Solveur linéaire haute-performance pour la thermo-hydro-mécanique avec régularisation par second gradient de dilatation

Ana Ordonez Egas

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Informatique et Télécommunication

Présentée et soutenue par :

Mme ANA ORDONEZ EGAS

le vendredi 25 novembre 2022

Titre :

Solveur linéaire haute-performance pour la thermo-hydro-mécanique avec régularisation par second gradient de dilatation

---

Ecole doctorale :

Mathématiques, Informatique, Télécommunications de Toulouse (MITT)

Unité de recherche :

Centre Européen de Recherche et Formation Avancées en Calcul Scientifique (CERFACS)

Directeur(s) de Thèse :

M. MICHEL DAYDE

MME CAROLA KRUSE

Rapporteurs :

MME CARMEN RODRIGO, UNIVERSIDAD DE ZARAGOZA

MME LAURA GRIGORI, INRIA

Membre(s) du jury :

M. IAIN DUFF, RUTHERFORD APPLETON LABORATORY OXFORD, Président

M. ALFREDO BUTTARI, TOULOUSE INP, Membre

M. DANIEL RUIZ, TOULOUSE INP, Membre

M. MASSIMILIANO FERRONATO, UNIVERSITA DEGLI STUDI DE PADOVA, Membre

MME CAROLA KRUSE, CERFACS, Invité(e)

M. MICHEL DAYDE, TOULOUSE INP, Membre

M. NICOLAS TARDIEU, EDF PALAISEAU, Invité(e)

Scalable linear solver for thermo-hydro-mechanics with a second gradient of dilation regularization problems

Ana Clara ORDONEZ EGAS

Director: Michel DAYDE
Supervisor: Carola KRUSE
Supervisor: Nicolas TARDIEU

Reviewers: Laura GRIGORI
Carmen RODRIGO
Examiners: Alfredo BUTTARI
Iain DUFF
Massimiliano FERRONATO
Daniel RUIZ

25 November 2022

Abstract

We are interested in the modelling of thermo-hydro-mechanical (THM) problems that describe the behaviour of a soil in which a weakly compressible fluid evolves. The soil is represented as a porous medium and the fluid is subjected to various mechanical, thermal and hydraulic stresses. This model is used for the simulation of the THM impact of the high level activity radioactive waste exothermicity within a deep geological disposal facility build in a clay-based host rock. To avoid the loss of uniqueness of the numerical solution and, more importantly, problems with the strain localization which are often encountered in soil computations, we shall consider non-locally regularized equations based on a second gradient theory. In this approach, a new primal unknown, modelling the trace of the displacement gradient, is introduced. The objective of this work is to find a parallel and scalable iterative solver for the system of equations after linearization. While extensive research has been carried on linear solvers for poroelasticity, this is not, to our knowledge, the case for the second gradient formulation. In this thesis, we shall present a block preconditioner for the fully coupled THM equations with a second gradient of dilation regularization. It is a block Gauss-Seidel approach, in which a multigrid method is used to precondition the blocks of the displacement, pressure, temperature and micro volume changes. Furthermore, we use a weighted mass matrix as preconditioner for the Lagrange multipliers block. We present numerical results that reflect the good performance of the proposed preconditioner in terms of iteration count of the iterative solver, the robustness of the preconditioner in terms of parameter variation and that is furthermore independent of the mesh size.

Keywords: Multiphysics, Preconditioning, Biot's Problem, Finite Elements, HPC, Functional Analysis.

Résumé

Nous nous intéressons à la modélisation de problèmes thermo-hydro-mécaniques (THM) qui décrivent le comportement d'un sol dans lequel évolue un fluide faiblement compressible. Le sol est représenté comme un milieu poreux et le fluide est soumis à diverses contraintes mécaniques, thermiques et hydrauliques. Ce modèle est utilisé pour la simulation de l'impact THM de l'exothermie des déchets radioactifs de haute activité dans une installation de stockage géologique profond construite dans une roche hôte à base d'argile. Pour éviter la perte d'unicité de la solution et, plus important encore, les problèmes de localisation des déformations, souvent rencontrés dans les calculs de sol, nous considérerons des équations non-localement régularisées basées sur une théorie du second gradient. Dans cette approche, une nouvelle inconnue primaire, modélisant la trace du gradient de déplacement, est introduite. L'objectif de ce travail est de trouver un solveur itératif parallèle et scalable pour le système d'équations après linéarisation. Bien que des recherches approfondies aient été menées sur les solveurs linéaires pour la poroélasticité, ce n'est pas, à notre connaissance, le cas pour la formulation du second gradient. Dans cette thèse, nous présenterons un préconditionneur par bloc pour les équations THM avec une régularisation par second gradient de dilatation. Il s'agit d'une approche Gauss-Seidel par blocs, dans laquelle une méthode multigrille est utilisée pour préconditionner les blocs de déplacement, de pression, de température et de déformation volumique microscopique. De plus, nous utilisons une matrice de masse pondérée comme préconditionneur pour le bloc de multiplicateurs de Lagrange. Nous présenterons des résultats numériques qui reflètent la robustesse du préconditionneur en termes de variation des paramètres, les bonnes performances du préconditionneur proposé en termes de nombre d'itérations du solveur itératif et qui est de plus indépendant de la taille du maillage.

Mots clés: Multiphysique, Préconditionnement, Problème de Biot, Éléments finis, Calcul intensif, Analyse fonctionnelle.

Acknowledgements

First of all, I would like to thank the reviewers, Laura Grigori and Carmen Rodrigo, for taking the time to carefully read this manuscript, as well as the other examiners, Alfredo Buttari, Iain Duff and Massimiliano Ferronato. Thank you for the valuable feedback and questions. I would also like to give a special thank you to Michel Dayde for his availability.

I am extremely grateful to my supervising team. I would like to thank Daniel Ruiz for his guidance throughout the thesis and for his crucial help on scaling techniques. I would also like to thank Carola Kruse for her knowledge and her answers to my many technical questions. I am also grateful for the friendly chats during meetings, as well as her continuous involvement and invaluable patience. Her guidance and advice kept me sane during some of the most difficult parts of this research. Finally, words can not express my gratitude to my mentor, Nicolas Tardieu. I could not have undertaken this journey without his encouragement, his immense knowledge and plentiful experience. I am thankful for his availability, willingness and enthusiasm to assist and guide me in any way he could during this journey. I will forever be grateful for his personal support in my professional and personal endeavours.

Thanks should also go to my EDF department ERMES as a whole, in particular the T6B team, as well as the PhD students team whose listening, permanent reassurance and coffee breaks have been of great help. Special thanks to the Algo team at Cerfacs that welcomed me warmly every time I went to Toulouse. I am also thankful for my friends in Toulouse, Paris and all over the world who always provided a shoulder to lean on.

My biggest thanks goes to Veronica Egas and Diego Ordonez, my parents. Without their tremendous understanding and encouragement in the past few years, it would have been impossible for me to complete this thesis. Their belief in me has kept my spirits and motivation high during this process, especially during the writing of the manuscript. Many thanks to my family in Ecuador, especially my grandparents, Carmen Reyes and Carlos Egas, for their unwavering support. Lastly, thanks to my love, Omar Rodriguez, for all the emotional support and helping me put things into perspective. A special mention to my little cat, Mistinguette, she was the only living being that I could stand to be around during the writing process, thank you for keeping me company.

Contents

Chapter 1

Introduction

Contents

The thesis, of which this manuscript is the main deliverable, is carried out under a CIFRE agreement (Industrial Conventions for Training through Research).

It is funded by the french utility EDF and supervised by IRIT (Institute for Research in Computer Science of Toulouse) and CERFACS (Center for Research and Advanced Training in Scientific Computing). The work is integrated in two EDF R&D projects, namely P-QUASI (Performance and Quality of Simulations) and GDR (Management of Radioactive Waste), the latter being a contributor to the Cigeo project, discussed below.

1.1 Industrial context

In 2006, France adopted a law on radioactive waste management that included a disposal solution for France's high-level and long-lived intermediate-level waste, *i.e.* the most radioactive waste. This led to the launch of the Cigeo project, the aim of which is to design and build a deep underground geological disposal in Meuse/Haute-Marne [3]. Waste disposal will take place for over 100 years and Cigeo will be expanded as space is needed. It will then be closed to ensure the containment of waste over very long periods of time without the need for human action. At this point in time, a disposal facility is the only solution for highly long-lived radioactive waste. However, if in the future a better solution is to be found, one of the other objectives of Cigeo is to be able to go back in there to retrieve radioactive waste. Even though the project is being conducted by Andra (Agence Nationale pour la gestion des Déchets

RAdioactifs), who is in charge of implementing safe management solutions for French radioactive waste, the safety and cost of radioactive waste management is still the responsibility of the producer, *i.e.* for the most part EDF. The safety of the site relies to a large extent on its geological formation. About 270 km of galleries and cells shall be excavated 500 m deep as illustrated in Figure 1.1. The safety concern in the

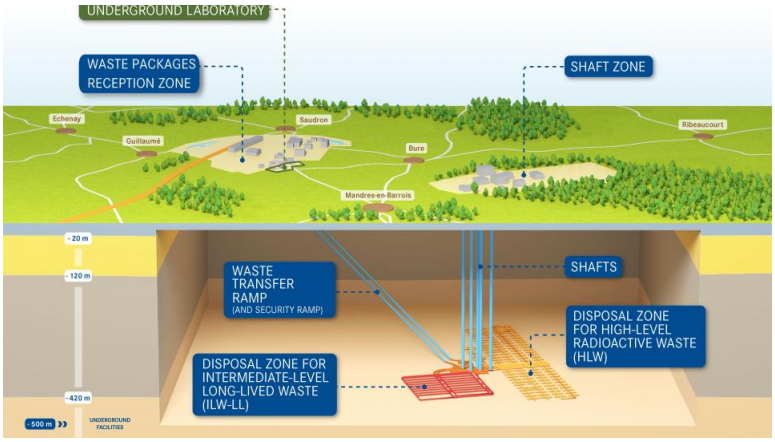


Figure 1.1: The Cigéo project[3]

more than 300 galleries is water infiltration bringing radiological contamination to the surface, once the concrete starts being porous and fissures develop in the vicinity of the galleries. Excavation, soil movement and material dilation due to heat from radioactive waste could generate such degradation. Thus, Cigeo is excavated in a stable geological clay layer with low permeability and radioactive waste is treated before its disposal. Even though risks are greatly reduced, accurate modelling is still necessary, particularly for the design of the gallery crossings, whose complex geometry requires 3D coupled Thermo-Hydro-Mechanical models.

The mechanics deals with the clay layer, the concrete and steel structures. The former are described as porous media saturated with a single-phased low-compressible fluid, which is taken care of by the hydraulics. At last, the thermics is mandatory to model the heat radiated by the nuclear wastes. Since geomaterials have complex mechanical behavior, non-linear constitutive laws are needed. Not only is the behavior of this material non-linear, but it also exhibits softening, that results in a decrease of the stress after the strain has reach a critical point. From a simulation point of view, this has a dramatic consequence on the equations of the problem, which loose the uniqueness of their solution. From a modeling point of view, this reflects the fact the classical continuum mechanics framework is not suited to deal with this type of problem, particularly it does not describe the medium at an adequate scale *i.e.* at the microscopic level. Indeed, experimentally, microscopic cracks are observed in the medium and they are not correctly described by a regular continuum mechanics model which leads to a pathological dependence of the simulation to the mesh and

consequently, unrealistic simulations results. In order to recover a realistic description of the medium and mesh independent simulations, it has been shown that the definition of a so-called internal length, characterizing the material, is an efficient solution [78, 90]. From a practical point of view, it is achieved by the use of regularization techniques, among which is the second gradient of dilation strategy [75, 73, 37].

When it comes to numerical simulations, the aforementioned characteristics lead to long and difficult problems to solve. Furthermore, due to the size of the problem in time (thousands of time steps) and space (hundreds of millions of degrees of freedom), the time-to-solution can reach months. Obviously this is not compatible with the engineers time constraints. Therefore, the design of optimal solvers is a main concern in order to reduce the time to solution.

1.2 Scientific context

When linear systems are of small to medium size, direct solvers are the solvers of choice, reason why MUMPS [2] is the default solver in `code_aster`. It is mainly due to its robustness and ease of use. However, when systems become large (say more than 10^7 degrees of freedom), time and memory consumption of direct solvers gets unfeasible and the use of iterative solvers becomes unavoidable. Nevertheless, in order to be efficient, they require to be tailored to the considered problem. It is then possible to dispose of very efficient methods, extremely well adapted to the architecture of massively parallel computers. This is the special goal of this work and we shall rely on Krylov and multigrid methods, in order to design a tailored solver for the considered problem, to which we now turn our attention.

Let us first introduce the fields needed to describe the behavior of the medium :

- $\underline{u}$ denotes the displacement of the solid matrix
- p denotes the pressure of the fluid, saturating the pores
- T denotes the temperature in the continuum
- χ denotes the dilation of the medium at the microscopic scale
- λ is a Lagrange multiplier, enforcing the equality of the dilation of the medium at the microscopic and macroscopic scales, namely between χ and $\mathrm{div}(\underline{u})$

We shall now briefly introduce the 5 equations of the Thermo-Hydro-Mechanics prob-

lem with a second gradient of dilation regularization that couple these fields.

$$
\begin{aligned}
-\operatorname{div}(\underline{\underline{\sigma}}(\underline{u})) + \nabla\lambda - r\nabla(\operatorname{div}(\underline{u})) + r\nabla\chi &= \underline{f}^e \\
\lambda - \operatorname{div}(\underline{\underline{S}}(\chi)) - r\operatorname{div}(\underline{u}) + r\chi &= 0 \\
\operatorname{div}(\underline{u}) - \chi &= 0 \\
-\operatorname{div}(\rho_f\lambda_H\nabla p) + \rho_f(\operatorname{div}(\underline{\dot{u}}) + \frac{\varphi}{K_l}\dot{p} - \alpha_m 3\dot{T}) &= 0 \\
-\operatorname{div}(\lambda_T\nabla T) - \operatorname{div}(\rho_f h_f \lambda_H \nabla p) & \\
+\rho_f h_f(\operatorname{div}(\underline{\dot{u}}) + \frac{\varphi}{K_l}\dot{p} - \alpha_m 3\dot{T}) & \\
+(3K_0\alpha_s\operatorname{div}(\underline{\dot{u}}) - 3\alpha_m\dot{p} - 9K_0\alpha_s^2\dot{T})T + C_\sigma^0\dot{T} &= \Theta
\end{aligned}
$$

The two first ones are balance equations of the mechanical stress $\underline{\underline{\sigma}}$ and of the second stress tensor $\underline{\underline{S}}$. They are respectively linked to $\underline{u}$, p, T and χ through constitutive equations. The third equation expresses the constraint of the equality of the microscopic and macroscopic volume changes. Then follows the water mass conservation equation and finally the energy conservation equation. We suppose that adequate Dirichlet and initial conditions are set in order for the problem to be well-posed. The detailed description of the equations and parameters is postponed to Chapter 3 and to Chapter 4.

An implicit Euler is applied for time discretization and $P2$-$P1$-$P1$-$P1$-$P1$ finite elements are considered for space discretization. Finally, the system is linearized using Newton's method, which results in a system of the following structure :

$$
\begin{bmatrix}
\boldsymbol{J}_{\underline{u}\underline{u}} & \boldsymbol{J}_{\underline{u}\chi} & \boldsymbol{J}_{\underline{u}\lambda} & \boldsymbol{J}_{\underline{u}p} & \boldsymbol{J}_{\underline{u}T} \\
\boldsymbol{J}_{\chi\underline{u}} & \boldsymbol{J}_{\chi\chi} & \boldsymbol{J}_{\chi\lambda} & 0 & 0 \\
\boldsymbol{J}_{\lambda\underline{u}} & \boldsymbol{J}_{\lambda\chi} & \boldsymbol{J}_{\lambda\lambda} & 0 & 0 \\
\boldsymbol{J}_{p\underline{u}} & 0 & 0 & \boldsymbol{J}_{pp} & \boldsymbol{J}_{pT} \\
\boldsymbol{J}_{T\underline{u}} & 0 & 0 & \boldsymbol{J}_{Tp} & \boldsymbol{J}_{TT}
\end{bmatrix}
\begin{bmatrix}
\boldsymbol{\delta}_{\underline{u}} \\ \boldsymbol{\delta}_\chi \\ \boldsymbol{\delta}_\lambda \\ \boldsymbol{\delta}_p \\ \boldsymbol{\delta}_T
\end{bmatrix}^{n+1}
= -
\begin{bmatrix}
\boldsymbol{R}_{\underline{u}} \\ \boldsymbol{R}_\chi \\ \boldsymbol{R}_\lambda \\ \boldsymbol{R}_p \\ \boldsymbol{R}_T
\end{bmatrix}
\tag{1.1}
$$

The block structure of the system highlights the couplings between the unknowns of the problem. Obviously, they can be split in two, $\underline{u}$, χ, λ on the one hand, $\underline{u}$, p, T on the other hand, since χ, λ are not directly coupled to p, T.

This fact has guided our approach of the global problem, where the THM and the second gradient systems will be studied independently, respectively in Chapter 3 and Chapter 4.

In order to design a robust preconditioner for the THM system, we shall give special attention to previous works for the poro-elasticity system, also known as the

Biot's problem. It enjoys an extensive literature, for which we shall present the state of the art methods. Starting from here, we focus in tailoring a preconditioner that will be scalable and robust for the range of parameters needed in Cigeo's gallery crossing simulations. Thus, Chapter 3, dedicated to the THM system, has a strong high performance computing component.

To our knowledge, no work has been done on preconditioning a system with a second gradient of dilation regularization. Fortunately, the matrix is a saddle-point and resembles, to some extent, the Stokes problem which enjoys extensive literature, that will be presented in Section 2.2.3. Therefore, we shall proceed to an in-depth analysis of the second gradient of dilation system, namely the proof of the *inf-sup* condition, in the continuous and in the discrete framework. Starting from here, we shall study the efficiency and robustness of several preconditioners. Thus, Chapter 4 has a strong theoretical component.

1.3 Outline of the manuscript

The Second Chapter presents a bibliographic review structured in three main parts. First, we briefly present an introduction to the main iterative methods and preconditioning techniques. In a second step, we focus on saddle-point problems. We start by introducing an existence and uniqueness framework followed by various block preconditioning techniques for saddle-point systems. Then, an illustrative case widely used in the literature, the Stokes system, is presented. In a third step, we focus on iterative methods for the Biot's problem. We introduce the governing equation and present the main Krylov and multigrid methods used in the literature.

The Third Chapter presents the Thermo-Hydro-Mechanics model with three distinct parts. In the first part, the framework, equations and final system of the THM problem are introduced. For the second part the preconditioning techniques used inside the THM solver are presented. Finally, the robustness and parallel scalability of the THM solver is shown.

The Forth Chapter deals with the second gradient of dilation system. A first part introduces the second gradient of dilatation model as developed in [37] as well as the evolution of its numerical formulation. The second part focuses on the proof of stability in the continuous and in the discrete framework. Several preconditioners are proposed and numerical results are shown. Lastly, the third part is the application to Thermo-Hydro-Mechanics of the second gradient of dilation model. Similarly to Chapter 3, the final preconditioner as well as numerical results are included.

Chapter 2

Numerical methods

Contents

First, we present an over view of iterative methods and preconditioning, followed by general preconditioning methods for saddle point systems and finally a bibliographic review for the Thermo-Hydro-Mechanics problem.

2.1 Iterative methods and preconditioning

To introduce iterative methods we follow [79] and the references therein. Iterative methods are used to solve linear system such as

$$\boldsymbol{A}\underline{x} = \underline{b}, \boldsymbol{A} \in \mathbb{R}^{n\times n} \tag{2.1}$$

The general principles of these methods are:

- We start with an initial vector $\underline{x}_0$ and repeat a certain number of mathematical operations to get to an approximate solution.
- The only information needed is the matrix vector product $\boldsymbol{A}\underline{x}_k$.
- The approximate solution satisfies one or multiple criteria, for example an absolute tolerance for the norm of the residual residual $\|\underline{r}_k\|_2 = \|\boldsymbol{A}\underline{x}_k - b\|_2 < \epsilon_a$ or a relative one for the scaled norm of the residual $\frac{\|\underline{r}_k\|_2}{\|\underline{r}_0\|_2} < \epsilon_r$.

2.1.1 Krylov and multigrid methods

There are three main types of iterative methods: stationary methods, Krylov methods and multigrid methods.

Classical iterative methods such as Jacobi,Gauss-Seidel or SOR are stationary. The iteration $\underline{x}_{k+1} = \boldsymbol{B}\underline{x}_k + \underline{c}$ is performed, where the iteration matrix $\boldsymbol{B}$ and vector $\underline{c}$ are independent of k. On the other hand, Krylov methods use information that changes from one iteration to another. Krylov methods are non-stationary and are usually one of the methods of choice for large-scale linear systems. Multigrid methods are one of the other methods of choice. There are two main multigrid methods, geometric multigrid methods based on a mesh hierarchy and algebraic multigrid that rely on purely algebraic components. Nevertheless, the same fundamental components are used for both. These methods are considered one of the most efficient, scalability wise.

Krylov methods

Krylov methods construct a succession of nested spaces $\mathcal{K}_k$ (approximation spaces) and $\mathcal{L}_k$(constraint spaces) for $k = 0, 1, ...$ where the solution $\underline{x}_k$ is to be found and

satisfy the Petrov-Galerkin condition i.e.

$$\underline{x}_k \in \underline{x}_0 + \mathcal{K}_k,\ \underline{r}_k = \underline{b} - \boldsymbol{A}\underline{x}_k \perp \mathcal{L}_k \tag{2.2}$$

We denote $\mathcal{K}_k = \mathcal{K}_k(\boldsymbol{A}, \underline{r}_0)$ the k-th Krylov subspace associated with $\boldsymbol{A}$ and $\underline{r}_0$ defined as $\mathcal{K}_k(\boldsymbol{A}, \underline{r}_0) = span\{\underline{r}_0, \boldsymbol{A}\underline{r}_0, \boldsymbol{A}^2\underline{r}_0, ..., \boldsymbol{A}^{n-1}\underline{r}_0\}$. From (2.2), we have :

- $\underline{x}_* - \underline{x}_k = p_k(\boldsymbol{A})(\underline{x}_* - \underline{x}_0)$ for the error, where $\underline{x}_*$ is the exact solution ;
- $\underline{r}_k = p_k(\boldsymbol{A})\underline{r}_0$ for the residual, where p_k stands for some polynomial of degree at most k satisfying $p_k(0) = 1$. P_k denotes the whole class of such polynomials.

There are three subclasses of Krylov subspace methods based on the choice of the spaces $\mathcal{L}_k$:

- **Orthogonal residual** methods: $\mathcal{L}_k = \mathcal{K}_k(\boldsymbol{A}, \underline{r}_0)$, (CG,FOM)
- **Minimal residual** methods: $\mathcal{L}_k = \boldsymbol{A}\mathcal{K}_k(\boldsymbol{A}, \underline{r}_0)$, (MINRES,GMRES)
- **Biorthogonalization** methods: $\mathcal{L}_k = \mathcal{K}_k(\boldsymbol{A}^t, \underline{r}_0)$, (BI-CG,QMR)

Symmetric systems
Two commonly used methods for symmetric systems are the Conjugate Gradient (CG) method for positive definite systems and the MINRES method for indefinite systems.

The Conjugate Gradient method generates the approximate solution satisfying the equivalent energy norm error minimization conditions

$$\underline{r}_k \perp \mathcal{K}_k(\boldsymbol{A}, \underline{r}_0) \Leftrightarrow \|\underline{x}_* - \underline{x}_k\|_{\boldsymbol{A}} = \min_{\underline{x} \in \underline{x}_0 + \mathcal{K}_k(\boldsymbol{A}, \underline{r}_0)} \|\underline{x}_* - \underline{x}\|_{\boldsymbol{A}} \tag{2.3}$$

$$\Leftrightarrow \big\|p_k(\boldsymbol{A})(\underline{x}_* - \underline{x}_0)\big\|_{\boldsymbol{A}} = \min_{p \in P_k, p(0)=1} \big\|p(\boldsymbol{A})(\underline{x}_* - \underline{x}_0)\big\|_{\boldsymbol{A}} \tag{2.4}$$

Using an expansion of the initial error $\underline{x}_* - \underline{x}_0$ into the orthogonal eigenvector basis of $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^t$ where $\boldsymbol{V}\boldsymbol{V}^t = \boldsymbol{V}^t\boldsymbol{V} = I$ and diagonal $\boldsymbol{D}$ contains the eigenvalues of $\boldsymbol{A}$, the following bound for the energy norm of the error is obtained

$$\begin{aligned} \|\underline{x}_* - \underline{x}_k\|_{\boldsymbol{A}} &= \big\|p_k(\boldsymbol{A})(\underline{x}_* - \underline{x}_0)\big\|_{\boldsymbol{A}} = \min_{p \in P_k} \big\|\boldsymbol{V}p(\boldsymbol{D})\boldsymbol{V}^t(\underline{x}_* - \underline{x}_0)\big\|_{\boldsymbol{A}} \\ &\leq \min_{p \in P_k} \big\|p(\boldsymbol{D})\big\| \big\|(\underline{x}_* - \underline{x}_0)\big\|_{\boldsymbol{A}} = \min_{p \in P_k} \max_{\lambda \in sp(\boldsymbol{A})} |p(\lambda)| \big\|(\underline{x}_* - \underline{x}_0)\big\|_{\boldsymbol{A}} \end{aligned} \tag{2.5}$$

where $sp(\boldsymbol{A})$ is the spectrum of the matrix A.
Therefore the convergence of the CG depends on the eigenvalue distribution of the matrix $\boldsymbol{A}$. Solving a polynomial approximation problem on the minimal interval

$[\lambda_{min}, \lambda_{max}]$ where $\lambda \in sp(\boldsymbol{A})$, we get a bound for the energy norm of the relative error in the CG method

$$\frac{\left\|(\underline{x}_* - \underline{x}_k)\right\|_{\boldsymbol{A}}}{\left\|(\underline{x}_* - \underline{x}_0)\right\|_{\boldsymbol{A}}} \leq 2\left(\frac{\sqrt{\kappa(\boldsymbol{A})}-1}{\sqrt{\kappa(\boldsymbol{A})}+1}\right)^k \tag{2.6}$$

Where $\kappa(\boldsymbol{A})$ is the condition number of $\boldsymbol{A}$. With this upper bound we can see how the convergence of CG depends on the conditioning of matrix $\boldsymbol{A}$. Both previous bounds highlight two of the main characteristics of iterative methods, their dependence on conditioning and eigenvalue distribution. Hence the search for a preconditioner that improves the conditioning and/or the eigenvalue distribution.

The MINRES method is the most frequently used method for symmetric indefinite systems. It computes the approximate solution satisfying the equivalent residual norm minimization conditions

$$\underline{r}_k \perp \boldsymbol{A}\mathcal{K}_k(\boldsymbol{A}, \underline{r}_0) \Leftrightarrow \|\underline{b} - \boldsymbol{A}\underline{x}_k\| = \min_{\underline{x} \in \underline{x}_0 + \mathcal{K}_k(\boldsymbol{A}, \underline{r}_0)} \|\underline{b} - \boldsymbol{A}\underline{x}_k\| \tag{2.7}$$

$$\Leftrightarrow \left\|p_k(\boldsymbol{A})\underline{r}_0\right\| = \min_{p \in P_k, p(0)=1} \left\|p(\boldsymbol{A})\underline{r}_0\right\| \tag{2.8}$$

Following the same reasoning as for the CG method, a bound for the residual norm in the MINRES method is

$$\begin{aligned} \|\underline{r}_k\| &= \min_{p \in P_k, p(0)=1} \left\|\boldsymbol{V}p(\boldsymbol{D})\boldsymbol{V}^t\underline{r}_0\right\| = \min_{p \in P_k, p(0)=1} \left\|p(\boldsymbol{D})\boldsymbol{V}^t\underline{r}_0\right\| \\ &\leq \min_{p \in P_k, p(0)=1} \left\|p(\boldsymbol{D})\right\| \left\|\boldsymbol{V}^t\underline{r}_0\right\| = \min_{p \in P_k, p(0)=1} \max_{\lambda \in sp(\boldsymbol{A})} |p(\lambda)| \|\underline{r}_0\| \end{aligned} \tag{2.9}$$

Thus the convergence rate of MINRES also depends on the eigenvalue distribution of the matrix $\boldsymbol{A}$. Since $\boldsymbol{A}$ is symmetric but indefinite, two disjoint intervals form the inclusion set for the spectrum. In this set, the polynomial approximation problem has always a unique solution but the optimal polynomial is analytically known only in special cases. This makes it significantly harder to get a practical bound for the MINRES method. Frequently one can estimate only the convergence rate using the asymptotic convergence factor satisfying the bound

$$\lim_{k \to \infty} \left(\frac{\|\underline{r}_k\|}{\|\underline{r}_0\|}\right)^{\frac{1}{k}} \leq \lim_{k \to \infty} \left(\min_{p \in P_k, p(0)=1} \max_{\lambda \in sp(\boldsymbol{A})} |p(\lambda)|\right)^{\frac{1}{k}} \tag{2.10}$$

Such bounds are quite descriptive when estimating the convergence rate of minimum residual methods for saddle-point problems that depend on asymptotically small parameters such as the mesh size.

Non-symmetric systems

For non-symmetric systems, the GMRES method is a direct generalization of the MINRES method [81] [72]. Non-symmetric iterative methods with short-term recurrences such as the biconjugate gradient (Bi-CG) method and the quasi-minimal residual (QMR) method, are also used.

For the GMRES method if we assume that the system matrix $\boldsymbol{A}$ is diagonalizable with $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^t$ and consider $\underline{r}_k = p_k(\boldsymbol{A}\underline{r}_0)$ with the expansion of $\underline{r}_0$ into the eignenvector basis $\boldsymbol{V}$, we get

$$\frac{\|\underline{r}_k\|}{\|\underline{r}_0\|} \leq \min_{p\in P_k, p(0)=1} \left\|\boldsymbol{V}p(\boldsymbol{D})\boldsymbol{V}^t\right\| \leq \kappa(\boldsymbol{V}) \min_{p\in P_k, p(0)=1} \max_{\lambda\in sp(\boldsymbol{A})} |p(\lambda)| \tag{2.11}$$

If the condition number $\kappa(\boldsymbol{V})$ of the eigenbasis $\boldsymbol{V}$ is reasonably bounded, the convergence rate is also determined by the eigenvalue distribution of the system matrix [80]. However, the system matrix is not necessarily diagonalizable and the bound seen above can not be applied. This is why a frequently used bound, independent of the spectrum, is based on the field of values of the matrix. Provided that the field of values does not contain the origin, we have for the relative residual norm the bound

$$\frac{\|\underline{r}_k\|}{\|\underline{r}_0\|} \leq \left[1 - \left(\frac{\min_{x\neq 0}(\boldsymbol{A}\underline{x}, \underline{x})}{\max_{x\neq 0}(\boldsymbol{A}\underline{x}, \underline{x})}\right)^2\right]^{\frac{k}{2}} \tag{2.12}$$

The convergence rate of the GMRES method for some application such as in [32] can be estimated using this bound. For a detailed convergence analysis of GMRES we refer to [61].

Since the GMRES method uses full-term recurrences that significantly limit its practical applicability [80], methods with short-term recurrences such as Bi-CG and QMR are used. Even thought these methods may not converge and are difficult to analyze, they usually work on real-world problems. There is no significant difference in the behavior of these two methods, a detailed analysis of the relation between them can be found in [51].

When a preconditioner is applied to Krylov subspace methods, the differences between each iterative method are even less notable and the efficiency of the solver is often determined by the choice of preconditioner.

