d}}$) and without ($\boldsymbol{z}_{\text{df}}$) dissipation to get the constrained minimization problem

$$\boldsymbol{z}_{\text{df}} = \arg\min_{\boldsymbol{s}_{\text{df}}} \{\mathcal{E}(\boldsymbol{s}_{\text{df}}) - \langle \mathcal{P}_{\text{ext,df}}, \boldsymbol{s}_{\text{df}} \rangle\} \tag{22}$$

$$(\partial_t \boldsymbol{z}_{\text{d}}, \mathcal{F}) = \arg\min_{\boldsymbol{s}_{\text{d}}, \boldsymbol{l}} \Big\{ \tilde{\mathcal{R}}(\boldsymbol{l}, \boldsymbol{z}_{\text{d}}) + \langle \mathcal{D}_{\boldsymbol{z}_{\text{d}}} \mathcal{E}(\boldsymbol{z}_{\text{d}}), \boldsymbol{s}_{\text{d}} \rangle - \langle \mathcal{P}_{\text{ext,d}}, \boldsymbol{s}_{\text{d}} \rangle \Big\} \tag{23}$$

subject to $\boldsymbol{s}_{\text{d}} + \nabla \cdot \boldsymbol{l} = \boldsymbol{S}$, where $\mathcal{R}(\partial_t \boldsymbol{z}_{\text{d}}, \boldsymbol{z}_{\text{d}}) = \tilde{\mathcal{R}}(\mathcal{F}, \boldsymbol{z}_{\text{d}})$, $\langle \cdot, \cdot \rangle$ is the canonical inner-product, and the balance law $\partial_t \boldsymbol{z}_d + \nabla \cdot \mathcal{F} = \boldsymbol{S}$ with flux $\mathcal{F}$, and source $\boldsymbol{S}$ holds.

For the Cahn–Hilliard–Biot system, consider the state variables $\boldsymbol{z} = (\varphi, \boldsymbol{u}, \theta)$, the energy $\mathcal{E}(\boldsymbol{z})$ from (4), and the state-dependent dissipation potential

$$\mathcal{R}(\boldsymbol{J}, \partial_t \boldsymbol{u}, \boldsymbol{q}, \varphi) := \mathcal{R}_{\text{ch}}(\boldsymbol{J}, \varphi) + \mathcal{R}_{\text{e}}(\partial_t \boldsymbol{u}) + \mathcal{R}_{\text{f}}(\boldsymbol{q}, \varphi), \tag{24}$$

with

$$\mathcal{R}_{\text{ch}}(\boldsymbol{J}, \varphi) := \int_\Omega \frac{1}{2m(\varphi)} |\boldsymbol{J}|^2 \, dx, \quad \mathcal{R}_{\text{e}}(\partial_t \boldsymbol{u}) := 0, \quad \text{and} \quad \mathcal{R}_{\text{f}}(\boldsymbol{q}, \varphi) := \int_\Omega \frac{1}{2\kappa(\varphi)} |\boldsymbol{q}|^2 \, dx$$

together with the conservation laws

$$\partial_t \varphi + \nabla \cdot \boldsymbol{J} = R \quad \text{and} \quad \partial_t \theta + \nabla \cdot \boldsymbol{q} = S_f. \tag{25}$$

As the deformation is assumed to be dissipation free, the generalized gradient flow reads: Find φ, $\boldsymbol{u}$, and θ such that

$$\boldsymbol{u} = \arg\min_{\boldsymbol{w}} \Big\{ \mathcal{E}(\varphi, \boldsymbol{w}, \theta) - \langle \mathcal{P}_{\text{ext,e}}, \boldsymbol{w} \rangle \Big\} \tag{26}$$

$$(\partial_t \varphi, \partial_t \theta, \boldsymbol{J}, \boldsymbol{q}) = \arg\min_{\eta, s, \boldsymbol{l}, \boldsymbol{v}} \Big\{ \mathcal{R}_{\text{ch}}(\boldsymbol{l}, \varphi) + \langle \mathcal{D}_\varphi \mathcal{E}(\varphi, \boldsymbol{u}, \theta), \eta \rangle + \mathcal{R}_{\text{f}}(\boldsymbol{v}, \varphi) + \langle \mathcal{D}_\theta \mathcal{E}(\varphi, \boldsymbol{u}, \theta), s \rangle + \langle \mathcal{P}_{\text{ext,f}}, s \rangle \Big\} \tag{27}$$

subject to $\eta + \nabla \cdot \boldsymbol{l} = R$ and $s + \nabla \cdot \boldsymbol{v} = S_f$ with balance laws (25), $\langle \mathcal{P}_{\text{ext,e}}, \boldsymbol{w} \rangle := \int_\Omega \boldsymbol{f} \cdot \boldsymbol{w} \, dx$ and $\mathcal{P}_{\text{ext,f}}$ corresponding to external forces related to the fluid (e.g., boundary conditions or gravitational force). Calculating optimality conditions, and substituting the phase-field flux $\boldsymbol{J}$ by the chemical potential μ through Fick's law (9), and the volumetric fluid content θ with the fluid pressure p through the relation (14), one obtains the variational form of the system (16)–(20).

4. Numerical example

Here, we present a numerical example that emphasizes the impact the flow has on the phase-field evolution in the Cahn–Hilliard–Biot model compared to a Cahn–Larché simulation (Cahn–Hilliard coupled with only elasticity). We apply a pressure boundary condition to the Cahn–Hilliard–Biot system that acts as an external force (in order to enforce flow in the domain), and compare it to both a simulation without the pressure condition and to a Cahn–Larché simulation. The example clearly shows that when the fluid flow is dominant, it also plays a crucial role in the evolution of the phase-field. However, in regimes with little, to no flow, the phase-field is unaffected compared to the Cahn–Larché model.

We consider a unit square domain where three circular shapes of phase $\varphi = 1$ are surrounded by phase $\varphi = -1$ initially, see Fig. 1(a),1(e),1(i). For both pressure and displacement, we apply zero initial data. The variational system (16)–(20) is discretized in time by a semi-implicit Euler method, where the deviation from fully implicit Euler is an application of the first order convex splitting method of the double-well potential

Table 1
Table of simulation parameters. Here, L denotes the unit of length, F force, and T time. Notice that the units are consistent in three spatial dimensions and that our example should be interpreted as a two-dimensional representation of a domain with thickness $1L$.

Parameter name	Symbol	Value	Unit	Parameter name	Symbol	Value	Unit
Chemical mobility	m	1	$\left[\frac{L^4}{FT}\right]$	Biot–Willis parameters	α_{-1}, α_1	1, 0.5	[–]
Interfacial tension	γ	1e−4	$[F]$	Permeabilities	κ_{-1}, κ_1	1, 0.1	$\left[\frac{L^4}{FT}\right]$
Compressibilities	M_{-1}, M_1	1, 0.1	$\left[\frac{F}{L^2}\right]$	Time step size	τ	1e−3	$[T]$
Swelling parameter	ξ	0.3	[–]	Mesh diameter	h	$\frac{\sqrt{2}}{65}$	$[L]$
Chemical energy density	E_Ψ	$\frac{1}{4}$	$\left[\frac{F}{L^2}\right]$	Elasticity tensors	$\mathbb{C}_{-1}, \mathbb{C}_1$	(28)	$\left[\frac{F}{L^2}\right]$

$\Psi(\varphi)$ as proposed in [18]. The three-way coupled nonlinear system is then solved by an iterative decoupling scheme, starting with the Cahn–Hilliard subsystem (16)–(17), then elasticity (18), and finally, flow (19)–(20), and the iterations are terminated when the (relative and absolute) residual and incremental values in the $L^2(\Omega)$-norm are smaller than a tolerance of 10^{-6}. The Cahn–Hilliard subsystem (16)–(17) is discretized in space with bilinear rectangular finite elements for both phase-field φ and chemical potential μ, and the nonlinear equations are solved by a Newton method in each iterative decoupling-iteration. As initial guess in both the Newton method and the iterative decoupling method, the solution at the previous time step (the initial value at the first time step) is chosen. The flow subsystem (19)–(20) is discretized in space by lowest-order Raviart–Thomas elements, RT0, for the flux and constant elements for pressures, and the elasticity equation (18) is discretized with bilinear finite elements. We have used modules from the DUNE project, specifically dune-functions [19], for the implementation.

The material parameters can be found in Table 1, and the permeability $\kappa(\varphi)$, compressibility $M(\varphi)$, Biot–Willlis coefficient $\alpha(\varphi)$ and elasticity tensor $\mathbb{C}(\varphi)$ are depending on the phase-field through the interpolation function $\pi(\varphi)$; $\kappa(\varphi) = \kappa_{-1} + \pi(\varphi)(\kappa_1 - \kappa_{-1})$, $M(\varphi) = M_{-1} + \pi(\varphi)(M_1 - M_{-1})$, $\alpha(\varphi) = \alpha_{-1} + \pi(\varphi)(\alpha_1 - \alpha_{-1})$ and $\mathbb{C}(\varphi) = \mathbb{C}_{-1} + \pi(\varphi)(\mathbb{C}_1 - \mathbb{C}_{-1})$. Here, we choose

$$\pi(\varphi) = \begin{cases} 0, & \varphi < -1 \\ \frac{1}{4}\left(-\varphi^3 + 3\varphi + 2\right), & \varphi \in [-1, 1] \\ 1, & \varphi > 1 \end{cases}, \qquad \mathbb{C}_{-1} = \begin{pmatrix} 4 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 8 \end{pmatrix}, \qquad \mathbb{C}_1 = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \tag{28}$$

with the two elasticity tensors written in Voigt notation in two spatial dimensions. Zero Neumann boundary conditions are applied to both the phase-field and the chemical potential, while the displacement is equipped with zero Dirichlet conditions on the entire boundary. For the flow subsystem, we enforce a pressure drop from $p = 0.25$ to $p = 0$ from top to bottom while no-flow conditions are applied on the left and right parts of the boundary. The reaction R, source S_f and body force $\boldsymbol{f}$ are all equal to 0.

In Fig. 1(a)–1(d), the phase-field function φ is plotted after a series of time steps for the Cahn–Hilliard–Biot model with a drop in pressure from $p = 0.25$ to $p = 0$ from top to bottom. In Fig. 1(e)–1(h) the solution is plotted at the same time steps, but with zero pressure on the entire boundary, and similarly in Fig. 1(i)–1(l) the plots are from a simulation of the Cahn–Larché system. We observe that when the flow is prominent in the simulation the phase-field is also significantly affected and takes a directional preference to that of the flow direction. When, on the other hand, the system merely is filled with a fluid that has no driving force in itself, the phase-field evolution is close to unaffected compared to the system without a fluid. We emphasize also that the system energies (including external forces) are decreasing over the scope of the simulation, as is expected from dissipative systems of gradient flow type. This is showed in Fig. 1, where the energy is a combination of the free energy of the system (4), and the external forces applied through the pressure boundary condition, $\mathcal{E}_{\text{Tot}} = \mathcal{E}(\varphi, \boldsymbol{u}, p) - \int_{\Gamma_{\text{Top}}} p_{\text{Top}}(\boldsymbol{q} \cdot \boldsymbol{n})\, dx$, with $\boldsymbol{n}$ being the outwards pointing normal vector. Moreover, notice that the simulation is only a redistribution of the phases, due to the lack of reaction/source terms.

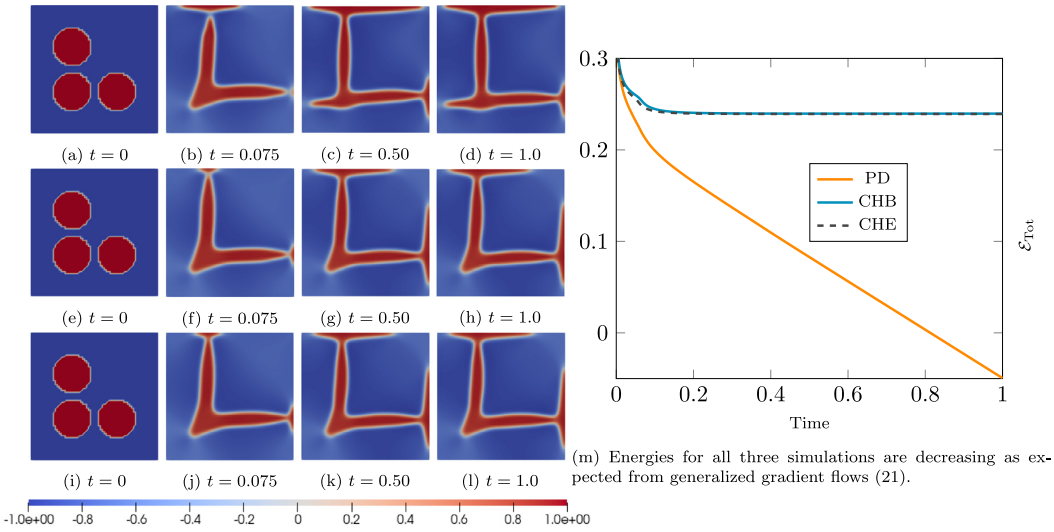


Fig. 1. (a)–(l): the solution at time t for the phase-field φ. (a)–(d): Cahn–Hilliard–Biot with $p = 0.25$ on the top, (e)–(h): Cahn–Hilliard–Biot with zero pressure BC, (i)–(l): Cahn–Larché. (m): system energy (with external contributions). PD is Cahn–Hilliard–Biot with $p = 0.25$ on the top, CHB is Cahn–Hilliard–Biot with zero pressure BC and CHE is Cahn–Larché.

5. Conclusions

The Cahn–Hilliard–Biot system was derived through balance laws and constitutive relations, i.e., Fick's law for the phase-field, and Darcy's law for the fluid flow. Key quantities are defined, following thermodynamical principles, as rates of change of the free energy. The equations feature a three-way coupling, and the impact from flow to the phase-field was showed to be significant through a numerical example; the phase-field does not only evolve as it would through the Cahn–Larché equations, but its evolution is aligned and magnified in the flow direction. Moreover, we showed that the system follows a generalized gradient flow framework and that the energy dissipates numerically as expected. By this, we lay the groundwork for a general model, showing numerical properties and highlighting important coupling terms, that can be further tailored and studied depending on the specific application in mind.

Acknowledgments

The work has in part been supported by the Research Council of Norway through the projects 294716 and 250223, and the FracFlow project funded by Equinor, Norway through Akademiaavtalen.