Multigrid

Multigrid methods allow for a fast convergence rate by employing grid of different mesh sizes in order to solve all wave length error components. High frequency error

components are smoothed on fine grid. Low frequency components, which are normally reduced very slowly, become high frequency components on coarser grids and are then consequently smoothed. The process is repeated until the coarsest grid is reached, where a direct solution is computed and then interpolated back to the fine grid to correct the solution. The methods take advantage of the problem usually being simpler to solve on coarser mesh than on finer mesh.

For example, to solve (2.1) using a standard V-cycle multigrid method with a fine grid of size h and a coarse grid of size $2h$, these steps are followed:

- ν_1 iterations are done on the fine grid to approximate $\underline{x}^h$ with a classical iterative method (Jacobi,Gauss-Seidel,SOR,...) called smoother;
- After the residue $\underline{r}^h = \underline{b}^h - \boldsymbol{A}^h \underline{x}^h$ is solved on the fine grid, a restriction $\underline{r}^{2h}$ is defined for the coarse grid;
- $\boldsymbol{A}^{2h} \underline{e}^{2h} = \underline{r}^{2h}$ is then solved on the coarse grid;
- $\underline{e}^{2h}$ is interpolated on the fine grid and used to correct $\underline{x}^h \rightarrow \underline{x}^h + \underline{e}^h$;
- To finish, ν_2 iterations are performed on the fine grid to approximate $\underline{x}^h$.

where ν_1 and ν_2 are typically 1,2 or 3. This is the definition of one V-cycle.

This example is the building foundation of multigrid methods. Furthermore other schemes types can be build, for example multiple V-cycles with multiple grids per cycle, W-cycle, μ-cycle or Full MultiGrid cycle (FMG-cycle).

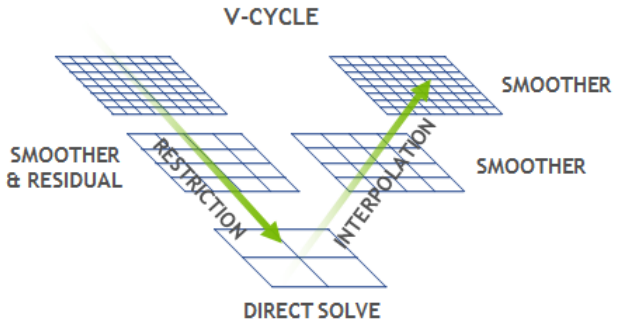


Figure 2.1: V-cycle multigrid algorithm [82]

The choice of smoothing procedure is essential for the convergence and efficiency of a multigrid method. The most common type of smoothers are classical stationary methods, since they have good smoothing properties over high frequency error components. In addition, there exists several classes of problem dependent smoothers,

for example inexact Uzawa methods and stationary methods based on symmetric indefinite splitting for saddle-point systems.

Until now the existence of a mesh has always being assumed, however purely algebraic multigrid methods exists, called Algebraic MultiGrid (AMG) [14]. The same fundamental components as for standard multigrid (which we refer to as geometric multigrid) are required. Except that for AMG all the components would be purely algebraic operators. For more detail on the construction of AMG methods refer to [15]. An advantage of AMG methods is that they can be used as "black-box" solvers, meaning that a same method can be applied to an array of problems. On the other hand, even though geometric multigrid need more information, it is a bit easier to build a tailored method and have excellent scalability and efficiency.

2.1.2 Preconditioning

In all practical cases, a preconditioner is used in order to make a linear system solvable by an iterative method. This step is crucial to the robustness of Krylov methods [80]. Roughly speaking, preconditioning is transforming the system into a system easier to solve. There are three main preconditioning approaches, left preconditioning, right preconditioning and two-sided preconditioning.

Let $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ be an invertible matrix. Then, to precondition a linear system we replace

$$\boldsymbol{A}\underline{x} = \underline{b},$$

by

$$\boldsymbol{M}^{-1}\boldsymbol{A}\underline{x} = \boldsymbol{M}^{-1}\underline{b}$$

for left preconditioning, or by

$$\begin{aligned} \boldsymbol{A}\boldsymbol{M}^{-1}\underline{y} &= \underline{b}, \\ \underline{x} &= \boldsymbol{M}^{-1}\underline{y}. \end{aligned}$$

for right preconditioning, where $\boldsymbol{M}^{-1}$ is the preconditioner.
Finally, there is also two-sided preconditioning provided that $\boldsymbol{M}$ can be factorized as $\boldsymbol{M} = \boldsymbol{M_1}\boldsymbol{M_2}$. This leads to the preconditioned system

$$\begin{aligned} \boldsymbol{M}_1^{-1}\boldsymbol{A}\boldsymbol{M}_2^{-1}\underline{y} &= \boldsymbol{M}_1^{-1}\underline{b}, \\ \underline{x} &= \boldsymbol{M}_2^{-1}\underline{y}. \end{aligned}$$

The ideal preconditioner is clearly $\boldsymbol{M}^{-1} = \boldsymbol{A}^{-1}$, this would be equivalent to solving the system and is of course of no practical use. There are two main classes of preconditioners, pure algebraic preconditioners based on an approximation of $\boldsymbol{A}^{-1}$ or algebraic multilevel approaches and problem dependent preconditioners that heavily use the information from the problem. For the last type of preconditioner, $\boldsymbol{M}^{-1}$ can be chosen such as $\boldsymbol{M}^{-1}\boldsymbol{A}$ has clusters of eigenvalues. This last property allows for a fast convergence using Krylov methods, especially for symmetric systems [80].

Symmetric positive definite systems and preconditioners

Some iterative methods can be applied only to symmetric positive definite systems. In that case, we have to make sure that the preconditioner is also symmetric positive definite. In this case, there exists a square root of $\boldsymbol{M}$, that is also symmetric positive definite. To keep symmetry two-sided preconditioning is used with $\boldsymbol{M} = \boldsymbol{M_1}\boldsymbol{M_2} = \boldsymbol{M}^{\frac{1}{2}}\boldsymbol{M}^{\frac{1}{2}}$. The matrix $\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{M}^{-\frac{1}{2}}$ is symmetric positive definite and the same methods as the unpreconditioned case would apply. we analyse the convergence behaviour on the example of the conjugate gradient method. Ideally, $\boldsymbol{M}$ would be spectrally equivalent to $\boldsymbol{A}$, i.e.there exist positive constants $0 < \hat{\beta} \leq \hat{\alpha}$ such that for every $\underline{x} \neq 0$, we have $\hat{\beta} \leq \frac{(\boldsymbol{A}\underline{x},\underline{x})}{(\boldsymbol{M}\underline{x},\underline{x})} \leq \hat{\alpha}$. Following (2.6), for the relative error in the preconditioned CG method we have

$$\frac{\left\|(\underline{x}_* - \underline{x}_k)\right\|_{\boldsymbol{A}}}{\left\|(\underline{x}_* - \underline{x}_0)\right\|_{\boldsymbol{A}}} \leq 2\left(\frac{\sqrt{\kappa(\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{M}^{-\frac{1}{2}})} - 1}{\sqrt{\kappa(\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{M}^{-\frac{1}{2}})} + 1}\right)^k \leq 2\left(\frac{\sqrt{\hat{\alpha}} - \sqrt{\hat{\beta}}}{\sqrt{\hat{\alpha}} + \sqrt{\hat{\beta}}}\right)^k \tag{2.13}$$

If $\hat{\alpha}$ and $\hat{\beta}$ are independent of the matrix dimension then also the bound does not depend on the matrix dimension [79].

Alternatively, we say that $\boldsymbol{M}$ is norm equivalent to $\boldsymbol{A}$, if there exist positive constants $0 < \hat{\beta} \leq \hat{\alpha}$ such that for every $\underline{x} \neq 0$, we have $\hat{\beta} \leq \frac{\|\boldsymbol{A}\underline{x}\|}{\|\boldsymbol{M}\underline{x}\|} \leq \hat{\alpha}$. The condition number of $\boldsymbol{A}\boldsymbol{M}^{-1}$ satisfies the bound

$$\kappa\left(\boldsymbol{A}\boldsymbol{M}^{-1}\right) = \left\|\boldsymbol{A}\boldsymbol{M}^{-1}\right\|\left\|\boldsymbol{M}^{-1}\boldsymbol{A}\right\| \leq \hat{\alpha}/\hat{\beta}, \tag{2.14}$$

Symmetric indefinite systems and symmetric positive definite preconditioners

Next we look into the case when A is symmetric indefinite but M stays symmetric positive definite (SPD). Since $\boldsymbol{M}$ is still symmetric positive definite but the system is symmetric indefinite, the preconditioned system $\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{M}^{-\frac{1}{2}}$ is symmetric indefinite. As seen above, MINRES can be used in this case. We have the following bound for

the relative residual norm in the preconditioned MINRES

$$\frac{\|\underline{r}_k\|_{\boldsymbol{M}^{-1}}}{\|\underline{r}_0\|_{\boldsymbol{M}^{-1}}} \leq \min_{p\in P_k, p(0)=1} \max_{\lambda\in sp(\boldsymbol{M}^{-1}\boldsymbol{A})} |p(\lambda)| \leq \min_{p\in P_k, p(0)=1} \max_{\lambda\in[-\hat{\alpha},-\hat{\beta}]\cup[\hat{\gamma},\hat{\delta}]} |p(\lambda)| \tag{2.15}$$

where $[-\hat{\alpha},-\hat{\beta}]\cup[\hat{\gamma},\hat{\delta}]$ is an inclusion set for all of the eigenvalues of $\boldsymbol{M}^{-1}\boldsymbol{A}$, for more details see [91].

Symmetric indefinite systems and preconditioners

Since $\boldsymbol{M}$ is symmetric indefinite, its square root does not exist in real arithmetic. Thus, in this case the methods of choice would be left or right preconditioning. Which leads to a non-symmetric preconditioned system. However since both $\boldsymbol{A}$ and $\boldsymbol{M}$ are symmetric, the preconditioned system is $\mathcal{H}$-symmetric,i.e. for left preconditioning it satisfies the relation

$$(\boldsymbol{A}\boldsymbol{M}^{-1})^t\mathcal{H} = \mathcal{H}(\boldsymbol{A}\boldsymbol{M}^{-1}) \tag{2.16}$$

where the matrix $\mathcal{H}$ is defined as $\mathcal{H} = \boldsymbol{M}^{-1}$.
Correspondingly, for right preconditioning it satisfies the relation

$$(\boldsymbol{M}^{-1}\boldsymbol{A})^t\mathcal{H} = \mathcal{H}(\boldsymbol{M}^{-1}\boldsymbol{A}) \tag{2.17}$$

where the matrix $\mathcal{H}$ is defined as $\mathcal{H} = \boldsymbol{M}$. In both cases, a common method of choice is the $\mathcal{H}$-symmetric variant of Bi-CG or QMR [42].

Non-symmetric systems and/or preconditioners

Since at least one of the matrices is non-symmetric, the preconditioned system is not symmetric. A non-symmetric iterative method must be applied, the most frequently used being GMRES. Despite its computational cost, with efficient preconditioning the method converges very quickly and its approximate solution reaches a desired tolerance level before it becomes too expensive. There are two main approaches for convergence analysis of the GMRES method. The first one is based on the field of values of the system $\boldsymbol{A}\boldsymbol{M}^{-1}$ and the other one based on the analysis of a degree of the minimal polynomial of the system $\boldsymbol{A}\boldsymbol{M}^{-1}$.

The first approach is mainly used for inexact versions of block preconditioners, where one of their diagonal blocks is negative definite. In this approach one looks for $\boldsymbol{M}$ equivalent to the matrix $\boldsymbol{A}$ with respect to the field of values, i.e. there exist positive constants $0 < \hat{\beta} \leq \hat{\alpha}$ such that for every $\underline{x} \neq 0$, we have $\hat{\beta}\|x\|^2 \leq (\boldsymbol{A}\boldsymbol{M}^{-1}\underline{x},\underline{x}) \leq \hat{\alpha}\|x\|^2$. The bound for the relative residual norm is then

$$\frac{\|\underline{r}_k\|}{\|\underline{r}_0\|} \leq \left[1 - \left(\frac{\min_{x\neq 0}(\boldsymbol{A}\boldsymbol{M}^{-1}\underline{x},\underline{x})}{\max_{x\neq 0}(\boldsymbol{A}\boldsymbol{M}^{-1}\underline{x},\underline{x})}\right)^2\right]^{\frac{k}{2}} \leq \left[1 - \left(\frac{\hat{\beta}}{\hat{\alpha}}\right)^2\right]^{\frac{k}{2}} \tag{2.18}$$

If $\hat{\alpha}$ and $\hat{\beta}$ are constants independent of the matrix dimension, then the bound will also be independent from the matrix dimension. If in addition $\hat{\alpha}$ and $\hat{\beta}$ are close, a fast convergence of GMRES can be expected.

The second approach is used for the exact version of block preconditioners or with some modification also for the constraint preconditioners. It is based on the analysis of a degree of the minimal polynomial for the matrix $\boldsymbol{A}\boldsymbol{M}^{-1}$,this degree can be very small, even equal to 2 or 3. Then one can expect that every Krylov subspace method terminates within this small number of steps.

Preconditioning techniques

To introduce preconditioning techniques, we follow [80] and the references therein. We draw a quick overview of some preconditioning techniques, keeping in mind that this list is not exhaustive.

One of the techniques to precondition a system is to perform an incomplete factorization of the matrix. This is done through a decomposition of the form :

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U} - \boldsymbol{R}$$

where $\boldsymbol{L}$ and $\boldsymbol{U}$ are upper and lower matrices with the same nonzero structure as $\boldsymbol{A}$ and $\boldsymbol{R}$ is the residual error of the factorization. The matrices $\boldsymbol{L}$ and $\boldsymbol{U}$ can then be used to form a preconditioner, for example $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{U}$. $\boldsymbol{L}$ and $\boldsymbol{U}$ are not unique, many pairs of matrices can satisfy these requirement leading to different preconditioners. For example, a preconditioner that results of this incomplete factorisation is Zero Fill-in ILU (ILU(0)). It can be define as any pair of $\boldsymbol{L}$ and $\boldsymbol{U}$ so that the elements of $\boldsymbol{L}\boldsymbol{U}$ has the same nonzero pattern as $\boldsymbol{A}$. More accurate incomplete factorizations can also be used by allowing $\boldsymbol{L}\boldsymbol{U}$ to have different levels of fill-in such as ILU(1) or ILU(p).

Another incomplete factorization that is often used as a preconditioner is the incomplete Cholesky factorization (IC) :

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^t - \boldsymbol{R}$$

where $\boldsymbol{L}$ is a lower matrix with the same nonzero structure as $\boldsymbol{A}$ and $\boldsymbol{R}$ is the residual error of the factorization. We then have $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^t$. Similarly as with ILU, different types of fill-ins can be used to create a IC type preconditioners with IC(0) having the same nonzero pattern as $\boldsymbol{A}$.

Other preconditioners are Jacobi, Gauss-Seidel and SOR. They rely on the splitting

$$\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{E} - \boldsymbol{F}$$

where $\boldsymbol{D}$ is the diagonal of $\boldsymbol{A}$, $-\boldsymbol{E}$ its strict lower part and $-\boldsymbol{F}$ its strict upper part. The Jacobi preconditioner corresponds to $\boldsymbol{M} = \boldsymbol{D}$, the lower Gauss-Seidel to $\boldsymbol{M} = \boldsymbol{D} - \boldsymbol{E}$ and the SOR preconditioner, thats stands for sucessive over relaxation, is $\boldsymbol{M} = \boldsymbol{D} - \omega\boldsymbol{E}$ where ω needs to be choosen. We can se that for $\omega = 1$, SOR and Gauss-seidel are equivalent. Upper and symmetric versions of Gauss-Seidel and SOR can also be considered.

Block preconditioners are often used for multi-physics, since they allow to treat each equation or group of equations separately. Block versions of Jacobi and Gauss-Seidel are described in Chapter 3, where as block preconditioners relying on a Schur complement approach are described in sub-section 2.2.2. Multigrid methods described in sub-section 2.1.1 are also available as preconditioners, they are used in particular inside block preconditioners throughout this manuscript.

2.2 Saddle-point problems

One of the many difficulties when preconditioning a problem with a second gradient of dilation regularization, is the existence of a saddle-point inside the system. It is brought by the use of Lagrange multipliers. A preconditioner has yet to be proposed for this particular saddle-point. Nevertheless extensive research has been done on preconditioning saddle-points, allowing us a starting point to find an appropriate preconditioner. In this section we start by introducing a framework to cover the existence and uniqueness question of saddle-point problems. Secondly, several saddle-point preconditioning techniques are presented, block preconditioners, constraint preconditioners and norm-dependent preconditioners. Finally, we present an illustrative case, the Stokes problem.

2.2.1 Existence and uniqueness

From here to the end of the manuscript we use the same notations for saddle-point problems. [12]

Continuous problem

We consider a basic saddle-point problem. Where V and Q are two Hilbert spaces and two continuous bilinear forms: $a(.,.)$ on $V \times V$ and $b(.,.)$ on $V \times Q$. We denote by A and B, respectively, the linear continuous operators associated with them. The linear continuous operators satisfy $A : V \to V'$ and $A^t : V \to V'$ such as

$$\langle Au, v\rangle_{V'\times V} = \langle u, A^t v\rangle_{V\times V'} = a(u, v), \forall v \in V, \forall u \in V.$$

and; $B : V \to Q'$ and $B^t : Q \to V'$ such as

$$\langle Bv, q\rangle_{Q'\times Q} = \langle v, B^t q\rangle_{V\times V'} = b(v, q), \forall v \in V, \forall q \in Q.$$

The problem is, given $f \in V'$ and $g \in Q'$, we want to find $(u, p) \in V \times Q$ solution of

$$\begin{cases} a(u,v) + b(v,p) = \langle f, v \rangle_{V' \times V}, \ \forall v \in V, \\ b(u,q) = \langle g, q \rangle_{Q' \times Q}, \ \forall q \in Q. \end{cases} \tag{2.19}$$

We now wish for a condition implying the existence and uniqueness of a solution to this problem. Therefore, we introduce a theorem that presents sufficient conditions on a and b in order to have a unique solution

Theorem 2.2.1. *Lets assume the same notations presented above.*
If there is a constant $\alpha > 0$, such that

$$a(u,u) \geq \alpha \|u\|_V^2, \ u \in KerB \tag{2.20}$$

and there is a constant $\beta > 0$, such that

$$\inf_{p \in Q} \sup_{u \in V} \frac{b(u,p)}{\|u\|_V \|p\|_Q} \geq \beta. \tag{2.21}$$

Then we can conclude that there exist a unique solution $(u, p) \in V \times Q$ for the system 2.19.

The theorem 2.2.1 is the one used in the rest of the manuscript. However, it does not give a necessary and sufficient condition. The coercivity condition 2.20 can be weakened by replacing it with two inf-sup conditions on a as follows. There exists $\alpha_1 > 0$ such that

$$\inf_{v \in KerB} \sup_{w \in KerB} \frac{a(v,w)}{\|v\|_V \|w\|_V} \geq \alpha_1, \tag{2.22}$$

$$\inf_{w \in KerB} \sup_{v \in KerB} \frac{a(v,w)}{\|v\|_V \|w\|_V} \geq \alpha_1, \tag{2.23}$$

$$\tag{2.24}$$

This results in necessary and sufficient conditions for the existence and uniqueness of a solution. In addition, the coercivity condition 2.20 can also be replaced by a stronger coercivity of $a(.,.)$ on the whole space V. Keeping in mind that the coercivity on $KerB$ may hold while there is no coercivity on V. As for the inf-sup condition on b 2.21, it is one of the most commonly used condition. Nevertheless equivalent formulations can be used, for example $ImB = Q'$.

Discrete problem

Now that the existence question has been treated in the continuous case, corresponding to an infinite dimensional case, we analyse the same question when working with

an approximation of the continuous case.
Let $V_h \subset V$ and $Q_h \subset Q$ be finite dimensional spaces. h refers to a mesh from which these approximations are derived. We consider the restriction of the bilinear forms a to $V_h \times V_h$ and b to $V_h \times Q_h$. In addition, we consider the restriction of the operators A and B that we denote A_h and B_h.
Therefore we consider the corresponding discrete problem to 2.19, we look for $(u_h, p_h) \in V_h \times Q_h$, solution of

$$\begin{cases} a(u_h, v_h) + b(v_h, p_h) = \langle f, v_h \rangle_{V' \times V}, \ \forall v_h \in V_h, \\ b(u_h, q_h) = \langle g, q_h \rangle_{Q' \times Q}, \ \forall q_h \in Q_h. \end{cases} \tag{2.25}$$

We now wish for a condition implying the existence and uniqueness of (u_h, p_h). Therefore, we introduce a theorem that presents sufficient conditions in order to have a unique solution.

Theorem 2.2.2. *Lets assume the same notations presented above.*
If there is a constant $\alpha > 0$, such that

$$a\left(u_h, u_h\right) \geq \alpha \|u_h\|_V^2 , \ u_h \in KerB_h \tag{2.26}$$

and there is a constant $\beta > 0$, such that

$$\inf_{p_h \in Q_h} \sup_{u_h \in V_h} \frac{b(u_h, p_h)}{\|u_h\|_V \|p_h\|_Q} \geq \beta. \tag{2.27}$$

Then we can conclude that there exist a unique solution $(u_h, p_h) \in V_h \times Q_h$ for the system (2.25) and since α and β are independent from h the discretization is stable.

The theorem 2.2.2 is the one used in the rest of the manuscript. However, as in the continuous case, it does not give necessary and sufficient conditions. The coercivity condition 2.26 can be weakened by replacing it with the discrete version of 2.22. In order to have necessary conditions, α and β could also depend on h, although it is not a desirable situation. In this case, an error estimate between the continuous solution and the discrete solution might still provide a convergence result but in general not an optimal one.
Condition 2.26 can also be replaced by the stronger coercivity on the whole space V_h, in this case the discrete version follows from the continuous case. As for the inf-sup discrete condition 2.27, it can be replaced by equivalent discrete formulations as in the continuous case.

2.2.2 Preconditioning techniques for block systems

Three main preconditioning techniques are presented. Block preconditioning, constraint preconditioning and norm-dependent preconditioning.

Block preconditioning

Block preconditioners for saddle point systems are based on the LDU block factorisation as follows

$$\boldsymbol{A} = \boldsymbol{LDU} = \begin{bmatrix} \boldsymbol{I} & 0 \\ \boldsymbol{BK}^{-1} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{K} & 0 \\ 0 & \boldsymbol{S} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{K}^{-1}\boldsymbol{B}^t \\ 0 & \boldsymbol{I} \end{bmatrix} \tag{2.28}$$

where

$$\boldsymbol{S} = -\boldsymbol{C} - \boldsymbol{BK}^{-1}\boldsymbol{B}^t$$

is the Schur complement. Their performance relies on the approximations of $\boldsymbol{K}$ and $\boldsymbol{S}$.
With the assumption that $\boldsymbol{K}$ and -$\boldsymbol{S}$ are SPD. Among the block diagonal preconditioners,the ideal one is

$$\boldsymbol{M}_d = \begin{bmatrix} \boldsymbol{K} & 0 \\ 0 & -\boldsymbol{S} \end{bmatrix}$$

The preconditioned system by $\boldsymbol{M}_d^{-1}$ results in

$$\mathcal{M} = \boldsymbol{M}_d^{-1}\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{K}^{-1}\boldsymbol{B}^t \\ -\boldsymbol{S}^{-1}\boldsymbol{B} & 0 \end{bmatrix}$$

a non-singular symmetrizable system that satisfies

$$(\mathcal{M} - I)\left(\mathcal{M} - \frac{1}{2}(1+\sqrt{5})I\right)\left(\mathcal{M} - \frac{1}{2}(1-\sqrt{5})I\right) = 0$$

$\mathcal{M}$ is diagonalizable with three distinct eigenvalues $1, \frac{1}{2}(1+\sqrt{5}), \frac{1}{2}(1-\sqrt{5})$, for a detailed analysis see [69]. This means, that the preconditioned MINRES with $\boldsymbol{M}_d^{-1}$ terminates in at most three iterations.

The same logic applies to block triangular preconditioner, where the ideal one is

$$\boldsymbol{M}_{tri} = \begin{bmatrix} \boldsymbol{K} & \boldsymbol{B}^t \\ 0 & \pm\boldsymbol{S} \end{bmatrix}.$$

A minus sign in front of the Schur complement results in a diagonalizable preconditioned matrix with only two distinct eigenvalues ± 1, whereas for a plus sign, all the eigenvalues are equal to 1. For either choice, GMRES will converge in at most 2 iterations.

Clearly, the ideal preconditioners $\boldsymbol{M}_d$ and $\boldsymbol{M}_{tri}$ are not practical. In practice, $\boldsymbol{K}$ and $\boldsymbol{S}$ are replaced by approximations i.e. $\tilde{\boldsymbol{K}}$ and $\tilde{\boldsymbol{S}}$, that are highly problem-dependent. Appropriately chosen approximations will cluster eigenvalues around those of the ideally preconditioned matrices.

Constraint preconditioning

The second type of preconditioners are of the form

$$\boldsymbol{M}_{cons} = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{B}^t \\ \boldsymbol{B} & 0 \end{bmatrix}$$

where $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ is an approximation of $\boldsymbol{K}$ [56]. For these type of systems the matrix blocks B and B^t are often associated with constraints. Since these blocks remain unchanged in the preconditioner, $\boldsymbol{M}_{cons}$ is called a constraint preconditioner. The inclusion of the constraints blocks into the preconditioner can lead to a more favorable distribution of the eigenvalues of the preconditioned system. If this is the case, convergence would be improved since we are working with Krylov methods. For example, in the case of $\boldsymbol{K}$ being nonsingular, $\boldsymbol{B} \in \mathbb{R}^{m\times m}$ will be of full rank, and a constraint preconditioner will cluster at least 2m eigenvalues at 1. The convergence will depend on the choice of $\boldsymbol{H}$. For a detailed proof of the example above as well as a more in depth analysis of eigenvalue distribution and bounds for other cases, see [56].

Norm dependent preconditioners

The final type of preconditioners that we present are norm dependent preconditioners. The technique consist in finding an appropriate preconditioner for the continuous problem and applying a stable discretization in order to identify a correct preconditioner for the discrete problem. Since we follow the theory developed in [64], it is easier to introduce the same notations.

We introduce the system 2.19 that can be rewritten in the form

$$\mathcal{A} = \begin{pmatrix} A & B^t \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}$$

Where A and B are bounded operators. Theorem 2.2.1 gives sufficient conditions for ensuring that the coefficient operator $\mathcal{A} : V \times Q \to V' \times Q'$ is an isomorphism. As we saw in the previous section if Theorem 2.2.1 holds, the saddle point system has a unique solution. In general, Krylov space methods are not going to be well defined for this type of problems. This is due to the fact that the operator $\mathcal{A}$ may map functions in $V \times Q$ out of the space. Therefore a preconditioner is needed in order to solve unbounded problems with a Krylov space method.

Next, we look for an appropriate preconditioner for the saddle point problem 2.19. Since $\mathcal{A} : V \times Q \to V' \times Q'$ is an isomorphism the corresponding preconditioners $\mathcal{B}$

should be isomorphisms mapping $V' \times Q'$ to $V \times Q$. Hence, the canonical choice should be a block diagonal operator $\mathcal{B}$ of the form

$$\mathcal{B} = \begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix}$$

where $M : V' \to V$ and $N : Q' \to Q$ are symmetric and positive definite isomorphisms. They are spectrally equivalent with the Riez mappings in the respective spaces.
As a consequence, the composition

$$\mathcal{BA} : \ V \times Q \xrightarrow{\mathcal{A}} V' \times Q' \xrightarrow{\mathcal{B}} V \times Q$$

is an isomorphism mapping $V \times Q$ to itself. Therefore, when using a Krylov method to solve the preconditioned system the convergence rate is bounded by $\kappa(\mathcal{BA}) = \|\mathcal{BA}\| \left\|(\mathcal{BA})^{-1}\right\|$.

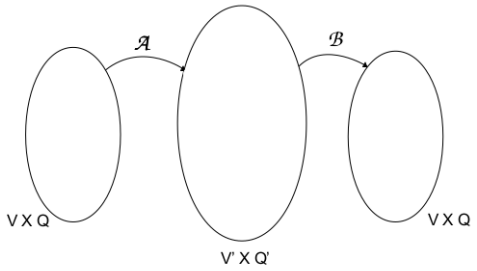


Figure 2.2: The mapping property of the operators $\mathcal{A}$ and $\mathcal{B}$ [64]

The next step is to apply a stable discretization with the spaces $V_h \subset V$ and $Q_h \subset Q$. In order for the discretization to be stable the theorem 2.2.2 gives sufficient conditions. If the theorem is verified and given that $\mathcal{B}$ is an appropriate preconditioner for the continuous problem, the canonical discrete preconditioner is then

$$\mathcal{B}_h = \begin{pmatrix} M_h & 0 \\ 0 & N_h \end{pmatrix}$$

where $M_h : V_h' \to V_h$ and $N_h : Q_h' \to Q_h$.

The last step is to construct efficient preconditioners for the discrete operators M_h and N_h. This step is completely problem dependent, although some operators are common. For example, for elliptic operators such as Δ and as $I - \Delta$ multigrid and domain decomposition methods are specially well suited[15, 87]. The multigrid

V-cycle algorithm is suited for reaction-diffusion operators such as $I - \epsilon\Delta$ where $\epsilon \geq 0$ is a small parameter [17]. Other operators can take extra care to construct appropriate preconditioners for them. For example for $I + curlcurl$ and $I - graddiv$ multigrid algorithms can be used but appropriate smoothers need to be constructed, more details can be fund in [5].

2.2.3 The illustrative case of the Stokes system

The Stokes problem is a fundamental model of viscous flow. Extensive research has been done on this problem, we choose to follow [31].

Let Ω be a d-dimensional domain, the Stokes model describes the evolution of the vector velocity field $\underline{u}(x)$ and the scalar pressure field $p(x)$. The first equation represents conservation of the momentum of the fluid and the second equation enforces conservation of mass. It is therefore also referred as the incompressibility constraint.

$$\begin{aligned} -\Delta\underline{u} + \nabla p &= \underline{f}^e \quad \text{in } \Omega \\ -\operatorname{div}(\underline{u}) &= g^e \quad \text{in } \Omega \\ \underline{u} &= \underline{0} \quad \text{on } \partial\Omega \end{aligned} \tag{2.29}$$

To solve the system, a finite element method is used. Let $\underline{u}, \underline{v} \in (H_0^1(\Omega))^d$, $p, q \in L^2(\Omega)$. Bilinear forms are defined as

$$\begin{aligned} \mathcal{A}(\underline{u}, \underline{v}) &= (\nabla\underline{u} : \nabla\underline{v}) \\ \mathcal{B}(\underline{u}, q) &= (q, \nabla\underline{u}) \end{aligned}$$

We can now write the variational formulation. Find $(\underline{u}, p) \in (H_0^1(\Omega))^d \times L^2(\Omega)$ such as for all $(\underline{v}, q) \in (H_0^1(\Omega))^d \times L^2(\Omega)$, we have

$$\begin{aligned} -\mathcal{A}(\underline{u}, \underline{v}) - \mathcal{B}(\underline{v}, p) &= f(\underline{v}) \\ -\mathcal{B}(\underline{u}, q) &= 0 \end{aligned} \tag{2.30}$$

The choice of a finite element family isn't straight forward but it is of central importance when it comes to find fast and reliable iterative solution methods. To construct fast convergent iterative methods the inf-sup stability condition is crucial.

Let $\{\phi_{\underline{v}_j}\}_{j=1}^{N_u}$ be a basis for the finite element space $U_h(\Omega) \subset (H_0^1(\Omega))^d$, for all $\underline{v}_h \in U_h(\Omega)$ we have

$$\underline{v}_h = \sum_{j=1}^{N_u} v_j \, \phi_{\underline{v}_j}$$

Let $\{\phi_{q_j}\}_{j=1}^{N_p}$ be a basis for the finite element space $P_h(\Omega) \subset L^2(\Omega)$, for all $q_h \in P_h(\Omega)$ we have

$$q_h = \sum_{j=1}^{N_p} q_j \, \phi_{q_j}$$

By injecting the above discretized fields in the variational formulation (2.30), the following linear system is obtained :

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B}^t \\ \boldsymbol{B} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_{\underline{u}} \\ \boldsymbol{b}_p \end{bmatrix} \tag{2.31}$$

where each matrix block is given by :

$$(\boldsymbol{A})_{ij} = -\int_\Omega \nabla(\phi_{v_i}) : \nabla(\phi_{\underline{v}_j}) \, dx, \qquad \forall i, j = 1, N_u$$
$$(\boldsymbol{B})_{ij} = -\int_\Omega \phi_{q_i} \, \nabla \phi_{v_j} \, dx, \qquad \forall i = 1, N_p, \forall j = 1, N_u$$

and each unknown array is given by :

$$\boldsymbol{u} = [u_i]_{i=1,N_u}, \boldsymbol{p} = [p_i]_{i=1,N_p}$$

One of the principal problems of Stokes is the indefiniteness of the system. If the mesh size is reduced, the discrete problem size will increase and so will the number of eigenvalues. The Stokes system then becomes highly indefinite. This can cause the CG method to fail in some cases, however the MINRES method is robust and is the method of choice. SYMMLQ could also be used, however it does not have a minimization property with respect to Krylov subspaces, and even though computational rounding errors can have a higher effect on MINRES than SYMMLQ, this only applies for ill-conditioned problems, see [72][86]. With multigrid methods there is the difficulty that simple smoothers such as Jacobi and Gauss-Seidel are not well-defined because of zero diagonals. However, multigrid methods can be used in conjunction with block preconditioners as we will see in the following. For a comparison of several different iterative solution methods using multigrid, see [32].

For this system, the discretization error is measured in the energy norm $(H^1(\Omega))^d$ for the velocity components and in the $L^2(\Omega)$ norm for the pressure.
Let $\boldsymbol{E}$ be

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{K} & 0 \\ 0 & \boldsymbol{Q} \end{bmatrix}.$$

The natural matrix norm of the error, $\left\|e^{(k)}\right\|_{\boldsymbol{E}}$ will then be

$$\left\|e^{(k)}\right\|_{\boldsymbol{E}}^2 = (\boldsymbol{E}e^{(k)}, e^{(k)}).$$

Since $\boldsymbol{A}e^{(k)} = r^{(k)}$, for the residual this becomes

$$\left\|e^{(k)}\right\|_{\boldsymbol{E}}^2 = (\boldsymbol{E}\boldsymbol{A}^{-1}r^{(k)}, \boldsymbol{A}^{-1}r^{(k)}) = \left\|r^{(k)}\right\|_{\boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{A}^{-1}}^2.$$

Since MINRES method reduces $\left\|r^{(k)}\right\|_{\boldsymbol{M}^{-1}}$, a good preconditioner could be

$$\boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{A}^{-1} = (\boldsymbol{A}\boldsymbol{E}^{-1}\boldsymbol{A})^{-1} = \begin{bmatrix} \boldsymbol{K} + \boldsymbol{B}^t\boldsymbol{Q}^{-1}\boldsymbol{B} & \boldsymbol{B}^t \\ \boldsymbol{B} & \boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{B}^t \end{bmatrix}$$

However, these matrix operators are not suitable since they do not satisfy the requirement concerning ease of solution. Because of that, effective and practical preconditioners derived from $(\boldsymbol{A}\boldsymbol{E}^{-1}\boldsymbol{A})^{-1}$ will be presented in the following. They will be good with respect to the norm being minimized by MINRES.
The form of $\boldsymbol{A}$ and the desired norm based on $\boldsymbol{E}$ suggests the importance of the block structure when preconditioning. Therefore we will consider block diagonal preconditioning matrices of the form

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{P} & 0 \\ 0 & \boldsymbol{T} \end{bmatrix}$$

where $\boldsymbol{P} \in \mathbb{R}^{n_u \times n_u}$ and $\boldsymbol{T} \in \mathbb{R}^{n_p \times n_p}$ are symmetric and positive-definite. If $\boldsymbol{P} = \boldsymbol{K}$, then $\lambda = 1$ is an eigenvalue, if also $\boldsymbol{T} = \boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{B}^t$, based on the block diagonal Schur based approach, then the remaining eigenvalues satisfy

$$(\lambda^2 - \lambda - 1)\boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{B}^t p = 0$$

Since the inf-sup stability in this case ensures that $\boldsymbol{B}\boldsymbol{K}^{-1}\boldsymbol{B}^t$ is positive-definite, we deduces that $\lambda = \frac{1}{2} \pm \frac{\sqrt{5}}{2}$ are the remaining eigenvalues. We have then 3 distinct eigenvalues, MINRES then terminates with the exact solution after three iterations independently of the size of the discrete system. This preconditioner is ideal, because it requires the action of the inverse of $\boldsymbol{K}$ and of the Schur complement that is a full matrix. Clearly this is completely unsuitable in practice. Thus we have to find suitable approximations for $\boldsymbol{P}$ and $\boldsymbol{T}$.
Since for Stokes the sparse pressure mass matrix $\boldsymbol{Q}$ is spectrally equivalent to the Schur complement, $\boldsymbol{Q}$ would be a good approximation for $\boldsymbol{T}$. The discrete inf-sup stability and boundedness imply $\gamma^2 \leq \lambda(\lambda - 1) \leq \tau^2$ where γ is the inf-sup constant and τ the upper bound on the pressure Schur complement. With a similar eigenvalue

analysis than above, every eigenvalue lies in

$$\{1\} \bigcup \left[\frac{1+\sqrt{1+4\gamma^2}}{2}, \frac{1+\sqrt{1+4\tau^2}}{2} \right] \bigcup \left[\frac{1-\sqrt{1+4\tau^2}}{2}, \frac{1-\sqrt{1+4\gamma^2}}{2} \right]. \tag{2.32}$$

For a detailed proof see [71]. The eigenvalue 1 is retained and the other eigenvalues are pairewise symmetric and lie in small intervals that are uniformly bounded from $\pm\infty$ and uniformly bounded away from the origin. Therefore, MINRES will not terminate in three iterations but convergence will be fast and independent of the size of the discret problem. Finally, this is the crucial point.
For $\boldsymbol{P}$, the "ideal" choice being $\boldsymbol{K}$ and $\boldsymbol{K}$ being the vector Laplacian, a convenient preconditioner for the Laplacian would be an effective approach. This would be the equivalent of preconditioning the Poisson equation. $\boldsymbol{P}$ could be derived from domain decomposition methods or a multigrid cycles. With multigrid, for example a V-cycle or a W-cycle preconditioner,we have the necessary bound conditions and with a single multigrid cycle to represent $\boldsymbol{P}$, MINRES would only have a few extra iterations compared to the more expensive choice $\boldsymbol{P} = \boldsymbol{K}$. If we chose a single V-cycle the convergence would be independent from the discrete problem size. The use of algebraic multigrid for $\boldsymbol{P}$ is also effective, it is described in [76]. For an in depth analysis of the preconditioning of the Poisson equation see chapter 1 from [31].

Other more traditional methods include methods that iterate between a segregated solution step for the velocity followed by a pressure update step. The principal method named after Uzawa is described in [40], but other segregated methods including a method that leads to a positive-definite preconditioned matrix which is self-adjoint in a nonstandard inner product for Stokes can be see in [95]. The issue with these methods is the selection of parameters. The use of block triangular preconditioners can be found in [57]. Multigrid methods are also very efficient for Stokes, a careful comparison between multigrid methods using V-cycle with distributive Gauss-Seidel (DGS) and incomplete factorization (ILU) as smoother, and Krylov methods using the preconditioner shown above can be seen in [32].

After the review of classical block-preconditioning techniques, we shall now focus on their application for the simplified Biot's poro-elasticity problem.

2.3 Iterative methods for Biot's poro-elasticity problem

2.3.1 Governing equations

We turn our attention to the behaviour of an elastic, homogeneous and isotropic porous medium filled with an almost incompressible viscous fluid, saturating the pores occupying an open subset $\Omega \in \mathbb{R}^d$, $d = 2, 3$. It is often refered to as the Biot's model[11]. The unknowns are the displacement $\underline{u}$ of the solid matrix and the pressure p of the fluid.

We shall briefly introduce the equations of the problem that constitutes the hydro-mechanical model of the considered medium. The detailed description is postponed to Chapter 3 in order to introduce the coupling with the thermics. We thus consider the solution of the coupled equations, namely the balance of linear momentum and the balance of liquid mass :

$$-\operatorname{div}(\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u})) + b\nabla p = \underline{f}^e \tag{2.33}$$

$$b\operatorname{div}(\underline{\dot{u}}) + s_0\dot{p} - \operatorname{div}(\lambda_H \nabla p) = g^e \tag{2.34}$$

Where $\underline{\underline{\underline{\underline{A}}}}$ is the forth-order tensor of elastic coefficients, $\underline{\underline{\varepsilon}}(\underline{u}) = \frac{1}{2}(\nabla \underline{u} + \nabla^T \underline{u})$ is the strain tensor, b is the Biot's modulus, s_0 is the fluid storage coefficient and λ_H is the hydraulic conductivity. The right-hand sides $\underline{f}^e$, g^e respectively denote the mechanical body forces and the fluid source. We suppose that adequate Dirichlet and initial conditions are set in order for the problem to be well-posed.
Though this 2 fields formulation is known to exhibit a pressure field with spurious oscillations [53], it is considered here since it is the one used in the industrial finite element software `code_aster`, that is used for the numerical illustrations in what follows.
We define the Sobolev spaces with obvious notations :

$$U(\Omega) = \{\underline{u} \in (H^1(\Omega))^d, \underline{u} = \underline{u}^e \text{ on } \partial\Omega^{\underline{u}}\},$$
$$P(\Omega) = \{p \in H^1(\Omega), p = p^e \text{ on } \partial\Omega^p\},$$

By considering the appropriate Sobolev spaces defined above and by integration by parts, we have the following weak form for $(\underline{u}, p) \in U(\Omega) \times P(\Omega)$ and $(\underline{v}, q) \in U(\Omega) \times P(\Omega)$:

$$\begin{aligned}\mathcal{A}(\underline{u},\underline{v}) &= (\underline{\underline{\underline{A}}} : \underline{\underline{\varepsilon}}(\underline{u}), \underline{\underline{\varepsilon}}(\underline{v}))\\ \mathcal{B}(\underline{v},q) &= (-bq,\ \mathrm{div}(\underline{v}))\\ \mathcal{C}(p,q) &= (s_0 p, q)\\ \mathcal{D}(p,q) &= (\lambda_H \nabla p, \nabla q)\end{aligned}$$

We can now express the problem as :
Find $(\underline{u}, p) \in U(\Omega) \times P(\Omega)$ such as for all $(\underline{v}, q) \in U(\Omega) \times P(\Omega)$, we have :

$$\begin{aligned}\mathcal{A}(\underline{u},\underline{v}) + \mathcal{B}(\underline{v},p) &= f(\underline{v})\\ -\mathcal{B}(\underline{\dot{u}},q) + \mathcal{C}(\dot{p},q) + \mathcal{D}(p,q) &= g(q)\end{aligned}$$

Where f, g are associated linear forms.

We now use a backward Euler scheme for the time discretization with the notations $\underline{u}^n := \underline{u}(t^n)$, $p^n := p(t^n)$ which denote the displacement and the pressure fields at $t^n = n\Delta t$, where Δt is a given time increment. The system now writes :

$$\begin{aligned}\mathcal{A}(\underline{u}^{n+1},\underline{v}) + \mathcal{B}(\underline{v},p^{n+1}) &= f(\underline{v})\\ -\mathcal{B}(\underline{u}^{n+1},q) + \mathcal{C}(p^{n+1},q) + \Delta t\mathcal{D}(p^{n+1},q) &= \Delta t g(q) - \mathcal{B}(\underline{u}^n,q) + \mathcal{C}(p^n,q)\end{aligned} \tag{2.35}$$

Let $U_h(\Omega)$ be the discrete Sobolev subspace of $U(\Omega)$ of dimension N_u with $h > 0$ a parameter that refers to the mesh size, the same formulation goes for $P_h(\Omega)$.
Let $\{\phi_{\underline{v}_j}\}_{j=1}^{N_u}$ be a basis for the finite element space $U_h(\Omega)$, for all $\underline{v}_h \in U_h(\Omega)$ we have

$$\underline{v}_h = \sum_{j=1}^{N_u} v_j\, \phi_{\underline{v}_j}$$

Let $\{\phi_{q_j}\}_{j=1}^{N_p}$ be a basis for the finite element space $P_h(\Omega)$, for all $q_h \in U_h(\Omega)$ we have

$$q_h = \sum_{j=1}^{N_p} q_j\, \phi_{q_j}$$

By injecting the above discretized fields in the variational formulation (2.35), the following linear system is obtained :

$$\begin{bmatrix} \boldsymbol{A}_{\underline{u}\underline{u}} & \boldsymbol{A}_{\underline{u}p} \\ \boldsymbol{A}_{p\underline{u}} & \boldsymbol{A}_{pp} \end{bmatrix} \begin{bmatrix} \underline{\boldsymbol{u}} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_{\underline{u}} \\ \boldsymbol{b}_p \end{bmatrix}$$

where the matrix blocks are given by :

$$
\begin{aligned}
(\boldsymbol{A}_{\underline{u}\underline{u}})_{ij} &= \int_\Omega -\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\phi_{\underline{v}_i}) : \underline{\underline{\varepsilon}}(\phi_{\underline{v}_j})\, dx, && \forall i,j = 1, N_u \\
(\boldsymbol{A}_{\underline{u}p})_{ij} &= -\int_\Omega b\phi_{q_j} \operatorname{div}(\phi_{\underline{v}_i})\, dx, && \forall i = 1, N_u, \forall j = 1, N_p \\
(\boldsymbol{A}_{p\underline{u}})_{ij} &= \int_\Omega b\phi_{q_i} \operatorname{div}(\phi_{\underline{v}_j})\, dx, && \forall i = 1, N_p, \forall j = 1, N_u \\
(\boldsymbol{A}_{pp})_{ij} &= -\int_\Omega s_0\phi_{q_i}\phi_{q_j} + \lambda_H \Delta t \operatorname{div}(\phi_{q_i}) \operatorname{div}(\phi_{q_j})\, dx, && \forall i = 1, N_p, \forall j = 1, N_p
\end{aligned}
$$

and each unknown array is given by :

$$\underline{\boldsymbol{u}} = [u_i]_{i=1,N_u}, \boldsymbol{p} = [p_i]_{i=1,N_p}$$

One can notice that multiplying the balance of liquid mass by -1 provides a symmetric expression of $\boldsymbol{A}$. Once again, since it is not the case in the finite element software `code_aster`, we have not them here and we keep the non-symmetric form of $\boldsymbol{A}$.

We shall now turn our attention to the solution techniques available in the literature for resolution of the Biot's poro-elasticity problem.

2.3.2 Solution strategies for coupled problems

When considering the numerical solution of a coupled system of PDE, sequential or monolithic approaches can be used.

For the sequential method, also called staggered or operator-splitting method, each balance equation is solved once at a time, thus requiring an update strategy in order to transfer the values of each field from a balance equation to the other. Thanks to an appropriate convergence criterion, the solution of the coupled system is recovered. The main interest of such approaches relies in the use of existing verified and robust simulators, dedicated to a particular problem, with the final goal to reduce coding efforts. However, they obviously require a special coupling algorithm, whose numerical stability and accuracy can be problemsome [84, 4] and sometimes also the use of a coupling software, which can penalize the reduction of the programming burden. [16].

For the monolithic method, all balance equations are solved simultaneously, requiring processing within the same computer software. In addition to the development of a dedicated software application, monolithic approaches require that the inf-sup condition is met [65], in order to avoid spurious oscillations in the pressure field. This topic benefited special attention in the literature and several strategies has been proposed to circumvent the problem. The latest works consider three-field formulations, with different choices of the extra-field (solid pressure [60, 23] or Darcy's velocity

[38, 19]) in order to alleviate the non-physical pressure oscillations at the interface between materials with different permeabilities.

In the sequel, a monolithic approach is considered. The next sections focus on state of the art scalable solution methods.

2.3.3 Krylov preconditioned methods

Block preconditioners are a natural solution for linear systems involving unknowns of different kinds. Based on block factorizations, diagonal or triangular preconditioners have been proposed. Algebraic multigrid (AMG), incomplete factorization and approximate inverses are often used for preconditioning the displacement block [92, 19]. The design of a preconditioner for the pressure block implies the approximation of the Schur complement, whose exact evaluation is computationally impossible even for moderate size problems due to its dense nature. When the media is fully saturated by the fluid, several approximations are proposed based on AMG, mass matrices and incomplete factorization [92, 93]. When phase change in the fluid are to be considered, the Constrained Pressure Residual (CPR) method is often considered as the preconditioner of choice for real-life problems in the oil reservoir community [18, 52, 88, 29].

It must be noticed that all block preconditioners do not rely on block factorization. Recent works based on the choice of physical parameter based norms have been proposed [60, 1]. They turn to be block diagonal or triangular preconditioners that show great independence from the material parameters, thus motivating the name of parameter-robust preconditioners. Other approaches are based on particular matrix decompositions which enjoy nice convergence properties when used as preconditioners [59].

In the following, three of these main preconditioning techniques are presented for the Biot's problem, namely block factorization preconditioner, norm-dependent preconditioners and constraint preconditioners.

Block factorization preconditioners

We start with block preconditioning inspired from the LDU factorisation in (2.28). Approximations are needed for $\boldsymbol{A}_{\underline{uu}}^{-1}$ and $\boldsymbol{S}^{-1}$.

$\boldsymbol{A}_{\underline{uu}}^{-1}$:
In Biot's problem, $\boldsymbol{A}_{\underline{uu}}$ is the stiffness matrix associated with the mechanical response of the porous medium. Algebraic preconditioners have been studied over the years. We focus on two main strategies, incomplete factorisation and algebraic multigrid (AMG).

- Algebraic multigrid such as ML [48] has shown to be very efficient, and we can use them to precondition $\boldsymbol{A}_{\underline{uu}}$ in two ways. The first one is to approximate the inverse of $\boldsymbol{A}_{\underline{uu}}$ with an AMG , we call this $\boldsymbol{A}_{\underline{uu}}^{-1}(AMG)$. The second one is to approximate $\boldsymbol{A}_{\underline{uu}}^{-1}$ with a Krylov method preconditioned by AMG, we'll call this $\boldsymbol{A}_{\underline{uu}}^{-1}(kry, AMG)$.
- Incomplete factorisation such as ILU(0) [92] or IC(0) [19] is also efficient. Following the same approach as above, $\boldsymbol{A}_{\underline{uu}}^{-1}(IC(0))$ applies IC(0) as the preconditioner and $\boldsymbol{A}_{\underline{uu}}^{-1}(kry, IC(0))$ solves $\boldsymbol{A}_{\underline{uu}}^{-1}$ with a krylov method preconditioned by IC(0).

$\mathbf{S^{-1}}$:
For Biot preconditioning the Schur complement $\boldsymbol{S} = -\boldsymbol{A}_{pp} - \boldsymbol{A}_{p\underline{u}}\boldsymbol{A}^{-1}\boldsymbol{A}_{\underline{u}p}$ is trickier. $\boldsymbol{S}$ is not sparse for saturated soil due to the presence of $\boldsymbol{A}_{\underline{uu}}^{-1}$. Because of this, the approach that is usually taken is to approximate the Schur complement and then choose a preconditioning strategy for the approximation.

1. Approximations of $\boldsymbol{S}$

   - Since $\boldsymbol{A}_{\underline{uu}}^{-1}$ in the second term leads to $\boldsymbol{S}$ being dense, a sparse approximation is $\boldsymbol{S}_d = -\boldsymbol{A}_{pp} - \boldsymbol{A}_{p\underline{u}}diag(\boldsymbol{A}^{-1})\boldsymbol{A}_{\underline{u}p}$ [92].
   - The second term of $\boldsymbol{S}$ can also be replaced by the pressure mass matrix. $\boldsymbol{S}_m = -\boldsymbol{A}_{pp} + \frac{b^2}{K}Mass_p$. This specific choice of $-\boldsymbol{A}_{p\underline{u}}\boldsymbol{A}_{\underline{uu}}^{-1}\boldsymbol{A}_{\underline{u}p} \approx \frac{b^2}{K}Mass_p$ physically represents the change in fluid storage volume due to mechanical dilation of the pore space, see [93].

   When soil is not saturated, $\boldsymbol{A}_{pp}$ is the dominant term in $\boldsymbol{S}$ and the second term approximation has little impact.Otherwise, in saturated soil, the mass matrix $\boldsymbol{S}$ approximation is the approximation of choice.

2. Approximations of the inverse $\boldsymbol{S}^{-1}$
   We now need to solve the inverse of the Schur approximation chosen. Same techniques as for $\boldsymbol{A}_{\underline{uu}}^{-1}$ are used.

   - In [93] a Krylov method preconditioned with an AMG is used, $\boldsymbol{S}^{-1}(kry, AMG)$.
   - They also point out that in the undrained limit, $\boldsymbol{A}_{pp} \to 0$ and $\boldsymbol{S}_m$ approaches a lumped mass matrix. In this case, a Jacobi preconditioner would be sufficient, $\boldsymbol{S}_m^{-1}(kry, Jacobi)$.

- In [92] a Krylov method preconditioned with ILU(0) is used, $\boldsymbol{S}^{-1}(kry, ILU(0))$.

Norm-dependent preconditioners

Another type of block diagonal preconditioners are parameter based preconditioners, they are also norm-equivalent. For example for the Biot's problem, in [60] a three field parameter-robust preconditioner is presented with robust results parameter wise. This preconditioner relies on an analysis made from the saddle point preconditioner of section 2.2.2. We use the same type of notations, new Hilbert spaces V_p and Q_p for the test functions, with parameter-dependent norms H^1, are defined as follows

$$\|\underline{v}\|_{V_p}^2 := \left\|\underline{\underline{\varepsilon}}(\underline{v})\right\|^2 + \lambda\left\|div(\underline{v})\right\|^2, \qquad \underline{v} \in H^1 \tag{2.36}$$

$$\|q\|_{Q_p}^2 := \frac{b^2}{\lambda}\|q\|^2 + (\kappa\nabla q, \nabla q), \qquad q \in H^1 \tag{2.37}$$

It can be shown that $\boldsymbol{K} : V_p \times Q_p \to V_p^* \times Q_p^*$ is an isomorphism with upper and lower bounds uniform in λ, κ and b. The parameter-robust preconditioner for the two field formulation will then be

$$\boldsymbol{M}_p^{-1} = \begin{bmatrix} (-\operatorname{div}(\underline{\underline{\varepsilon}}(\underline{v})) - \lambda\nabla(\operatorname{div}(\underline{v})))^{-1} & 0 \\ 0 & (\frac{b^2}{\lambda}I + \kappa\nabla)^{-1} \end{bmatrix} \tag{2.38}$$

where the operator I should not be thought of as the identity operator on the Hilbert space L^2, but as the Riesz map between this space and its dual. The corresponding discrete operator is typically represented by the mass matrix.$(-\operatorname{div}(\underline{\underline{\varepsilon}}) - \lambda\nabla(\operatorname{div}()))^{-1}$ and $(\frac{b^2}{\lambda}I + \kappa\nabla)^{-1}$ can be replaced by any spectrally equivalent operator. In [60] they chose corresponding algebraic multigrid preconditioners implemented in the library Hypre [34]. However, for $(\frac{b^2}{\lambda}I + \kappa\nabla)^{-1}$ these preconditioners do not perform well when λ is large. In order to overcome this difficulty, the authors introduce a new unknown called total pressure, $p_T := -\lambda\operatorname{div}(\underline{u}) + bp$. With this new field, $(\frac{b^2}{\lambda}I + \kappa\nabla)^{-1}$ isn't necessary anymore and a new three field preconditioner is proposed. For more on this subject see [60].

A parameter robust preconditioner for Biot is also studied in [1], although the formulation is different as seen before , instead of the stabilization term $c_0 p$ isn't present in the continuous problem, instead they use two stable $\mathcal{P}_1 - \mathcal{P}_1$ elements and the Mini elements with stabilization. This yields a different discrete matrix $\boldsymbol{A}$, but they find their robust preconditioner with the same analysis as [60] and a similar preconditioner as (2.38) but with slightly different parameters. However they use GMRES with a tolerance of 10^{-2} to solve the diagonal blocks of the preconditioner instead of AMG. Their numerical results prove that the convergences with this preconditioner is independent of mesh size, time step,and the physical parameters of the model.

Constraint preconditioners

For an efficient implementation of constraint preconditioners, factorisations of the preconditioner can be applied and $\boldsymbol{A}_{\underline{uu}}$ and $\boldsymbol{S}$ approximations are needed [8, 7].

- Exact constraint preconditioner. For this an approximate inverse AINV [6] is used, where $\boldsymbol{Z}$ is upper triangular:

  $$\boldsymbol{A}_{\underline{uu}}^{-1} = \boldsymbol{Z}\boldsymbol{Z}^t$$

  $\boldsymbol{A}_{\underline{uu}}^{-1}$ is replaced by $\boldsymbol{Z}\boldsymbol{Z}^t$, this approximation is $\boldsymbol{A}_{\underline{uu}}^{-1}(\boldsymbol{Z}\boldsymbol{Z}^t)$. This also leads to $\boldsymbol{S}$ approximation $\boldsymbol{S}_{AINV} = \boldsymbol{A}_{pp} - \boldsymbol{A}_{p\underline{u}}\boldsymbol{Z}\boldsymbol{Z}^t\boldsymbol{A}_{\underline{u}p}$. Even though factorisations don't have good performance scalability wise, algebraic multilevel methods for an approximate factorisation of the inverse can be used [67].

- Inexact constraint preconditioner. On top of the methods seen above, an incomplete triangular factorisation of $\boldsymbol{S}_{AINV}$ can be done

  $$\boldsymbol{S}_{AINV} = \boldsymbol{L}_s\boldsymbol{L}_s^t.$$

  The inexact constraint preconditioner would then be

  $$\boldsymbol{M}_{ICP}^{-1} = \boldsymbol{U}\boldsymbol{L} = \begin{bmatrix} \boldsymbol{Z} & -\boldsymbol{Z}\boldsymbol{Z}^t\boldsymbol{A}_{up}\boldsymbol{L}_s^t \\ 0 & \boldsymbol{L}_s^{-t} \end{bmatrix} \begin{bmatrix} \boldsymbol{Z}^t & 0 \\ \boldsymbol{L}_s^t\boldsymbol{A}_{pu}\boldsymbol{Z}\boldsymbol{Z}^t & -\boldsymbol{L}_s^{-1} \end{bmatrix}$$

- Mixed constraint preconditioner. In this case, $\boldsymbol{S}_{AINV}$ with the incomplete triangular factorisation is used but with an incomplete Cholesky factorisation to approximate $\boldsymbol{A}_{\underline{uu}}$. The idea here is to use a sparse approximation for the construction of the Schur complement and a more precise approximation for $\boldsymbol{A}_{\underline{uu}}$.This preconditioner is presented for poroelasticity in [7].

2.3.4 Multigrid

In the multigrid framework, the choice of the smoother is a key point to ensure the convergence and the performance of the method. Several strategies have proposed such as Vanka-type smoothers for two and three-field non-linear poroelasticity, Uzawa-type smoothers obtained by splitting the discrete operators or also parameter dependent smoothers based on a fixed-stress scheme [62, 63, 46]. An efficient distributive smoother for staggered grids is proposed and analyzed in [94, 45].

We present Vanka-type smoothers, Uzawa-type smoothers and fixed-stress smoothers for the Biot's problem,.

Vanka-type smoothers

Vanka-type smoothers can be considered as block Gauss-Seidel methods, where a block corresponds to all degrees of freedom that are connected with one element. We denote by $\boldsymbol{A}_T$ the block of the matrix A that is connected with the degrees of freedom of the element T.

$$\boldsymbol{A}_T = \begin{bmatrix} \boldsymbol{A}_{\underline{u}\underline{u}_T} & \boldsymbol{A}_{\underline{u}p_T} \\ \boldsymbol{A}_{p\underline{u}_T} & \boldsymbol{A}_{pp_T} \end{bmatrix}$$

The Vanka smoother computes new displacement and pressure values in each element by,

$$\begin{bmatrix} u \\ p \end{bmatrix}_T = \begin{bmatrix} u \\ p \end{bmatrix}_T + \omega \boldsymbol{A}_T^{-1} \left[\begin{bmatrix} f \\ g \end{bmatrix} - A \begin{bmatrix} u \\ p \end{bmatrix} \right]_T$$

where ω is a relaxation parameter.
In [62], Vanka-type smoother are compared for 3 field non-linear poroelasticity problems using a Newton multigrid approach. A smoother called point-wise collective Gauss-Seidel(PGS), where a 3×3 system is solved at each grid point, is compared to the Vanka smoother proposed by Vanka in [89],where a 5X5 system is solved at each grid point. They show that Vanka is more efficient and robust when the values of the coefficient are extremely small or the system is discretized on a very fine grid.

Uzawa smoother

In [63] an Uzawa-type smoother is compared to the Vanka smoother in [89] for the Biot's problem. Is obtained by splitting the discrete operator as follows

$$\begin{bmatrix} \boldsymbol{A}_{\underline{u}\underline{u}} & \boldsymbol{A}_{\underline{u}p} \\ \boldsymbol{A}_{p\underline{u}} & \boldsymbol{A}_{pp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_{\boldsymbol{A}_{\underline{u}\underline{u}}} & 0 \\ \boldsymbol{A}_{p\underline{u}} & -\omega^{-1} I \end{bmatrix} - \begin{bmatrix} \boldsymbol{M}_{\boldsymbol{A}_{\underline{u}\underline{u}}} - \boldsymbol{A}_{\underline{u}\underline{u}} & -\boldsymbol{A}_{\underline{u}p} \\ 0 & -\boldsymbol{A}_{pp} - \omega^{-1} I \end{bmatrix}$$

where $\boldsymbol{M}_{\boldsymbol{A}_{\underline{u}\underline{u}}}$ is a typical smoother for $\boldsymbol{A}_{\underline{u}\underline{u}}$ and ω some positive parameter. If $(\underline{u}, p)$ is an approximation of the solution, the relaxed approximation $(\hat{\underline{u}}, \hat{p})$ is

$$\begin{bmatrix} \boldsymbol{M}_{\boldsymbol{A}_{\underline{u}\underline{u}}} & 0 \\ \boldsymbol{A}_{p\underline{u}} & -\omega^{-1} I \end{bmatrix} \begin{bmatrix} \hat{\underline{u}} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_{\boldsymbol{A}_{\underline{u}\underline{u}}} - \boldsymbol{A}_{\underline{u}\underline{u}} & -\boldsymbol{A}_{\underline{u}p} \\ 0 & -\boldsymbol{A}_{pp} - \omega^{-1} I \end{bmatrix} \begin{bmatrix} \underline{u} \\ p \end{bmatrix} + \begin{bmatrix} \underline{g} \\ f \end{bmatrix}$$

Here, the symmetric Gauss-Seidel (SGS) method is considered as $\boldsymbol{M}_{\boldsymbol{A}_{\underline{u}\underline{u}}}$ and based on a local fourier analysis (LFA) $\omega = \frac{h^2(\lambda+2\mu)}{5\kappa(\lambda+2\mu)+h^2}$
Numerical results with different parameters for a uniform grid of cell of size $h = \frac{1}{256}$ and a relaxation parameter $\omega = 0.7$ used for the Vanka smoother, because it provides the best multigrid convergence with this smoother, show that whereas the number of iterations is comparable, the Uzawa smoother has a lower computational cost.