References

[1] E.A.B.F. Lima, J.T. Oden, D.A. Hormuth, T.E. Yankeelov, R.C. Almeida, Selection, calibration, and validation of models of tumor growth, Math. Models Methods Appl. Sci. 26 (12) (2016) 2341–2368.
[2] G. Cheng, J. Tse, R. Jain, L.L. Munn, Micro-environmental mechanical stress controls tumor spheroid size and morphology by suppressing proliferation and inducing apoptosis in cancer cells, PLoS One 4 (2) (2009) e4632.
[3] M Milosevic, S.J. Lunt, E. Leung, J. Skliarenko, P.A. Shaw, A Fyles, R.P. Hill, Interstitial permeability and elasticity in human cervix cancer, Microvasc Res. 75 (3) (2008) 381–390.
[4] J.W. Cahn, J.E. Hilliard, Free energy of a nonuniform system. I. Interfacial free energy, J. Chem. Phys. 28 (2) (1958) 258–267.
[5] F.C. Larché, J.W. Cahn, The effect of self-stress on diffusion in solids, Acta. Metall. 30 (10) (1982) 1835–1845.

[6] P. Areias, E. Samaniego, T. Rabczuk, A staggered approach for the coupling of Cahn–Hilliard type diffusion and finite strain elasticity, Comput. Mech. 57 (2) (2016) 339–351.
[7] H. Garcke, K.F. Lam, A. Signori, On a phase field model of Cahn–Hilliard type for tumour growth with mechanical effects, Nonlinear Anal-Real. 57 (2021) 103192.
[8] M. Fritz, C. Kuttler, M.L. Rajendran, L. Scarabosio, B. Wohlmuth, On a subdiffusive tumour growth model with fractional time derivative, IMA J. Appl. Math. 86 (2021) 688–729.
[9] O. Coussy, Poromechanics, John Wiley & Sons, 2004.
[10] M.A. Peletier, Variational modelling: Energies, gradient flows, and large deviations, 2014, arXiv preprint arXiv:1402.1990.
[11] C. Cancès, T.O. Gallouët, L. Monsaingeon, The gradient flow structure for incompressible immiscible two-phase flows in porous media, C. R. Math. 353 (11) (2015) 985–989.
[12] P. Colli, On some doubly nonlinear evolution equations in Banach spaces, Jpn. J. Ind. Appl. Math. 9 (2) (1992) 181–203.
[13] R.H. Nochetto, G. Savaré, C. Verdi, A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations, Commun. Pure Appl. Anal. 53 (5) (2000) 525–589.
[14] S. Bartels, R.H. Nochetto, A.J. Salgado, Discrete total variation flows without regularization, SIAM J. Numer. Anal. 52 (1) (2014) 363–385.
[15] J.W. Both, K. Kumar, J.M. Nordbotten, F.A. Radu, The gradient flow structures of thermo-poro-visco-elastic processes in porous media, 2019, arXiv preprint arXiv:1907.03134.
[16] A. Jüngel, U. Stefanelli, L. Trussardi, Two structure-preserving time discretizations for gradient flows, Appl. Math. Optim. 80 (3) (2019) 733–764.
[17] P. Fratzl, O. Penrose, J.L. Lebowitz, Modeling of phase separation in alloys with coherent elastic misfit, J. Stat. Phys. 95 (5) (1999) 1429–1503.
[18] D.J. Eyre, Unconditionally gradient stable time marching the Cahn–Hilliard equation, Mater. Res. Soc. Symp. Proc 529 (1998).
[19] C. Engwer, C. Gräser, S. Müthing, O. Sander, The interface for functions in the dune-functions module, 2015, arXiv preprint arXiv:1512.06136.

Paper E

A robust solution strategy for the Cahn-Larché equations

Storvik, E., Both, J.W., Nordbotten, J.M., and Radu, F.A.
In review.
arXiv:2206.01541[math.NA]

A robust solution strategy for the Cahn-Larché equations

Erlend Storvik*[1], Jakub Wiktor Both[1], Jan Martin Nordbotten[1], and Florin Adrian Radu[1]

[1]Center for Modeling of Coupled Subsurface Dynamics, Department of Mathematics, University of Bergen, Allégaten 44, 5007 Bergen, Norway

Abstract

In this paper we propose a solution strategy for the Cahn-Larché equations, which is a model for linearized elasticity in a medium with two elastic phases that evolve subject to a Ginzburg-Landau type energy functional. The system can be seen as a combination of the Cahn-Hilliard regularized interface equation and linearized elasticity, and is non-linearly coupled, has a fourth order term that comes from the Cahn-Hilliard subsystem, and is non-convex and nonlinear in both the phase-field and displacement variables. We propose a novel semi-implicit discretization in time that uses a standard convex-concave splitting method of the nonlinear double-well potential, as well as special treatment to the elastic energy. We show that the resulting discrete system is equivalent to a convex minimization problem, and propose and prove the convergence of alternating minimization applied to it. Finally, we present numerical experiments that show the robustness and effectiveness of both alternating minimization and the monolithic Newton method applied to the newly proposed discrete system of equations. We compare it to a system of equations that has been discretized with a standard convex-concave splitting of the double-well potential, and implicit evaluations of the elasticity contributions and show that the newly proposed discrete system is better conditioned for linearization techniques.

1 Introduction

The Cahn-Larché system models elastic deformation within a two-phase solid material. Here, the solid phases evolve subject to a Ginzburg-Landau type energy functional, as proposed in the work of Cahn and Hilliard [1, 2], additively coupled with the elastic energy of the system. The equations are credited to the work of Cahn and Larché [3, 4] which considered stress effects related to diffusion in solids. More recently, the equations were studied experimentally and verified in [5] as a model for the connection between chemical and mechanical processes in alloys. Additionally the Cahn-Larché system has been applied in relation to tumor modelling [6, 7, 8], diffusional corsening in solders [9, 10], and to model the process of intercalation of lithium ions into silicon [11]. Moreover, in [12] a phase-field model, closely related to the Cahn-Hilliard equation was proposed to account for unsaturated flow through porous materials. Extensions to a Cahn-Larché setting could be considered to model flow through swelling deformable porous media.

Over the last two decades there has been extensive research on the well-posedness and analysis of both the continuous and discrete counterparts of Cahn-Larché systems. In [13, 14] existence and uniqueness results are obtained for the weak system of equations, in [6] similar results are obtained for the coupling of Cahn-Larché to transport in the context of tumor growth, and on the same model an optimal control problem is analyzed in [7]. In [10], existence and uniqueness of a discretized Cahn-Larché system is provided, and in [15, 16] the sharp interface limit of the equations is showed to be equivalent to a modified Hele-Shaw system coupled with elasticity. There are several published works on numerical discretization techniques for the system. In [10, 17], adaptive mesh refinement techniques are discussed and [18, 19] consider spatial discretization with linear finite elements together with the implicit Euler and Crank-Nicholson time discretizations.

In this work, we propose a novel semi-implicit time-discretization that corresponds to the optimality conditions of a convex minimization problem, and therefore is suitable for nonlinear solvers. The semi-implicit time discretization is related to the unconditionally gradient stable convex-concave splitting method that Eyre proposed in [20] for the double-well potential of the Cahn-Hilliard equation. Here, that treatment is adopted and applied to the Cahn-Larché equations, in two different settings; when the elasticity tensor is independent of, and dependent on the phase-field. In the former case, the coupling between phase-field and elasticity is linear and by evaluating the terms from the elasticity subsystem implicitly the discrete system of equations is identified with a convex minimization problem, similar to the treatment in [18]. Furthermore, the system of equations is showed to be unconditionally gradient stable, and that an alternating minimization technique, alternating between solving for phase-field and displacement, applied to the proposed minimization problem converges. In the second case, however, implicit evaluation in time of the terms corresponding to the elasticity

*Corresponding author: erlend.storvik@uib.no

subsystem does not lead to a convex minimization problem when the elasticity tensor depends on the phase-field, even with the convex-concave splitting method applied to the double-well potential [20]. We show through numerical examples that the Newton method fails to converge in several instances in this case and propose a way to carefully evaluate some terms explicitly in time, such that the corresponding minimization problem is convex. This leads to a system that is better conditioned for solution algorithms, and a theoretical proof of convergence for the alternating minimization method is provided. Moreover, convergence is experienced for the Newton method in all numerical examples.

When solving the coupled discrete system of equations there exists two common choices: Either, to solve the entire system monolithically, using some linearization procedure, or to apply an iterative decoupling method. A beneficial trait of decoupling methods is the possibility to use readily available solvers for each subsystem. For the discrete system of equations that we present in this paper that corresponds to solving an extended Cahn-Hilliard equation with well-behaving nonlinearities, due to the convex-concave splitting method, and an elasticity equation with heterogeneous elasticity tensor subsequently. For the Cahn-Hilliard subsystem some linearization technique (e.g., Newton's method) is still needed to handle the nonlinearities corresponding to the modified double-well potential and terms that arise from the elasticity contribution. The elasticity subsystem, on the other hand, reduces to a standard elasticity equation with, possibly, heterogeneous elasticity tensor. Any readily available solvers and preconditioners for these subproblems can be applied, and combining the decoupling method with the linearization of the nonlinear Cahn-Hilliard subsystem (doing only one linearization iteration in each decoupling iteration) as discussed in [21, 22] is possible as well. Decoupling techniques are often also known as staggered solution strategies, splitting schemes or alternating minimization for symmetric problems with an underlying minimization structure, and have been widely adopted to solve equations related to phase-field modelling of brittle fracture propagation [23, 24, 25, 26, 27], and poroelasticity equations where flow and elasticity is coupled [28, 29, 30, 31]. Moreover, a staggered solution strategy was used to solve finite-strain elasticity coupled with the Cahn-Hilliard equation in [32].

Here, we investigate the properties of both monolithical solvers and decoupling methods for the Cahn-Larché equations. Moreover, we properly address the theoretical convergence properties of alternating minimization. To do this we formulate the discretized system of equations as a minimization problem and utilize an abstract convergence result for alternating minimization provided in [33]. This framework requires at least convexity of the minimization problem in each variable, and Lipschitz continuity of its gradients. We prove that this holds true for the discretized Cahn-Larché equations and obtain convergence rates that we investigate through numerical examples. Moreover, it can be useful to apply the Anderson acceleration [34] post-processing technique (as done in e.g., [25, 22]) to enhance the convergence speed of the alternating minimization method. This is particularly useful for staggered solution methods as the Anderson acceleration is known to be accelerating for linearly convergent fixed-point schemes [35].

To summarize, the main contributions of the paper are:

- We propose a new, semi-implicit time discretization of the Cahn-Larché equations that leads to a nonlinear system which is suitable for linearization and decoupling methods.
- Identification of the proposed discretized equations with a convex minimization problem.
- A proof of convergence for alternating minimization as an iterative solver, including convergence rates.
- Numerical experiments showing the efficiency of the proposed time-discretization and iterative solver with comparison to monolithic methods and acceleration.

Moreover, we stress that the time-discretization and decoupling procedures that we apply here, can be extended and applied to similar models, e.g., the Cahn-Hilliard-Biot model [36], tumor growth models with transport effects [6], phase-field models for precipitation and dissolution processes [37] and the two-phase two fluxes Cahn-Hilliard model [38].

The paper is structured as follows: The mathematical model and assumptions on the model parameters are presented in Section 2. In Section 3, we discuss the discrete problem associated with the Cahn-Larché system both for constant and phase-field-dependent elasticity tensor. Moreover, we show equivalence between the discrete model and a minimization problem, and prove convergence of alternating minimization applied to this problem. In Section 4, we present several numerical experiments and show the benefits of the proposed discretization and linearization/decoupling method compared to standard choices. Finally, in Section 5 we make concluding remarks.

2 The mathematical problem and assumptions on model parameters

The Cahn-Larché system is a combination of a Cahn-Hilliard phase-field model and linearized elasticity with infinitesimal strains and displacements [3, 18]. We consider the domain $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary, where d is the spatial dimension, and the time interval $[0, T]$ with final time T. Let $\varphi : \Omega \times [0, T] \to [-1, 1]$ be the phase-field variable, where pure phases are attained for $\varphi = -1$, $\varphi = 1$. Moreover, let $\boldsymbol{u} : \Omega \times [0, T] \to \mathbb{R}^d$ be the infinitesimal displacement.

2.1 Balance laws and constitutive relations

We assume that the phase-field φ follows the balance law

$$\partial_t \varphi + \nabla \cdot \boldsymbol{J} = R,$$

where $\boldsymbol{J}$ is the phase-field flux and R accounts for reactions. Moreover, the stress follows quasi-static linear momentum balance (ignoring inertial effects)

$$-\nabla \cdot \boldsymbol{\sigma} = \boldsymbol{f},$$

where $\boldsymbol{\sigma}$ is the stress-tensor and $\boldsymbol{f}$ corresponds to external forces. The free energy $\mathcal{E}(\varphi, \boldsymbol{u})$ of the system is assumed to be an additive combination of the regularized interface energy $\mathcal{E}_{\text{ch}}(\varphi)$ and the potential elastic energy $\mathcal{E}_{\text{e}}(\varphi, \boldsymbol{u})$

$$\mathcal{E}(\varphi, \boldsymbol{u}) := \mathcal{E}_{\text{ch}}(\varphi) + \mathcal{E}_{\text{e}}(\varphi, \boldsymbol{u}). \tag{1}$$

The regularized chemical energy of the system is defined as

$$\mathcal{E}_{\text{ch}}(\varphi) := \int_\Omega \gamma \left(\frac{1}{\ell} \Psi(\varphi) + \frac{\ell}{2} |\nabla \varphi|^2 \right) \, dx, \tag{2}$$

where $\Psi(\varphi)$, often chosen as $\Psi(\varphi) = \left(1 - \varphi^2\right)^2$, is a double-well potential that penalizes non-pure phase-field values ($|\varphi| \neq 1$), and $\frac{|\nabla \varphi|^2}{2}$ regularizes the transition between phases by penalizing rapid changes (in space) of the phase-field. The parameter γ is related to the interfacial tension between the two phases, and can be considered to account for adhesive/cohesive forces between the phases, and ℓ is related to the width of the regularization region. The elastic potential energy is

$$\mathcal{E}_{\text{e}}(\varphi, \boldsymbol{u}) := \frac{1}{2} \int_\Omega \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \left(\varphi - \tilde{\varphi} \right) \boldsymbol{I} \right) : \mathbb{C}(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \left(\varphi - \tilde{\varphi} \right) \boldsymbol{I} \right) \, dx \tag{3}$$

where $\boldsymbol{\varepsilon}(\boldsymbol{u}) := \frac{\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top}{2}$ is the linearized symmetric strain tensor, $\mathbb{C}(\varphi)$ is the fourth order elasticity tensor, the term $\xi \left(\varphi - \tilde{\varphi} \right) \boldsymbol{I}$ accounts for swelling effects where $\tilde{\varphi}$ is a reference phase-field, and $\boldsymbol{I}$ is the identity tensor in $\mathbb{R}^{d \times d}$. For the rest of the paper, we assume that $\tilde{\varphi} = 0$ to make the notation more simplistic. All the theory and numerical examples can trivially be extended to account for $\tilde{\varphi} \in [-1, 1]$.