Fixed-stress smoother

In [46] a fixed-stress smoother dependent on the physics of the problem is presented for solving Biot's consolidation problem. The smoother is based on the fixed-stress scheme

$$\begin{bmatrix} \boldsymbol{A}_{\underline{uu}} & \boldsymbol{A}_{\underline{up}} \\ \boldsymbol{A}_{p\underline{u}} & \boldsymbol{A}_{pp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{\underline{uu}} & \boldsymbol{A}_{\underline{up}} \\ 0 & \boldsymbol{A}_{pp} + \frac{b^2}{K}\boldsymbol{M}_p \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ -\boldsymbol{A}_{p\underline{u}} & \frac{b^2}{K}\boldsymbol{M}_p \end{bmatrix}$$

where $\boldsymbol{M}_p$ is the pressure mass matrix. The two physics-based smoothers related to the splitting algorithm described above are

$$\boldsymbol{M}_{upperFS} = \begin{bmatrix} \boldsymbol{M}_{\boldsymbol{A}_{\underline{uu}}} & \boldsymbol{A}_{\underline{up}} \\ 0 & \boldsymbol{M}_S \end{bmatrix} \tag{2.39}$$

and

$$\boldsymbol{M}_{diagFS} = \begin{bmatrix} \boldsymbol{M}_{\boldsymbol{A}_{\underline{uu}}} & 0 \\ 0 & \boldsymbol{M}_S \end{bmatrix} \tag{2.40}$$

where $\boldsymbol{M}_{\boldsymbol{A}_{\underline{uu}}}$ and $\boldsymbol{M}_S$ are suitable smoothers for $\boldsymbol{A}_{\underline{uu}}$ and $\boldsymbol{S} = \boldsymbol{A}_{pp} + \frac{b^2}{K}\boldsymbol{M}_p$. They use a two iterations SGS method for $\boldsymbol{M}_{\boldsymbol{A}_{\underline{uu}}}$ and one iteration SGS method for $\boldsymbol{M}_S$. The performance of the smoothers is very satisfactory independently of the values of the physical problem.

Chapter 3

Thermo-Hydro-Mechanics

Contents

This chapter covers some of the points and results from the article "Scalable block preconditioners for saturated Thermo-Hydro-Mechanics problems", that is currently accepted with major reviews in the journal Advanced Modeling and Simulation in Engineering Sciences (AMSES).

We consider a porous deformable solid material saturated with an almost incompressible fluid with non-isothermal effects. Though the material is considered elastic and the fluid obeys Darcy's law, the energy balance equation is nonlinear and drives ourselves to a the use of the Newton's method. We use a monolithic solution approach and focus on the use of block preconditioned Krylov methods in order to solve the linearized system, whose unknowns are the displacement, the pressure and the temperature of the continuum. In the present work, we consider a two field, displacement and pressure, formulation with respective quadratic and linear interpolations. It has indeed shown to be stable on several severe tests [54]. The temperature field has linear interpolation and this choice will be discussed in the sequel.

We start by presenting the general framework of the THM system in some detail in order to show how the couplings with temperature through diffusion and convection mechanisms are handled. After writing the weak formulation, time and space discretizations are presented followed by the linearization of the residual. The block preconditioning strategy is discussed and its robustness with respect to the number of DoF and to the physical parameters are assessed on a test case. The weak and strong scaling are also evaluated. The solution of a large industrial problem by the proposed framework is discussed. The paper ends with conclusions and outlooks.

3.1 The THM equations

Due to the large quantity of parameters, all symbols and units used in the article are listed in Table 3.1.

3.1.1 General framework and system

An isotropic saturated mono-phased porous media is considered in the context of small perturbations. According to Biot's theory, it is modeled as a linear elastic solid skeleton with pores containing a freely moving fluid. Due to the presence of the pores, an essential characteristics of the medium is the porosity, named φ in the sequel. It is the ratio between the volume of the void and the total volume of the medium. The latter is considered on a macroscopic scale within the framework of continuum mechanics and we therefore assume that the representative elementary volume includes a sufficient volume of grains and void space to verify this assumption. The above volumes are expressed in the current configuration so that φ is often referred to as the Eulerian porosity.

Another essential parameter of the medium is the Biot's coefficient b. It is the ratio of the volume of fluid gained or lost in a specimen under load to the change in volume of that specimen, when the pore pressure remains constant [10]. Given the solid matrix bulk modulus K_s and the bulk modulus of the drained medium K_0, it

Table 3.1: Parameters

Symbol	Definition	Unit
$\underline{\underline{\underline{\underline{A}}}}$	Forth order Hooke's tensor	Pa
$\underline{\underline{\varepsilon}}$	Strain tensor	-
E	Young's modulus	Pa
ν	Poisson's ratio	-
K_0	Drained bulk modulus of the continuum	Pa
K_l	Bulk modulus of the fluid	Pa
K_s	Bulk modulus of the solid matrix	Pa
K_{int}	Intrinsic permeability	
φ	Porosity	-
μ_l	Fluid dynamic viscosity	Pa.s
h_f	Specific enthalpy of the fluid	$\mathrm{J.kg^{-1}}$
h_{f0}	Initial specific enthalpy of the fluid	$\mathrm{J.kg^{-1}}$
p_{atm}	Atmospheric pressure	Pa
C_s	Specific heat of the solid	$\mathrm{J.kg^{-1}.K^{-1}}$
C_f	Specific heat of the fluid	$\mathrm{J.kg^{-1}.K^{-1}}$
C_f^p	Specific heat of the fluid with constant pressure	$\mathrm{J.kg^{-1}.K^{-1}}$
C_ϵ^0	Specific heat of the medium to constant deformation	$\mathrm{J.K^{-1}.m^{-3}}$
C_σ^0	Specific heat of the medium to constant constraint	$\mathrm{J.K^{-1}.m^{-3}}$
ρ_s	Solid density	$\mathrm{kg.m^{-3}}$
ρ_f	Fluid density	$\mathrm{kg.m^{-3}}$
ρ_m	Medium density	$\mathrm{kg.m^{-3}}$
λ_H	Hydraulic conductivity	$\mathrm{Pa^{-1}.m^{\mathbf{2}}.s^{-1}}$
λ_T	Thermal conductivity	$\mathrm{W.m^{-1}.K}$
T_0	Temperature of reference	K
α_s	Dilation coefficient of the solid	$\mathrm{K^{-1}}$
α_l	Dilation coefficient of the fluid	$\mathrm{K^{-1}}$
α_m	Homogenized dilation coefficient of the medium	$\mathrm{K^{-1}}$

expresses as $b = 1 - \frac{K_0}{K_s}$. We shall suppose that the solid matrix does not undergo significant volume changes, which is the case for the soft soils we consider here ; it results in $b = 1$, that will be used in the sequel.

Besides the aforementioned bulk moduli, the medium is also characterized from a material point of view by the following parameters :

- the hydraulic permeability λ_H. It measures the medium's ability to transmit a given fluid. It is the ratio between the intrinsic permeability named K_{int} and the fluid viscosity μ_l.
- the thermal conductivity named λ_T. It measures the medium's ability to conduct heat.
- the specific enthalpy of the water h_f represents the enthalpy of the fluid per unit mass. It is the sum of the specific internal energy of the fluid and the product of the pressure and the specific volume.

These parameters, some of which appear explicitly therein, are of major importance in the balance equations. They are three in number since the medium is saturated and mono-phased : the linear momentum, the mass of fluid and the energy of the medium.

Balance equations

We introduce the three balance equations of the problem that constitutes the thermo-hydromechanical model of the ground. This model follows the work of Coussy [28].

- The linear momentum equation:

$$-\operatorname{div}(\underline{\underline{\sigma}}) = \underline{f}^e$$

  - $\underline{\underline{\sigma}}$ denotes the total Cauchy stress tensor
  - $\underline{f}^e$ denotes the total external forces

- The water mass conservation:

$$\dot{m}_f + \operatorname{div}(\underline{\psi}) = 0$$

  - m_f denotes the fluid mass of the continuum
  - $\underline{\psi}$ is the fluid mass flux

- The energy conservation;

$$h_f \dot{m}_f + \dot{Q}' + \operatorname{div}(h_f \underline{\psi}) + \operatorname{div}(\underline{q}) = \Theta$$

- h_f denotes the specific fluid enthalpy
- $\underline{q}$ denotes the heat flux
- Θ denotes the source/sink of heat
- Q' denotes the heat in the medium that is not convected, the heat input that doesn't come from an outside source.

Let us now detail the above balance equation in order to reveal the couplings between the phenomena involved.

The balance of linear momentum
The mechanics equilibrium equations is applied on the total stress $\underline{\underline{\sigma}}$

$$-\operatorname{div}(\underline{\underline{\sigma}}) = \underline{f}^e \tag{3.1}$$

where $\underline{f}^e$ denotes the total volume external forces. Biot's definition of effective stress $\underline{\underline{\sigma}}' = \underline{\underline{\sigma}} + p\underline{\underline{I}}$ with the tension positive-sign convention is used in this manuscript (we recall that $b = 1$ in the previous equation). After including it in equation (3.1), we get:

$$-\operatorname{div}(\underline{\underline{\sigma}}') + \nabla p = \underline{f}^e \tag{3.2}$$

We shall now use the expression of the constitutive equation while taking into account the thermal expansion of the medium $\underline{\underline{\varepsilon}}^{th}$, which is a function of T :

$$\underline{\underline{\sigma}}' = \underline{\underline{\underline{\underline{A}}}} : (\underline{\underline{\varepsilon}}(\underline{u}) - \underline{\underline{\varepsilon}}^{th}(T)) \tag{3.3}$$

$$= \underline{\underline{\underline{\underline{A}}}} : (\underline{\underline{\varepsilon}}(\underline{u}) - \alpha_s(T - T_0)\underline{\underline{I}}) \tag{3.4}$$

$$= \underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u}) - 3K_s\alpha_s(T - T_0)\underline{\underline{I}} \tag{3.5}$$

$\underline{\underline{\underline{\underline{A}}}}$ is the forth order Hooke's tensor (which is a function of E and ν), $K_s = \frac{E}{3(1-2\nu)}$ is the bulk modulus of the solid matrix, α_s is the thermal expansion coefficient and T_0 is the reference temperature (temperature at equilibrium).

If we inject the constitutive law into equation (3.2), we obtain the expression, where all couplings become explicit :

$$-\operatorname{div}(\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u})) + \nabla p + 3K_s\alpha_s\nabla T = \underline{f}^e \tag{3.6}$$

The conservation of water mass
The water mass conservation equation is

$$\dot{m}_f + \operatorname{div}(\underline{\psi}) = 0 \tag{3.7}$$

We shall now inject in the above equation several hypothesis on the fluid behaviour. First, we call the domain's initial porosity φ^0 and the fluid's initial density ρ_f^0. Then the total fluid mass expresses with respect to this initial state :

$$m_f = (1 + \operatorname{div}(\underline{u}))\rho_f\varphi - \rho_f^0\varphi^0$$

The time derivative of the fluid mass is then given by :

$$\dot{m}_f = \rho_f\varphi \operatorname{div}(\dot{\underline{u}}) + (1 + \operatorname{div}(\underline{u}))\varphi\dot{\rho}_f + \rho_f(1 + \operatorname{div}(\underline{u}))\dot{\varphi} \tag{3.8}$$

$$= \rho_f\varphi \operatorname{div}(\dot{\underline{u}}) + \varphi\dot{\rho}_f + \rho_f\dot{\varphi} \tag{3.9}$$

where we used $\operatorname{div}(\underline{u}) << 1$ since we make the assumption of small displacements. Next, We use the definition of the time derivative of the fluid's density [28] :

$$\dot{\rho}_f = \rho_f(\frac{1}{K_l}\dot{p} - 3\alpha_l\dot{T}) \tag{3.10}$$

We now turn our attention to the evolution of the porosity. By using the definition of the Eulerian porosity φ related to the Lagrangian porosity by $\phi = (1 + \operatorname{div}(\underline{u}))\varphi$, its expression $\phi = \operatorname{div}(\underline{u}) + (1-\varphi)\frac{p}{K_s} + 3(1-\varphi)\alpha_s T$ in the THM context [25] and the incompressibility of the solid matrix, we have :

$$\dot{\varphi} = (1-\varphi)(\operatorname{div}(\dot{\underline{u}}) - 3\alpha_s\dot{T} + \frac{\dot{p}}{K_s}) \tag{3.11}$$

$$= (1-\varphi)(\operatorname{div}(\dot{\underline{u}}) - 3\alpha_s\dot{T}) \tag{3.12}$$

The fluid mass apport is then

$$\begin{aligned}\dot{m}_f &= \rho_f(\varphi \operatorname{div}(\dot{\underline{u}}) + \frac{\varphi}{K_l}\dot{p} - 3\alpha_l\varphi\dot{T} + (1-\varphi)(\operatorname{div}(\dot{\underline{u}}) - 3\alpha_s\dot{T})) \\ &= \rho_f(\operatorname{div}(\dot{\underline{u}}) + \frac{\varphi}{K_l}\dot{p} - 3\alpha_m\dot{T})\end{aligned}$$

Where $3\alpha_m = (3\alpha_l\varphi + 3(1-\varphi)\alpha_s)$.
Finally, we take into consideration Darcy's law where the effect of gravity is neglected in coherence with the targeted application (the theory does not need this assumption which is just a short simplification)

$$\underline{\psi} = -\rho_f\lambda_H\nabla p$$

The final water mass conservation equation (3.7) is then

$$\rho_f(\operatorname{div}(\dot{\underline{u}}) + \frac{\varphi}{K_l}\dot{p} - 3\alpha_m\dot{T}) - \operatorname{div}(\rho_f\lambda_H\nabla p) = 0 \tag{3.13}$$

The energy conservation
The energy conservation equation is

$$h_f \dot{m}_f + \operatorname{div}(h_f \underline{\psi}) + \operatorname{div}(\underline{q}) + \dot{Q}' = \Theta \tag{3.14}$$

Θ denotes the total sources of heat and it equals the four terms on the left-hand side : the heat coming from the fluid enthalpy, the energy convected by the fluid, the heat flux and the non-convective heat. We shall start by detailing the latter.
The non-convective heat Q' is the thermal input received by the system excluding the enthalpy contribution of the fluid. It is the sum of three terms of heat input due respectively to the deformation of the solid matrix, to the fluid compression and to temperature variation. It is a non-linear term whose expression is :

$$\dot{Q}' = 3K_0\alpha_s \operatorname{div}(\underline{\dot{u}})T - 3\alpha_l \dot{p} T + C_\epsilon^0 \dot{T}$$

By developing the specific heat of the medium to constant deformation, we get $C_\epsilon^0 = C_\sigma^0 - 9TK_0\alpha_s^2$ [26, p78] .

$$\dot{Q}' = (3K_0\alpha_s \operatorname{div}(\underline{\dot{u}}) - 3\alpha_l \dot{p} - 9K_0\alpha_s^2 \dot{T})T + C_\sigma^0 \dot{T}$$

By replacing $\dot{Q}'$, $\dot{m}_f$, $\underline{\psi}$ and using the fact that the heat diffusion follows Fourier's law $\underline{q} = -\lambda_T \nabla T$, equation (3.14) becomes

$$\begin{aligned}\rho_f h_f(\operatorname{div}(\underline{\dot{u}}) + \frac{\varphi}{K_l}\dot{p} - 3\alpha_m \dot{T}) - \operatorname{div}(\rho_f h_f \lambda_H \nabla p) \\ + (3K_0\alpha_s \operatorname{div}(\underline{\dot{u}}) - 3\alpha_l \dot{p} - 9K_0\alpha_s^2 \dot{T})T + C_\sigma^0 \dot{T} - \operatorname{div}(\lambda_T \nabla T) = \Theta\end{aligned}$$

The final system

After detailing each balance equation in order to reveal the detailed coupling between the phenomena involved, the final system is obtained.

Let Ω be a d dimensional domain, $1 \leq d \leq 3$, and t_f the final time of the simulation. The THM model describes the evolution of 3 primal unknowns: the vector displacement field, $\underline{u}(x,t)$, the fluid pressure field, $p(x,t)$, the temperature field $T(x,t)$.

The coupled system consists of , $\forall x \in \Omega$ and $\forall t > 0 \in [0, t_f]$:

$$-\operatorname{div}(\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u})) + \nabla p + 3K_s\alpha_s\nabla T = \underline{f}^e \qquad \text{in } \Omega \times (0, t_f)$$

$$-\operatorname{div}(\rho_f\lambda_H\nabla p) + \rho_f(\operatorname{div}(\underline{\dot{u}}) + \frac{\varphi}{K_l}\dot{p} - \alpha_m 3\dot{T}) = 0 \qquad \text{in } \Omega \times (0, t_f)$$

$$\begin{aligned} -\operatorname{div}(\lambda_T\nabla T) - \operatorname{div}(\rho_f h_f \lambda_H \nabla p) \\ +\rho_f h_f(\operatorname{div}(\underline{\dot{u}}) + \frac{\varphi}{K_l}\dot{p} - \alpha_m 3\dot{T}) \\ +(3K_0\alpha_s \operatorname{div}(\underline{\dot{u}}) - 3\alpha_m\dot{p} - 9K_0\alpha_s^2\dot{T})T + C_\sigma^0\dot{T} = \Theta \end{aligned} \qquad \text{in } \Omega \times (0, t_f)$$

The boundary of Ω is denoted $\partial\Omega$ and 6 different partitions are needed to define the boundary conditions. For each primal unknown, we may define Dirichlet and Neumann boundary conditions, say the displacement $\underline{u}$ and the stress $\underline{\underline{\sigma}}$, the pressure p and the fluid flux q, the temperature T and the thermal flux Ψ.
We thus have, respectively, the boundary conditions on the displacement unknowns, on the pressure unknowns and on the temperature unknowns such as :

$$\partial\Omega = \partial\Omega^{\underline{u}} \cup \partial\Omega^{\underline{t}} \text{ with } \partial\Omega^{\underline{u}} \cap \partial\Omega^{\underline{t}} = \emptyset$$
$$\partial\Omega = \partial\Omega^{p} \cup \partial\Omega^{q} \text{ with } \partial\Omega^{p} \cap \partial\Omega^{q} = \emptyset$$
$$\partial\Omega = \partial\Omega^{T} \cup \partial\Omega^{\Psi} \text{ with } \partial\Omega^{T} \cap \partial\Omega^{\Psi} = \emptyset$$

The boundary conditions are given by :

$$\begin{aligned}
\underline{\underline{\sigma}}(\underline{u}) \cdot \underline{n} &= \underline{t}^e && \text{on } \partial\Omega^{\underline{t}} \times (0, t_f) \\
-\lambda_H\nabla p \cdot \underline{n} &= q^e && \text{on } \partial\Omega^{q} \times (0, t_f) \\
-\lambda_T\nabla T \cdot \underline{n} &= \Psi^e && \text{on } \partial\Omega^{\Psi} \times (0, t_f) \\
\underline{u} &= \underline{u}^e && \text{on } \partial\Omega^{\underline{u}} \times (0, t_f) \\
p &= p^e && \text{on } \partial\Omega^{p} \times (0, t_f) \\
T &= T^e && \text{on } \partial\Omega^{T} \times (0, t_f) \\
\underline{u}(\,, 0) &= \underline{u}_0 && \text{in } \Omega \\
p(\,, 0) &= p_0 && \text{in } \Omega \\
T(\,, 0) &= T_0 && \text{in } \Omega
\end{aligned}$$

where $\underline{n}$ is the outward normal.
Furthermore, the material parameters' definitions are given in Table 3.1.

3.1.2 Linearization and discretization

The next step to solve the non-linear time-dependent THM system is to do the time and space discretization, as well as the linearization.

Variational formulation

We define the Sobolev spaces

$$\begin{aligned} U(\Omega) &= \{\underline{u} \in (H^1(\Omega))^d, \underline{u} = \underline{u}^e \quad \text{on} \quad \partial\Omega^{\underline{u}}\}, \\ P(\Omega) &= \{p \in H^1(\Omega), p = p^e \quad \text{on} \quad \partial\Omega^p\}, \\ T(\Omega) &= \{T \in H^1(\Omega), T = T^e \quad \text{on} \quad \partial\Omega^T\}, \end{aligned}$$

By considering the appropriate Sobolev spaces defined above and by integration by parts, we have the following weak form:
Find $(\underline{u}, p, T) \in U(\Omega) \times P(\Omega) \times T(\Omega)$ such as for all $(\underline{v}, q, W) \in U(\Omega) \times P(\Omega) \times T(\Omega)$, we have

$$\int_\Omega (-\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v}) + p \ \mathrm{div}(\underline{v}) + 3K_s\alpha_s T \ \mathrm{div}(\underline{v}))dx = \int_\Omega \underline{f}^e \, \underline{v} \, dx + \int_{\partial\Omega^{\underline{t}}} \underline{t}^e \, \underline{v} \, ds$$

$$\int_\Omega \rho_f(-\lambda_H \nabla p \nabla q + \ \mathrm{div}(\dot{\underline{u}}) \, q + \frac{\varphi}{K_l} \dot{p} \, q - \alpha_m 3 \dot{T} \, q)dx = \int_{\partial\Omega^q} \rho_f \, q^e \, q \, ds \tag{3.15}$$

$$\begin{aligned} \int_\Omega (&-\lambda_T \nabla T \nabla W + C_\sigma^0 \dot{T} W \\ &+ \rho_f h_f(-\lambda_H \nabla p \nabla W + \ \mathrm{div}(\dot{\underline{u}}) \, W + \frac{\varphi}{K_l} \dot{p} \, W - \alpha_m 3 \dot{T} \, W) \\ &+ T(3K_0\alpha_s \ \mathrm{div}(\dot{\underline{u}}) \, W - 3\alpha_m \dot{p} \, W - 9K_0\alpha_s^2 \dot{T} \, W))dx = \int_\Omega \Theta \, W \, dx + \int_{\partial\Omega^\Psi} \Psi^e \, W \, ds \end{aligned}$$

Time discretization

To solve the THM time-dependent problem, we chose an implicit Euler method to discretize the problem in time and solve a static problem at each time step.
We use the notations $\underline{u}^n(x) := \underline{u}(x, t^n)$, $p^n(x) := p(x, t^n)$ and $T^n(x) := T(x, t^n)$ which denote the displacement field, the pressure field and the temperature field at $t^n = n\Delta t$, where Δt is a given time increment.
To apply Euler's implicit method, $\partial_t \mathbf{u}$(the same goes for $\dot{p}$ and $\dot{T}$) is replaced by:

$$\dot{\underline{u}}(x, t^{n+1}) = \frac{\underline{u}(x, t^{n+1}) - \underline{u}(x, t^n)}{t^{n+1} - t^n} = \frac{\underline{u}^{n+1}(x) - \underline{u}^n(x)}{\Delta t} \tag{3.16}$$

The THM semi-discrete weak formulation becomes

$$\int_\Omega (\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u}^{n+1}) : \underline{\underline{\varepsilon}}(\underline{v}) - \operatorname{div}(\underline{v})\, p^{n+1} - 3K_s\alpha_s \operatorname{div}(\underline{v})\, T^{n+1})dx = \int_\Omega \underline{f}^e\, \underline{v}\, dx + \int_{\partial\Omega^{\underline{t}}} \underline{t}^e\, \underline{v}\, ds$$

$$\int_\Omega \rho_f \operatorname{div}(\underline{u}^{n+1})\, q + \frac{\varphi}{K_l}\, p^{n+1}\, q - \alpha_m 3\, T^{n+1}\, q + \Delta t \lambda_H\, \nabla p^{n+1}\, \nabla q dx = \int_{\partial\Omega^q} \rho_f\, q^e\, q\, ds + \int_\Omega \rho_f\, (\operatorname{div}(\underline{u}^n)\, q + \frac{\varphi}{K_l}\, p^n\, q - \alpha_m 3\, T^n\, q) dx$$

$$\begin{aligned}
&\int_\Omega -\Delta t \lambda_T\, \nabla T^{n+1}\, \nabla W + C_\sigma^0\, T^{n+1}\, W \\
&+\rho_f h_f(-\Delta t \lambda_H\, \nabla p^{n+1}\, \nabla W + \operatorname{div}(\underline{u}^{n+1})\, W + \frac{\varphi}{K_l}\, p^{n+1}\, W - \alpha_m 3\, T^{n+1}\, W) \\
&+T^{n+1}\Big(3K_0\alpha_s \operatorname{div}(\underline{u}^{n+1})\, W - 3\alpha_m\, p^{n+1}\, W - 9K_0\alpha_s^2\, T^{n+1}\, W\Big) dx \\
&= \Delta t \int_\Omega \Theta\, W\, dx + \int_{\partial\Omega^\Psi} \Psi^e\, W\, ds \\
&+ \int_\Omega C_\sigma^0\, T^n\, W + \rho_f h_f \Big(\operatorname{div}(\underline{u}^n)\, W + \frac{\varphi}{K_l}\, p^n\, W - \alpha_m 3\, T^n\, W\Big) \\
&+T^n\Big(3K_0\alpha_s \operatorname{div}(\underline{u}^n)\, W - 3\alpha_m\, p^n\, W - 9K_0\alpha_s^2\, T^n\, W\Big) dx
\end{aligned}$$

Linearization and Newton's method

The system is linearized using Newton's method and requires the partial derivatives of each equation residue with respect to $\underline{u}^{n+1}$, p^{n+1} and T^{n+1}.
Let's introduce the residue notation for the displacement

$$R_{\underline{u}} := \int_\Omega (\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u}^{n+1}) : \underline{\underline{\varepsilon}}(\underline{v}) - \operatorname{div}(\underline{v})\, p^{n+1} - 3K_s\alpha_s \operatorname{div}(\underline{v})\, T^{n+1})dx - \int_\Omega \underline{f}^e\, \underline{v}\, dx + \int_{\partial\Omega^{\underline{t}}} \underline{t}^e\, \underline{v}\, ds$$

Where $R_{\underline{u}}$ is a function of $((\underline{u}^{n+1}, p^{n+1}, T^{n+1}), (\underline{u}^n, p^n, T^n), \underline{v})$. The same goes for the pressure residue R_p and the temperature residue R_T.
Newton's method requires to find a correction $(\delta_{\underline{u}}, \delta_p, \delta_T)$ solution of

$$\boldsymbol{J} \begin{bmatrix} \delta_{\underline{u}} \\ \delta_p \\ \delta_T \end{bmatrix} = \begin{bmatrix} \frac{\partial R_{\underline{u}}}{\partial \underline{u}} & \frac{\partial R_{\underline{u}}}{\partial p} & \frac{\partial R_{\underline{u}}}{\partial T} \\ \frac{\partial R_p}{\partial \underline{u}} & \frac{\partial R_p}{\partial p} & \frac{\partial R_p}{\partial T} \\ \frac{\partial R_T}{\partial \underline{u}} & \frac{\partial R_T}{\partial p} & \frac{\partial R_T}{\partial T} \end{bmatrix} \begin{bmatrix} \delta_{\underline{u}} \\ \delta_p \\ \delta_T \end{bmatrix} = - \begin{bmatrix} R_{\underline{u}} \\ R_p \\ R_T \end{bmatrix} \tag{3.17}$$

where $\boldsymbol{J}$ is the residue's Jacobian.
The solution is then updated,

$$\underline{u}_k^{n+1} = \underline{u}_{k-1}^{n+1} + \delta_{\underline{u}} \tag{3.18}$$
$$p_k^{n+1} = p_{k-1}^{n+1} + \delta_p \tag{3.19}$$
$$T_k^{n+1} = T_{k-1}^{n+1} + \delta_T \tag{3.20}$$

until the stopping criterion is reached

$$\frac{\left\|\underline{r_k}\right\|}{\left\|\underline{r_0}\right\|} < 10^{-6}$$

where $\underline{\delta} = \begin{bmatrix} \delta_{\underline{u}} \\ \delta_p \\ \delta_T \end{bmatrix}$ and $\underline{r_k} = \begin{bmatrix} R_{\underline{u}} \\ R_p \\ R_T \end{bmatrix}$ at the k^{th} iteration.
In order to solve the system (3.17), we need to find $\boldsymbol{J}$ by linearizing the three residues. Since the first two equations of the system are linear, the first two equations of the linearized system are

$$\begin{aligned} &\frac{\partial R_{\underline{u}}}{\partial \underline{u}^{n+1}}\,\delta_{\underline{u}} + \frac{\partial R_{\underline{u}}}{\partial p^{n+1}}\,\delta_p + \frac{\partial R_{\underline{u}}}{\partial T^{n+1}}\,\delta_T = \\ &\int_\Omega (\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\delta_{\underline{u}}) : \underline{\underline{\varepsilon}}(\underline{v}) - \ \mathrm{div}(\underline{v})\,\delta_p - 3K_s\alpha_s\ \mathrm{div}(\underline{v})\,\delta_T)dx \end{aligned} \tag{3.21}$$

$$\begin{aligned} &\frac{\partial R_p}{\partial \underline{u}^{n+1}}\,\delta_{\underline{u}} + \frac{\partial R_p}{\partial p^{n+1}}\,\delta_p + \frac{\partial R_p}{\partial T^{n+1}}\,\delta_T = \\ &\int_\Omega \rho_f\,(\,\mathrm{div}(\delta_{\underline{u}})\,q + \frac{\varphi}{K_l}\,\delta_p\,q - \alpha_m 3\,\delta_T\,q + \Delta t\lambda_H\,\nabla\delta_p\,\nabla q))dx \end{aligned} \tag{3.22}$$

We linearize each term of the third equation using $\frac{\partial h_f}{\partial p} = \frac{(1-3\alpha_l T)}{\rho_f}$ and $\frac{\partial h_f}{\partial T} = C_f^p$ [26, 27].

$$\begin{aligned}
\frac{\partial R_T}{\partial \underline{u}^{n+1}}\,\delta_{\underline{u}} &= \int_\Omega (\rho_f h_f\ \mathrm{div}(\delta_{\underline{u}})\,W + T_{k-1}^{n+1}\,3K_0\alpha_s\ \mathrm{div}(\delta_{\underline{u}})\,W)dx \\
\frac{\partial R_T}{\partial p^{n+1}}\,\delta_p &= \int_\Omega ((1-3\alpha_l T_{k-1}^{n+1})\delta_p(-\Delta t\lambda_H\,\nabla p_{k-1}^{n+1}\,\nabla W + \ \mathrm{div}(\underline{u}_{k-1}^{n+1})\,W + \frac{\varphi}{K_l}\,p_{k-1}^{n+1}\,W - \alpha_m 3\,T_{k-1}^{n+1}\,W) \\
&\quad + \rho_f h_f(-\Delta t\lambda_H\,\nabla\delta_p\,\nabla W + \frac{\varphi}{K_l}\,\delta_p\,W) - T_{k-1}^{n+1}3\alpha_m\,\delta_p\,W)dx \\
\frac{\partial R_T}{\partial T^{n+1}}\,\delta_T &= \int_\Omega (-\Delta t\lambda_T\,\nabla\delta_T\,\nabla W + C_\sigma^0\,\delta_T\,W \\
&\quad + \rho_f C_f^p\delta_T(-\Delta t\lambda_H\,\nabla p_{k-1}^{n+1}\,\nabla W + \ \mathrm{div}(\underline{u}_{k-1}^{n+1})\,W + \frac{\varphi}{K_l}\,p_{k-1}^{n+1}\,W - \alpha_m 3\,T_{k-1}^{n+1}\,W) \\
&\quad - \rho_f h_f(\alpha_m 3\,\delta_T\,W) + \delta_T(3K_0\alpha_s\ \mathrm{div}(\underline{u}_{k-1}^{n+1})\,W - 3\alpha_m\,p_{k-1}^{n+1}\,W - 18K_0\alpha_s^2\,T_{k-1}^{n+1}\,W))dx
\end{aligned}$$

Space discretization

The finite element method is used for space discretization and Taylor-Hood $P2$-$P1$-$P1$ finite elements are considered. This translates into using continuous piecewise quadratics to approximate the displacement and continuous piecewise linears to approximate the pressure and the temperature. In [33], these elements where studied for poroelasticity and having the polynomial interpolation for the displacement be one degree higher than for the pressure, equilibrates the convergence rate of all terms in the energy norm. Furthermore the convergence is robust with respect to the mesh size. As mentioned in the introduction, this choice has also shown to be stable on several severe tests [54].
The choice of the $P1$ interpolation of the temperature relies on the fact that this field is directly used in (3.3) for the evaluation of the thermal expansion of the medium. Since the latter is substracted to the mechanical strain computed as the symmetric gradient of the displacement, this choice insures consistency of the interpolations and avoid non-physical artefacts in the case where the temperature is interpolated with the same shape functions as the displacement, as pointed in [30, p.104].