As constitutive relations we assume that the phase-field flux $\boldsymbol{J}$ is diffusive and follows Fick's law

$$\boldsymbol{J} = -m(\varphi) \nabla \mu,$$

where $m(\varphi)$ is the chemical mobility, which we will assume to be constant in this work, and μ is the chemical potential, which is defined as the rate of change, variational derivative, of the free energy of the system with respect to the phase-field. Here, we denote the variational derivative of $\mathcal{E}$ with respect to y by $\delta_y \mathcal{E}$, and standard computations yield

$$\begin{aligned} \mu := \delta_\varphi \mathcal{E}(\varphi, \boldsymbol{u}) \quad &= \quad \gamma \left(\frac{1}{\ell} \Psi'(\varphi) - \ell \Delta \varphi \right) - \xi \boldsymbol{I} : \mathbb{C}(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right) \\ &\quad + \frac{1}{2} \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right) : \mathbb{C}'(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right), \end{aligned}$$

where, we have utilized that the normal derivative of the phase-field vanishes on the boundary ($\nabla \varphi \cdot \boldsymbol{n} = 0$ at $\partial \Omega$). The stress tensor $\boldsymbol{\sigma}$ is defined as the rate of change of the free energy with respect to strain $\boldsymbol{\varepsilon}$

$$\boldsymbol{\sigma} := \delta_{\boldsymbol{\varepsilon}} \mathcal{E}(\varphi, \boldsymbol{\varepsilon}(\boldsymbol{u})) = \mathbb{C}(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right).$$

In total, we search for the triplet $(\varphi, \mu, \boldsymbol{u})$ such that

$$\partial_t \varphi - \nabla \cdot (m \nabla \mu) = R \qquad \text{in} \quad \Omega \times [0, T], \tag{4}$$

$$\mu + \gamma \left(\ell \Delta \varphi - \frac{1}{\ell} \Psi'(\varphi) \right) - \delta_\varphi \mathcal{E}_{\text{e}}(\varphi, \boldsymbol{u}) = 0 \qquad \text{in} \quad \Omega \times [0, T], \tag{5}$$

$$-\nabla \cdot \left(\mathbb{C}(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right) \right) = \boldsymbol{f} \qquad \text{in} \quad \Omega \times [0, T], \tag{6}$$

with the boundary conditions $\nabla \varphi \cdot \boldsymbol{n} = \nabla \mu \cdot \boldsymbol{n} = 0$ and $\boldsymbol{u} = \boldsymbol{u}_b$ on $\partial \Omega \times [0, T]$, and initial condition $\varphi = \varphi_0$ in $\Omega \times \{0\}$. For completeness, we mention that

$$\delta_\varphi \mathcal{E}_{\text{e}}(\varphi, \boldsymbol{u}) = \frac{1}{2} \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right) : \mathbb{C}'(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right) - \xi \boldsymbol{I} : \mathbb{C}(\varphi) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right), \tag{7}$$

where the elasticity tensor $\mathbb{C}(\varphi)$ is depending on the phase-field through the interpolation function $\pi(\varphi)$; $\mathbb{C}(\varphi) = \mathbb{C}_{-1} + \pi(\varphi)(\mathbb{C}_1 - \mathbb{C}_{-1})$, and we assume for simplicity to have homogeneous Dirichlet boundary conditions for the elasticity subproblem, i.e., $\boldsymbol{u}_b = 0$.

2.2 Phase-field independent elasticity tensor

A simplified model is obtained in the special case of phase-field independent elasticity tensor $\mathbb{C}(\varphi) = \mathbb{C}$. We consider it as a special case here because it is a popular simplification to the system, and the analysis of it will make the foundation for the numerical solution strategies for the situations where the elasticity tensor depends on the phase-field. The system (4)–(6) now becomes: Find $(\varphi, \mu, \boldsymbol{u})$ such that

$$\partial_t \varphi - \nabla \cdot (m \nabla \mu) = R \quad \text{in} \quad \Omega \times [0,T], \tag{8}$$

$$\mu + \gamma \left(\ell \Delta \varphi - \frac{1}{\ell} \Psi'(\varphi) \right) + \xi \boldsymbol{I} : \mathbb{C} \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right) = 0 \quad \text{in} \quad \Omega \times [0,T], \tag{9}$$

$$-\nabla \cdot \left(\mathbb{C} \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi \varphi \boldsymbol{I} \right) \right) = \boldsymbol{f} \quad \text{in} \quad \Omega \times [0,T], \tag{10}$$

with the boundary conditions $\nabla \varphi \cdot \boldsymbol{n} = \nabla \mu \cdot \boldsymbol{n} = 0$ and $\boldsymbol{u} = 0$ on $\partial\Omega \times [0,T]$, and initial condition $\varphi = \varphi_0$ in $\Omega \times \{0\}$.

Remark 1. *Notice that the equations* (4) *and* (8) *imply that the total phase-field is balanced in time by the reaction term*

$$\partial_t \int_\Omega \varphi \, dx = \int_\Omega R \, dx \tag{11}$$

due to the homogeneous Neumann boundary conditions on μ.

2.3 Assumptions on material parameters

In this paper we will use the following assumptions on the model:

(A1) We require that the double-well potential has a convex-concave splitting

$$\Psi(\varphi) = \Psi_c(\varphi) - \Psi_e(\varphi),$$

where $\Psi_c(\varphi)$ and $\Psi_e(\varphi)$ are convex functions, and that the derivative of the convex part $\Psi_c'(\varphi)$ is Lipschitz continuous

$$\left(\Psi_c'(\varphi_1) - \Psi_c'(\varphi_2)\right)(\varphi_1 - \varphi_2) \leq L_{\Psi_c}(\varphi_1 - \varphi_2)^2, \quad \forall \varphi_1, \varphi_2 \in \mathbb{R},$$

with Lipschitz constant L_{Ψ_c}. The convex-concave splitting of the classical double-well potential does not satisfy this assumption, since the Lipschitz constant of the convex part is not bounded. To rectify this situation, we modify the double-well potential outside the interval $(-\theta, \theta)$, for some choice of $\theta > 1$, in the following way:

$$\Psi(\varphi) = \begin{cases} 2\left(\theta^2 - 1\right)\varphi^2 - \left(\theta^4 - 1\right), & \varphi \geq \theta, \\ (1 - \varphi^2)^2, & \varphi \in (-\theta, \theta), \\ 2\left(\theta^2 - 1\right)\varphi^2 - \left(\theta^4 - 1\right), & \varphi \leq -\theta, \end{cases}$$

which is split into the convex functions

$$\Psi_c(\varphi) = \begin{cases} 2\theta^2\varphi^2 - \left(\theta^4 - 1\right), & \varphi \geq \theta, \\ \varphi^4 + 1, & \varphi \in (-\theta, \theta), \\ 2\theta^2\varphi^2 - \left(\theta^4 - 1\right), & \varphi \leq -\theta, \end{cases}$$

and

$$\Psi_e(\varphi) = 2\varphi^2.$$

This modification ensures the uniformly bounded Lipschitz continuity of Ψ_c', with bound $L_{\Psi_c} = 2\theta^2$, without altering the solution to the problem, since the phase-field rarely takes values outside $[-1, 1]$.

(A2) There exist constants $c_{\mathbb{C}} > 0$ and $C_{\mathbb{C}} > 0$ such that

$$c_{\mathbb{C}} \|\boldsymbol{e}\|_{L^2(\Omega)}^2 \leq (\mathbb{C}(s)\boldsymbol{e}; \boldsymbol{e}) \leq C_{\mathbb{C}} \|\boldsymbol{e}\|_{L^2(\Omega)}^2 \tag{12}$$

for all symmetric second order tensor functions $\boldsymbol{e} \in L^2(\Omega)$ and scalar functions $s \in L^\infty(\Omega)$, where $(\cdot;\cdot)$ is the $L^2(\Omega)$ tensor inner-product. It follows that $(\boldsymbol{e}, \boldsymbol{w}) \mapsto (\mathbb{C}(s)\boldsymbol{e}; \boldsymbol{w})$ defines an inner-product on $L^2(\Omega)$, hence we have the Cauchy-Schwarz'-type inequality

$$(\mathbb{C}(s)\boldsymbol{e}; \boldsymbol{w}) \leq (\mathbb{C}(s)\boldsymbol{e}; \boldsymbol{e})^{\frac{1}{2}} \, (\mathbb{C}(s)\boldsymbol{w}; \boldsymbol{w})^{\frac{1}{2}}. \tag{13}$$

3 Numerical solution strategies for the Cahn-Larché equations

We now consider numerical solution strategies for the Cahn-Larché equations with the aim of establishing an efficient and robust solver. At first, in Section 3.2, a solution strategy for the system with phase-field independent elasticity tensor (8)–(10) is proposed. Then, in Section 3.3, the equations with phase-field dependent elasticity tensor (4)–(6) are considered.

3.1 Notation, variational system of equations and discrete function spaces

Throughout the paper $(\cdot,\cdot)$ will denote the $L^2(\Omega)$ inner product for scalar- and vector-valued functions, $\langle\cdot,\cdot\rangle$ is the duality pairing, and $\langle\cdot,\cdot\rangle_X$ represents specific inner products defined on the Hilbert space X. We consider the following continuous variational formulation of the system (4)–(6): Find $(\varphi,\mu,\boldsymbol{u}) \in H^1\left([0,T],H^1(\Omega)\right) \times L^2\left([0,T],H^1(\Omega)\right) \times L^2\left([0,T],\left(H_0^1(\Omega)\right)^d\right)$ such that

$$
\begin{aligned}
(\partial_t\varphi, q^\varphi) + (m\nabla\mu, \nabla q^\varphi) - (R, q^\varphi) &= 0 && (14)\\
(\mu, q^\mu) - \gamma\ell\,(\nabla\varphi, \nabla q^\mu) - \frac{\gamma}{\ell}\,(\Psi'(\varphi), q^\mu) - (\delta_\varphi\mathcal{E}_{\mathrm{e}}(\varphi,\boldsymbol{u}), q^\mu) &= 0 && (15)\\
(\mathbb{C}(\varphi)\,(\boldsymbol{\varepsilon}\,(\boldsymbol{u}) - \xi\varphi\boldsymbol{I})\,;\boldsymbol{\varepsilon}(\boldsymbol{v})) - (\boldsymbol{f},\boldsymbol{v}) &= 0, && (16)
\end{aligned}
$$

for all $(q^\varphi, q^\mu, \boldsymbol{v}) \in H^1(\Omega)\times H^1(\Omega)\times\left(H_0^1(\Omega)\right)^d$, and almost all $t\in[0,T]$.

As notation for the discrete equations, let τ be a uniform time-step size, defined by $\tau := \frac{T}{N}$, where N is the number of time steps. Moreover, the index n will refer to the time step, h the mesh diameter, and i the iteration number. Let $Q_h \subseteq H^1(\Omega)$ and $\boldsymbol{V}_h \subseteq \left(H_0^1(\Omega)\right)^d$ be conforming finite element function spaces, where Q_h is the solution space for phase-field and chemical potential, and $\boldsymbol{V}_h$ is the solution space for the displacement. Furthermore, we define $Q_{h,0} = \left\{q_h \in Q_h : \int_\Omega q_h\,dx = 0\right\}$, and consider the dual space of $(Q_{h,0}, \|\cdot\|_{h,m})$ where $\|q_h\|_{h,m} := \|m^{\frac{1}{2}}\nabla q_h\|_{L^2(\Omega)}$ as $Q_{h,m}^*$ with canonical dual norm $\|\cdot\|_{Q_{h,m}^*}$. Notice that the space $Q_{h,m}^*$ is a discrete superspace of $H^{-1}(\Omega)$.

Due to the Lax-Milgram lemma there exists a unique $v_h \in Q_{h,0}$ for all $s_h \in Q_{h,m}^*$ such that

$$
\langle s_h, q_h\rangle = (m\nabla v_h, \nabla q_h)\,, \quad \forall q_h \in Q_{h,0}. \tag{17}
$$

Thereby, we have

$$
\|s_h\|_{Q_{h,m}^*} := \sup_{\substack{q_h\in Q_{h,0}\\ \|q_h\|_{h,m}\neq 0}} \frac{\langle s_h, q_h\rangle}{\|q_h\|_{h,m}} = \sup_{\substack{q_h\in Q_{h,0}\\ \|q_h\|_{h,m}\neq 0}} \frac{(m\nabla v_h, \nabla q_h)}{\|m^{\frac{1}{2}}\nabla q_h\|_{L^2(\Omega)}} = \|m^{\frac{1}{2}}\nabla v_h\|_{L^2(\Omega)}, \tag{18}
$$

where v_h satisfies (17). Moreover, we identify the $Q_{h,m}^*$ inner-product for $s_h, l_h \in Q_{h,0}$ as

$$
\langle s_h, l_h\rangle_{Q_{h,m}^*} := (s_h, v_h) \tag{19}
$$

where $v_h \in Q_h$ is a solution to the variational equation

$$
(l_h, q_h) = (m\nabla v_h, \nabla q_h), \quad \forall q_h \in Q_{h,0}. \tag{20}
$$

We then have that

$$
\langle s_h, s_h\rangle_{Q_{h,m}^*}^{\frac{1}{2}} = (s_h, r_h)^{\frac{1}{2}} = (m\nabla r_h, \nabla r_h)^{\frac{1}{2}} = \|m^{\frac{1}{2}}\nabla r_h\|_{L^2(\Omega)} = \|s_h\|_{Q_{h,m}^*} \tag{21}
$$

where $r_h \in Q_h$ satisfies $(s_h, q_h) = (m\nabla r_h, \nabla q_h)$ for all $q_h \in Q_{h,0}$.

Remark 2. *Notice that, as $l_h \in Q_{h,0}$, equation (20) holds for all $q_h \in Q_h$, and uniqueness of v_h can be imposed by prescribing its mean. Choosing different values for the mean of v_h does not alter the value of the inner-product (s_h, v_h) as $s_h \in Q_{h,0}$.*

3.2 Solution strategy for Cahn-Larché with phase-field-independent elasticity tensor

Here, we present a robust solution strategy for the Cahn-Larché equations in the special case where the elasticity tensor is independent of the phase-field, (8)–(10). First, we discretize the equations by the convex-concave splitting of the double-well potential (A1), i.e., we evaluate the convex part implicitly in time and the expansive part explicitly to make the discrete system more suitable for linearization techniques. Moreover, we show that the discrete system of equations are equivalent to a minimization problem and utilize its structure to show unconditional gradient stability of the discretization (the free energy of the system does not increase without the presence of external contributions). Then, we prove convergence of alternating minimization applied to the minimization problem.

3.2.1 Discrete system of equations

Using the convex-concave splitting method in time for the double-well potential (A1), and evaluating other terms implicitly, we get the discretized (in time and space) system of equations corresponding to (14)–(16) with

phase-field independent elasticity tensor as: Given $\varphi_h^{n-1} \in Q_h$, find $\varphi_h^n, \mu_h^n \in Q_h$ and $\boldsymbol{u}_h^n \in \boldsymbol{V}_h$, such that

$$\left(\frac{\varphi_h^n - \varphi_h^{n-1}}{\tau}, q_h^\varphi\right) + (m\nabla\mu_h^n, \nabla q_h^\varphi) - (R^n, q_h^\varphi) = 0 \tag{22}$$

$$(\mu_h^n, q_h^\mu) - \gamma\ell\,(\nabla\varphi_h^n, \nabla q_h^\mu) - \frac{\gamma}{\ell}\left(\Psi_c'(\varphi_h^n) - \Psi_e'(\varphi_h^{n-1}), q_h^\mu\right) + (\mathbb{C}\,(\boldsymbol{\varepsilon}\,(\boldsymbol{u}_h^n) - \xi\varphi_h^n\boldsymbol{I})\,;q_h^\mu\xi\boldsymbol{I}) = 0 \tag{23}$$

$$(\mathbb{C}\,(\boldsymbol{\varepsilon}\,(\boldsymbol{u}_h^n) - \xi\varphi_h^n\boldsymbol{I})\,;\boldsymbol{\varepsilon}(\boldsymbol{v}_h)) - (\boldsymbol{f}^n, \boldsymbol{v}_h) = 0, \tag{24}$$

for all $q_h^\varphi, q_h^\mu \in Q_h$, and all $\boldsymbol{v}_h \in \boldsymbol{V}_h$. Similar discretizations have been considered in [10] for a phase-field dependent elasticity tensor, and in [18] without a convex-concave splitting of the double-well potential.

Proposition 1. *The solution to the discrete problem* (22)–(24) *is equivalent to the solution of the minimization problem: Given* $\varphi_h^{n-1} \in Q_h$*, solve*

$$(\varphi_h^n, \boldsymbol{u}_h^n) = \underset{s_h \in \bar{Q}_h^n, \boldsymbol{w}_h \in \boldsymbol{V}_h}{\arg\min} \mathcal{H}_\tau^n(s_h, \boldsymbol{w}_h) \tag{25}$$

where the admissible space for the phase-field is defined as

$$\bar{Q}_h^n := \left\{ s_h \in Q_h \;\middle|\; \int_\Omega \frac{s_h - \varphi_h^{n-1}}{\tau}\,dx = \int_\Omega R^n\,dx \right\} \tag{26}$$

and

$$\mathcal{H}_\tau^n(s_h, \boldsymbol{w}_h) := \frac{\|s_h - \varphi_h^{n-1} - \tau R^n\|_{Q_{h,m}^*}^2}{2\tau} + \mathcal{E}_c(s_h, \boldsymbol{w}_h) - \frac{\gamma}{\ell}\left(\Psi_e'(\varphi_h^{n-1}), s_h\right) - (\boldsymbol{f}^n, \boldsymbol{w}_h)\,,$$

where

$$\mathcal{E}_c(s_h, \boldsymbol{w}_h) := \int_\Omega \frac{\gamma}{\ell}\Psi_c(s_h) + \gamma\ell\frac{|\nabla s_h|^2}{2} + \frac{1}{2}\left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h) - \xi s_h\boldsymbol{I}\right) : \mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h) - \xi s_h\boldsymbol{I}\right)\,dx.$$

Proof. We derive the optimality conditions of the minimization problem which are similar to (22)–(24), but over restricted spaces. By employing canonical extensions, we establish the equivalence. Let $\delta_\varphi\mathcal{H}_\tau^n$ and $\delta_{\boldsymbol{u}}\mathcal{H}_\tau^n$ represent the variational derivatives with respect to the first and second argument of the potential $\mathcal{H}_\tau^n$ respectively. Then the optimality conditions to the minimization problem (25) reads: Find $\varphi_h^n, \boldsymbol{u}_h^n \in \bar{Q}_h^n \times \boldsymbol{V}_h$ such that

$$0 = \langle\delta_\varphi\mathcal{H}_\tau^n(\varphi_h^n, \boldsymbol{u}_h^n), q_h\rangle = \left\langle \frac{\varphi_h^n - \varphi_h^{n-1}}{\tau} - R^n, q_h \right\rangle_{Q_{h,m}^*} + \left(\delta_\varphi\mathcal{E}_c(\varphi_h^n, \boldsymbol{u}_h^n) - \frac{\gamma}{\ell}\Psi_e'(\varphi_h^{n-1}), q_h\right) \tag{27}$$

$$0 = \langle\delta_{\boldsymbol{u}}\mathcal{H}_\tau^n(\varphi_h^n, \boldsymbol{u}_h^n), \boldsymbol{w}_h\rangle = \left(\delta_{\boldsymbol{\varepsilon}(\boldsymbol{u})}\mathcal{E}_c(\varphi_h^n, \boldsymbol{u}_h^n); \boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right) - (\boldsymbol{f}^n, \boldsymbol{w}_h)\,, \tag{28}$$

for all $q_h \in Q_{h,0}$ and $\boldsymbol{w}_h \in \boldsymbol{V}_h$ where

$$\delta_\varphi\mathcal{E}_c(\varphi_h^n, \boldsymbol{u}_h^n) = \frac{\gamma}{\ell}\Psi_c'(\varphi_h^n) - \gamma\ell\Delta\varphi_h^n - \xi\boldsymbol{I} : \mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n) - \xi\varphi_h^n\boldsymbol{I}\right)$$

and

$$\delta_{\boldsymbol{\varepsilon}(\boldsymbol{u})}\mathcal{E}_c(\varphi_h^n, \boldsymbol{u}_h^n) = \mathbb{C}\big(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n) - \xi\varphi_h^n\boldsymbol{I}\big).$$

Using the definition of $\langle\cdot,\cdot\rangle_{Q_{h,m}^*}$ equation (27) is equivalent to

$$0 = (-\mu_h^n, q_h) + \left(\delta_\varphi\mathcal{E}_c(\varphi_h^n, \boldsymbol{u}_h^n) - \frac{\gamma}{\ell}\Psi_e'(\varphi_h^{n-1}), q_h\right), \quad \forall q_h \in Q_{h,0} \tag{29}$$

where μ_h^n is the solution to the problem

$$-(m\nabla\mu_h^n, \nabla l_h) = \left(\frac{\varphi_h^n - \varphi_h^{n-1}}{\tau} - R^n, l_h\right), \quad \forall l_h \in Q_{h,0}, \tag{30}$$

with mean fixed as

$$\int_\Omega \mu_h^n\,dx = \int_\Omega \delta_\varphi\mathcal{E}_c(\varphi_h^n, \boldsymbol{u}_h^n) - \frac{\gamma}{\ell}\Psi_e'(\varphi_h^{n-1})\,dx, \tag{31}$$

in accordance with Remark 2. The constraint $\varphi_h^n \in \bar{Q}_h^n$ and (30) are equivalent to requiring that equality (30) holds for all $l_h \in Q_h$. Due to (31), equation (29) holds for all $q_h \in Q_h$, and we have that the solutions to (28), (29) and (30) are equivalent to the solutions of the discrete problem (22)–(24). □

Remark 3 (Affine structure of the admissible set)**.** *The admissible set for the phase-field in the optimization problem* (25)*,* $\bar{Q}_h^n$*, is an affine space. For any two* $s_h^1,\ s_h^2 \in \bar{Q}_h^n$ *it holds that* $s_h^1 - s_h^2 \in Q_{h,0}$*.*

Theorem 1. *The discretization scheme (22)–(24) is unconditionally gradient stable, i.e., the free energy*

$$\mathcal{E}(\varphi, \boldsymbol{u}) = \int_\Omega \gamma \left(\frac{1}{\ell}\Psi(\varphi) + \frac{\ell}{2}|\nabla\varphi|^2 \right) + \frac{1}{2}\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\varphi\boldsymbol{I}\right) : \mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \xi\varphi\boldsymbol{I}\right)\, dx$$

dissipates over the time-steps assuming the absence of external contributions ($R = 0$ and $\boldsymbol{f} = 0$).

Proof. Exploiting the equivalence between the discrete system of equations (22)–(24) and the minimization problem in Proposition 1, we get that

$$\mathcal{H}_\tau^n(\varphi_h^n, \boldsymbol{u}_h^n) - \mathcal{H}_\tau^n(\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}) \quad \leq \quad 0,$$

due to the fact that $\varphi^{n-1} \in \bar{Q}_h^n$ when $R = 0$. It follows that

$$\frac{\|\varphi_h^n - \varphi_h^{n-1}\|^2_{Q^*_{h,m}}}{2\tau} + \mathcal{E}_c(\varphi_h^n, \boldsymbol{u}_h^n) - \frac{\gamma}{\ell}\left(\Psi_e'(\varphi_h^{n-1}), \varphi_h^n\right) - \left[\mathcal{E}_c(\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}) - \frac{\gamma}{\ell}\left(\Psi_e'(\varphi_h^{n-1}), \varphi_h^{n-1}\right)\right] \quad \leq \quad 0,$$

and by rearrangement and application of the convexity of Ψ_e we get

$$\Psi_e(\varphi_h^n) - \Psi_e(\varphi_h^{n-1}) \geq \Psi_e'(\varphi_h^{n-1})(\varphi_h^n - \varphi_h^{n-1}).$$

Recalling that $\Psi(s) = \Psi_c(s) - \Psi_e(s)$ we get the inequality

$$\frac{\|\varphi_h^n - \varphi_h^{n-1}\|^2_{Q^*_{h,m}}}{2\tau} + \mathcal{E}(\varphi_h^n, \boldsymbol{u}_h^n) - \mathcal{E}(\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}) \leq 0.$$

Hence,

$$\mathcal{E}(\varphi_h^n, \boldsymbol{u}_h^n) \leq \mathcal{E}(\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1})$$

for all τ and n. □

3.2.2 Alternating minimization for the Cahn-Larché equations with phase-field-independent elasticity tensor

There exists several ways to solve the nonlinear discrete system of equations (22)–(24), and due to the convexity of the related minimization problem (see Proposition 1) we expect the Newton method to be a viable and efficient choice. However, we propose here to solve the system with an alternating minimization method. The main benefit of this is that it allows for the use of readily available solvers, as it corresponds to solving a Cahn-Hilliard equation and an elasticity equation subsequently. In each time step we initialize the solver with the solution at the previous time step

$$\varphi_h^{n,0} = \varphi_h^{n-1}, \quad \text{and} \quad \boldsymbol{u}_h^{n,0} = \boldsymbol{u}_h^{n-1},$$

and minimize the potential $\mathcal{H}_\tau^n$ sequentially

$$\varphi_h^{n,i} \quad = \quad \underset{s_h \in \bar{Q}_h^n}{\arg\min}\, \mathcal{H}_\tau^n(s_h, \boldsymbol{u}_h^{n,i-1}), \tag{32}$$

$$\boldsymbol{u}_h^{n,i} \quad = \quad \underset{\boldsymbol{w}_h \in \boldsymbol{V}_h}{\arg\min}\, \mathcal{H}_\tau^n(\varphi_h^{n,i}, \boldsymbol{w}_h) \tag{33}$$

where i is the iteration index. The corresponding variational system of equations in the i-th iteration reads: Given $(\varphi_h^{n-1}, \boldsymbol{u}_h^{n,i-1}) \in Q_h \times \boldsymbol{V}_h$, find $(\varphi_h^{n,i}, \mu_h^{n,i}, \boldsymbol{u}_h^{n,i}) \in Q_h \times Q_h \times \boldsymbol{V}_h$ such that

$$\left(\frac{\varphi_h^{n,i} - \varphi_h^{n-1}}{\tau}, q_h^\varphi\right) + \left(m\nabla\mu_h^{n,i}, \nabla q_h^\varphi\right) - \left(R^n, q_h^\varphi\right) \quad = \quad 0 \tag{34}$$

$$\begin{aligned}\left(\mu_h^{n,i}, q_h^\mu\right) - \gamma\ell\left(\nabla\varphi_h^{n,i}, \nabla q_h^\mu\right) - \frac{\gamma}{\ell}\left(\Psi_c'\left(\varphi_h^{n,i}\right) - \Psi_e'\left(\varphi_h^{n-1}\right), q_h^\mu\right) & \\ + \left(\mathbb{C}\left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i-1}\right) - \xi\varphi_h^{n,i}\boldsymbol{I}\right); q_h^\mu\xi\boldsymbol{I}\right) \quad &= \quad 0 \end{aligned} \tag{35}$$

$$\left(\mathbb{C}\left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right) - \xi\varphi_h^{n,i}\boldsymbol{I}\right); \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\right) - \left(\boldsymbol{f}^n, \boldsymbol{v}_h\right) \quad = \quad 0 \tag{36}$$

for all $(q_h^\varphi, q_h^\mu, \boldsymbol{v}_h) \in Q_h \times Q_h \times \boldsymbol{V}_h$. Here, the space Q_h appears in the discrete system instead of $\bar{Q}_h^n$ due to the same argumentation as in the proof of Proposition 1.