Let $U_h(\Omega)$ be the discrete Sobolev subspace of $U(\Omega)$ of dimension N_u with $h > 0$ a parameter that refers to the mesh size, the same formulation goes for $P_h(\Omega)$ and $T_h(\Omega)$.
Let $\{\phi_{\underline{v}_j}\}_{j=1}^{N_u}$ be a basis for the finite element space $U_h(\Omega)$, for all $\underline{v}_h \in U_h(\Omega)$ we have

$$\underline{v}_h = \sum_{j=1}^{N_u} v_j \, \phi_{\underline{v}_j}$$

Let $\{\phi_{q_j}\}_{j=1}^{N_p}$ be a basis for the finite element space $P_h(\Omega)$, for all $q_h \in U_h(\Omega)$ we have

$$q_h = \sum_{j=1}^{N_p} q_j \, \phi_{q_j}$$

Let $\{\phi_{w_j}\}_{j=1}^{N_T}$ be a basis for the finite element space $T_h(\Omega)$. Then for all $W_h \in T_h(\Omega)$ we have

$$W_h = \sum_{j=1}^{N_T} W_j \, \phi_{w_j}$$

The discrete problem is, find $(\delta_{\underline{u}_h}, \delta_{p_h}, \delta_{T_h}) \in U_h(\Omega) \times P_h(\Omega) \times T_h(\Omega)$ such as for all $(\underline{v}_h, q_h, W_h) \in U_h(\Omega) \times P_h(\Omega) \times \mathcal{T}_h(\Omega)$ and the linear system becomes By injecting the above discretized fields in the variational formulation, the following linear system is obtained :

$$\begin{bmatrix} \boldsymbol{J}_{\underline{u}\underline{u}} & \boldsymbol{J}_{\underline{u}p} & \boldsymbol{J}_{\underline{u}T} \\ \boldsymbol{J}_{p\underline{u}} & \boldsymbol{J}_{pp} & \boldsymbol{J}_{pT} \\ \boldsymbol{J}_{T\underline{u}} & \boldsymbol{J}_{Tp} & \boldsymbol{J}_{TT} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_{\underline{u}_h} \\ \boldsymbol{\delta}_{p_h} \\ \boldsymbol{\delta}_{T_h} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{R}_{\underline{u}_h} \\ \boldsymbol{R}_{p_h} \\ \boldsymbol{R}_{T_h} \end{bmatrix} \tag{3.23}$$

where the matrix blocks are given by

$$(\boldsymbol{J}_{\underline{u}\underline{u}})_{ij} = \int_\Omega -\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\phi_{\underline{v}_i}) : \underline{\underline{\varepsilon}}(\phi_{\underline{v}_j})\, dx, \qquad \forall i,j = 1, N_u$$

$$(\boldsymbol{J}_{\underline{u}p})_{ij} = \int_\Omega \phi_{q_j} \operatorname{div}(\phi_{\underline{v}_i})\, dx, \qquad \forall i = 1, N_u, \forall j = 1, N_p$$

$$(\boldsymbol{J}_{uT})_{ij} = \int_\Omega 3K_s\alpha_s\, \phi_{w_j} \operatorname{div}(\phi_{\underline{v}_i})\, dx, \qquad \forall i = 1, N_u, \forall j = 1, N_T$$

$$(\boldsymbol{J}_{p\underline{u}})_{ij} = \int_\Omega \rho_f \phi_{q_i} \operatorname{div}(\phi_{\underline{v}_j})\, dx, \qquad \forall i = 1, N_p, \forall j = 1, N_u$$

$$(\boldsymbol{J}_{pp})_{ij} = \int_\Omega \frac{\varphi}{K_l}\, \rho_f \phi_{q_i}\phi_{q_j} - \rho_f \lambda_H\, \Delta t \nabla\phi_{q_i} \nabla\phi_{q_j}\, dx, \qquad \forall i,j = 1, N_p$$

$$(\boldsymbol{J}_{pT})_{ij} = \int_\Omega -\rho_f\, \alpha_m 3\, \phi_{q_i}\phi_{w_j}\, dx, \qquad \forall i = 1, N_p, \forall j = 1, N_T$$

$$(\boldsymbol{J}_{Tu})_{ij} = \int_\Omega \rho_f h_f\, \phi_{\underline{v}_j} \operatorname{div}(\phi_{w_i}) + T_{k_h}^{n+1} 3K_0\alpha_s\, \phi_{\underline{v}_j} \operatorname{div}(\phi_{w_i})\, dx, \qquad \forall i = 1, N_T, \forall j = 1, N_u$$

$$\begin{aligned} (\boldsymbol{J}_{Tp})_{ij} = \int_\Omega & (1 - 3\alpha_l T_{k_h}^{n+1})\phi_{q_j}(-\Delta t\lambda_H\, \nabla p_{k_h}^{n+1} \nabla\phi_{w_i} + \operatorname{div}(\underline{u}_{k_h}^{n+1})\, \phi_{w_i} \\ & + \frac{\varphi}{K_l}\, p_{k_h}^{n+1}\, \phi_{w_i} - \alpha_m 3\, T_{k_h}^{n+1}, \phi_{w_i}) \\ & + \rho_f h_f(-\Delta t\lambda_H\, \nabla\phi_{q_j}\, \nabla\phi_{w_i} + \frac{\varphi}{K_l}\, \phi_{q_j}\, \phi_{w_i}) \\ & - T_{k_h}^{n+1} 3\alpha_m\, \phi_{q_j}\, \phi_{w_i}\, dx, \end{aligned} \qquad \forall i = 1, N_T, \forall j = 1, N_p$$

$$\begin{aligned} (\boldsymbol{J}_{TT})_{ij} = \int_\Omega & - \Delta t\lambda_T\, \nabla\phi_{w_i} \nabla\phi_{w_j} + C_\sigma^0\, \phi_{w_i}\phi_{w_j} \\ & + \rho_f C_f^p \phi_{w_i}(-\Delta t\lambda_H\, \nabla p_{k_h}^{n+1} \nabla\phi_{w_j} + \operatorname{div}(\underline{u}_{k_h}^{n+1})\phi_{w_j} \\ & + \frac{\varphi}{K_l}\, p_{k_h}^{n+1}\phi_{w_j} - \alpha_m 3\, T_{k_h}^{n+1}\phi_{w_j}) \\ & - \rho_f h_f \alpha_m 3\, \phi_{w_i}\phi_{w_j} \\ & + \phi_{w_i}(3K_0\alpha_s\, \operatorname{div}(\underline{u}_{k_h}^{n+1})\phi_{w_j} - 3\alpha_m\, p_{k_h}^{n+1}\phi_{w_j} \\ & - 18K_0\alpha_s^2\, T_{k_h}^{n+1}\phi_{w_j})\, dx, \end{aligned} \qquad \forall i,j = 1, N_T$$

3.2 Preconditioning

As was seen in chapter 2, iterative solvers often suffer from bad conditioning of the linear system matrix and require preconditioning to achieve satisfactory performance in terms of iteration count and simulation time. This is especially true for THM problems, that are in general ill-conditioned due to the properties of each physical component which are included via parameters into the linear system and the right-hand side.

The linear system to be solved is of the structure

$$\begin{bmatrix} \boldsymbol{J}_{\underline{u}\underline{u}} & \boldsymbol{J}_{\underline{u}p} & \boldsymbol{J}_{\underline{u}T} \\ \boldsymbol{J}_{p\underline{u}} & \boldsymbol{J}_{pp} & \boldsymbol{J}_{pT} \\ \boldsymbol{J}_{T\underline{u}} & \boldsymbol{J}_{Tp} & \boldsymbol{J}_{TT} \end{bmatrix} \begin{bmatrix} \delta_{\underline{u}_h} \\ \delta_{p_h} \\ \delta_{T_h} \end{bmatrix} = - \begin{bmatrix} R_{\underline{u}_h} \\ R_{p_h} \\ R_{T_h} \end{bmatrix}. \tag{3.24}$$

This system is ill-conditioned. Furthermore, there is a significant differences in the order of magnitudes of each parameter (see Table 3.3 in the sequel for illustrative values) which translates into different orders of magnitude between the 2-norms of each physics-based block in the matrix in (3.24)

$$\boldsymbol{S} := \begin{bmatrix} \left\|\boldsymbol{J}_{\underline{u}\underline{u}}\right\|_2 & \left\|\boldsymbol{J}_{\underline{u}p}\right\|_2 & \left\|\boldsymbol{J}_{uT}\right\|_2 \\ \left\|\boldsymbol{J}_{p\underline{u}}\right\|_2 & \left\|\boldsymbol{J}_{pp}\right\|_2 & \left\|\boldsymbol{J}_{pT}\right\|_2 \\ \left\|\boldsymbol{J}_{Tu}\right\|_2 & \left\|\boldsymbol{J}_{Tp}\right\|_2 & \left\|\boldsymbol{J}_{TT}\right\|_2 \end{bmatrix} \approx \begin{bmatrix} 1.\,\mathrm{e}{+13} & 1.\,\mathrm{e}{+01} & 1.\,\mathrm{e}{+06} \\ 1.\,\mathrm{e}{+04} & 1.\,\mathrm{e}{-08} & 1.\,\mathrm{e}{-02} \\ 1.\,\mathrm{e}{+08} & 1.\,\mathrm{e}{-03} & 1.\,\mathrm{e}{+05} \end{bmatrix}. \tag{3.25}$$

We have a maximal scaling difference of 10^{21} between the displacement and pressure block. Solving this system naively could lead to cancellation effects in the solution. Prior scaling or preconditioning of the matrix are thus compulsory.

In this section, we define a preconditioner tailored for our application and discuss the impact of the difference in orders of magnitude between the parameters.

3.2.1 Definition of block preconditioners

Note that the matrix $\boldsymbol{J}$ is non-symmetric. We thus need an iterative solver for non-symmetric systems and choose the *flexible GMRES* (FGMRES) method. We will next define a preconditioner that can be applied to the THM system. The idea behind preconditioning is to construct a matrix $\boldsymbol{P}$ that is a good enough approximation of $\boldsymbol{J}$ but that is easily invertible. In our computations, the preconditioner is applied from the right, which means that we solve the system

$$\boldsymbol{J}\boldsymbol{P}^{-1}y = r,$$

with $y = \boldsymbol{P}x$. The solution x of the system remains the same but if $\boldsymbol{P}$ is a good approximation of $\boldsymbol{J}$, then $\boldsymbol{J}\boldsymbol{P}^{-1}$ becomes 'closer' to the identity and the iterative

method will converge faster than for the unpreconditioned system.

The linear system in (3.24) is of block structure, where each diagonal block corresponds to one of the three physical models. It thus seems natural to choose a block preconditioner, as it has been for example decribed in the reference [64]. The simplest preconditioner is probably the Block Jacobi preconditioner given by

$$\boldsymbol{P}_{Jac} = \begin{bmatrix} \boldsymbol{J}_{\underline{uu}} & 0 & 0 \\ 0 & \boldsymbol{J}_{pp} & 0 \\ 0 & 0 & \boldsymbol{J}_{TT} \end{bmatrix}.$$

The application of a standard Jacobi preconditioner is simple, as it is trivial to invert a diagonal matrix. For the Block Jacobi preconditioner we need the inverses of the three separate individual physics blocks, i.e. $\boldsymbol{J}_{\underline{uu}}^{-1}$, $\boldsymbol{J}_{pp}^{-1}$ and $\boldsymbol{J}_{TT}^{-1}$. This is costly and thus we search for a good approximation of each block, that can be more easily inverted. Before we discuss this further, we introduce our second and third choice for a preconditioner. These are the lower and upper Block Gauss-Seidel preconditioners, denoted by $\boldsymbol{P}_{LGS}$ and $\boldsymbol{P}_{UGS}$, respectively, given by

$$\boldsymbol{P}_{LGS} = \begin{bmatrix} \boldsymbol{J}_{\underline{uu}} & 0 & 0 \\ \boldsymbol{J}_{p\underline{u}} & \boldsymbol{J}_{pp} & 0 \\ \boldsymbol{J}_{T\underline{u}} & \boldsymbol{J}_{Tp} & \boldsymbol{J}_{TT} \end{bmatrix}, \qquad \boldsymbol{P}_{UGS} = \begin{bmatrix} \boldsymbol{J}_{\underline{uu}} & \boldsymbol{J}_{\underline{u}p} & \boldsymbol{J}_{\underline{u}T} \\ 0 & \boldsymbol{J}_{pp} & \boldsymbol{J}_{pT} \\ 0 & 0 & \boldsymbol{J}_{TT} \end{bmatrix}$$

Even though these preconditioners use the rectangular lower (or upper) triangular blocks of the system, when applying their inverse we still only need to compute the inverses of the three diagonal blocks $\boldsymbol{J}_{\underline{uu}}^{-1}$, $\boldsymbol{J}_{pp}^{-1}$ and $\boldsymbol{J}_{TT}^{-1}$. Let $\mathbf{I}$ be the identity matrix of appropriate size for each block. For ease of notation, we do not add the size in the index. The inverse of the lower Gauss-Seidel preconditioner is given by

$$\boldsymbol{P}_{LGS}^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \boldsymbol{J}_{TT}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -\boldsymbol{J}_{T\underline{u}} & -\boldsymbol{J}_{Tp} & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & \boldsymbol{J}_{pp}^{-1} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ -\boldsymbol{J}_{p\underline{u}} & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \boldsymbol{J}_{\underline{uu}}^{-1} & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & I \end{bmatrix}$$

and the inverse of $\boldsymbol{P}_{UGS}$ is given by

$$\boldsymbol{P}_{UGS}^{-1} = \begin{bmatrix} \boldsymbol{J}_{\underline{uu}}^{-1} & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & -\boldsymbol{J}_{\underline{u}p} & -\boldsymbol{J}_{\underline{u}T} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & \boldsymbol{J}_{pp}^{-1} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & -\boldsymbol{J}_{pT} \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \boldsymbol{J}_{TT}^{-1} \end{bmatrix}$$

As mentioned above, for each one of these three preconditioners, we need, in theory, the inverse matrices of the diagonal blocks. In practice, these inverses are never explicitly computed, as this is too costly. Incomplete factorisations of the original matrices are for example well suited but lack scalability. Furthermore, it is not necessary

to have an explicit or precomputed representation of the preconditioner. In iterative methods like for example CG or FGMRES, it is indeed enough to apply the preconditioner in form of matrix-vector products. It is for example possible to completely replace a particular inverse approximation by a linear process. In our later numerical experiments, we will use one V-cycle of an algebraic multigrid solver (AMG) as preconditioner for each one of the three diagonal block matrices. Furthermore, the preconditioner can be replaced by another iterative method, for example GMRES itself, but then the preconditioner becomes non-linear. In this case, we need *flexible* versions of the iterative solvers, as FGMRES, that allow a variable preconditioner at each iteration. A nested approach can also be used, where the preconditioner is implemented in form of some iterations of an iterative solver that is preconditioned itself. In any case, the choice of each approximation will ultimately come down to the specific characteristics of the block to invert. In the following experiments, we use a nested preconditioning approach, where we apply the block-preconditioners $\boldsymbol{P}_{Jac}, \boldsymbol{P}_{LGS}, \boldsymbol{P}_{UGS}$ by using some iterations of the FGMRES method preconditioned by one V-cycle of AMG for each block $\boldsymbol{J}_{\underline{uu}}, \boldsymbol{J}_{pp}, \boldsymbol{J}_{TT}$. This provides the possibility to control the quality of the inverse for each block by defining a stopping tolerance or a fixed number of FGMRES iterations. In this strategy, we have an interplay between the number of outer iterations of FGMRES on the block system with the number of inner FGMRES iterations applied to each diagonal block. A stricter tolerance for the inner FGMRES solvers might lead to a smaller number of outer FGMRES iterations and vice-versa. This choice is guided by the numerical experiments described in the following, where the trade-off between performance and robustness was a primary goal.

The choice of an AMG preconditioner for each inner FGMRES iteration can be explained as follows. The diagonal blocks of the discretized system (see Section 3.1.2) involve elliptic and non-degenerate parabolic operators, for which V-cycle multigrid preconditioners are especially suited [60, 64]. Note that in our particular case, we could as well use the GMRES method, since one V-cycle is a constant linear preconditioner and thus does not require a flexible version. This would come with a small memory gain.

3.2.2 Numerical experiments for scaling issues

Let us first look into matrix scaling and then run numerical experiments to evaluate the impact of the scaling when using the preconditioners defined above.

Matrix scaling

We follow the algorithm given in [58]. The main features of this algorithm are that the scaled matrix becomes diagonally dominant and the scaling becomes block-symmetric. For these kinds of matrices, iterative solvers often converge more easily. The scaling

algorithm takes the 3×3 matrix $\boldsymbol{S}$ in (3.25) and computes the following block-diagonal matrices

$$\boldsymbol{D}_r = \begin{bmatrix} 10^{-7}\,\mathbf{I}_{N_u} & 0 & 0 \\ 0 & 10^{-2}\,\mathbf{I}_{N_p} & 0 \\ 0 & 0 & 10^{-4}\,\mathbf{I}_{N_T} \end{bmatrix},\ \boldsymbol{D}_l = \begin{bmatrix} 10^{-8}\,\mathbf{I}_{N_u} & 0 & 0 \\ 0 & 10^{4}\,\mathbf{I}_{N_p} & 0 \\ 0 & 0 & 10^{-2}\,\mathbf{I}_{N_T}, \end{bmatrix}$$

where $\mathbf{I}_{N_u}, \mathbf{I}_{N_p}, \mathbf{I}_{N_T}$ are the identity matrices of size N_u, N_p, N_T. The THM system is then scaled using $\boldsymbol{D}_r$ and $\boldsymbol{D}_l$ by

$$\boldsymbol{J}^{sc} = \boldsymbol{D}_r\,\boldsymbol{J}\,\boldsymbol{D}_l x_s = b_s$$

with $x_s = \boldsymbol{D}_l^{-1}x$ and $b_s = \boldsymbol{D}_r^{-1}b$ The entries in the scaled system are now of the following magnitudes

$$\begin{bmatrix} \left\|\boldsymbol{J}^{sc}_{\underline{u}\underline{u}}\right\|_2 & \left\|\boldsymbol{J}^{sc}_{\underline{u}p}\right\|_2 & \left\|\boldsymbol{J}^{sc}_{uT}\right\|_2 \\ \left\|\boldsymbol{J}^{sc}_{p\underline{u}}\right\|_2 & \left\|\boldsymbol{J}^{sc}_{pp}\right\|_2 & \left\|\boldsymbol{J}^{sc}_{pT}\right\|_2 \\ \left\|\boldsymbol{J}^{sc}_{Tu}\right\|_2 & \left\|\boldsymbol{J}^{sc}_{Tp}\right\|_2 & \left\|\boldsymbol{J}^{sc}_{TT}\right\|_2 \end{bmatrix} \approx \begin{bmatrix} 1.\,\mathrm{e}{-01} & 1.\,\mathrm{e}{-01} & 1.\,\mathrm{e}{-03} \\ 1.\,\mathrm{e}{-01} & 1.\,\mathrm{e}{-01} & 1.\,\mathrm{e}{-02} \\ 1.\,\mathrm{e}{-03} & 1.\,\mathrm{e}{-02} & 1.\,\mathrm{e}{-01} \end{bmatrix}.$$

Numerical tests

We present numerical results for the above defined nested solvers for the scaled and unscaled linear system. The tolerance of the outer FGMRES solver is set to $\epsilon = 10^{-6}$. This rather large tolerance is used since the linear system is the linearized problem in a Newton fixed point iteration. The Newton iterations are required to converge at a tolerance of $\epsilon_N = 10^{-6}$, so that a stricter tolerance ϵ would be more costly than useful. For the nested preconditioner, we use FGMRES precondioned by one V-cycle of AMG. We have found empirically that using a fixed number of 10 iterations for the displacement block, and three iterations for the pressure and temperature blocks gives a good compromise between the cost of inner and outer iterations with respect to the global computation time. We use the algebraic multigrid solver BoomerAMG from the hypre library through PETSc with its default parameters [34].

Since an analytic solution is not available available for the test problem, we solve the scaled system as precise as possible by using the sparse direct solver MUMPS [2] and use this solution as reference solution. Before computing the errors, the solution x_s of the scaled system is brought back to the original scaling x using the formula $x = \boldsymbol{D}_l x_s$. We use the following notations for the different solution strategies :

xs_0: direct solution of the scaled system,
xs: Iterative solver solution of the scaled system,
x: Iterative solver solution of the initial system.

The relative error with respect to the reference solution $\boldsymbol{D}_l xs_0$ is computed for each physical unknown. With obvious notation, three errors are computed as follows :

- for x

$$\mathrm{err}_{\underline{u}} = \frac{\left\|\boldsymbol{D}_l xs_{0_{\underline{u}}} - x_{\underline{u}}\right\|_2}{\left\|\boldsymbol{D}_l xs_{0_{\underline{u}}}\right\|_2} \quad \mathrm{err}_p = \frac{\left\|\boldsymbol{D}_l xs_{0_p} - x_p\right\|_2}{\left\|\boldsymbol{D}_l xs_{0_p}\right\|_2} \quad \mathrm{err}_T = \frac{\|\boldsymbol{D}_l xs_{0_T} - x_T\|_2}{\|\boldsymbol{D}_l xs_{0_T}\|_2}$$

- for xs

$$\mathrm{err}_{\underline{u}} = \frac{\left\|\boldsymbol{D}_l xs_{0_{\underline{u}}} - \boldsymbol{D}_l xs_{\underline{u}}\right\|_2}{\left\|\boldsymbol{D}_l xs_{0_{\underline{u}}}\right\|_2} \quad \mathrm{err}_p = \frac{\left\|\boldsymbol{D}_l xs_{0_p} - \boldsymbol{D}_l xs_p\right\|_2}{\left\|\boldsymbol{D}_l xs_p\right\|_2} \quad \mathrm{err}_T = \frac{\|\boldsymbol{D}_l xs_{0_T} - \boldsymbol{D}_l xs_T\|_2}{\|\boldsymbol{D}_l xs_T\|_2}$$

We present the simulation time, the number of outer FGMRES iterations and the above defined errors for each one of the three preconditioner $\boldsymbol{P}_{Jac}, \boldsymbol{P}_{LGS}, \boldsymbol{P}_{UGS}$ in Table 3.2. Comparing the effect of the scaling on the solution, there is no clear winner. Indeed, the errors for each unknown are less variable across the preconditioners for the scaled system. In case of the $\boldsymbol{P}_{UGS}$, scaling is even compulsory. Here, solving the unscaled system does not lead to satisfactory results in the displacement and pressure variables. This can be explained by the difference in order of magnitudes in the entries of the matrix blocks and the order in which these are applied in the solution process. In terms of iteration count, FGMRES shows mesh independent convergence for $\boldsymbol{P}_{LGS}$ and we expect this behavior also for the Jacobi preconditioner once the problem size is further increased. In general, FGMRES needs fewer iterations to reduce the residual below the required tolerance for the scaled system. This leads however to higher errors in (almost) each variable when the matrix is preconditioned by $\boldsymbol{P}_{Jac}$ and $\boldsymbol{P}_{LGS}$. The lower iteration count thus does not necessarily present an advantage when interested in the actual error and not the residual.

This numerical experiment suggests that the use of $\boldsymbol{P}_{Jac}$ and $\boldsymbol{P}_{LGS}$ as preconditioners on the unscaled system results in a precise and robust solution strategy. The latter is therefore used in the robustness and scalability studies in the following.

3.3 Solver performance

The robustness and efficiency of the proposed solver are crucial for industrial applications. We thus turn to the presentation of the results of a illustrative test case, challenging the preconditioner's robustness by varying some parameters. The parallel efficiency is also evaluated by weak and strong scalability tests. The method is implemented in code_aster, the massively parallel open source general purpose finite element solver developed at EDF R&D [41].

Table 3.2: Error analysis

	Size	Errors using the scaled system : xs				
		time	**it**	**err**$_{\underline{u}}$	**err**$_p$	**err**$_T$
$\boldsymbol{P}_J$	**429**	1.40e-01	15	3.71e-05	2.96e-04	6.44e-09
	2 437	4.69e-01	15	3.78e-05	2.55e-04	9.82e-09
	117 637	4.55e+01	15	1.34e-04	1.39e-04	7.01e-08
$\boldsymbol{P}_{LGS}$	**429**	1.13e-01	9	7.41e-06	3.47e-04	1.08e-09
	2 437	3.53e-01	10	1.64e-05	2.19e-04	1.83e-08
	117 637	2.95e+01	10	2.31e-04	1.40e-04	1.83e-08
$\boldsymbol{P}_{UGS}$	**429**	1.17e-01	9	1.49e-05	8.85e-05	1.46e-10
	2 437	4.28e-01	13	1.51e-06	8.24e-06	1.57e-09
	117 637	3.24e+01	12	4.26e-06	3.40e-05	6.29e-09

	Size	Errors using the initial system : x				
		time	**it**	**err**$_{\underline{u}}$	**err**$_p$	**err**$_T$
$\boldsymbol{P}_J$	**429**	1.66e-01	18	1.19e-03	1.44e-03	1.13e-07
	2 437	1.02e+00	34	6.24e-05	4.74e-05	3.79e-08
	117 637	1.15e+02	40	4.36e-05	1.25e-05	9.01e-08
$\boldsymbol{P}_{LGS}$	**429**	1.40e-01	12	5.81e-07	2.20e-06	1.36e-10
	2 437	5.39e-01	14	3.49e-06	7.14e-07	4.01e-10
	117 637	4.21e+01	14	1.93e-06	4.53e-07	2.59e-10
$\boldsymbol{P}_{UGS}$	**429**	1.07e-01	3	8.67e-01	1.08e+00	1.70e-05
	2 437	3.29e-01	8	1.78e-01	1.79e-01	6.06e-07
	117 637	2.45e+01	8	3.39e-01	2.95e-01	6.85e-07

Model problem

The test case needs to be simple enough so that the mesh can be easily refined but complex enough to resemble the industrial problem in consideration. For this purpose, a 3D rectangular sample is modelled as seen in Figure 3.1, with a 0.1 m length following x, a 0.1 m height following y and 0.05 m large following z. The tetrahedral mesh was generated using Gmsh 4.4.1.
The displacement was set to 0 on the bottom surface ($y = 0$), a mechanical pressure of 5 MPa was applied on the top surface ($y = 0.1$) and a temperature of 80°C was imposed on the whole surface of the sample.
Depending on the assessment under consideration, the sample consists of a single material or of 2 different materials. For the robustness experiments, it consists of clay only, while for the scalability experiments, it consists of clay and concrete. The 2 different subdomains are illustrated in Figure 3.1. We emphasize that the order of magnitude of the material parameters are of great importance in the industrial applications. The values of the material parameters, displayed in Table 3.3, are representative of a typical industrial problem of geological waste disposal [50].

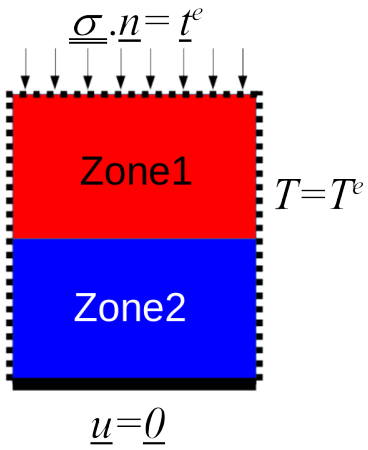


Figure 3.1: Test case

The tests were solved with code_aster using the THM framework presented in the section above, corresponding to an isotropic saturated single-phased THM medium using P2-P1-P1 finite elements.

3.3.1 Robustness

The robustness of the preconditioners is evaluated by varying the values of the Young's modulus E, the intrinsic permeability k_{int} and the thermal conductivity λ_T. These parameters are chosen since they appear respectively in each balance equation and have a major influence therein. The tests are done using the test case of Figure 3.1 with both Zones 1 and 2 made of clay, using the material parameters in Table 3.3. The results are compiled in Table 3.4 for P_J and Table 3.5 for $\boldsymbol{P}_{LGS}$. The maximum number of outer FGMRES iterations during the Newton iterations are displayed first, followed by the total number of Newton iterations in parentheses. The very large range of variation of the mesh size and of each parameter (up to 6 orders of magnitude) is emphasized.

In order to analyse the results in Tables 3.4 and 3.5, we propose first a row-wise reading then a column-wise reading.
The row-wise reading provides information on the influence of the mesh size, the material parameters being fixed. An excellent independence with respect to the mesh size is observed. The outer number of FGMRES iterations remains constant even though the size of the system is multiplied by 20, except for P_J with the set of parameters (E=1.e+9, k_{int}=4.e-21, λ_T=2.3) where the number of iterations goes from 70 to 44 but seems to stabilized by reaching 40 in the biggest mesh.
The column-wise reading provides information on the influence of the material parameters, the mesh size being fixed. A moderate variation of the outer number of

Table 3.3: The test case parameters

Symbol	Value	Unit
μ_l	10^{-3}	Pa.s
K_l	2.10^9	Pa
C_s	1000	J.kg^{-1}.K^{-1}
C_f	4180	J.kg^{-1}.K^{-1}
C_f^p	4180	J.kg^{-1}.K^{-1}
ρ_f	1000	kg.m^{-3}
λ_T	1.6	W.m^{-1}.K
T_0	273	K
p_{atm}	10^5	Pa
α_s	10^{-5}	K^{-1}
α_l	10^{-4}	K^{-1}
h_{f0}	$\frac{p_{atm}}{\rho_f}$	J.kg^{-1}
K_s	$\frac{E}{3(1-2\nu)}$	Pa
K_0	K_s	Pa
λ_H	K_{int}/μ_l	Pa^{-1}.m^2.s^{-1}
C_σ^0	$C_s\rho_s(1-\varphi)+C_l\rho_f\varphi$	J.K^{-1}.m^{-3}
ρ_s	$(\rho_m-\varphi\rho_f)/(1-\varphi)$	kg.kg^{-3}
α_m	$\varphi\alpha_l+(1-\varphi)\alpha_s$	K^{-1}

Clay		
Symbol	Value	Unit
E	6.10^9	Pa
ν	0.3	-
ρ_m	2410	kg.m^{-3}
K_{int}	4.10^{-21}	
φ	0.18	-
Concrete		
Symbol	Value	Unit
E	15.10^9	Pa
ν	0.2	-
ρ_m	2500	kg.m^{-3}
K_{int}	10^{-11}	
φ	0.2	-

FGMRES iterations is observed, that remains mostly under 12 except for the "worst" set of parameters (E=1.e+09, k_{int}=4.e-21), where it reaches up to 70 iterations for P_J and 22 for $\boldsymbol{P}_{LGS}$. This particular result tends to show a better robustness of the Block Gauss-Seidel variant compared to the Jacobi variant, which is further analyzed in the next section. In spite of this, both preconditioners appear to be very robust as they achieve convergence at each run and the increase in Krylov iterations remains moderate compared to the large variations in material parameters.
Finally, we highlight the excellent robustness with respect to the Newton iterations, which remain between 2 and 4 for every run.