Remark 4. *The Cahn-Hilliard subsystem (34)–(35) is still nonlinear due to $\Psi_c'(\varphi^{n,i})$. In this work, we solve it with the Newton method which is known to converge for this problem [39].*

We apply the abstract theory available in [33] to prove that the alternating minimization algorithm converges and summarize the appropriate result as a lemma (using the notation of the present article):

Lemma 1. *Assume that there exist norms* $\|(\cdot,\cdot)\| : Q_{h,0} \times \boldsymbol{V}_h \to \mathbb{R}^+$, $\|\cdot\|_{\mathrm{ch}} : Q_{h,0} \to \mathbb{R}^+$ *and* $\|\cdot\|_{\mathrm{e}} : \boldsymbol{V}_h \to \mathbb{R}^+$, *related by the inequalities*

$$\|(s_h, \boldsymbol{w}_h)\|^2 \geq \beta_{\mathrm{ch}} \|s_h\|_{\mathrm{ch}}^2, \quad \text{and} \quad \|(s_h, \boldsymbol{w}_h)\|^2 \geq \beta_{\mathrm{e}} \|\boldsymbol{w}_h\|_{\mathrm{e}}^2, \quad \forall (s_h, \boldsymbol{w}_h) \in Q_{h,0} \times \boldsymbol{V}_h, \tag{37}$$

for some $\beta_{\mathrm{ch}}, \beta_{\mathrm{e}} \geq 0$, *and let the potential* $\mathcal{H} : \bar{Q}_h^n \times \boldsymbol{V}_h \to \mathbb{R}$ *be given. If*

- $\mathcal{H}$ *is convex with respect to the norm* $\|(\cdot,\cdot)\|$ *with convexity constant* $\sigma \geq 0$, *i.e.,*

$$\left\langle \delta\mathcal{H}\left(s_h^1, \boldsymbol{w}_h^1\right) - \delta\mathcal{H}\left(s_h^2, \boldsymbol{w}_h^2\right), \left(s_h^1 - s_h^2, \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2\right)\right\rangle \geq \sigma \left\| \left(s_h^1 - s_h^2, \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2\right)\right\|^2, \tag{38}$$

  for all $\left(s_h^1, s_h^2, \boldsymbol{w}_h^1, \boldsymbol{w}_h^2\right) \in \bar{Q}_h^n \times \bar{Q}_h^n \times \boldsymbol{V}_h \times \boldsymbol{V}_h$,

and

- *the variational derivatives of* $\mathcal{H}$ *with respect to the first and second arguments are Lipschitz continuous in the norm* $\|\cdot\|_{\mathrm{ch}}$ *with constant* L_{ch} *and* $\|\cdot\|_{\mathrm{e}}$ *with constant* L_{e}, *respectively, i.e., there exist* $L_{\mathrm{ch}} > 0$, $L_{\mathrm{e}} > 0$ *such that*

$$\left\langle \delta_\varphi \mathcal{H}\left(s_h^1, \boldsymbol{w}_h\right) - \delta_\varphi \mathcal{H}\left(s_h^2, \boldsymbol{w}_h\right), s_h^1 - s_h^2 \right\rangle \leq L_{\mathrm{ch}} \left\| s_h^1 - s_h^2 \right\|_{\mathrm{ch}}^2, \quad \forall \left(s_h^1, s_h^2, \boldsymbol{w}_h\right) \in \bar{Q}_h^n \times \bar{Q}_h^n \times \boldsymbol{V}_h, \tag{39}$$

  and

$$\left\langle \delta_{\boldsymbol{u}} \mathcal{H}\left(s_h, \boldsymbol{w}_h^1\right) - \delta_{\boldsymbol{u}} \mathcal{H}\left(s_h, \boldsymbol{w}_h^2\right), \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2 \right\rangle \leq L_{\mathrm{e}} \left\| \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2 \right\|_{\mathrm{e}}^2, \quad \forall \left(\boldsymbol{w}_h^1, \boldsymbol{w}_h^2, s_h\right) \in \boldsymbol{V}_h \times \boldsymbol{V}_h \times \bar{Q}_h^n, \tag{40}$$

then the alternating minimization scheme (as proposed in (32)–(33) *with* $\mathcal{H}_\tau^n = \mathcal{H}$*) converges in the sense that*

$$\mathcal{H}\left(\varphi_h^{n,i}, \boldsymbol{u}_h^{n,i}\right) - \mathcal{H}\left(\varphi_h^n, \boldsymbol{u}_h^n\right) \leq \left(1 - \frac{\sigma\beta_{\mathrm{ch}}}{L_{\mathrm{ch}}}\right)\left(1 - \frac{\sigma\beta_{\mathrm{e}}}{L_{\mathrm{e}}}\right)\left(\mathcal{H}\left(\varphi_h^{n,i-1}, \boldsymbol{u}_h^{n,i-1}\right) - \mathcal{H}\left(\varphi_h^n, \boldsymbol{u}_h^n\right)\right),$$

where $(\varphi_h^n, \boldsymbol{u}_h^n) \in \bar{Q}_h^n \times \boldsymbol{V}_h$ *is the minimizer of* $\mathcal{H}$.

Remark 5. *Notice that* $\frac{\sigma\beta_{\mathrm{ch}}}{L_{\mathrm{ch}}} \leq 1$ *and* $\frac{\sigma\beta_{\mathrm{e}}}{L_{\mathrm{e}}} \leq 1$ *due to* (37)–(40).

We are also going to take advantage of the following inverse inequality:

Lemma 2. *There exists a constant* $C_{\mathrm{inv}} > 0$ *such that*

$$C_{\mathrm{inv}} h^{-1} \|s_h\|_{Q_{m,h}^*} \geq \|s_h\|_{L^2(\Omega)},$$

for all $s_h \in Q_{h,0}$.

Proof. From standard finite element text books, e.g., Theorem 4.5.11 in [40], one can find the inverse inequality

$$\|s_h\|_{H^1(\Omega)} \leq \tilde{C} h^{-1} \|s_h\|_{L^2(\Omega)}, \tag{41}$$

for some $\tilde{C} > 0$. By the definition of the $Q_{h,m}^*$-norm (18) we have for $s_h \in Q_{h,0}$ and $\|s_h\|_{h,m} \neq 0$

$$\|s_h\|_{Q_{h,m}^*} \geq \frac{\langle s_h, s_h \rangle}{\|m^{\frac{1}{2}} \nabla s_h\|_{L^2(\Omega)}},$$

which implies

$$m^{\frac{1}{2}} \|s_h\|_{H^1(\Omega)} \|s_h\|_{Q_{m,h}^*} \geq \|s_h\|_{L^2(\Omega)}^2.$$

Using (41) we get by choosing $C_{\mathrm{inv}} = \tilde{C} m^{\frac{1}{2}}$ the desired inequality

$$C_{\mathrm{inv}} h^{-1} \|s_h\|_{L^2(\Omega)} \|s_h\|_{Q_{h,m}^*} \geq \|s_h\|_{L^2(\Omega)}^2.$$

□

Theorem 2. *The alternating minimization algorithm* (32)–(33) *converges linearly in the sense that*

$$\mathcal{H}_\tau^n\left(\varphi_h^{n,i}, \boldsymbol{u}_h^{n,i}\right) - \mathcal{H}_\tau^n\left(\varphi_h^n, \boldsymbol{u}_h^n\right) \leq \left(1 - \frac{\beta_{\mathrm{ch}}}{L_{\mathrm{ch}}}\right)(1 - \beta_{\mathrm{e}})\left(\mathcal{H}_\tau^n\left(\varphi_h^{n,i-1}, \boldsymbol{u}_h^{n,i-1}\right) - \mathcal{H}_\tau^n\left(\varphi_h^n, \boldsymbol{u}_h^n\right)\right), \tag{42}$$

where $\beta_{\mathrm{ch}} = \beta_{\mathrm{e}} = 1 - \left(\frac{h^2}{\tau C_{\mathrm{inv}}^2 \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}} + \frac{\gamma\ell}{C_\Omega^2 \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}} + 1\right)^{-1}$, *and* $L_{\mathrm{ch}} = 1 + L_\Psi \left(\frac{h^2}{\tau C_{\mathrm{inv}}^2} + \frac{\gamma\ell}{C_\Omega^2} + \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}\right)^{-1}$.

Proof. We apply Lemma 1. Let $\mathcal{H} = \mathcal{H}_\tau^n$, $\bar{Q}_h = \bar{Q}_h^n$, and define the norms

$$\begin{aligned}
\|(s_h, \boldsymbol{w}_h)\|^2 &:= \frac{\|s_h\|^2_{Q^*_{h,m}}}{\tau} + \gamma\ell\|\nabla s_h\|^2_{L^2(\Omega)} + \left(\mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h) - \xi s_h \boldsymbol{I}\right); \boldsymbol{\varepsilon}(\boldsymbol{w}_h) - \xi s_h \boldsymbol{I}\right), \\
\|s_h\|^2_{\mathrm{ch}} &:= \frac{\|s_h\|^2_{Q^*_{h,m}}}{\tau} + \gamma\ell\|\nabla s_h\|^2_{L^2(\Omega)} + \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}\|s_h\|^2_{L^2(\Omega)}, \\
\|\boldsymbol{w}_h\|^2_{\mathrm{e}} &:= \left(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{w}_h); \boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right),
\end{aligned}$$

for $(s_h, \boldsymbol{w}_h) \in Q_{h,0} \times \boldsymbol{V}_h$. Notice that $\|(\cdot,\cdot)\|$ and $\|\cdot\|_{\mathrm{e}}$ are norms due to (12).

Relation (37) *between norms.* We have that for $(s_h, \boldsymbol{w}_h) \in Q_{h,0} \times \boldsymbol{V}_h$

$$\begin{aligned}
\|(s_h, \boldsymbol{w}_h)\|^2 &= \frac{\|s_h\|^2_{Q^*_{h,m}}}{\tau} + \gamma\ell\|\nabla s_h\|^2_{L^2(\Omega)} + \left(\mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right); \boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right) \\
&\quad + \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}\|s_h\|^2_{L^2(\Omega)} - 2\left(\mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right); \xi s_h \boldsymbol{I}\right),
\end{aligned} \tag{43}$$

and by the Cauchy-Schwarz' inequality (13) and Young's inequality on the last term we obtain

$$\begin{aligned}
2\left(\mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right); \xi s_h \boldsymbol{I}\right) &\leq \delta\left(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{w}_h); \boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right) + \frac{k_1 \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}}{\delta}\|s_h\|^2_{L^2(\Omega)} \\
&\quad + \frac{k_2 \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}}{\delta}\|s_h\|^2_{L^2(\Omega)} + \frac{k_3 \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}}{\delta}\|s_h\|^2_{L^2(\Omega)}
\end{aligned}$$

where $1 \geq k_i \geq 0$, $k_1 + k_2 + k_3 = 1$, and $\delta > 0$ are free to be chosen. Using Lemma 2 and the Poincaré inequality, with constant C_Ω, we get

$$\begin{aligned}
2\left(\mathbb{C}\left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right); \xi s_h \boldsymbol{I}\right) &\leq \delta\left(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{w}_h); \boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right) + \frac{k_1 C_{\mathrm{inv}}^2 h^{-2} \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}}{\delta}\|s_h\|^2_{Q^*_{h,m}} \\
&\quad + \frac{k_2 C_\Omega^2 \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}}{\delta}\|\nabla s_h\|^2_{L^2(\Omega)} + \frac{k_3 \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}}{\delta}\|s_h\|^2_{L^2(\Omega)}.
\end{aligned}$$

Hence, we have from (43) that

$$\begin{aligned}
\|(s_h, \boldsymbol{w}_h)\|^2 &\geq (1-\delta)\left(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{w}_h); \boldsymbol{\varepsilon}(\boldsymbol{w}_h)\right) + \left(\frac{1}{\tau} - \frac{k_1 C_{\mathrm{inv}}^2 h^{-2} \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}}{\delta}\right)\|s_h\|^2_{Q^*_{h,m}} \\
&\quad + \left(\gamma\ell - \frac{k_2 C_\Omega^2 \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}}{\delta}\right)\|\nabla s_h\|^2_{L^2(\Omega)} + \left(1 - \frac{k_3}{\delta}\right)\xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I}\|s_h\|^2_{L^2(\Omega)}.
\end{aligned} \tag{44}$$

Choosing $\delta = 1$, $\beta_{\mathrm{ch}} = 1 - \left(\frac{h^2}{\tau C_{\mathrm{inv}}^2 \xi^2 \boldsymbol{I}:\mathbb{C}\boldsymbol{I}} + \frac{\gamma\ell}{C_\Omega^2 \xi^2 \boldsymbol{I}:\mathbb{C}\boldsymbol{I}} + 1\right)^{-1}$, $k_1 = (1-\beta_{\mathrm{ch}})\frac{h^2}{\tau C_{\mathrm{inv}}^2 \xi^2 \boldsymbol{I}:\mathbb{C}\boldsymbol{I}}$, $k_2 = (1-\beta_{\mathrm{ch}})\frac{\gamma\ell}{C_\Omega^2 \xi^2 \boldsymbol{I}:\mathbb{C}\boldsymbol{I}}$, and $k_3 = 1 - \beta_{\mathrm{ch}}$ we get the desired bound

$$\|(s_h, \boldsymbol{w}_h)\|^2 \geq \beta_{\mathrm{ch}}\|s_h\|^2_{\mathrm{ch}}, \quad \forall (s_h, \boldsymbol{w}_h) \in Q_{h,0} \times \boldsymbol{V}_h.$$

Choosing now $\delta = \left(\frac{h^2}{\tau C_{\mathrm{inv}}^2 \xi^2 \boldsymbol{I}:\mathbb{C}\boldsymbol{I}} + \frac{\gamma\ell}{C_\Omega^2 \xi^2 \boldsymbol{I}:\mathbb{C}\boldsymbol{I}} + 1\right)^{-1}$, $k_1 = \frac{\delta h^2}{\tau C_{\mathrm{inv}}^2 \xi^2 \boldsymbol{I}:\mathbb{C}\boldsymbol{I}}$, $k_2 = \frac{\gamma\ell\delta}{C_\Omega^2 \xi^2 \boldsymbol{I}:\mathbb{C}\boldsymbol{I}}$, $k_3 = \delta$, and $\beta_{\mathrm{e}} = 1 - \delta$ in equation (44) we obtain

$$\|(s_h, \boldsymbol{w}_h)\|^2 \geq \beta_{\mathrm{e}}\|\boldsymbol{w}_h\|^2_{\mathrm{e}}, \quad \forall (s_h, \boldsymbol{w}_h) \in Q_{h,0} \times \boldsymbol{V}_h.$$

Strong convexity. By assumption (A2)

$$\begin{aligned}
&\left\langle \delta\mathcal{H}_\tau^n(s_h^1, \boldsymbol{w}_h^1) - \delta\mathcal{H}_\tau^n(s_h^2, \boldsymbol{w}_h^2), \left(s_h^1 - s_h^2, \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2\right)\right\rangle \\
&= \left\langle \delta_\varphi\mathcal{H}_\tau^n(s_h^1, \boldsymbol{w}_h^1) - \delta_\varphi\mathcal{H}_\tau^n(s_h^2, \boldsymbol{w}_h^2), s_h^1 - s_h^2\right\rangle + \left\langle \delta_{\boldsymbol{u}}\mathcal{H}_\tau^n(s_h^1, \boldsymbol{w}_h^1) - \delta_{\boldsymbol{u}}\mathcal{H}_\tau^n(s_h^2, \boldsymbol{w}_h^2), \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2\right\rangle \\
&= \left\|\left(s_h^1 - s_h^2, \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2\right)\right\|^2 + \frac{\gamma}{\ell}\left(\Psi_c'(s_h^1) - \Psi_c'(s_h^2), s_h^1 - s_h^2\right) \\
&\geq \left\|\left(s_h^1 - s_h^2, \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2\right)\right\|^2,
\end{aligned}$$

for all $\left(s_h^1, s_h^2, \boldsymbol{w}_h^1, \boldsymbol{w}_h^2\right) \in \bar{Q}_h^n \times \bar{Q}_h^n \times \boldsymbol{V}_h \times \boldsymbol{V}_h$, we have that $\mathcal{H}_\tau^n(s_h, \boldsymbol{w}_h)$ is convex in $\|(s_h, \boldsymbol{w}_h)\|$ with convexity constant $\sigma = 1$.

Lipschitz continuity of the partial gradients. We have

$$\left\langle \delta_\varphi\mathcal{H}_\tau^n(s_h^1, \boldsymbol{w}_h) - \delta_\varphi\mathcal{H}_\tau^n(s_h^2, \boldsymbol{w}_h), s_h^1 - s_h^2\right\rangle = \|s_h^1 - s_h^2\|^2_{\mathrm{ch}} + \frac{\gamma}{\ell}\left(\Psi_c'(s_h^1) - \Psi_c'(s_h^2), s_h^1 - s_h^2\right)$$

for all $\left(s_h^1, s_h^2, \boldsymbol{w}_h\right) \in \bar{Q}_h^n \times \bar{Q}_h^n \times \boldsymbol{V}_h$. Assumption (A2) gives

$$\left(\Psi_c'(s_h^1) - \Psi_c'(s_h^2), s_h^1 - s_h^2\right) \leq L_{\Psi_c}\|s_h^1 - s_h^2\|^2_{L^2(\Omega)}$$

and by Lemma 2 we get

$$\left\langle \delta_\varphi \mathcal{H}_\tau^n(s_h^1, \boldsymbol{w}_h) - \delta_\varphi \mathcal{H}_\tau^n(s_h^2, \boldsymbol{w}_h), s_h^1 - s_h^2 \right\rangle \leq L_{\mathrm{ch}} \|s_h^1 - s_h^2\|_{\mathrm{ch}}^2,$$

where $L_{\mathrm{ch}} = 1 + L_\Psi \left(\frac{h^2}{\tau C_{\mathrm{inv}}^2} + \frac{\gamma \ell}{C_\Omega^2} + \xi^2 \boldsymbol{I} : \mathbb{C}\boldsymbol{I} \right)^{-1}$. Finally, $\delta_{\boldsymbol{u}} \mathcal{H}_\tau^n$ is Lipschitz continuous with respect to $\|\cdot\|_{\mathrm{e}}$ with constant $L_{\mathrm{e}} = 1$, since

$$\left\langle \delta_{\boldsymbol{u}} \mathcal{H}_\tau^n(s_h, \boldsymbol{w}_h^1) - \delta_{\boldsymbol{u}} \mathcal{H}_\tau^n(s_h, \boldsymbol{w}_h^2), \boldsymbol{w}_h^1 - \boldsymbol{w}_h^2 \right\rangle = \|\boldsymbol{w}_h^1 - \boldsymbol{w}_h^2\|_{\mathrm{e}}^2, \quad \forall \left(s_h, \boldsymbol{w}_h^1, \boldsymbol{w}_h^2\right) \in \bar{Q}_h^n \times \boldsymbol{V}_h, \times \boldsymbol{V}_h,$$

and the convergence result (42) is obtained through Lemma 1. □

3.3 Solution strategy for the Cahn-Larché equations with phase-field-dependent elasticity tensor

When the elasticity tensor depends on the phase-field, $\mathbb{C}(\varphi)$, the situation is slightly more involved because a naive implicit discretization, using the convex-concave splitting of the double-well potential Ψ leads to a discrete system that is related to a nonconvex minimization problem (similar treatment as in Proposition 1). It reads: Given $\varphi_h^{n-1} \in Q_h$, find $\varphi_h^n, \mu_h^n \in Q_h$ and $\boldsymbol{u}_h^n \in \boldsymbol{V}_h$, such that

$$\left(\frac{\varphi_h^n - \varphi_h^{n-1}}{\tau}, q_h^\varphi \right) + (m \nabla \mu_h^n, \nabla q_h^\varphi) - (R^n, q_h^\varphi) = 0, \tag{45}$$

$$(\mu_h^n, q_h^\mu) - \gamma \ell \, (\nabla \varphi_h^n, \nabla q_h^\mu) - \frac{\gamma}{\ell} \left(\Psi_c'(\varphi_h^n) - \Psi_e'(\varphi_h^{n-1}), q_h^\mu \right) - (\delta_\varphi \mathcal{E}_{\mathrm{e}}(\varphi_h^n, \boldsymbol{u}_h^n), q_h^\mu) = 0, \tag{46}$$

$$\left(\mathbb{C}\left(\varphi_h^n\right) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n) - \xi \varphi_h^n \boldsymbol{I} \right); \boldsymbol{\varepsilon}(\boldsymbol{v}_h) \right) - (\boldsymbol{f}^n, \boldsymbol{v}_h) = 0, \tag{47}$$

for all $(q_h^\varphi, q_h^\mu, \boldsymbol{v}_h) \in Q_h \times Q_h \times \boldsymbol{V}_h$ with $\delta_\varphi \mathcal{E}_{\mathrm{e}}(\varphi_h^n, \boldsymbol{u}_h^n)$ from (7). To mitigate the nonconvexity of the related minimization problem one could evaluate the entire term related to the elastic energy explicitly, $\delta_\varphi \mathcal{E}_{\mathrm{e}}(\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1})$. Then one could show, using the same technique as in Theorem 2 that an alternating minimization type method would converge. Instead, we propose a semi-implicit evaluation of the term $\delta_\varphi \mathcal{E}_{\mathrm{e}}(\cdot, \cdot)$, which corresponds to a convex minimization problem. The discretization reads: Given $\varphi_h^{n-1} \in Q_h$, find $\varphi_h^n, \mu_h^n \in Q_h$ and $\boldsymbol{u}_h^n \in V_h$, such that

$$\left(\frac{\varphi_h^n - \varphi_h^{n-1}}{\tau}, q_h^\varphi \right) + (m \nabla \mu_h^n, \nabla q_h^\varphi) - (R^n, q_h^\varphi) = 0, \tag{48}$$

$$(\mu_h^n, q_h^\mu) - \gamma \ell \, (\nabla \varphi_h^n, \nabla q_h^\mu) - \frac{\gamma}{\ell} \left(\Psi_c'(\varphi_h^n) - \Psi_e'(\varphi_h^{n-1}), q_h^\mu \right) - \left(\mathcal{E}_{\mathrm{e},\varphi}^{\mathrm{si}}(\varphi_h^n, \boldsymbol{u}_h^n; \varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}), q_h^\mu \right) = 0, \tag{49}$$

$$\left(\mathbb{C}\left(\varphi_h^{n-1}\right) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n) - \xi \varphi_h^n \boldsymbol{I} \right) \right); \boldsymbol{\varepsilon}(\boldsymbol{v}_h) \right) - (\boldsymbol{f}^n, \boldsymbol{v}_h) = 0, \tag{50}$$

for all $(q_h^\varphi, q_h^\mu, \boldsymbol{v}_h) \in Q_h \times Q_h \times \boldsymbol{V}_h$ where

$$\begin{aligned} \mathcal{E}_{\mathrm{e},\varphi}^{\mathrm{si}}(\varphi_h^n, \boldsymbol{u}_h^n; \varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}) := \; & \frac{1}{2} \left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n-1}\right) - \xi \varphi_h^{n-1} \boldsymbol{I} \right) \mathbb{C}'\left(\varphi_h^{n-1}\right) \left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n-1}\right) - \xi \varphi_h^{n-1} \boldsymbol{I} \right) \\ & - \xi \boldsymbol{I} : \mathbb{C}(\varphi_h^{n-1}) \left(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n) - \xi \varphi_h^n \boldsymbol{I} \right). \end{aligned}$$

Notice here, that

$$\delta_\varphi \mathcal{E}_{\mathrm{e}}(\varphi, \boldsymbol{u}) = \mathcal{E}_{\mathrm{e},\varphi}^{\mathrm{si}}(\varphi, \boldsymbol{u}, \varphi, \boldsymbol{u}).$$

Analogous to Proposition 1 we can prove that (48)–(50) is related to a minimization problem.

Proposition 2. *The solution to the discrete system of equation* (48)–(50) *are equivalent to the solution of the minimization problem: Given* $\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1} \in Q_h \times \boldsymbol{V}_h$ *solve*

$$(\varphi_h^n, \boldsymbol{u}_h^n) = \underset{s_h \in \bar{Q}_h^n, \boldsymbol{w}_h \in \boldsymbol{V}_h}{\arg\min} \; \mathcal{F}_\tau^n(s_h, \boldsymbol{w}_h) \tag{51}$$

for

$$\begin{aligned} \mathcal{F}_\tau^n(s_h, \boldsymbol{w}_h) := \; & \frac{\|s_h - \varphi_h^{n-1} - \tau R^n\|_{\bar{Q}_{h,m}^*}^2}{2\tau} + \mathcal{E}_c^{\mathrm{c}}(s_h, \boldsymbol{w}_h, \varphi_h^{n-1}) - \left(\mathcal{E}_{\mathrm{e}}^{\mathrm{e}}(\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}), s_h \right) \\ & - \frac{\gamma}{\ell} \left(\Psi_e'(\varphi_h^{n-1}), s_h \right) - (\boldsymbol{f}^n, \boldsymbol{w}_h), \end{aligned}$$

where

$$\mathcal{E}_c^c(s_h, \boldsymbol{w}_h, \varphi_h^{n-1}) := \int_\Omega \frac{\gamma}{\ell} \Psi_c(s_h) + \gamma \ell \frac{|\nabla s_h|^2}{2} + \frac{1}{2} \left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h) - \xi s_h \boldsymbol{I} \right) : \mathbb{C}\left(\varphi_h^{n-1}\right) \left(\boldsymbol{\varepsilon}(\boldsymbol{w}_h) - \xi s_h \boldsymbol{I} \right) \, dx,$$

and

$$\mathcal{E}_e^e(\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}) := \frac{1}{2} \left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n-1}\right) - \xi \varphi_h^{n-1} \boldsymbol{I} \right) \mathbb{C}'\left(\varphi_h^{n-1}\right) \left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n-1}\right) - \xi \varphi_h^{n-1} \boldsymbol{I} \right).$$

3.3.1 Alternating minimization for Cahn-Larché with phase-field-dependent elasticity tensor

Similarly to Section 3.2.2 we propose an alternating minimization algorithm, which again naturally is formulated as a block Gauss-Seidel method, to solve the discrete system of equations (48)–(50). Given $(\varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}, \boldsymbol{u}_h^{n,i-1}) \in Q_h \times \boldsymbol{V}_h \times \boldsymbol{V}_h$, find $(\varphi_h^{n,i}, \mu_h^{n,i}, \boldsymbol{u}_h^{n,i}) \in Q_h \times Q_h \times \boldsymbol{V}_h$ such that

$$\left(\frac{\varphi_h^{n,i} - \varphi_h^{n-1}}{\tau}, q_h^{\varphi}\right) + \left(m\nabla\mu_h^{n,i}, \nabla q_h^{\varphi}\right) - \left(R^n, q_h^{\varphi}\right) = 0, \tag{52}$$

$$\begin{aligned}\left(\mu_h^{n,i}, q_h^{\mu}\right) - \gamma\ell\left(\nabla\varphi_h^{n,i}, \nabla q_h^{\mu}\right) - \frac{\gamma}{\ell}\left(\Psi_c'\left(\varphi_h^{n,i}\right) - \Psi_e'\left(\varphi_h^{n-1}\right), q_h^{\mu}\right) \\ + \left(\mathcal{E}_{\mathrm{e},\varphi}^{\mathrm{si}}\left(\varphi_h^{n,i}, \boldsymbol{u}_h^{n,i-1}, \varphi_h^{n-1}, \boldsymbol{u}_h^{n-1}\right), q_h^{\mu}\right) = 0,\end{aligned} \tag{53}$$

$$\left(\mathbb{C}\left(\varphi_h^{n-1}\right)\left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}_h^{n,i}\right) - \xi\varphi_h^{n,i}\boldsymbol{I}\right); \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\right) - \left(\boldsymbol{f}^n, \boldsymbol{v}_h\right) = 0, \tag{54}$$

for all $(q_h^{\varphi}, q_h^{\mu}, \boldsymbol{v}_h) \in Q_h \times Q_h \times \boldsymbol{V}_h$.

Corollary 1. *The alternating minimization decoupling scheme* (52)–(54) *converges in each time-step n, with convergence rate*

$$\mathcal{F}_\tau^n\left(\varphi_h^{n,i}, \boldsymbol{u}_h^{n,i}\right) - \mathcal{F}_\tau^n\left(\varphi_h^n, \boldsymbol{u}_h^n\right) \leq \left(1 - \frac{\beta_{\mathrm{ch}}}{L_{\mathrm{ch}}}\right)(1 - \beta_{\mathrm{e}})\left(\mathcal{F}_\tau^n\left(\varphi_h^{n,i-1}, \boldsymbol{u}_h^{n,i-1}\right) - \mathcal{F}_\tau^n\left(\varphi_h^n, \boldsymbol{u}_h^n\right)\right), \tag{55}$$

where $\beta_{\mathrm{ch}} = \beta_{\mathrm{e}} = 1 - \left(\frac{h^2}{\tau C_{\mathrm{inv}}^2 \xi^2 \boldsymbol{I}:\boldsymbol{I} C_{\mathbb{C}}} + \frac{\gamma\ell}{C_\Omega^2 \xi^2 \boldsymbol{I}:\boldsymbol{I} C_{\mathbb{C}}} + 1\right)^{-1}$, *and* $L_{\mathrm{ch}} = 1 + L_\Psi\left(\frac{h^2}{\tau C_{\mathrm{inv}}^2} + \frac{\gamma\ell}{C_\Omega^2} + \xi^2 \boldsymbol{I} : \boldsymbol{I} c_{\mathbb{C}}\right)^{-1}$.

Proof. This proof is analogous that of Theorem 2. Simply replace $\mathbb{C}$ with $\mathbb{C}(\varphi_h^{n-1})$ and apply the bounds from assumption (A2). □

Remark 6. *Notice that as the discrete system of equations* (48)–(50) *corresponds to a convex minimization problem, we also expect a Newton-type solver to be rather robust, and have a higher convergence rate than the alternating minimization method.*

4 Numerical experiments

In this section, we present experiments to numerically investigate the performance and robustness of both the Newton method and alternating minimization applied to the semi-implicit time-discretized Cahn-Larché equations (48)–(50) compared with applying them to the implicit-in-time discretizaton (45)–(47). In all numerical experiments, the unit square in two spatial dimensions with a quadrilateral mesh is considered, and we apply bilinear conforming finite elements to all subproblems; phase-field, potential, and displacement.

When the elasticity tensor depends on the phase-field it is through the C^1 interpolation function

$$\pi(\varphi) = \begin{cases} 0, & \varphi < -1 \\ \frac{1}{4}\left(-\varphi^3 + 3\varphi + 2\right), & \varphi \in [-1, 1] \\ 1, & \varphi > 1 \end{cases}, \tag{56}$$

and the relation $\mathbb{C}(\varphi) = \mathbb{C}_{-1} + \pi(\varphi)\left(\mathbb{C}_1 - \mathbb{C}_{-1}\right)$, where $\mathbb{C}_{-1}$ and $\mathbb{C}_1$ are the elasticity tensors corresponding to the pure phases at $\varphi = -1$ and $\varphi = 1$, respectively.