Spectral analysis

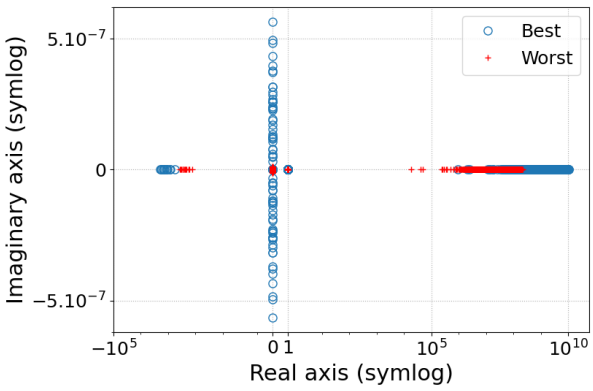
In the previous section, a better robustness of the Lower Block Gauss-Seidel variant $\boldsymbol{P}_{LGS}$ compared to the Jacobi variant P_J was observed. In order to further analyze this, let us denote by :

- "best" case, the set of parameters ($E = 5.\,\text{e+10}$, $k_{int} = 5.\,\text{e}{-15}$, $\lambda_T = 0.4$)
- "worst" case, the set of parameters (E=1.e+9, k_{int}=4.e-21, λ_T=2.3)

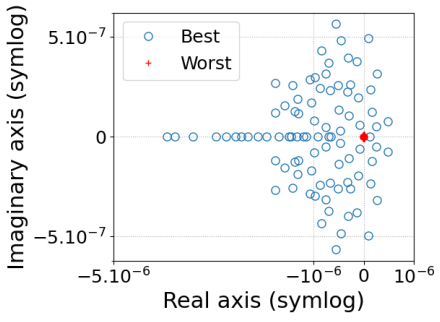
In the "best" case, the proposed Krylov method and preconditioners reach converge in six iterations while in the "worst" case, 70 iterations are needed.

Table 3.4: Block Jacobi Parameter Robustness

P_J					
Parameters			DoF		
E	k_{int}	λ_T	10 000	60 000	200 000
1.e+09	4.e-15	4.e-01	6 (3)	6 (2)	6 (4)
		2.3e+00	5 (3)	5 (2)	4 (2)
	4.e-18	4.e-01	8 (3)	8 (3)	9 (3)
		2.3e+00	8 (3)	7 (3)	8 (3)
	4.e-21	4.e-01	48 (3)	46 (3)	50 (3)
		2.3e+00	70 (3)	44 (3)	40 (3)
2.5e+10	4.e-15	4.e-01	6 (2)	6 (2)	7 (3)
		2.3e+00	6 (2)	6 (2)	4 (2)
	4.e-18	4.e-01	7 (2)	6 (2)	8 (2)
		2.3e+00	6 (2)	6 (2)	7 (2)
	4.e-21	4.e-01	12 (2)	12 (2)	11 (2)
		2.3e+00	11 (2)	11 (2)	11 (2)
5.0e+10	4.e-15	4.e-01	6 (2)	6 (2)	7 (3)
		2.3e+00	6 (2)	6 (2)	5 (2)
	4.e-18	4.e-01	6 (2)	6 (2)	8 (2)
		2.3e+00	6 (2)	6 (2)	6 (2)
	4.e-21	4.e-01	10 (2)	11 (2)	10 (2)
		2.3e+00	10 (2)	9 (2)	9 (2)



(a) All eigenvalues



(b) Zoom around zero

Figure 3.2: Eigenvalue distribution for the "best" and "worst" cases

Table 3.5: Block Gauss-Seidel Parameter Robustness

$\boldsymbol{P}_{LGS}$					
Parameters			DoF		
E	k_{int}	λ_T	10 000	60 000	200 000
1.e+09	4.e-15	4.e-01	4 (3)	5 (2)	5 (4)
		2.3e+00	4 (3)	4 (2)	3 (2)
	4.e-18	4.e-01	6 (3)	5 (3)	6 (3)
		2.3e+00	5 (3)	5 (3)	6 (3)
	4.e-21	4.e-01	21 (3)	22 (3)	20 (3)
		2.3e+00	18 (3)	20 (3)	18 (3)
2.5e+10	4.e-15	4.e-01	5 (2)	6 (2)	5 (3)
		2.3e+00	5 (2)	5 (2)	4 (2)
	4.e-18	4.e-01	5 (2)	6 (2)	6 (2)
		2.3e+00	5 (2)	5 (2)	6 (2)
	4.e-21	4.e-01	8 (2)	8 (2)	8 (2)
		2.3e+00	7 (2)	7 (2)	7 (2)
5.0e+10	4.e-15	4.e-01	5 (2)	5 (2)	5 (3)
		2.3e+00	5 (2)	5 (2)	4 (2)
	4.e-18	4.e-01	5 (2)	5 (2)	7 (2)
		2.3e+00	5 (2)	5 (2)	7 (2)
	4.e-21	4.e-01	7 (2)	7 (2)	7 (2)
		2.3e+00	6 (2)	7 (2)	6 (2)

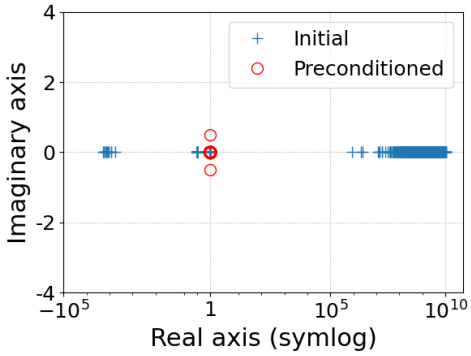
Let us begin by commenting the spectrum of the initial system, displayed in Figure 3.2. In both cases, apart from some differences in the magnitude of the extreme values, the real parts of the eigenvalues are organised in 4 blocks :

- few percents are negative, lying around -10^3
- few percents lie around zero
- few percents lie around one
- the most part lie around 10^5 and 10^8 (roughly speaking 80%)

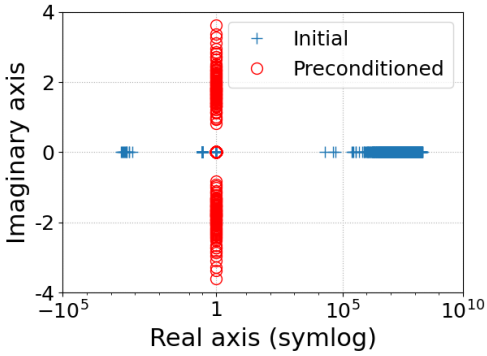
In fact, the main difference lies around the origin. As can be seen from the zoom around the origin in Figure 3.2, the "worst" case exhibit a clear cluster of almost zero eigenvalues while they are much more scattered in the "best" case.

We shall now evaluate the effect of both preconditioners on the spectrum of the Jacobian matrix. The eigenvalues of both the initial (*i.e.* not preconditioned) and the preconditioned system are displayed in Figure 3.3.

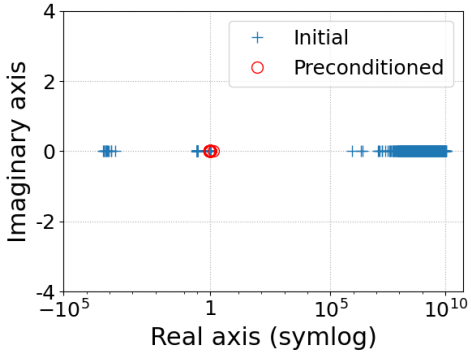
Once a preconditioner is applied, the real part of all eigenvalues is clustered around 1. $\boldsymbol{P}_J$ completely clusters the real part of the eigenvalues to 1 but distributes the



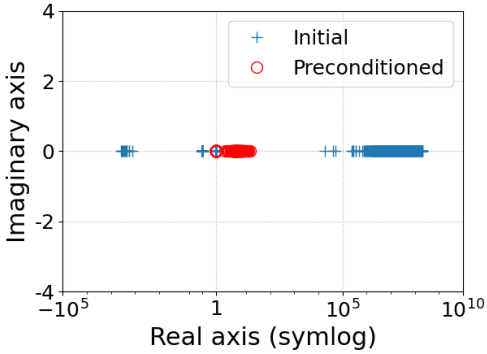
(a) $\boldsymbol{J}$ and $\boldsymbol{J}\boldsymbol{P}_J^{-1}$ eigenvalues for the "best" case



(b) $\boldsymbol{J}$ and $\boldsymbol{J}\boldsymbol{P}_J^{-1}$ eigenvalues for the "worst" case



(c) $\boldsymbol{J}$ and $\boldsymbol{J}\boldsymbol{P}_{LGS}^{-1}$ eigenvalues for the "best" case



(d) $\boldsymbol{J}$ and $\boldsymbol{J}\boldsymbol{P}_{LGS}^{-1}$ eigenvalues for the "worst" case

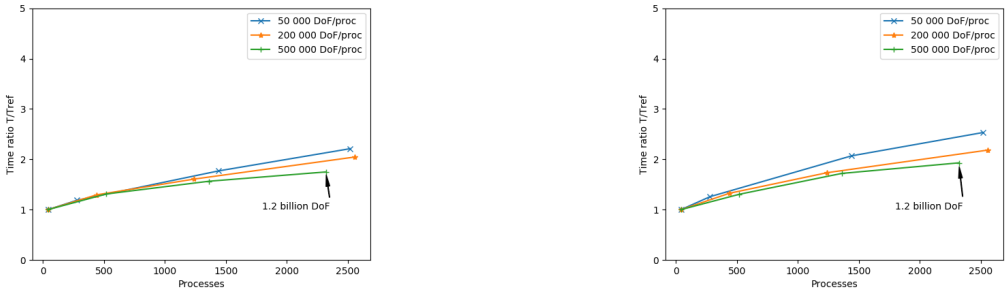
Figure 3.3: Eigenvalue distribution

imaginary part to of the eigenvalues between -1 and 1 in the "best" case and between -4 and 4 in the "worst" case. In any case, the eigenvalues belong to 3 differents clusters, more scattered in the "worst" case.
$\boldsymbol{P}_{LGS}$ generate almost real eigenvalues, with a tight cluster around 1 for the "best" case and between 1 and 10^2 in the "worst" case.
The difference in clustering the eigenvalues (3 blocks for $\boldsymbol{P}_J$ and a single one for $\boldsymbol{P}_{LGS}$) might explain the better results of the latter.

3.3.2 Parallel scalability

A good scalability of the proposed preconditioner is essential to keep the resolution time reasonable when switching to bigger systems. Weak and strong scalability tests are considered using the bi-material case from Figure 3.1 with Zone 1 made of clay and Zone 2 made of concrete. Realistic parameter values where chosen from Table 3.3. Both of the scalability tests are run on EDF's cluster Cronos. It consists in 1272 nodes, equipped with 2 Xeon Platinum 8260 24C 2.4 GHz processors with 24 cores each.

A weak scalability test consists in setting a fixed number of degrees of freedom (DoF) by processor and increasing the size of the problem by increasing the number of processes. In other words, we set the size of a sub-domain and make the problem bigger by increasing the total number of sub-domains. Our goal is to investigate if the solution algorithm needs the same resolution time whether we solve N DoF on 1 process or 1000xN DoF on 1000 processes. In case of perfect weak scalability, the time should remain constant when increasing the number of processes.



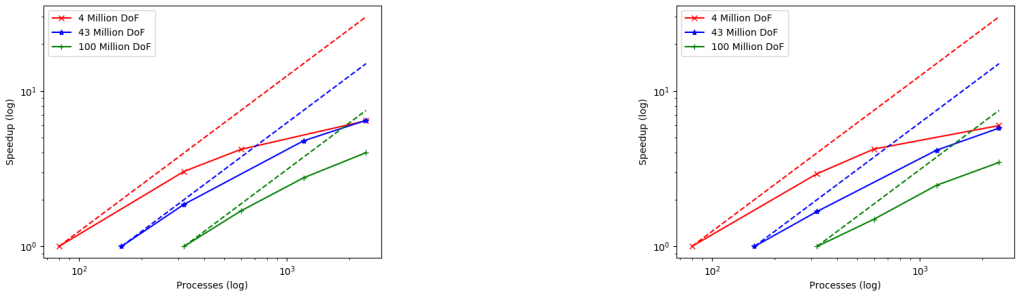
(a) $\boldsymbol{P}_J$ (b) $\boldsymbol{P}_{LGS}$

Figure 3.4: Weak scalability

As can be seen in the Figure 3.4, the number of DoF per process is fixed to 50 000 (blue line), 200 000 (orange line) and 500 000 (green line) and the test case is run from 40 processes to 2500 processes. The ratio between the solution time to the

40 processes time is presented. For small numbers of DoF per process, it remains between 1. and 2.5 for $\boldsymbol{P}_{LGS}$ and between 1. and 2.2 for P_J. Whereas for 500 000 DoF per process, it remains between 1. and 1.9 for $\boldsymbol{P}_{LGS}$ and between 1. and 1.7 for P_J. This sub-optimal behavior for small sub-domains is often due to latency of the cluster's network. When sub-domains are large and there is more work per process, the computation dominates the cost associated with communication. Even though P_J scales slightly better, the resolution time is higher than with $\boldsymbol{P}_{LGS}$ due to higher number of iterations that range between 8 and 16 for P_J and 7 and 11 for $\boldsymbol{P}_{LGS}$. We highlight that using $\boldsymbol{P}_{LGS}$ for 500 000 DoF per processor (green line), the size of the linear system ranges from 20 million with a solving time of 465 seconds to more than 1.2 billion DoF with a solving time of 891 seconds. The size of the problem is multiplied by 60 whereas the solving time only increases by 1.9. This is a very good scalability result since the test case is rather complex especially due to the variation of material parameters between clay and concrete.

Let us switch to the strong scalability test, which consists in fixing the size of the problem and increasing the number of processors. The goal is to solve the system faster by adding resources. For example, if we solve a system of a given size using N processes when using $N \times M$ processes the solving time should be dived by M. In case of perfect strong scalability, the solving time decreases proportionally to the increase of the number of processes.



(a) $\boldsymbol{P}_J$ (b) $\boldsymbol{P}_{LGS}$

Figure 3.5: Strong scalability

The strong scalability tests are presented in Figure 3.5. The size of the problem is fixed to 4 millions DoF (red line), 43 millions DoF (blue line) and 100 millions DoF(green line). The number of processes were increased from 80 to 2 400. The dashed line represents the ideal strong scalability. The speedup with respect to the number of processes is presented. For all cases, when increasing the processes from 80 to 320 with 4 million DoF, the strong scalability remains satisfactory with an efficiency of 76% for P_J and of 73% for $\boldsymbol{P}_{LGS}$. Then the efficiency starts declining until it reaches 2 400 process and is at 20% for P_J and for $\boldsymbol{P}_{LGS}$. For the 100 million

DoF case, from 320 to 600 processes the efficiency for P_J is 90% and for $\boldsymbol{P}_{LGS}$ 79%. When increasing the processes to 2400, the efficiency for P_J is 53% and for $\boldsymbol{P}_{LGS}$ 46%. Similarly to the weak scalability, the difference in efficiency between each problem is often due to latency of the cluster's network, since at 4 millions DoF there is less work per process. P_J scales slightly better than $\boldsymbol{P}_{LGS}$ but $\boldsymbol{P}_{LGS}$ remains faster for all the tested cases. Given the complexity of the test case and the fact that we started at 80 processes in order to be able to solve size-wise representative problems, the proposed preconditioner strong scalability is satisfactory.

Conclusion

This chapter deals with the assessment of the robustness and the weak and strong scalability of a preconditioner dedicated to coupled THM problems, which relies on the block structure of the Jacobian of the linearized system. A Block Jacobi and a Block Gauss-Seidel variants are investigated, sharing the same tailored sub-solvers (Krylov methods preconditioned by AMG preconditioners).

Coupled systems often exhibit very bad scaling due to the presence of parameters of different orders of magnitude. It can be addressed by the use of a dedicated scaling algorithm that efficiently re-balances the Jacobian. It is nevertheless not mandatory in our case since it is shown that the proposed block preconditioners can handle naturally the unbalance of the different blocks. For the case of the Block Gauss-Seidel variant, special attention is needed to eliminate the unknowns in a well-chosen order. Finally, both variants show excellent mesh size independence, good robustness with respect to parameters variation and good scalability on a simple yet representative test case.

Though established in the linear regime, these results a very valuable when considering to move to nonlinear constitutive laws. This point is being investigated and encouraging results have already been obtained.

Chapter 4

Second gradient of dilation regularisation

Contents

4.1 The second gradient of dilation model

In the considered application, the excavation occurs in a geological clay layer, so that the numerical model must handle constitutive laws of geomaterials. These laws possess a common feature : beyond a given level of strains, the stress in the medium decreases. This phenomenon is called strain-softening and these laws are called of softening type.

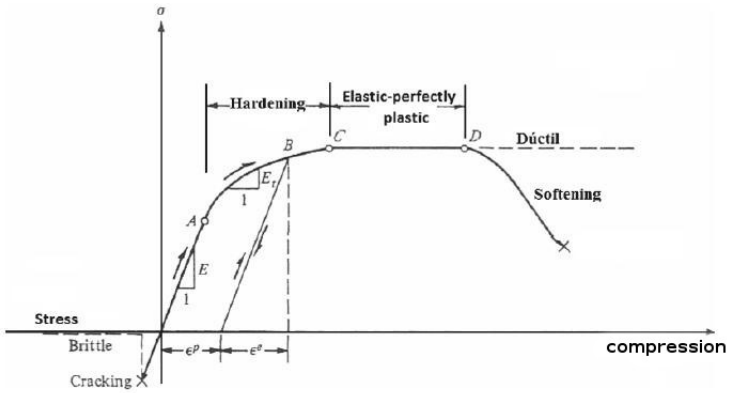


Figure 4.1: Softening constitutive law [70].

This property is responsible for the localization of the mechanical field (strain, damage, plastic strain,...) inside narrow bands. They are called strain localization bands and appear at a *microscopic* scale. Since classical continuum mechanics models reproduce the mechanical behaviour of a medium at a *macroscopic* scale, they will not follow the initiation and the evolution of the strain localization bands. Ultimately, it is the behavior of the material inside the localization bands that will govern the response of the medium. Hence the importance to predict the time of appearance of these localization bands, their quantity, their orientation and their width.

Among the characteristics of the bands, their width plays a particularly important role. Studies have shown the dependence of the width to the microstructure [90, 78]. Therefore, it is essential to correctly take into account the micro-mechanical phenomena to predict precisely the damaged areas in the medium. However, the separation of the two scales is a fundamental assumption in order to apply the usual homogenisation techniques. In addition, classical constitutive laws come from these homogenisation techniques and are no longer representative. Mathematically, this translates in a loss of uniqueness and strong mesh dependency when softening material laws are used. The loss of uniqueness can be easily observed in a 1D example that is the topic of the next section.

4.1.1 Simple 1D example

The following example is available at [39], we only present the main results hereafter. The rod of length L, represented in 4.2, is submitted to an increasing displacement U at one end and clamped at the other end. Since it follows an elasto-plastic softening law, represented in 4.2, when the critical load is reached the stress can not increase anymore. The rod behaves in two different ways. In the damaged region of length l, the stress follows a damaging unloading law and in the rest of the region of length $L - l$, it follows an elastic unloading. It can then be shown that the global response of the rod depends of l. The issue is that the length of the damaged zone l remains unknown, resulting in an infinity of solutions, hence the loss of uniqueness.

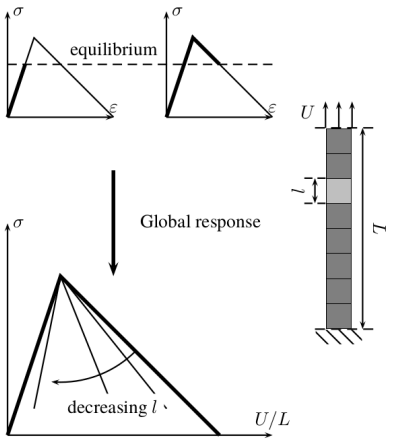


Figure 4.2: 1D rod example [39]

The width of the damaged zone also comes into play when considering the solution of the problem using finite elements. For example, when using piece-wise constant elements, the length of the damaged zone is necessarily a multiple of the size of a single element. When taking into account round-off errors or geometrical imperfections, one element reaches the critical load slightly earlier than the others. This results in the width of the damaged zone coinciding with the size of one element. Therefore the strain localizes inside bands of the width of one element making the solution mesh dependent. Mesh dependency appears in simple simulations such as the 1D rod example as well as more complex problems such as the case of gallery excavation simulations [37]. In order to solve this problem, a richer model that correctly predicts the strain localization is necessary. A few regularization methods have been proposed and lead in one way or another to the introduction of an internal length in the model.

For the THM model, the second gradient of dilation regularization is a possible way [55][35]. This regularization forces the equality of microscopic and macroscopic volume changes. An enlargement of the previous microscopic localization is created

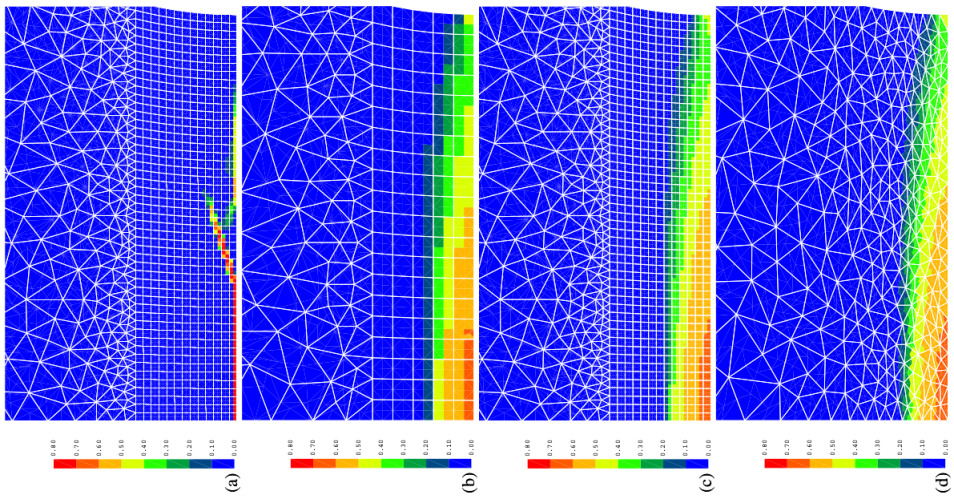


Figure 4.3: Strain localization before and after a regularization [39]

and the bands appear in the discrete model. In Figure 4.3 is presented a model of a strain localization band (a) without regularization and (b,c,d) with regularization. In (a), the width of the strain localization band is inferior to the cells of the mesh making our model and the strain localization band mesh dependent. After the regularization, we see in (b) that the strain localization band is enlarged. Then a finer mesh is chosen in (c), we see that the discretization doesn't affect the strain localization band when the regularization is present. Finally in (d), a non-uniform mesh is chosen and it still does not affect the localization band. We conclude that after the regularization, the model is no longer mesh dependent. Next the second gradient of dilation model is introduced as well as some regularization techniques.

4.1.2 Regularization techniques

Regularization techniques add an internal length to the model which results in the well-posedness of the problem and in the well characterization of the width of the band in the macroscopic level. Some of these methods can be classified into two large categories. Non-local interaction regularization techniques, where the behaviour of the material at a certain point takes into account the behaviour of neighbour points under a certain distance. This can be done trough a non-local integral approach [75], an implicit gradient approach [73] or an explicit gradient approach [74]. In the other category are the regularization techniques that take into account the microstructure. For example the method of virtual power where the rotation of the microstructure is introduced in the model [24, 49]. An in-depth bibliography review of these regularization techniques is presented in [44]. Second gradient models are part of the second category, they impose an equality between the microscopic deformation and

the macroscopic displacement gradient.

The second gradient of dilation formulation can be obtained by two different ways : through dilation microstructure models where the macroscopic dilation is equal to the microscopic dilation, or through a simplified elasto-plastic second gradient model with a particular choice for the constitutive law of the second gradient. In order to develop a deep understanding of the second gradient of dilation formulation, the two different models and their connection with the second gradient of dilation are presented below.

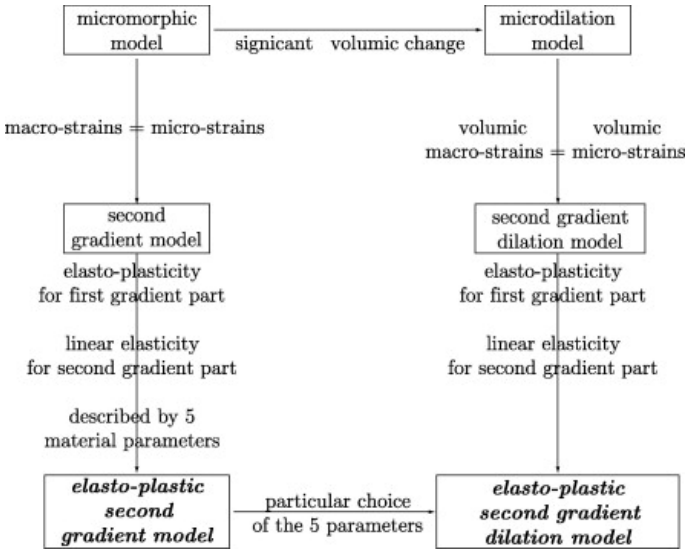


Figure 4.4: Definition of the second gradient of dilation [37]

A dilation microstructure model with an additional constraint

Microdilation models take into account the microstructure of the medium by incorporating the relative micro deformation of the microstructure (with respect to the macrodeformation) $\epsilon_V - \chi$, into the formulation, with $\epsilon_V = tr(\underline{\underline{\varepsilon}})$ representing the macroscopic volume changes and χ representing the microscopic volume changes. This results into two equilibrium equations : the first one is the balance equation on the mechanical stress and the second one is the balance equation on the double dilation stress $\underline{S}$, related to the microscopic volume changes χ. The two equilibrium equations with the two unknowns $(\underline{u}, \chi)$ are

$$\begin{aligned} \mathrm{div}(\underline{\underline{\sigma}}(\underline{u})) + \nabla\kappa &= 0 \\ \kappa + \mathrm{div}(\underline{S}(\chi)) &= 0 \end{aligned}$$

Here, κ is the microscopic dilation stress that is conjugated with the relative micro deformation $\epsilon_V - \chi$. Starting from this formulation, we introduce an equality constraint

between the macroscopic volume changes ϵ_V and the microscopic volume changes χ. If we interpret κ as a Lagrange multiplier for weakly imposing the equality constraint, the equilibrium equations become :

$$\begin{aligned} \mathrm{div}(\underline{\underline{\sigma}}(\underline{u})) + \nabla\kappa &= 0 \\ \kappa + \mathrm{div}(\underline{S}(\chi)) &= 0 \\ \epsilon_V &= \chi \end{aligned}$$

where the unknowns are $(\underline{u}, \chi, \kappa)$.
We then choose an elasto-plastic law for the macroscopic stress $\underline{\underline{\sigma}}$ and the linear elastic law especially developed for the second gradient of dilation in [37] for the microscopic volumetric stress $\underline{S}$. The result is an elasto-plastic simplified second gradient of dilation model obtained trough a microstructure of dilation model. The second gradient of dilation can also be obtained trough a second gradient model.

A simplified elasto-plastic second gradient model

Second gradient models take into account the microstructure by imposing an equality between the microdeformation $\underline{\underline{f}}$ and the macroscopic gradient $\nabla\underline{u}$ directly in the balance equations :

$$\underline{\underline{f}} = \nabla\underline{u}$$

It is shown in [21] that this results in the balance equation :

$$\mathrm{div}(\underline{\underline{\sigma}}(\underline{u}) - \mathrm{div}(\underline{\underline{\underline{\Sigma}}}(\underline{u}))) = 0$$

where $\underline{\underline{\sigma}}$ is the first gradient stress that describes the behaviour of the displacement at a macroscopic level and $\underline{\underline{\underline{\Sigma}}}$ is the second gradient double dilation stress that is conjugated with the behaviour of the microdeformations. The assumption to separate both constitutive laws is generally made and will be follow for the rest of the manuscript. This means that theoretically any usual constitutive law could be used for $\underline{\underline{\sigma}}$. As for $\underline{\underline{\underline{\Sigma}}}$, a few laws have been proposed. In [22], the authors chose a elasto-plastic law that relies on an additive decomposition of the deformation gradient for $\underline{\underline{\underline{\Sigma}}}$. Nevertheless, an isotropic linear elastic behavior is generally retained [21].

By following previous work on elastic second gradient medium [68], it can be shown that for 2D isotropic materials, the second gradient constitutive law can be written in the form :

$$\begin{pmatrix} \Sigma_{111} \\ \Sigma_{112} \\ \Sigma_{121} \\ \Sigma_{122} \\ \Sigma_{211} \\ \Sigma_{212} \\ \Sigma_{221} \\ \Sigma_{222} \end{pmatrix} = \begin{pmatrix} a_{12345} & 0 & 0 & a_{23} & 0 & a_{12} & a_{12} & 0 \\ 0 & a_{145} & a_{145} & 0 & a_{25} & 0 & 0 & a_{12} \\ 0 & a_{145} & a_{145} & 0 & a_{25} & 0 & 0 & a_{12} \\ a_{23} & 0 & 0 & a_{34} & 0 & a_{25} & a_{25} & 0 \\ 0 & a_{25} & a_{25} & 0 & a_{34} & 0 & 0 & a_{23} \\ a_{12} & 0 & 0 & a_{25} & 0 & a_{145} & a_{145} & 0 \\ a_{12} & 0 & 0 & a_{25} & 0 & a_{145} & a_{145} & 0 \\ 0 & a_{12} & a_{12} & 0 & a_{23} & 0 & 0 & a_{12345} \end{pmatrix} \begin{pmatrix} \frac{\partial^2 u_1}{\partial^2 x_1} \\ \frac{\partial^2 u_1}{\partial x_1 \partial x_2} \\ \frac{\partial^2 u_1}{\partial x_2 \partial x_1} \\ \frac{\partial^2 u_1}{\partial^2 x_2} \\ \frac{\partial^2 u_2}{\partial^2 x_1} \\ \frac{\partial^2 u_2}{\partial x_1 \partial x_2} \\ \frac{\partial^2 u_2}{\partial x_2 \partial x_1} \\ \frac{\partial^2 u_2}{\partial^2 x_2} \end{pmatrix}$$

Where all the terms depend on the 5 constants $a_i, i = 1, 5$.

$$\begin{cases} a_{12345} &= 2(a_1 + a_2 + a_3 + a_4 + a_5) \\ a_{23} &= a_2 + 2a_3 \\ a_{12} &= a_1 + \frac{1}{2}a_2 \\ a_{145} &= \frac{1}{2}a_1 + a_4 + \frac{1}{2}a_5 \\ a_{25} &= \frac{1}{2}a_2 + a_5 \\ a_{34} &= 2(a_3 + a_4) \end{cases} \tag{4.1}$$

However, from a experimental point of view, identifying five constants can be challenging. In [37], Fernandes shows that if the double stress verifies the following equations, then the second gradient model degenerates into a second gradient of dilation model.

$$\begin{cases} S_1 &= \Sigma_{111} = 2\Sigma_{212} = 2\Sigma_{221} = 2\Sigma_{313} = 2\Sigma_{331} \\ S_2 &= \Sigma_{222} = 2\Sigma_{112} = 2\Sigma_{121} = 2\Sigma_{323} = 2\Sigma_{332} \\ S_3 &= \Sigma_{333} = 2\Sigma_{113} = 2\Sigma_{131} = 2\Sigma_{223} = 2\Sigma_{232} \\ \Sigma_{122} &= \Sigma_{123} = \Sigma_{132} = \Sigma_{133} = 0 \\ \Sigma_{211} &= \Sigma_{213} = \Sigma_{231} = \Sigma_{233} = 0 \\ \Sigma_{311} &= \Sigma_{312} = \Sigma_{321} = \Sigma_{322} = 0 \end{cases} \tag{4.2}$$

In order to verify 4.2, we impose the following conditions to 4.1 :

$$\begin{cases} a_{12345} &= 2a_{12} \\ a_{23} &= 0 \\ a_{12} &= 2a_{145} \\ a_{25} &= 0 \\ a_{34} &= 0 \end{cases}$$

This results into the equality :

$$\operatorname{div}(\operatorname{div}(\underline{\underline{\Sigma}}(\underline{u}))) = \nabla(\operatorname{div}(\underline{S}(\operatorname{div}(\underline{u}))))$$

with

$$\underline{S}(\operatorname{div}(\underline{u})) = 3a_1 \nabla \operatorname{div}(\underline{u})$$

$\underline{S}$ follows an isotropic and linear constitutive law that is defined in function of a single stiffness parameter a_1. To generalise $\underline{S}$ for 3D problems, Fernandes shows that $\underline{S}$ is of the following form :

$$\underline{S}(\operatorname{div}(\underline{u})) = (n+1)a_1 \nabla \operatorname{div}(\underline{u})$$

where n is the problem's dimension, so 2 for 2D problems and 3 for 3D problems. This is the second gradient constitutive law that is retained for the rest of the manuscript.