Four different solution strategies to the Cahn-Larché equations are tested. For the discrete system (45)–(47) we test both the monolithic Newton method (marked by "Imp. Mono." in figure legends) and a staggered solution scheme, solving the Cahn-Hilliard subsystem (45)–(46) and the elasticity subsystem (47) sequentially (marked by "Imp. Split." in figure legends). The same is done for the discrete system (48)–(50) and mark the monolithic Newton method as "Semi-Imp. Mono." and the alternating minimization method (52)–(54) as "Semi-Imp. Split.". For both the monolithic and the decoupling solvers, the iterative procedures are terminated when the absolute and relative residuals and increments (iteration $i-1$ subtracted from iteration i), in the $L^2(\Omega)$-norm, reach a prescribed tolerance, i.e.,

$$\left\|\mathrm{Res}\left(\varphi_h^{n,i}, \mu_h^{n,i}, \boldsymbol{u}_h^{n,i}\right)\right\|_2 \leq \mathrm{Tol}_{\mathrm{res,abs}},$$

$$\frac{\left\|\mathrm{Res}\left(\varphi_h^{n,i}, \mu_h^{n,i}, \boldsymbol{u}_h^{n,i}\right)\right\|_2}{\left\|\mathrm{Res}\left(\varphi_h^{n,0}, \mu_h^{n,0}, \boldsymbol{u}_h^{n,0}\right)\right\|_2} \leq \mathrm{Tol}_{\mathrm{res,rel}},$$

$$\left\|\varphi_h^{n,i} - \varphi_h^{n,i-1}\right\|_{L^2(\Omega)} + \left\|\mu_h^{n,i} - \mu_h^{n,i-1}\right\|_{L^2(\Omega)} + \left\|\boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^{n,i-1}\right\|_{L^2(\Omega)} \leq \mathrm{Tol}_{\mathrm{inc,abs}},$$

$$\frac{\left\|\varphi_h^{n,i} - \varphi_h^{n,i-1}\right\|_{L^2(\Omega)}}{\left\|\varphi_h^{n,1} - \varphi_h^{n,0}\right\|_{L^2(\Omega)}} + \frac{\left\|\mu_h^{n,i} - \mu_h^{n,i-1}\right\|_{L^2(\Omega)}}{\left\|\mu_h^{n,1} - \mu_h^{n,0}\right\|_{L^2(\Omega)}} + \frac{\left\|\boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^{n,i-1}\right\|_{L^2(\Omega)}}{\left\|\boldsymbol{u}_h^{n,1} - \boldsymbol{u}_h^{n,0}\right\|_{L^2(\Omega)}} \leq \mathrm{Tol}_{\mathrm{inc,rel}},$$

where $\text{Res}\left(\varphi_h^{n,i}, \mu_h^{n,i}, \boldsymbol{u}_h^{n,i}\right)$ is the algebraic residual corresponding to the discretized system of equations. For all test cases that we run in this paper, $\text{Tol}_{\text{res,abs}}$, $\text{Tol}_{\text{res,rel}}$, $\text{Tol}_{\text{inc,abs}}$, and $\text{Tol}_{\text{inc,rel}}$ are set to $1e-6$. Moreover, the parameter θ in the modification to the standard double-well potential and the related convex-concave splitting, see Assumption (A1), is chosen as $\theta = 2$.

Remark 7. *The Cahn-Hilliard subproblem is nonlinear even though the alternating minimization method is applied. We use the Newton method and iterate until similar tolerances as for the full problem are reached (1e−6). One could, however, consider to only perform a single iteration of the Newton method in each alternating minimization iteration instead of iterating until the prescribed tolerance is reached, as done in [21], in order to speed up the convergence of the total iterative solver.*

4.1 Test case with phases separated along the middle

In this test case we initialize the simulation by separating the phases along the middle of the domain, see Figure 1a. We take $\boldsymbol{u}_h^{0,0} = 0$ as initial guess for displacement in the first time step and impose zero Dirichlet boundary conditions for it on the entire boundary. The model parameters can be found in Table 1, with

$$\mathbb{C}_{-1} = \begin{pmatrix} 100 & 20 & 0 \\ 20 & 100 & 0 \\ 0 & 0 & 200 \end{pmatrix}, \quad \text{and} \quad \mathbb{C}_1 = \begin{pmatrix} 1 & 0.1 & 0 \\ 0.1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

where the elasticity tensors are given in Voigt notation. First, we test with different values for the interfacial tension $\gamma = 1, 5, 10, 50, 100$, and then for different values of the swelling parameter $\xi = 0.1, 0.5, 1, 1.5, 2$. Simulation results for different values of γ are plotted in Figure 1a–1d ($\gamma = 5$), Figure 1e–1h ($\gamma = 10$), and Figure 1i–1l ($\gamma = 100$). Moreover, in Figure 1m we see that the energy decays over time, using both the semi-implicit time discretization (48)–(50), and the implicit one (45)–(47), for different time-step sizes, $\gamma = 5$ and $\xi = 1$.

Parameter name	Symbol	Value	Unit
Chemical mobility	m	1	$\left[\frac{L^4}{FT}\right]$
Interfacial tension	γ	–	$[F]$
Time step size	τ	1e-5	$[T]$
Final time	T	0.01	$[T]$
Swelling parameter	ξ	–	$[-]$
Mesh diameter	h	$\frac{\sqrt{2}}{65}$	$[L]$
Regularization parameter	ℓ	0.02	$[-]$
Elasticity tensors	$\mathbb{C}_{-1}, \mathbb{C}_1$	-	$\left[\frac{F}{L^2}\right]$

Table 1: Table of simulation parameters. Here, L denotes the unit of length, F force, and T time.

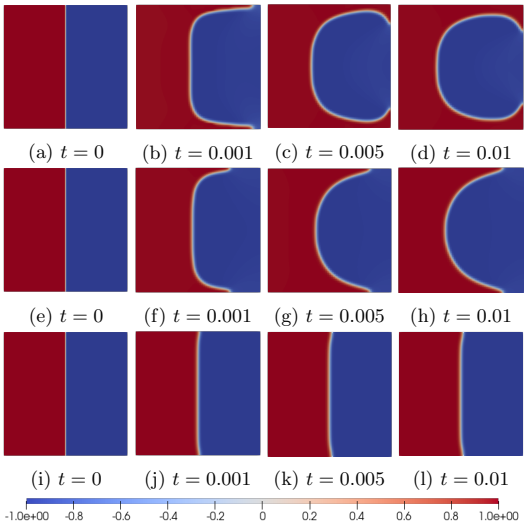
4.1.1 Dependence on interfacial tension

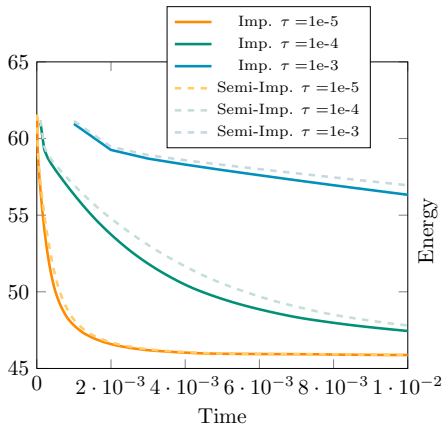
We run several simulations with different values for the interfacial tension $\gamma = 1, 5, 10, 50, 100$, while counting the number of iterations the different solution strategies take to achieve satisfactory precision. The other parameters are found in Table 1, and the swelling parameter is chosen to be $\xi = 1$.

In Figure 2, we see that the monolithic Newton method converges in fewer iterations than the alternating minimization algorithms. However, for the smallest value of interfacial tension, $\gamma = 1$, (when the coupling strength is highest) the monolithic Newton method with implicit-in-time evaluation of the elastic energy (45)–(47) does not converge at all, and is therefore not a robust choice as a solution strategy. The monolithic Newton method applied to the semi-implicitly discretized system of equations (48)–(50) seems to be a robust choice of linearization procedure, which is due to the convex nature of the related minimization problem, see Proposition 1. Moreover, as expected from Corollary 1, the number of iterations for the alternating minimization method (52)–(54) decreases with increasing interfacial tension. This is in fact true for all of the solution strategies as the relative coupling strength between Cahn-Hilliard and elasticity is decreasing for increasing interfacial tension.

4.1.2 Dependence on swelling parameter

A similar test is considered for several values of the swelling parameter, $\xi = 0.01, 0.1, 0.5, 1, 1.5, 2$ and a fixed interfacial tension $\gamma = 5$, see Figure 3. Here, we observe, as is expected from the theory, Corollary 1, that the coupled problems become more difficult to solve (require more iterations of either the Newton method or alternating minimization) when the swelling parameter increases. This is natural as the swelling parameter is directly connected to the coupling strength between the phase-field and elasticity equations. Another important observation is that for large values of the swelling parameter ($\xi = 1.5$ and $\xi = 2$) the monolithic Newton method applied to the discrete system of equations (45)–(47) does not converge at all. On the other hand, alternating minimization converges for these cases as well, which (although we have no theoretical proof for it) suggests that





(m) Energy decay over time for different time-step sizes, $\gamma = 5$ and $\xi = 1$.

Figure 1: (a) – (l): the solution at time t for the phase-field φ. (a) – (d): $\gamma = 5$, (e) – (h): $\gamma = 10$, (i) – (l): $\gamma = 100$. (m): Total energy (1) for both the implicit (in the elastic energy) time discretization (45)–(47) and the semi-implicit one (48)–(50) with different time step sizes and $\gamma = 10$.

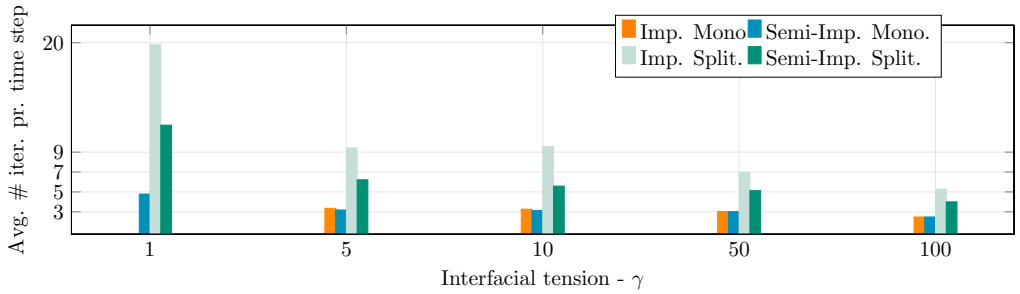


Figure 2: Test case with phases separated along the middle: Total number of iterations for different values of the interfacial tension parameter γ. Here, "Imp." refers to the discrete system of equation (45)–(47), whereas "Semi.-Imp." corresponds to the discrete system of equations (48)–(50). Moreover, "Mono." refers to the monolithic full Newton method applied to the discrete system of equations and the alternating minimization algorithm is labeled with "Split.". The numerical scheme (52)–(54) corresponds to "Semi-Imp. Split.". Notice that "Imp. Mono." failed to converge for $\gamma = 1$ and, therefore, it is not marked above that value in the plot.

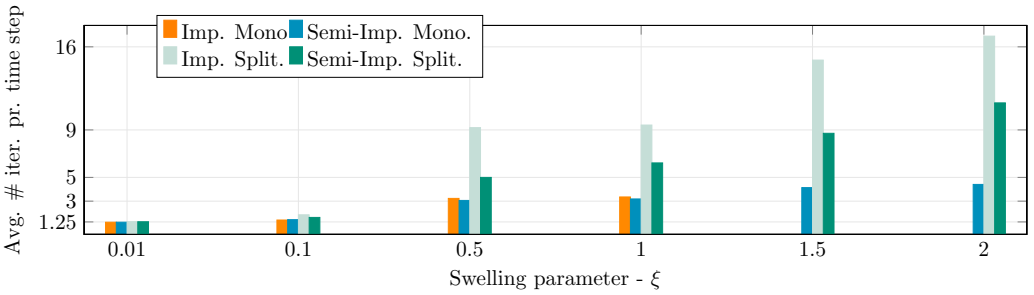


Figure 3: Test case with phases separated along the middle: Total number of iterations for different values of the swelling parameter ξ. Here, "Imp." refers to the discrete system of equation (45)–(47), whereas "Semi.-Imp." corresponds to the discrete system of equations (48)–(50). Moreover, "Mono." refers to the monolithic full Newton method applied to the discrete system of equations and the alternating minimization algorithm is labeled with "Split.". The numerical scheme (52)–(54) corresponds to "Semi-Imp. Split.". Notice that "Imp. Mono." failed to converge for $\xi = 1.5$ and $\xi = 2$ and, therefore, it is not marked above those values in the plot.

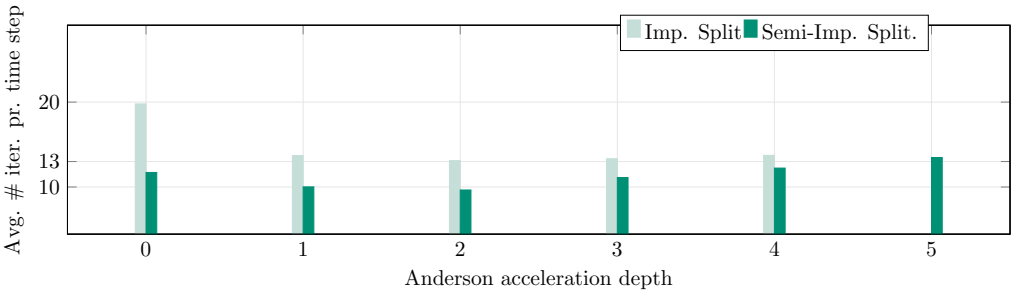


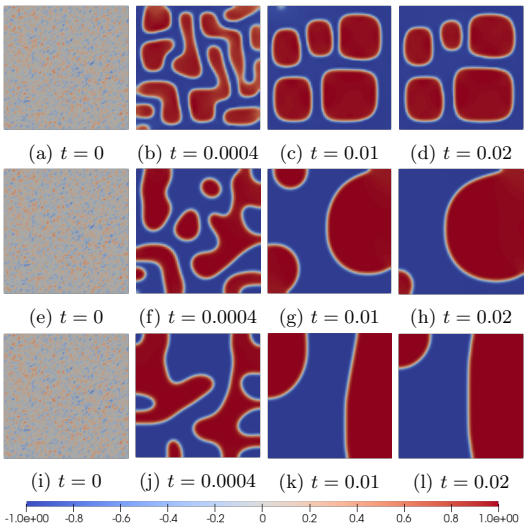
Figure 4: Test case with phases segregated in the middle: Total number of iterations for different Anderson acceleration depths. Here, "Imp." refers to the discrete system of equation (45)–(47), whereas "Semi.-Imp." corresponds to the discrete system of equations (48)–(50). Notice that "Semi-Imp. Split." failed to converge for depth 5 and, therefore, it is not marked above that value in the plot.