The balance equation of the second gradient of dilation now writes :

$$\operatorname{div}(\underline{\underline{\sigma}}(\underline{u})) - \nabla(\operatorname{div}(\underline{S}(\operatorname{div}(\underline{u})))) = 0$$

where $\underline{u}$ has to be twice differentiable. This adds a real difficulty since in order to solve the problem using finite elements, $C1$ elements would be needed. Such elements are used in [20] for a 1D second gradient model, however when solving 2D or 3D problems they lead to a significant increase in the number of degrees of freedom. A solution is to simplify the model by weakly imposing the equality constraint between the macroscopic volume changes $\operatorname{div}(\underline{u})$ and the microscopic volume changes χ trough a Lagrange multiplier [66, 85]. For the second gradient of dilation, Fernandes shows in [37] that the equations become :

$$\begin{aligned} -\operatorname{div}(\underline{\underline{\sigma}}(\underline{u})) + \nabla\lambda &= 0 \\ \lambda - \operatorname{div}(\underline{S}(\chi)) &= 0 \\ \epsilon_V &= \chi \end{aligned}$$

where λ are the Lagrange multipliers and χ the microscopic volume changes. Since this formulation only requires the unknowns to be once differentiable, $C0$ elements can be used thus reducing the size of the problem. In [37], Fernandes shows that, for 2D problems, stress and strain oscillations can appear at Gauss points inside a single element. He then introduces an augmented Lagrangian formulation in order to eliminate any oscillations. The simplified second gradient of dilation then becomes :

$$\begin{aligned} -\operatorname{div}(\underline{\underline{\sigma}}(\underline{u})) + \nabla\lambda - r\nabla(\operatorname{div}(\underline{u})) + r\nabla\chi &= 0 \\ \lambda - \operatorname{div}(\underline{S}(\chi)) - r\operatorname{div}(\underline{u}) + r\chi &= 0 \\ \epsilon_V &= \chi \end{aligned}$$

where the unknowns are u, χ, λ and r is the penalization parameter of the augmented Lagrangian. This formulation prevents any oscillations of the fields at Gauss points. However, with the augmented Lagrangian, comes an extra parameter, r, that needs to be calibrated. This choice has to be made carefully since large values result in badly conditioned displacement stiffness matrix [37].

4.1.3 Difficulties when implementing the model for industrial applications

The second gradient of dilation model was found to be effective for geomaterials to remove the mesh dependence of the solution. Nevertheless, even with the regularization, there are multiple solutions when used for softening material laws. The second gradient of dilation regularization turns the number of solutions finite but not unique. In [37], Fernandes introduces a bifurcation algorithm that finds every solution when non-linear laws are used. However the multiplicity of solutions is not within the scope of this work.

Two main issues are addressed in this chapter. First, the well-posedness of the model when using linear material laws. To our knowledge, this has never been investigated in the literature. Since $\underline{\underline{\sigma}}$ is independent from $\underline{S}$, any constitutive law can be used. For the rest of the chapter, we place ourselves in a simplified case by using linear elasticity for $\underline{\underline{\sigma}}$. Not only does this allow us to prove the well-posedness of the continuous and discrete problems, but these proofs will be usefull to propose a preconditioner for the second gradient of dilation model. This is the second issue of this chapter, namely, find a preconditioner that deals with the saddle point nature of the system, with the micro-volume changes χ and the Lagrange multiplier λ as unknowns. Furthermore, since large scale 3D simulations are targeted, the scalability of the preconditioner must be addressed.

4.2 Application to mechanics

Let us now start by doing an in-depth analysis of the second gradient of dilation regularization applied to mechanics. According to the work of Fernandes on the second gradient of dilation, the model describes the evolution of the vector displacement field $\underline{u}(x)$, the scalar micro-volume change field $\chi(x)$ and λ as the Lagrange multipliers $\forall x \in \Omega$ [37]. An augmented Lagrangian regularisation can be used trough the penalization constant $r > 0$. The use of an augmented Lagrangian is optional : for $r = 0$ the regularization is only imposed through the Lagrange multipliers.

The balance equations are :

$$-\operatorname{div}(\underline{\underline{\sigma}}(\underline{u})) + \nabla\lambda - r\nabla(\operatorname{div}(\underline{u})) + r\nabla\chi = 0 \qquad \text{in } \Omega \tag{4.3}$$

$$\lambda - \operatorname{div}(\underline{S}(\chi)) - r\operatorname{div}(\underline{u}) + r\chi = 0 \qquad \text{in } \Omega \tag{4.4}$$

$$\operatorname{div}(\underline{u}) - \chi = 0 \qquad \text{in } \Omega \tag{4.5}$$

$$\underline{u} = 0 \qquad \text{on } \Gamma_1 \tag{4.6}$$

$$\frac{\partial \underline{u}}{\partial \underline{n}} = 0 \qquad \text{on } \Gamma_2 \tag{4.7}$$

$$\underline{S}(\chi)\,\underline{n} = 0 \qquad \text{on } \partial\Omega \tag{4.8}$$

Where $\underline{\underline{\sigma}}(\underline{u}) = 2\mu\underline{\underline{\varepsilon}}(\underline{u}) + \lambda_m \operatorname{div}(\underline{u})I$. $\underline{S}(\chi) = 3a^1\nabla\chi$ where a^1 is a given material parameter and $r \geq 0$. (4.3) and (4.4) are balance equations of the mechanical stress $\underline{\underline{\sigma}}$ and of the second stress tensor $\underline{S}$; (4.5) is a constraint equation that forces the equality of microscopic and macroscopic volume changes. The boundary of Ω is denoted $\partial\Omega$ and 2 different partitions are needed to define the boundary conditions. The two different partitions are defined as follows $\partial\Omega = \Gamma_1 \cup \Gamma_2$ with $\Gamma_1 \cap \Gamma_2 = \emptyset$.

Remark 1. *For the sake of simplicity, we chose the external double microscopic forces to be equal to 0 which results in $\underline{S}(\chi)\,\underline{n} = 0$ on the boundary. Given that our computations aim to characterise the localization bands at a macroscopic level, the effect of this boundary condition is proven to be limited [44].*

Variational formulation

We define the Sobolev space :

$$H^1_{\Gamma_1}(\Omega) = \{v \in H^1(\Omega) | v = 0 \text{ on } \Gamma_1\}$$

We also introduce the norm :

$$\|v\|_{H^1_{\Gamma_1}\times H^1} = (\|v\|^2_{H^1_{\Gamma_1}} + \|v\|^2_{H^1})^{\frac{1}{2}}$$

Let $\underline{u}, \underline{v} \in (H^1_{\Gamma_1}(\Omega))^d$, $\chi, \psi \in H^1(\Omega)$ and $\lambda, \tau \in L^2(\Omega)$. By considering the appropriate Sobolev spaces defined above and by integration by parts, we have the following weak form :

Find $(\underline{u}, \chi, \lambda) \in (H^1_{\Gamma_1}(\Omega))^d \times H^1(\Omega) \times L^2(\Omega)$ such as for all $(\underline{v}, \psi, \tau) \in (H^1_{\Gamma_1}(\Omega))^d \times H^1(\Omega) \times L^2(\Omega)$, we have :

$$\int_\Omega \Big(\underline{\underline{\sigma}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v}) - \lambda \operatorname{div}(\underline{v}) + r \operatorname{div}(\underline{u}) \operatorname{div}(\underline{v}) - r\,\chi \operatorname{div}(\underline{v})\Big)\, dx = 0$$

$$\int_\Omega \big(\underline{S}(\chi)\,\nabla\psi + \lambda\,\psi + r\,\chi\,\psi - r\,\psi \operatorname{div}(\underline{u})\big)\, dx = 0$$

$$\int_\Omega \big(\tau \operatorname{div}(\underline{u}) - \tau\,\chi\big)\, dx = 0$$

where $r \geq 0$ is the penalization term.

Remark 2. *Another way to write the variational formulation is :*

$$\int_\Omega (\underline{\underline{\sigma}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v}) + \underline{S}(\chi)\nabla\psi - \lambda(\ div(\underline{v}) - \psi) + \tau(\ div(\underline{u}) - \chi) \\ + r(\ div(\underline{u}) - \chi)(\ div(\underline{v}) - \psi)dx = 0$$

This is the way it is written in `code_aster`*'s documentation.*

Following the Brezzi theory outlined in sub-section 2.2.1, a way to ensure the well-posedness of the problem is by proving Theorem 2.2.1.
For that, we define the bilinear forms :

$$\begin{aligned} a\left((\underline{u},\chi),(\underline{v},\psi)\right) = \underline{\underline{\sigma}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v}) + r\ \mathrm{div}(\underline{u})\ \mathrm{div}(\underline{v}) - r\,\chi\ \mathrm{div}(\underline{v}) \\ + \underline{S}(\chi)\,\nabla\psi + r\,\chi\,\psi - r\,\psi\ \mathrm{div}(\underline{u}) \end{aligned} \tag{4.9}$$

and

$$b\left((\underline{u},\chi),\tau\right) = \tau\ \mathrm{div}(\underline{u}) - \tau\,\chi \tag{4.10}$$

We denote by A and B, respectively, the linear continuous operators associated with the bilinear form $a(.,.)$ and $b(.,.)$ as in 2.2.1. The kernel of B is denoted $KerB$.

The system can then be written as :

$$\begin{aligned} a\left((\underline{u},\chi),(\underline{v},\psi)\right) - b\left((\underline{v},\psi),\lambda\right) = 0 \\ b\left((\underline{u},\chi),\tau\right) = 0 \end{aligned} \tag{4.11}$$

Clearly we see that it is an anti-symmetric saddle point system. Therefore, Theorem 2.2.1 applies. Inspiring ourselves from that theorem, we define a new Theorem 4.2.1 for the problem 4.11.

Theorem 4.2.1. *Let $V := (H^1_{\Gamma_1})^d \times H^1$ and $Q := L^2$.*
There is a constant $\alpha > 0$, such that

$$a\left((\underline{u},\chi),(\underline{u},\chi)\right) \geq \alpha\left\|(\underline{u},\chi)\right\|^2_V,\ (\underline{u},\chi) \in KerB \tag{4.12}$$

and there is a constant $\beta > 0$, such that

$$\inf_{\lambda\in Q}\sup_{(\underline{u},\chi)\in V}\frac{b((\underline{u},\chi),\lambda)}{\left\|(\underline{u},\chi)\right\|_V\|\lambda\|_Q} \geq \beta. \tag{4.13}$$

We can conclude that there exist a unique solution $(\underline{u},\chi,\lambda) \in V \times Q$ for the system.

Proof. In order to prove the Theorem 4.2.1, we first prove the coercivity of the bilinear form $a(.,.)$ on the Kernel of B, then we prove the *inf-sup* condition of the bilinear form $b(.,.)$.

To prove (4.12), we need to prove it for $r = 0$ and for $r > 0$. For $r = 0$, the coercivity of $a(.,.)$ holds on the whole V space, the detailed proof is left to the reader. For $r > 0$, the proof is more complex since the coercivity holds on the Kernel of B. In order to prove it we start with

$$a\left((\underline{u},\chi),(\underline{u},\chi)\right) = \int_\Omega \underline{\underline{\sigma}}(\underline{u})\underline{\underline{\varepsilon}}(\underline{u}) + 3a^1\nabla\chi^2 dx. \tag{4.14}$$

where $a1 > 0$. Then, since the mechanics follows linear elasticity and $\underline{u} \in H^1_\Gamma$, we can use Korn's second inequality :

$$\int_\Omega \underline{\underline{\sigma}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{u}) dx \geq C_1 \|\underline{u}\|^2_{H^1_{\Gamma_1}} \tag{4.15}$$

where $C_1 > 0$ [13]. This leads to the inequality :

$$a\left((\underline{u},\chi),(\underline{u},\chi)\right) \geq C_1\|\nabla\underline{u}\|^2_{L^2} + C_1\|\underline{u}\|^2_{L^2} + \int_\Omega 3a^1\nabla\chi^2 dx, \tag{4.16}$$

$$\geq \frac{C_1}{2}\|\nabla\underline{u}\|^2_{L^2} + \frac{C_1}{2}\|\nabla\underline{u}\|^2_{L^2} + C_1\|\underline{u}\|^2_{L^2} + 3a^1\|\nabla\chi\|^2_{L^2}. \tag{4.17}$$

Next, we use the following elementary formulas obtained by two fold integration by parts and $-\nabla = \operatorname{curl}\operatorname{curl} - \operatorname{grad}\operatorname{div}$

$$\|\nabla\underline{u}\|^2_{L^2} = \|\operatorname{curl}\underline{u}\|^2_{L^2} + \left\|\operatorname{div}(\underline{u})\right\|^2_{L^2}, \tag{4.18}$$

$$\|\nabla\underline{u}\|^2_{L^2} \geq \left\|\operatorname{div}(\underline{u})\right\|^2_{L^2}. \tag{4.19}$$

This results in

$$a\left((\underline{u},\chi),(\underline{u},\chi)\right) \geq \frac{C_1}{2}\|\nabla\underline{u}\|^2_{L^2} + C_1\|\underline{u}\|^2_{L^2} + \frac{C_1}{2}\left\|\operatorname{div}(\underline{u})\right\|^2_{L^2} + 3a^1\|\nabla\chi\|^2_{L^2} \tag{4.20}$$

Until now we did not need for $(\underline{u},\chi) \in \mathrm{Ker}B$. Nevertheless in order for the $\|\chi\|^2_{L^2}$ part of the norm $\left\|(\underline{u},\chi)\right\|^2_V$ to appear in the lower bound of the coercivity, we use $\left\|\operatorname{div}(\underline{u})\right\|^2_{L^2} = \|\chi\|^2_{L^2}$ brought by $(\underline{u},\chi) \in \mathrm{Ker}B$. Consequently, for $(\underline{u},\chi) \in \mathrm{Ker}B$, the inequality 4.20 becomes :

$$a\left((\underline{u},\chi),(\underline{u},\chi)\right) \geq \frac{C_1}{2}\|\nabla\underline{u}\|^2_{L^2} + C_1\|\underline{u}\|^2_{L^2} + \frac{C_1}{2}\|\chi\|^2_{L^2} + 3a^1\|\nabla\chi\|^2_{L^2}. \tag{4.21}$$

Finally, with $C_2 = min(\frac{C_1}{2}, 3a^1)$ and for $(\underline{u},\chi) \in \mathrm{Ker}B$

$$a\left((\underline{u},\chi),(\underline{u},\chi)\right) \geq C_2\|\underline{u}\|^2_{H^1_\Gamma} + C_2\|\chi\|^2_{H^1} \tag{4.22}$$

$$\geq C_2\left\|(\underline{u},\chi)\right\|^2_V \tag{4.23}$$

$a(.,.)$ is coercive on $\operatorname{Ker} B$, (4.12) is proven with $\alpha = C_2$.

Secondly, we prove the *inf-sup* condition of the bilinear form $b(.,.)$

$$\inf_{\lambda \in Q} \sup_{(\underline{u},\chi)\in V} \frac{b((\underline{u},\chi),\lambda)}{\left\|(\underline{u},\chi)\right\|_V \|\lambda\|_Q} \geq \beta. \tag{4.24}$$

In order to do that, we prove an equivalent formulation :

For every $\lambda \in Q$,

$$\sup_{(\underline{u},\chi)\in V} \frac{b((\underline{u},\chi),\lambda)}{\left\|(\underline{u},\chi)\right\|_V} \geq \beta \|\lambda\|_Q . \tag{4.25}$$

This allows us to focus on finding a lower bound to the sup part of the *inf-sup* condition. By only working with the supremum, if we find any $((\underline{u}_0,\chi_0) \in V | (\underline{u}_0,\chi_0) \neq (0,0))$ that verifies $\forall \lambda \in Q$,

$$\frac{b((\underline{u}_0,\chi_0),\lambda)}{\left\|(\underline{u}_0,\chi_0)\right\|_V} \geq \beta \|\lambda\|_Q$$

then (4.25) will automatically follow.
Therefore, we use the resemblance of our problem to the Stokes problem by reducing χ to 0 as such [47], $\forall \lambda \in Q$,

$$\sup_{(\underline{u},\chi)\in V} \frac{b((\underline{u},\chi),\lambda)}{\left\|(\underline{u},\chi)\right\|_V} \geq \sup_{\underline{u}\in H^1_{\Gamma_1}} \frac{b((\underline{u},0),\lambda)}{\|\underline{u}\|_{H^1}} = \sup_{\underline{u}\in H^1_{\Gamma_1}} \frac{|\int_\Omega \lambda \operatorname{div}(\underline{u}) dx|}{\|\underline{u}\|_{H^1}}.$$

So if we prove $\forall \lambda \in Q$,

$$\sup_{\underline{u}\in H^1_{\Gamma_1}} \frac{|\int_\Omega \lambda \operatorname{div}(\underline{u}) dx|}{\|\underline{u}\|_{H^1}} \geq \beta \|\lambda\|_Q \tag{4.26}$$

(4.13) will be proven. We clearly see thanks to Subsection 2.2.3 that proving (4.26) is equivalent to proving the Stokes *inf-sup* condition on $b(.,.)$ with $\underline{u} \in H^1_{\Gamma_1}$ and $\lambda \in L^2$. The majority of proofs for Stokes *inf-sup* condition on $b(.,.)$ are done using $\underline{u} \in H^1_0$ and $\lambda \in L^2_0$, this is not our case. However in [9], Bertoluzza proves the Stokes *inf-sup* condition with $\beta = 1$ using $\underline{u}_0 \in \{v \in H^1_{\Gamma_1} | v.\overrightarrow{\tau} = 0 \in \Gamma_2\}$ and $\lambda_0 \in L^2$ where $\overrightarrow{\tau}$ is the unitary tangential vector. Which results in $\forall \lambda_0 \in L^2$,

$$\sup_{\underline{u}_0} \frac{|\int_\Omega \lambda_0 \operatorname{div}(\underline{u}_0) dx|}{\|\underline{u}_0\|_{H^1}} \geq \|\lambda_0\|_{L^2} \tag{4.27}$$

Since $\underline{u}_0 \in H^1_{\Gamma_1}$ we can use (4.27) to prove (4.26) $\forall \lambda \in Q$

$$\sup_{\underline{u} \in H^1_{\Gamma_1}} \frac{|\int_\Omega \lambda \operatorname{div}(\underline{u}) dx|}{\|\underline{u}\|_{H^1}} \geq \sup_{\underline{u}_0} \frac{|\int_\Omega \lambda_0 \operatorname{div}(\underline{u}_0) dx|}{\|\underline{u}_0\|_{H^1}} \geq \|\lambda_0\|_{L^2} = \|\lambda\|_Q \tag{4.28}$$

Finally, the *inf-sup* condition (4.13) is proved with $\beta = 1$

$$\inf_{\lambda \in Q} \sup_{(\underline{u},\chi) \in V} \frac{b((\underline{u},\chi),\lambda)}{\|(\underline{u},\chi)\|_V \|\lambda\|_Q} \geq 1. \tag{4.29}$$

□

Theorem 3.2.1 is proven, we have the existence and uniqueness of a solution for the mechanics with a second gradient of dilation regularization continuous system.

Space discretion

For space discretization, we use the finite element method with Taylor-Hood $P2$-$P1$-$P1$ elements in 3D and $P2$-$P1$-$P0$ elements in 2D. This translates into using continuous piecewise quadratics functions to approximate the displacement and continuous piecewise linear functions to approximate the microscopic dilation and the Lagrange multipliers in 3D and continuous piecewise constant functions for the Lagrange multipliers in 2D. In [37] and [44], $P1$ and $P2$ elements were also tested for the Lagrange multipliers in 2D. Nevertheless $P0$ remains the best one time wise, with the same quality of solution. To verify the stability of the discretization, the *inf-sup* condition has to be fulfilled for the discrete problem as well.
Let $H^1_{\Gamma_{1h}}(\Omega)$ be the discrete $P2$ finite element space such as $H^1_{\Gamma_{1h}}(\Omega) \subset H^1_{\Gamma_1}(\Omega)$ of dimension N_u with $h > 0$ a parameter that refers to the mesh size. $H^1_h(\Omega)$ be the discrete $P1$ finite element space such as $H^1_h(\Omega) \subset H^1(\Omega)$ of dimension N_χ. $L^2_h(\Omega)$ be the discrete $P1$ finite element space such as $L^2_h(\Omega) \subset L^2(\Omega)$ of dimension N_λ.

Let $\{\phi_{\underline{v}_j}\}_{j=1}^{N_u}$ be a basis for the finite element space $(H^1_{\Gamma_{1h}}(\Omega))^d$, for all $v_h \in (H^1_{\Gamma_{1h}}(\Omega))^d$ we have

$$\underline{v}_h = \sum_{j=1}^{N_u} v_j \, \phi_{\underline{v}_j}$$

Let $\{\phi_{\psi_j}\}_{j=1}^{N_\chi}$ be a basis for the finite element space $H^1_h(\Omega)$, for all $\psi_h \in H^1_h(\Omega)$ we have

$$\psi_h = \sum_{j=1}^{N_\chi} \psi_j \, \phi_{\psi_j}$$

Let $\{\phi_{\tau_j}\}_{j=1}^{N_\lambda}$ be a basis for the finite element space $L^2_h(\Omega)$, for all $\tau_h \in L^2_h(\Omega)$ we have

$$\tau_h = \sum_{j=1}^{N_\lambda} \tau_j \, \phi_{\tau_j}$$

We consider the restriction of the bilinear forms a and b onto the discrete spaces. In addition, we consider the restriction of the operators A and B that we denote A_h and B_h.

The discrete problem is, find $(\underline{u}_h, \chi_h, \lambda_h) \in (H^1_{\Gamma_1})^d_h \times H^1_h \times L^2_h$ such as for all $(\underline{v}_h, \psi_h, \tau_h) \in (H^1_{\Gamma_1})^d_h \times H^1_h \times L^2_h$, we have

$$a\left((\underline{u}_h, \chi_h), (\underline{v}_h, \psi_h)\right) - b\left((\underline{v}_h, \psi_h), \lambda_h\right) = 0 \tag{4.30}$$

$$b\left((\underline{u}_h, \chi_h), \tau_h\right) = 0 \tag{4.31}$$

The linear system is then

$$\begin{bmatrix} \boldsymbol{A}_{\underline{u}\underline{u}} & \boldsymbol{A}_{\underline{u}\chi} & \boldsymbol{A}_{\underline{u}\lambda} \\ \boldsymbol{A}_{\chi\underline{u}} & \boldsymbol{A}_{\chi\chi} & \boldsymbol{A}_{\chi\lambda} \\ \boldsymbol{A}_{\lambda\underline{u}} & \boldsymbol{A}_{\lambda\chi} & \boldsymbol{A}_{\lambda\lambda} \end{bmatrix} \begin{bmatrix} \underline{\boldsymbol{u}} \\ \boldsymbol{\chi} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_{\underline{u}} \\ \boldsymbol{b}_{\chi} \\ \boldsymbol{b}_{\lambda} \end{bmatrix} \tag{4.32}$$

where the matrix blocks are given by

$$(\boldsymbol{A}_{\underline{u}\underline{u}})_{ij} = \int_\Omega \underline{\underline{\sigma}}(\phi_{v_i}) : \underline{\underline{\varepsilon}}(\phi_{\underline{v}_j}) + r\left(\operatorname{div}(\phi_{v_i}) \operatorname{div}(\phi_{\underline{v}_j})\right) dx, \qquad \forall i, j = 1, N_u$$

$$(\boldsymbol{A}_{\underline{u}\chi})_{ij} = \int_\Omega -r\phi_{\psi_j} \operatorname{div}(\phi_{v_i}) \, dx, \qquad \forall i = 1, N_u, \forall j = 1, N_\chi$$

$$(\boldsymbol{A}_{\chi\underline{u}})_{ij} = \int_\Omega -r\phi_{\psi_i} \operatorname{div}(\phi_{\underline{v}_j}) \, dx, \qquad \forall i = 1, N_\chi, \forall j = 1, N_u$$

$$(\boldsymbol{A}_{\chi\chi})_{ij} = \int_\Omega \underline{S}(\phi_{\psi_i}) \nabla\phi_{\psi_j} + r\phi_{\psi_i}\phi_{\psi_j} \, dx, \qquad \forall i, j = 1, N_\chi$$

$$(\boldsymbol{A}_{u\lambda})_{ij} = \int_\Omega -\phi_{\tau_j} \operatorname{div}(\phi_{v_i}) \, dx, \qquad \forall i = 1, N_u, \forall j = 1, N_\lambda$$

$$(\boldsymbol{A}_{\lambda\underline{u}})_{ij} = \int_\Omega \phi_{\tau_i} \operatorname{div}(\phi_{\underline{v}_j}) \, dx, \qquad \forall i = 1, N_\lambda, \forall j = 1, N_u$$

$$(\boldsymbol{A}_{\chi\lambda})_{ij} = \int_\Omega \phi_{\tau_j} \operatorname{div}(\phi_{\psi_i}) \, dx, \qquad \forall i = 1, N_\chi, \forall j = 1, N_\lambda$$

$$(\boldsymbol{A}_{\lambda\chi})_{ij} = \int_\Omega -\phi_{\tau_i} \operatorname{div}(\phi_{\psi_j}) \, dx, \qquad \forall i = 1, N_\lambda, \forall j = 1, N_\chi$$

$$(\boldsymbol{A}_{\lambda\lambda})_{ij} = \int_\Omega 0 \, dx, \qquad \forall i, j = 1, N_\lambda$$

and each unknown array is given by :

$$\underline{\boldsymbol{u}} = [u_i]_{i=1,N_u}, \boldsymbol{\chi} = [\chi_i]_{i=1,N_\chi}, \boldsymbol{\lambda} = [\lambda]_{i=1,N_\lambda}$$

Remark 3. *The system easily can be symmetric, if* (4.5) *is multiplied by -1. Then* (4.31) *would be multiplied by -1 and*

$$(\boldsymbol{A}_{\lambda\underline{u}}) = \int_\Omega -\phi_{\tau_i}\, div(\phi_{\underline{v}_j})dx, \qquad \forall i = 1, N_\lambda, \forall j = 1, N_u$$
$$(\boldsymbol{A}_{\lambda\chi}) = \int_\Omega \phi_{\tau_i}\, div(\phi_{\psi_j})dx, \qquad \forall i = 1, N_\lambda, \forall j = 1, N_\chi$$

The well-posedness of the discrete problem does not automatically follow from the continuous case. For this, we can use the same reasoning as for the continuous case but with the discrete spaces.

Theorem 4.2.2. *Let* $V_h := (H^1_{\Gamma_{1h}}(\Omega))^d \times H^1_h(\Omega)$, $Q_h := L^2_h(\Omega)$, $a(.,.)$ *be defined as in* (4.9) *and* $b(.,.)$ *be defined as in* (4.10).
There is a constant $\alpha > 0$ *, such that*

$$a\left((\underline{u}_h, \chi_h), (\underline{u}_h, \chi_h)\right) \geq \alpha \left\|(\underline{u}_h, \chi_h)\right\|^2_V, \ (\underline{u}_h, \chi_h) \in KerB_h \tag{4.33}$$

and there is a constant $\beta > 0$*, such that*

$$\inf_{\lambda_h \in Q_h} \sup_{(\underline{u}_h, \chi_h) \in V_h} \frac{b((\underline{u}_h, \chi_h), \lambda_h)}{\left\|(\underline{u}_h, \chi_h)\right\|_V \|\lambda_h\|_Q} \geq \beta. \tag{4.34}$$

We can conclude that there exist a unique solution $(\underline{u}_h, \chi_h, \lambda_h) \in V_h \times Q_h$ *for the system* (4.11) *and the discretization is stable.*

Proof. In order to prove the Theorem 4.2.2, we first prove the coercivity of the bilinear form $a(.,.)$ on the Kernel of B_h, then we prove the *inf-sup* condition of the bilinear form $b(.,.)$.

To prove (4.33), we need to prove it for $r = 0$ and for $r > 0$. Since for $r = 0$, the coercivity of $a(.,.)$ holds on the whole V space for the continuous case, it automatically follows in the discrete case. For $r > 0$, the continuous coercivity condition does not hold on the whole V space but on the Kernel of B_h, the discrete coercivity condition does not automatically follows from it. In order to prove (4.33), we follow similar steps as the continuous proof. We start with :

$$a\left((\underline{u}_h, \chi_h), (\underline{u}_h, \chi_h)\right) = \int_\Omega \underline{\underline{\sigma}}(\underline{u}_h) : \underline{\underline{\varepsilon}}(\underline{u}_h) + 3a^1 \nabla \chi_h^2 dx \tag{4.35}$$

Since $H^1_{\Gamma_{1_h}}(\Omega) \subset H^1_{\Gamma_1}(\Omega)$ and Korn's second inequality holds on the whole $H^1_{\Gamma_1}(\Omega)$ continuous space, the discrete form will automatically follow with $C_1 > 0$.

$$\int_\Omega \underline{\underline{\sigma}}(\underline{u}_h) : \underline{\underline{\varepsilon}}(\underline{u}_h) dx \geq C_1 \|\underline{u}_h\|^2_{H^1_{\Gamma_1}} \ , \ u_h \in (H^1_{\Gamma_{1_h}})^d$$

This leads to inequality :

$$a\left((\underline{u}_h, \chi_h), (\underline{u}_h, \chi_h)\right) \geq C_1 \|\nabla \underline{u}_h\|^2_{L^2} + C_1 \|\underline{u}_h\|^2_{L^2} + \int_\Omega 3a^1 \nabla \chi_h^2 dx \tag{4.36}$$

$$\begin{aligned} &\geq \frac{C_1}{2} \|\nabla \underline{u}_h\|^2_{L^2} + C_1 \|\underline{u}_h\|^2_{L^2} + \frac{C_1}{2} \|\nabla \underline{u}_h\|^2_{L^2} \\ &+ 3a^1 \|\nabla \chi_h\|^2_{L^2} \, dx \end{aligned} \tag{4.37}$$

And by the same reasoning as with Korn's inequality, since $H^1_{\Gamma_{1_h}}(\Omega) \subset H^1_{\Gamma_1}(\Omega)$ the discrete form of inequality 4.18 follows for the continuous case such as :

$$\|\nabla \underline{v}_h\|^2_{L^2} \geq \left\| \operatorname{div}(\underline{v}_h) \right\|^2_{L^2}. \tag{4.38}$$

This results in :

$$\begin{aligned} a\left((\underline{u}_h, \chi_h), (\underline{u}_h, \chi_h)\right) \geq \frac{C_1}{2} \|\nabla \underline{u}_h\|^2_{L^2} + C_1 \|\underline{u}_h\|^2_{L^2} + \frac{C_1}{2} \left\| \operatorname{div}(\underline{v}_h) \right\|^2_{L^2} \\ + 3a^1 \|\nabla \chi_h\|^2_{L^2} \, dx \end{aligned} \tag{4.39}$$

As for the continuous form, in order for the $\|\chi_h\|^2_{L^2}$ part of the norm $\left\|(\underline{u}_h, \chi_h)\right\|^2_V$ to appear in the lower bound of the coercivity, we use $\left\| \operatorname{div}(\underline{v}_h) \right\|^2_{L^2} = \|\chi_h\|^2_{L^2}$ brought by $(\underline{u}_h, \chi_h) \in \operatorname{Ker} B_h$. Consequently, for $(\underline{u}_h, \chi_h) \in \operatorname{Ker} B_h$,

$$\begin{aligned} a\left((\underline{u}_h, \chi_h), (\underline{u}_h, \chi_h)\right) \geq \frac{C_1}{2} \|\nabla \underline{u}_h\|^2_{L^2} + C_1 \|\underline{u}_h\|^2_{L^2} + \frac{C_1}{2} \|\chi_h\|^2_{L^2} \\ + 3a^1 \|\nabla \chi_h\|^2_{L^2} \, dx \end{aligned} \tag{4.40}$$

Finally, with $C_2 = min(\frac{C_1}{2}, 3a^1)$

$$a\left((\underline{u}, \chi), (\underline{u}, \chi)\right) \geq C_2 \left\|(\underline{u}_h, \chi_h)\right\|^2_V \tag{4.41}$$

$a(.,.)$ is coercive on $\operatorname{Ker} B_h$.