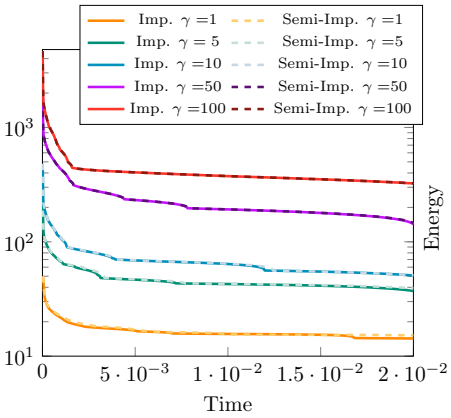
the alternating minimization method is more robust than the Newton method for this problem. Notice also that for the smallest value of swelling parameter $\xi = 0.01$, the problem is almost decoupled, and convergence of the linearization/decoupling methods is reached in approximately one iteration (in some time-steps two iterations are required).

4.1.3 Anderson acceleration applied to the decoupling algorithms

As mentioned in the introduction, the Anderson acceleration [34] has been successfully applied to accelerate decoupling/splitting schemes, as alternating minimization previously [25, 22], or linearly convergence methods like the Picard algorithm for Navier-Stokes [41]. The scheme is applied as a post-process to fixed-point iterations and updates the current iterate as a linear combination of the m (called depth of the acceleration) previous iterates. More careful explanation of the method can be found in e.g., [25, 22, 41].

Here we applied the Anderson acceleration to accelerate the alternating minimization method (52)–(54) ("Semi-Imp. Split."), and the staggered scheme applied to (45)–(47) ("Imp. Split."). Simulation parameters from Table 1 with $\gamma = 1$ and $\xi = 1$ are used, similar to the first column in Figure 2, and we test for acceleration depths ranging from $m = 0$ (no acceleration) to $m = 5$. The results are displayed in Figure 4. We observe that for the staggered scheme applied to (45)–(47) ("Imp. Split."), the postprocessing accelerates the convergence quite significantly, however, it fails to converge for the largest depth ($m = 5$). For the the alternating minimization method (52)–(54) ("Semi-Imp. Split."), it only accelerates slightly, and actually decelerates the convergence for larger values of depths ($m = 4, 5$). Therefore, using the Anderson acceleration to solve the alternating minimization problem might be beneficial for smaller depths. Moreover, there are several ways of improving the convergence of the Anderson acceleration, e.g, periodically restart it from depth $m = 0$ or turn it on and off using some safeguard mechanics (see [25]), but this is outside the scope of the current paper to investigate.





(m) Energy (1) decay over time for both the discrete method (45)–(47) ("Imp."), and (48)–(50) ("Semi-Imp.") for different values of interfacial tension parameter. Notice the logarithmic scale of the y-axis.

Figure 5: (a) – (l): the solution at time t for the phase-field φ. (a) – (d): $\gamma = 5$, (e) – (h): $\gamma = 10$, (i) – (l): $\gamma = 100$. (m): Total energy (1) for both the implicit (in the elastic energy) time discretization (45)–(47) and the semi-implicit one (48)–(50) with different time step sizes and $\gamma = 10$.

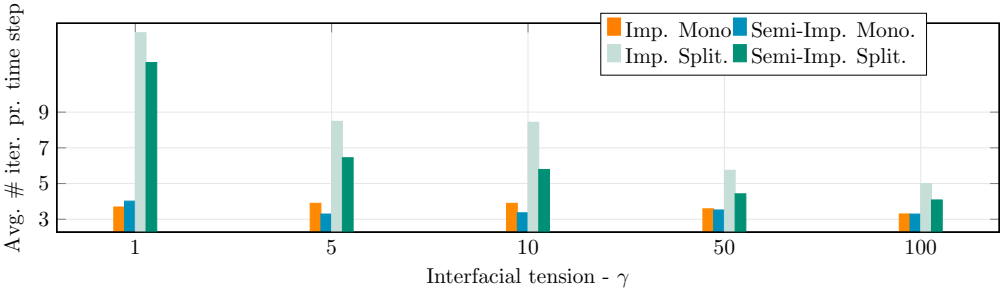


Figure 6: Test case with random initial data: Total number of iterations for different values of the interfacial tension parameter γ. Here, "Imp." refers to the discrete system of equation (45)–(47), whereas "Semi.-Imp." corresponds to the discrete system of equations (48)–(50). Moreover, "Mono." refers to the monolithic full Newton method applied to the discrete system of equations and the alternating minimization algorithm is labeled with "Split.". The numerical scheme (52)–(54) corresponds to "Semi-Imp. Split.".

4.2 Random initial conditions: Spinodal decomposition

We provide another numerical experiment here, with randomized initial conditions, where the initial "mixture" decomposes into pure phases and we observe a coarsening effect that resembles spinodal decomposition. This effect has been studied for the Cahn-Larché equations previously in e.g., [10, 19]. In Figure 5 we present simulation results using parameters from Table 1, $\xi = 1$ and $\gamma = 5$, $\gamma = 10$, and $\gamma = 100$. In Figure 5m, we plot the total energy (1) of the system for both the discrete system of equations (45)–(47) ("Imp.") and (48)–(50) ("Semi-Imp.") for different values of the interfacial tension parameter. We observe that there is close to no difference between the free energy over the simulation for the two time-discretizations and that both of them are decreasing over time.

In Figure 6, the total number of iterations for the different solution strategies are presented for different values of the interfacial tension $\gamma = 1, 5, 10, 50, 100$. We see that, as in Section 4.1.1, the number of decoupling/linearization iterations decrease for increasing values of the interfacial tension, exactly as the theory for alternating minimization predicts, Corollary 1. Again the Newton method outperforms the alternating minimization method in terms of numbers of iterations, although the difference shrinks significantly for lower relative coupling strengths (γ increasing). Moreover, we stress that the alternating minimization method has the added benefit of allowing for the use of readily available implementations and solvers for Cahn-Hilliard and elasticity with only small modifications.

5 Conclusions

In this paper, we proposed a semi-implicit time discretization to the Cahn-Larché equations and showed that it is equivalent to a convex minimization problem. Then convergence of alternating minimization applied to this problem was proved, and several numerical experiments to study its convergence properties in comparison to the monolithic Newton method were provided. Additionally, the alternating minimization (splitting method) and the monolithic Newton method applied to the newly proposed semi-implicit time-discretization were compared numerically to the same iterative methods applied to a more standard choice of time-discretization with implicit-in-time evaluations of the elastic contributions and a convex-concave split of the double-well potential. We observed that the convergence properties of the iterative methods (Newton's method and alternating minimization) applied to the newly proposed time-discretization are superior to those that are applied to the standard discretization, and in several cases we get convergence of the Newton method for the former and not for the latter. Moreover, for the special case of phase-field-independent elasticity tensor we proved that the discretization is unconditionally gradient stable, by exploiting its minimization structure. For the phase-field dependent elasticity tensor, numerical experiments show that the free energy of the system decreases over time. The newly proposed time-discretization is shown to be well suited for iterative solution schemes and provides a needed alternative to the standard implicit methods.

Acknowledgments

The work has been partly supported by the Centre for Sustainable Subsurface Resources, funded by the Norwegian Research council, as well as the FracFlow project funded by Equinor, Norway through Akademiaavtalen.

References

[1] JW Cahn and JE Hilliard. Free energy of a nonuniform system. I. Interfacial free energy. *J. Chem. Phys.*, 28(2):258–267, 1958.

[2] JW Cahn. Free energy of a nonuniform system. II. thermodynamic basis. *J. Chem. Phys.*, 30(5):1121–1124, 1959.

[3] FC Larché and JW Cahn. A linear theory of thermochemical equilibrium of solids under stress. *Acta. Metall.*, 21(8):1051–1063, 1973.

[4] FC Larché and JW Cahn. The effect of self-stress on diffusion in solids. *Acta. Metall.*, 30(10):1835–1845, 1982.

[5] S Shi, J Markmann, and J Weissmüller. Verifying Larché–Cahn elasticity, a milestone of 20th-century thermodynamics. *P. Natl. A. Sci.*, 115(43):10914–10919, 2018.

[6] H Garcke, KF Lam, and A Signori. On a phase field model of Cahn–Hilliard type for tumour growth with mechanical effects. *Nonlinear Anal-Real*, 57:103192, 2021.

[7] H Garcke, KF Lam, and A Signori. Sparse optimal control of a phase field tumor model with mechanical effects. *SIAM. J. Control. Optim.*, 59(2):1555–1580, 2021.

[8] M Fritz, C Kuttler, ML Rajendran, L Scarabosio, and B Wohlmuth. On a subdiffusive tumour growth model with fractional time derivative. *IMA J. Appl. Math.*, 86:688 – 729, 2021.

[9] W Dreyer and WH Müller. Modeling diffusional coarsening in eutectic tin/lead solders: a quantitative approach. *Int. J. Solids. Struct.*, 38(8):1433–1458, 2001.

[10] C Gräser, R Kornhuber, and U Sack. Numerical simulation of coarsening in binary solder alloys. *Comp. Mater. Sci.*, 93:221–233, 2014.

[11] E Meca, A Münch, and B Wagner. Sharp-interface formation during lithium intercalation into silicon. *E. J. Appl. Math.*, 29(1):118–145, 2018.

[12] L Cueto-Felgueroso and R Juanes. A phase field model of unsaturated flow. *Water Resour. Res.*, 45(10), 2009.

[13] E Bonetti, P Colli, W Dreyer, G Gilardi, G Schimperna, and J Sprekels. On a model for phase separation in binary alloys driven by mechanical effects. *Physica D.*, 165(1-2):48–65, 2002.

[14] H Garcke. On Cahn–Hilliard systems with elasticity. *P. Roy. Soc. Edinb. A.*, 133(2):307, 2003.

[15] H Abels and S Schaubeck. Sharp interface limit for the Cahn–Larché system. *Asymptotic Anal.*, 91(3-4):283–340, 2015.

[16] H Garcke and DJC Kwak. On asymptotic limits of cahn-hilliard systems with elastic misfit. In *Analysis, modeling and simulation of multiscale problems*, pages 87–111. Springer, 2006.

[17] WM Feng, P Yu, Shenyang Y Hu, Zi-Kui Liu, Q Du, and LQ Chen. A fourier spectral moving mesh method for the Cahn-Hilliard equation with elasticity. *Commun. Comput. Phys*, 5(2-4):582–599, 2009.

[18] H Garcke and U Weikard. Numerical approximation of the Cahn-Larché equation. *Numer. Math.*, 100(4):639–662, 2005.

[19] H Garcke, M Rumpf, and U Weikard. The Cahn-Hilliard equation with elasticity-finite element approximation and qualitative studies. *Interface. Free. Bound.*, 3(1):101–118, 2001.

[20] DJ Eyre. Unconditionally gradient stable time marching the Cahn-Hilliard equation. *Mater. Res. Soc. Symp. Proc.*, 529, 1998.

[21] D. Illiano, IS Pop, and FA Radu. Iterative schemes for surfactant transport in porous media. *Computat. Geosci.*, 25(2):805–822, 2021.

[22] JW Both, K Kumar, JM Nordbotten, and FA Radu. Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media. *Comput. Math. Appl.*, 77(6):1479–1502, 2019.

[23] T Gerasimov and L De Lorenzis. A line search assisted monolithic approach for phase-field computing of brittle fracture. *Comput. Method. Appl. M.*, 312:276–303, 2016.

[24] P Farrell and C Maurini. Linear and nonlinear solvers for variational phase-field models of brittle fracture. *Int. J. Numer. Meth. Eng.*, 109(5):648–667, 2017.

[25] E Storvik, JW Both, JM Sargado, JM Nordbotten, and FA Radu. An accelerated staggered scheme for variational phase-field models of brittle fracture. *Comput. Method. Appl. M.*, 381:113822, 2021.

[26] MK Brun, T Wick, I Berre, J; Nordbotten, and FA Radu. An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters. *Comput. Meth. Appl. M.*, 361:112752, 2020.

[27] T Wick. *Multiphysics Phase-Field Fracture: Modeling, Adaptive Discretizations, and Solvers*. De Gruyter, 2020.

[28] JW Both, M Borregales, JM Nordbotten, K Kumar, and FA Radu. Robust fixed stress splitting for biot's equations in heterogeneous media. *Appl. Math. Lett.*, 68:101–108, 2017.

[29] E Storvik, JW Both, K Kumar, JM Nordbotten, and FA Radu. On the optimization of the fixed-stress splitting for biot's equations. *Int. J. Numer. Meth. Eng.*, 120(2):179–194, 2019.

[30] A Mikelić and MF Wheeler. Convergence of iterative coupling for coupled flow and geomechanics. *Computat. Geosci.*, 17(3):455–461, 2013.

[31] JW Both, K Kumar, JM Nordbotten, and FA Radu. The gradient flow structures of thermo-poro-visco-elastic processes in porous media. *arXiv preprint arXiv:1907.03134*, 2019.

[32] P Areias, E Samaniego, and T Rabczuk. A staggered approach for the coupling of Cahn–Hilliard type diffusion and finite strain elasticity. *Comput. Mech.*, 57(2):339–351, 2016.

[33] JW Both. On the rate of convergence of alternating minimization for non-smooth non-strongly convex optimization in Banach spaces. *Optim. Lett.*, pages 1–15, 2021.

[34] DG Anderson. Iterative procedures for nonlinear integral equations. *J. ACM*, 12(4):547–560, 1965.

[35] C Evans, S Pollock, LG Rebholz, and M Xiao. A proof that anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically). *SIAM J. Numer. Anal.*, 58(1):788–810, 2020.

[36] E Storvik, JW Both, JM Nordbotten, and FA Radu. A Cahn–Hilliard–Biot system and its generalized gradient flow structure. *Appl. Math. Lett.*, 126:107799, 2022.

[37] C Bringedal, L von Wolff, and IS Pop. Phase field modeling of precipitation and dissolution processes in porous media: Upscaling and numerical experiments. *Multiscale Model. Sim.*, 18(2):1076–1112, 2020.

[38] C Cancès and F Nabet. Finite volume approximation of a two-phase two fluxes degenerate Cahn–Hilliard model. *ESAIM-Math. Model. Num.*, 55(3):969–1003, 2021.

[39] F Guillén-González and G Tierra. Second order schemes and time-step adaptivity for Allen–Cahn and Cahn–Hilliard models. *Comput. Math. Appl.*, 68(8):821–846, 2014.

[40] SC Brenner and LR Scott. *The mathematical theory of finite element methods*, volume 3. Springer, 2008.

[41] S Pollock, LG Rebholz, and M Xiao. Anderson-accelerated convergence of Picard iterations for incompressible Navier–Stokes equations. *SIAM J. Numer. Anal.*, 57(2):615–637, 2019.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230843062 (print)
9788230848012 (PDF)