Secondly we prove the condition (4.34)

$$\inf_{\lambda_h \in Q_h} \sup_{(\underline{u}_h, \chi_h) \in V_h} \frac{b((\underline{u}_h, \chi_h), \lambda_h)}{\left\|(\underline{u}_h, \chi_h)\right\|_V \|\lambda_h\|_Q} \geq \beta.$$

As in the continuous proof, we use an equivalent formulation, $\forall \lambda_h \in Q_h$

$$\sup_{(\underline{u}_h,\chi_h)\in V_h} \frac{b((\underline{u}_h,\chi_h),\lambda_h)}{\left\|(\underline{u}_h,\chi_h)\right\|_V} \geq \beta \|\lambda_h\|_Q \,.$$

and reduce χ_h to 0 [47],$\forall \lambda_h \in Q_h$,

$$\sup_{(\underline{u}_h,\chi_h)\in V_h} \frac{b((\underline{u}_h,\chi_h),\lambda_h)}{\left\|(\underline{u}_h,\chi_h)\right\|_V} \geq \sup_{\underline{u}_h\in H^1_{\Gamma_{1h}}} \frac{b((\underline{u}_h,0),\lambda_h)}{\|\underline{u}_h\|_{H^1}} = \sup_{\underline{u}_h\in H^1_{\Gamma_{1h}}} \frac{|\int_\Omega \lambda_h \operatorname{div}(\underline{u}_h)dx|}{\|\underline{u}_h\|_{H^1}}.$$

Finally, in [9] Bertoluzza also proved the Stokes discrete *inf-sup* condition on $b(.,.)$ using $\beta = 1$ using $\underline{u}_{h0} \in \{v_h \in H^1_{\Gamma_{h1}} | v_h.\overrightarrow{\tau} = 0 \in \Gamma_2\}$ and $\lambda_{h0} \in L^2_h$ where$\overrightarrow{\tau}$ is the unitary tangential vector. Which results in $\forall \lambda_{h0} \in L^2_h$,

$$\sup_{\underline{u}_{h0}} \frac{|\int_\Omega \lambda_{h0} \operatorname{div}(\underline{u}_{h0})dx|}{\left\|\underline{u}_{h0}\right\|_{H^1}} \geq \|\lambda_{h0}\|_{L^2} \tag{4.42}$$

Since $\underline{u}_{h0} \in (H^1_{\Gamma_{h1}})^d$ we can use (4.42) to prove $\forall \lambda_h \in Q_h$

$$\begin{aligned}\sup_{\underline{u}_h\in H^1_{\Gamma_{h1}}} \frac{|\int_\Omega \lambda_h \operatorname{div}(\underline{u}_h)dx|}{\|\underline{u}_h\|_{H^1}} &\geq \sup_{\underline{u}_{h0}} \frac{|\int_\Omega \lambda_{h0} \operatorname{div}(\underline{u}_{h0})dx|}{\left\|\underline{u}_{h0}\right\|_{H^1}} \\ &\geq \|\lambda_{h0}\|_{L^2} = \|\lambda_h\|_Q\end{aligned} \tag{4.43}$$

The discrete *inf-sup* condition (4.34) is proven with $\beta = 1$

$$\inf_{\lambda_h\in Q_h} \sup_{(\underline{u}_h,\chi_h)\in V_h} \frac{b((\underline{u}_h,\chi_h),\lambda_h)}{\left\|(\underline{u}_h,\chi_h)\right\|_V \|\lambda_h\|_Q} \geq 1. \tag{4.44}$$

□

Theorem 4.2.2 is proven, we have the existence and uniqueness of a solution for the mechanics with a second gradient of dilation regularization system discretized with $P2$-$P1$-$P1$ finite elements.

4.2.1 Preconditioner

In order to construct a preconditioner for the mechanics with a second gradient of dilation regularization problem, we follow the approach developed by Mardal and Winther, explained in the first chapter[64]. The technique consist in finding an appropriate preconditioner for the continuous problem and applying a stable finite element discretization in order to identify a correct preconditioner for the discrete problem.

Since we follow the theory developed in [64], it is easier to introduce the same notations, in the same manner that was done in the first chapter.

First, we propose a preconditioner for the continuous problem. In order to do so we start by introducing the coefficient operator corresponding to the mechanics with a second gradient of dilation regularization system. Since the well-posedness proofs where done with $r = 0$, we start with that case :

$$\mathcal{A}_0 = \begin{pmatrix} -\Delta & 0 & \operatorname{grad} \\ 0 & -\Delta & I \\ \operatorname{div} & -I & 0 \end{pmatrix}.$$

As a consequence of the two Brezzi conditions being met, $\mathcal{A}_0$ is an isomorphism mapping $H^1_{\Gamma_1} \times H^1 \times L^2$ onto $H^{-1} \times H^{-1} \times L^2$ where H^{-1} denotes the dual of H^1. The canonical choice of a preconditioner is then the block diagonal operator:

$$\mathcal{B} = \begin{pmatrix} (-\Delta)^{-1} & 0 & 0 \\ 0 & (-\Delta)^{-1} & 0 \\ 0 & 0 & I \end{pmatrix}$$

mapping the space $H^{-1} \times H^{-1} \times L^2$ onto $H^1_{\Gamma_1} \times H^1 \times L^2$. Unfortunately, the second gradient of dilation system depends on the penalization parameter r. Consequently a parameter dependent approach could be more suited. In order to do so we introduce the coefficient operator corresponding to the problem with $r \in [0, 1.\text{e+}14]$.

$$\mathcal{A}_r = \begin{pmatrix} -\Delta - r \operatorname{grad}\operatorname{div} & r \operatorname{grad} & \operatorname{grad} \\ r \operatorname{div} & -\Delta - rI & I \\ \operatorname{div} & -I & 0 \end{pmatrix}.$$

We can now propose a parameter dependent preconditioner for the continuous problem $\mathcal{A}_r$ where each operator can be replaced by spectrally equivalent ones.

$$\mathcal{B}_r = \begin{pmatrix} (-\Delta - r \operatorname{grad}\operatorname{div})^{-1} & 0 & 0 \\ 0 & (-\Delta - r\, I)^{-1} & 0 \\ 0 & 0 & I^{-1} \end{pmatrix}$$

$\mathcal{B}_r$ maps isomorphically $H^{-1} \times H^{-1} \times L^2$ onto $H^1_{\Gamma_1} \times H^1 \times L^2$. This is due to the operators $\operatorname{grad}\operatorname{div} : H(\operatorname{div}) \to H(\operatorname{div})^*$ and $I : L^2 \to L^2$ mapping weaker spaces than the operator $\Delta : H^1 \to H^{-1}$ such as $H^1 \subset H(\operatorname{div}) \subset L^2$. Additionally, since the operator Δ does not depend on r, it remains present inside $-\Delta - r \operatorname{grad}\operatorname{div}$ and $-\Delta - r\, I$ for $r \in [0, 1.\text{e+}14]$. Resulting in $\mathcal{B}_r$ mapping isomorphically $H^{-1} \times H^{-1} \times L^2$ onto $H^1_{\Gamma_1} \times H^1 \times L^2$ for $r \in [0, 1.\text{e+}14]$.

Remark 4. *Since, even with the inclusion of r in the preconditioner, $\mathcal{B}_r$ continues to map isomorphically $H^{-1}\times H^{-1}\times L^2$ onto $H^1_{\Gamma_1}\times H^1\times L^2$, the introduction of parameter dependent norms were not needed. Nevertheless, as future work, we would like to include other parameters in the preconditioners and consider a parameter-dependent approach.*

We consider now a $P2$-$P1$-$P1$ discretization, with the spaces $V_h \times Q_h$ defined above. The discretization is stable in the norm of $H^1_{\Gamma_1} \times H^1 \times L^2$. The proposed discrete preconditioner is then

$$\mathcal{B}_{r,h} = \begin{pmatrix} (-\Delta - r\,\mathrm{grad}\,\mathrm{div})_h^{-1} & 0 & 0 \\ 0 & (-\Delta - r\,I)_h^{-1} & 0 \\ 0 & 0 & I_h^{-1} \end{pmatrix}$$

From here, we introduce the preconditioners we use in our computations, based on $\mathcal{B}_{r,h}$. Inside the preconditioners, we replace each discrete operator $(-\Delta - r\,\mathrm{grad}\,\mathrm{div})_h^{-1}$, $(-\Delta - r\,I)_h^{-1}$ and I_h^{-1} by corresponding efficient preconditioners. We start by choosing spectrally equivalent operators for $(-\Delta - r\,\mathrm{grad}\,\mathrm{div})_h$, $(-\Delta - r\,I)_h$ and I_h to afterwards choose an appropriate method to invert each operator.

First $(-\Delta - r\,\mathrm{grad}\,\mathrm{div})_h$ and $(-\Delta - r\,I)_h$ are represented in the preconditioner by $\boldsymbol{A}_{\underline{uu}}$ and $\boldsymbol{A}_{\chi\chi}$ respectively. $\boldsymbol{A}_{\underline{uu}}$ and $\boldsymbol{A}_{\chi\chi}$ are essentially equivalent to the operators $(-\Delta - r\,\mathrm{grad}\,\mathrm{div})_h$ and $(-\Delta - r\,I)_h$ with the addition of physical parameters. Here I_h is the Riez operator mapping L^2 onto L^2 and is represented by the discrete mass matrix and not the identity matrix. We have found empirically that scaling the discrete mass matrix by $-\frac{1}{r}$ adds robustness to the preconditioner. Finally, since the inverses are never explicitly computed, we chose methods to approximate them.

We shall use one V-cycle of an algebraic multigrid solver (AMG) as a preconditioner. In Chapter 3, we saw that this method is especially suited for elliptic and non-degenerate parabolic operators which is the case here with the lower values of r. For $\boldsymbol{A}_{\underline{uu}}^{-1}$ and $\boldsymbol{A}_{\chi\chi}^{-1}$, we also test the adding of an accelerator, the result is the inverse approximation computed trough an FGMRES method with one V-cycle AMG preconditioner. For each approximation we chose a certain tolerance for FGMRES, this allows for a better quality of the inverse of each block and we can have a decrease in the number of outer iterations. As for the outer iterative method, since the second gradient of dilation system is anti-symmetric and the preconditioner can vary at each iteration when we add the inner accelerator, we need a flexible version of the iterative solver, namely FGMRES.

Four preconditioners are tested, two diagonal ones $\boldsymbol{P}_{JH}$ and $\boldsymbol{P}_{JFg}$, where J stand

for a Block Jacobi type preconditioner,

$$\boldsymbol{P}_{JH}^{-1} = \begin{bmatrix} \tilde{\boldsymbol{A}}_{\underline{uu}}^{-1} & 0 & 0 \\ 0 & \tilde{\boldsymbol{A}}_{\chi\chi}^{-1} & 0 \\ 0 & 0 & \tilde{\boldsymbol{M}}_{\lambda}^{-1} \end{bmatrix}, \qquad \boldsymbol{P}_{JFg}^{-1} = \begin{bmatrix} \hat{\boldsymbol{A}}_{\underline{uu}}^{-1} & 0 & 0 \\ 0 & \hat{\boldsymbol{A}}_{\chi\chi}^{-1} & 0 \\ 0 & 0 & \tilde{\boldsymbol{M}}_{\lambda}^{-1} \end{bmatrix}.$$

where the $\tilde{}$ symbol represents the inverse approximation trough one AMG V-cycle and the $\hat{}$ symbol trough an FGMRES method with one V-cycle AMG preconditioner where the stopping criterion is a relative tolerance set to 10^{-3}.
We also test two lower triangular ones $\boldsymbol{P}_{GH}$ and $\boldsymbol{P}_{GFg}$, where G stand for a Block Gauss-Seidel type preconditioner,

$$\boldsymbol{P}_{GH}^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \tilde{\boldsymbol{M}}_{\lambda}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -\boldsymbol{A}_{\lambda\underline{u}} & -\boldsymbol{A}_{\lambda\chi} & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & \tilde{\boldsymbol{A}}_{\chi\chi}^{-1} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ \boldsymbol{A}_{\chi\underline{u}} & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{A}}^{-1} & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & I \end{bmatrix}$$

$$\boldsymbol{P}_{GFg}^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \tilde{\boldsymbol{M}}_{\lambda}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -\boldsymbol{A}_{\lambda\underline{u}} & -\boldsymbol{A}_{\lambda\chi} & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & \hat{\boldsymbol{A}}_{\chi\chi}^{-1} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ \boldsymbol{A}_{\chi\underline{u}} & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{A}}_{\underline{uu}}^{-1} & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & I \end{bmatrix}$$

4.2.2 Numerical results

The robustness and efficiency of the proposed solver are crucial for industrial applications. We thus turn to the presentation of the results of an illustrative test case, challenging the preconditioner's robustness by varying some parameters. The method is implemented in Firedrake [77]

Test Case

A similar test case as the model problem for the THM is used in this chapter. The test case needs to be simple enough so that the mesh can be easily refined but complex enough to resemble the industrial problem in consideration. For this purpose, a 3D rectangular sample is modelled as seen in Figure 4.5, with a 0.5 m length in x-direction, a 0.5 m height in y-direction and 1 m large in z-direction. The tetrahedral mesh was generated using the BoxMesh function of Firedrake.
The displacement was set to 0 on the bottom surface ($y = 0$), a mechanical pressure of 5 MPa was applied on the top surface ($y = 0.1$).
The sample consists of clay only. We emphasize that the value of the material parameters are of great importance in the industrial applications. They are displayed in Table 4.1, and are representative of a typical industrial problem of geological waste disposal [44].
The tests were solved with Firedrake using the mechanics with a second gradient of dilation regularization framework presented in the previous section, where $P2$-$P1$-$P1$ finite elements are used.

Figure 4.5: Test case

Table 4.1: The test case parameters

Clay		
Symbol	Value	Unit
E	6.10^9	Pa
ν	0.3	-
a_1	500	-
r	10^{10}	-

Preconditioner's performance

Next, the two diagonal and two lower triangular preconditioners presented above are tested on the model problem. The solution time of each preconditioner is compared to the solution time of the direct solver MUMPS. We use the hierarchy function of firedrake to discretize the mesh. The number of iterations of the external FGMRES solver of each preconditioner is computed for every mesh between 97 DoF and 895 749 DoF.

First, the results from the block Jacobi preconditioners are presented in Table 4.2. Once the DoF are greater than 429, the number of iterations $\boldsymbol{P}_{JFg}$ and $\boldsymbol{P}_{JH}$ are independent of the mesh size. By adding the internal FGMRES accelerators in $\boldsymbol{P}_{JFg}$ there is a decrease of 10 iterations. Time wise once 895 749 DoF are reached, both preconditioners beat MUMPS, with $\boldsymbol{P}_{JH}$ being the fastest one which is to be expected since $\boldsymbol{P}_{JFg}$ has two internal FGMRES loops.

Then, the results from the lower block Gauss-Seidel preconditioners are presented in Table 4.3. Overall, the number of iterations $\boldsymbol{P}_{GFg}$ and $\boldsymbol{P}_{GH}$ are independent of the

Table 4.2: Block Jacobi preconditioners results

	MUMPS	$\boldsymbol{P}_{JFg}$		$\boldsymbol{P}_{JH}$	
size	**time**	**it**	**time**	**it**	**time**
97	2.38e-01	20	1.15e-01	28	6.71e-02
429	9.01e-02	28	1.85e-01	37	3.07e-01
2437	2.48e-01	30	1.33e+00	39	3.47e-01
16197	1.86e+00	30	1.70e+01	40	3.55e+00
117637	2.58e+01	30	1.61e+02	40	3.65e+01
895749	2.77e+03	30	1.54e+03	41	3.43e+02

mesh size even with the smallest meshes. By adding the internal FGMRES accelerators in $\boldsymbol{P}_{GFg}$ there is a decrease of 10 iterations. Time wise once 895 749 are reached, both preconditioners beat MUMPS, with $\boldsymbol{P}_{GH}$ being the fastest one beating MUMPS even at 117 637 DoF.

Table 4.3: Lower Block Gauss-Seidel preconditioners results

	MUMPS	$\boldsymbol{P}_{GFg}$		$\boldsymbol{P}_{GH}$	
size	**time**	**it**	**time**	**it**	**time**
97	2.38e-01	10	6.54e-02	15	7.81e-02
429	9.01e-02	11	1.05e-01	18	2.27e-01
2437	2.48e-01	12	6.88e-01	19	2.34e-01
16197	1.86e+00	12	7.27e+00	21	2.28e+00
117637	2.58e+01	12	7.02e+01	21	2.27e+01
895749	2.77e+03	12	6.75e+02	22	2.18e+02

As for the comparison between both types of preconditioners, the lower block Gauss-Seidel preconditioners divided by 2 the number of iterations in comparison to the block Jacobi ones. Even if overall the preconditioners are robust in regard to the mesh size, $\boldsymbol{P}_{GFg}$ is extremely robust since it goes from 10 iteration for 97 DoF to only 12 iterations for 895 749 DoF. Since the next step is to test the preconditioner's parameter robustness, we chose to continue with the one with the lowest number of iterations and the most robust with respect to the mesh size, $\boldsymbol{P}_{GFg}$.

Robustness

The robustness of $\boldsymbol{P}_{GFg}$ is evaluated by varying the values of the Young's modulus E and the penalization parameter r. These parameters are chosen since they appear in the first two balance equations and can have large variations. The tests are done using the test case of Figure 4.5 made of clay with the parameters in Table 4.1.
The results are compiled in Table 4.4. The number of outer FGMRES iterations are

displayed. The very large range of variation of the mesh size and of each parameter (up to 14 orders of magnitude) is emphasized. We chose this large variation in r in order to test all of the values evaluated by Fernandes in [36] where $r \in [1.e{+}8, 1.e{+}14]$ and by Gantier in [44] where $r \in 0, 1.e+10, 4.e+11$. When $r > 0$, the value of r is chosen as a ratio with respect to E in order to properly evaluate the influence of the penalty term which explains the large values.
In order to analyse the results in Table 4.4, we propose first a row-wise reading then

Table 4.4: Lower Block Gauss-Seidel parameter robustness

Parameters		DoF		
E	r	16 197	117 637	895 749
1.e+9	0	-	-	-
	1.e+8	6	8	13
	1.e+10	25	25	25
	4.e+11	96	106	108
	1.e+14	-	-	-
2.5e+10	0	-	-	-
	1.e+8	5	6	10
	1.e+10	7	7	7
	4.e+11	30	30	30
	1.e+14	-	-	-
5.0e+10	0	-	-	-
	1.e+8	4	5	9
	1.e+10	6	6	6
	4.e+11	22	22	22
	1.e+14	-	-	-

a column-wise reading.

The row-wise reading provides information on the influence of the mesh size, the material parameters being fixed. A good independence with respect to the mesh size is observed. The outer number of FGMRES iterations remains constant enough even though the size of the system is multiplied by 500, with a slight exception for r=1.e+8 where the number of iterations tends to increase for the biggest mesh.
The column-wise reading provides information on the influence of the material parameters, the mesh size being fixed. We start with the cases where the preconditioner was successful . A variation of the outer number of FGMRES iterations is observed, remaining between 4 and 30, except for the "worst" set of parameters (E=1.e+9, r=4.e+11), where it reaches up to 108 iterations. The preconditioner tends to show better robustness when r is closer in order of magnitude to E. This is clearly emphasised by the fact that for the two outlier values of r, 0 and 1.e+14, the preconditioner did not converge. This is due to two different reasons.

First, since the operator for the displacement block is $-\nabla - r\,\mathrm{grad}\,\mathrm{div}$, when r=1.e+14 the grad div operator becomes dominant. This results in the operator no longer being elliptic and the AMG preconditioner is no longer adapted [64]. Secondly, in the continuous proof, when r=0 the bilinear form $a(.,.)$ is no longer coercive on the whole space but only on the Kernel of B. The loss of coercivity of the bilinear form a is due to the loss of coercivity of the matrix micro-volume changes χ, where the mass matrix part of the matrix disappears when r=0.

4.3 Application to linear Thermo-Hydro-Mechanics

Now we test the preconditioners on the whole problem, i.e. Thermo-Hydro-Mechanics with a second gradient of dilation regularization. The specific fluid enthalpy, h_f, has been neglected in order to simplify the energy conservation equation. In addition, the non-time-dependent T inside the non-convective heat, Q', is replaced by the reference temperature T_0. As a consequence of these strong simplifications, the problem becomes linear. Nevertheless the preconditioning strategy is not compromised in the general case.
Let Ω be a d dimensional domain, $1 \leq d \leq 3$, and t_f the final time of the simulation. The coupled system consists of , $\forall x \in \Omega$ and $\forall t > 0 \in [0, t_f]$,

$$
\begin{aligned}
-\operatorname{div}(\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u})) + \nabla p + 3K_s\alpha_s\nabla T + \nabla\lambda - r\nabla(\operatorname{div}(\underline{u})) + r\nabla\chi &= \underline{f}^e \quad \text{in } \Omega \times (0,t_f) \\
-\operatorname{div}(\rho_f\lambda_H\nabla p) + \rho_f(\operatorname{div}(\underline{\dot{u}}) + \frac{\varphi}{K_l}\dot{p} - \alpha_m 3\dot{T}) &= 0 \quad \text{in } \Omega \times (0,t_f) \\
-\operatorname{div}(\lambda_T\nabla T) + (3K_0\alpha_s\operatorname{div}(\underline{\dot{u}}) - 3\alpha_m\dot{p} - 9K_0\alpha_s^2\dot{T})T_0 + C_\sigma^0\dot{T} &= \Theta \quad \text{in } \Omega \times (0,t_f) \\
\lambda - \operatorname{div}(\underline{S}(\chi)) - r\operatorname{div}(\underline{u}) + r\chi &= 0 \quad \text{in } \Omega \times (0,t_f) \\
\operatorname{div}(\underline{u}) - \chi &= 0 \quad \text{in } \Omega \times (0,t_f)
\end{aligned}
$$

The boundary of Ω is denoted $\partial\Omega$ and six different partitions are needed to define the boundary conditions. We apply Dirichlet boundary conditions to the displacement, pressure and temperature. Neumann boundary conditions for the the displacement $\underline{u}$ follow the stress $\underline{\sigma}$, the ones for the pressure $\boldsymbol{P}$ follow the fluid flux q and the ones for the temperature T follow the thermal flux Ψ.
We thus have, respectively, the boundary conditions on the displacement unknowns,

on the pressure unknowns and on the temperature unknowns such as:

$$\partial\Omega = \partial\Omega^{\underline{u}} \cup \partial\Omega^{\underline{t}} \text{ with } \partial\Omega^{\underline{u}} \cap \partial\Omega^{\underline{t}} = \emptyset$$
$$\partial\Omega = \partial\Omega^{p} \cup \partial\Omega^{q} \text{ with } \partial\Omega^{p} \cap \partial\Omega^{q} = \emptyset$$
$$\partial\Omega = \partial\Omega^{T} \cup \partial\Omega^{\Psi} \text{ with } \partial\Omega^{T} \cap \partial\Omega^{\Psi} = \emptyset$$

The boundary conditions are given by :

$$\begin{aligned}
\underline{\underline{\sigma}}(\underline{u}) \cdot \underline{n} &= \underline{t}^e && \text{on } \partial\Omega^{\underline{t}} \times (0, t_f) \\
\underline{\underline{S}}(\chi) \cdot \underline{n} &= 0 && \text{on } \partial\Omega \times (0, t_f) \\
-\lambda_H \nabla p \cdot \underline{n} &= q^e && \text{on } \partial\Omega^{q} \times (0, t_f) \\
-\lambda_T \nabla T \cdot \underline{n} &= \Psi^e && \text{on } \partial\Omega^{\Psi} \times (0, t_f) \\
\underline{u} &= \underline{u}^e && \text{on } \partial\Omega^{\underline{u}} \times (0, t_f) \\
p &= p^e && \text{on } \partial\Omega^{p} \times (0, t_f) \\
T &= T^e && \text{on } \partial\Omega^{T} \times (0, t_f) \\
\underline{u}(\,,0) &= \underline{u}_0 && \text{in } \Omega \\
p(\,,0) &= p_0 && \text{in } \Omega \\
T(\,,0) &= T_0 && \text{in } \Omega
\end{aligned}$$

where $\underline{n}$ is the outward normal.
For the discretization of the problem, the same steps are followed as in the THM chapter and the section above. An implicit Euler is applied for time discretization and P2-P1-P1-P1-P1 finite elements are considered for space discretization. The linear system to solve is of the structure

$$\begin{bmatrix}
\boldsymbol{A}_{\underline{u}\underline{u}} & \boldsymbol{A}_{\underline{u}\chi} & \boldsymbol{A}_{\underline{u}\lambda} & \boldsymbol{A}_{\underline{u}p} & \boldsymbol{A}_{\underline{u}T} \\
\boldsymbol{A}_{\chi\underline{u}} & \boldsymbol{A}_{\chi\chi} & \boldsymbol{A}_{\chi\lambda} & 0 & 0 \\
\boldsymbol{A}_{\lambda\underline{u}} & \boldsymbol{A}_{\lambda\chi} & \boldsymbol{A}_{\lambda\lambda} & 0 & 0 \\
\boldsymbol{A}_{p\underline{u}} & 0 & 0 & \boldsymbol{A}_{pp} & \boldsymbol{A}_{pT} \\
\boldsymbol{A}_{T\underline{u}} & 0 & 0 & \boldsymbol{A}_{Tp} & \boldsymbol{A}_{TT}
\end{bmatrix}
\begin{bmatrix}
\underline{\boldsymbol{u}} \\ \boldsymbol{\chi} \\ \boldsymbol{\lambda} \\ \boldsymbol{p} \\ \boldsymbol{T}
\end{bmatrix}^{n+1}
=
\begin{bmatrix}
\boldsymbol{b}_{\underline{u}} \\ \boldsymbol{b}_{\chi} \\ \boldsymbol{b}_{\lambda} \\ \boldsymbol{b}_{p} \\ \boldsymbol{b}_{T}
\end{bmatrix} \tag{4.45}$$

4.3.1 Preconditioner

The reason we decided to separate the THM from the mechanics with a second gradient of dilation regularization is because there is no coupling between (p, T) and (χ, λ). This allowed us to treat the difficulties inside each sub-problem individually in order to propose a suitable preconditioner. Since there is no coupling between both sub-problems, the final preconditioner is a merger between both sub-preconditioner and should keep the positive qualities of each one. The best preconditioner for the THM problem is $\boldsymbol{P}_{LGS}$ and for the second gradient of dilation $\boldsymbol{P}_{Gfg}$. The final preconditioner

is

$$\boldsymbol{P}_{Final} = \begin{bmatrix} \boldsymbol{A}_{\underline{u}\underline{u}} & 0 & 0 & 0 & 0 \\ \boldsymbol{A}_{\chi\underline{u}} & \boldsymbol{A}_{\chi\chi} & 0 & 0 & 0 \\ \boldsymbol{A}_{\lambda\underline{u}} & \boldsymbol{A}_{\lambda\chi} & \boldsymbol{M}_{\lambda} & 0 & 0 \\ \boldsymbol{A}_{p\underline{u}} & 0 & 0 & \boldsymbol{A}_{pp} & 0 \\ \boldsymbol{A}_{Tu} & 0 & 0 & \boldsymbol{A}_{Tp} & \boldsymbol{A}_{TT} \end{bmatrix}.$$

$\boldsymbol{P}_{Final}^{-1}$ is applied as a lower block Gauss-Seidel preconditioner. In order to apply it, we only need to approximate the inverses of the diagonal blocks. The same approximation as in $\boldsymbol{P}_{LGS}$ and $\boldsymbol{P}_{Gfg}$ are used. This translates into the inverses of $\boldsymbol{A}_{\underline{u}\underline{u}}^{-1}$,$\boldsymbol{A}_{\chi\chi}^{-1}$,$\boldsymbol{A}_{pp}^{-1}$,and $\boldsymbol{A}_{TT}^{-1}$ approximated with an FGMRES method with a single V-cycle of an AMG. The stopping criterion of each inner FGMRES are, 10 iterations for $\boldsymbol{A}_{\underline{u}\underline{u}}^{-1}$, three iterations for $\boldsymbol{A}_{pp}^{-1}$, three iterations for $\boldsymbol{A}_{TT}^{-1}$ and a relative tolerance of 10^{-3} for $\boldsymbol{A}_{\chi\chi}^{-1}$. The inverse of $\boldsymbol{M}_{\lambda}^{-1}$ is approximated with one V-cycle of an AMG.

4.3.2 Numerical results

Test case

We use the same test as in the section above, Figure 4.5, where we add Dirichlet boundary conditions to the temperature. It is a 3D rectangular sample is modelled

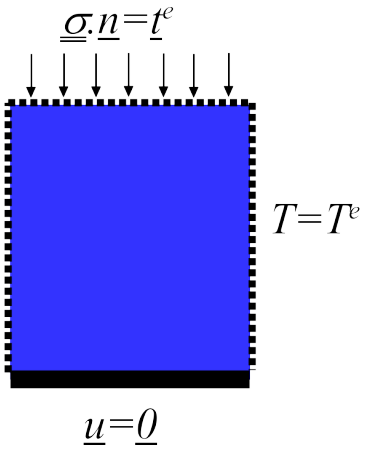


Figure 4.6: Test case

as seen in Figure 4.6, with a 0.5 m length in x-direction, a 0.5 m height in y-direction and 1 m large in z-direction. The tetrahedral mesh was generated using the BoxMesh function of Firedrake.
The displacement was set to 0 on the bottom surface ($y = 0$), a mechanical pressure of 5 MPa was applied on the top surface ($y = 0.1$) and a temperature of 80°C is

imposed on the whole surface of the sample.
The sample consists of clay only. We emphasize that the value of the material parameters are of great importance in the industrial applications. They are displayed in Table 4.1 and in Table 3.3, and are representative of a typical industrial problem of geological waste disposal [44].
The tests were solved with Firedrake using the Thermo-Hydro-Mechanics with a second gradient of dilation regularization framework presented above, where $P2$-$P1$-$P1$-$P1$-$P1$ finite elements are used.

Robustness

Since the performance of each sub-preconditioner was already tested, for the final problem we directly test the robustness of $\boldsymbol{P}_{Final}$ by varying the values of the Young's modulus E, the intrinsic permeability K_{int} and the penalization parameter r. Since the second gradient of dilation is no longer adapted when $r = 0$ and when $r = 1.e{+}14$, we test the final preconditioner with $r \in [1.e+8, 1.e+10]$. The results are compiled in Table 4.5. The maximum number of outer FGMRES iterations are displayed.

Table 4.5: Lower Block Gauss-Seidel parameter robustness

Parameters			DoF		
E	K_{int}	r	17 655	127 463	967 623
1.e+9	4.e-15	1.e+8	16	18	24
		1.e+10	22	23	23
	4.e-18	1.e+8	21	29	49
		1.e+10	10	14	17
	4.e-21	1.e+8	19	24	37
		1.e+10	9	9	10
2.5e+10	4.e-15	1.e+8	6	7	10
		1.e+10	8	7	7
	4.e-18	1.e+8	8	9	12
		1.e+10	7	7	7
	4.e-21	1.e+8	8	8	12
		1.e+10	7	7	7
5.0e+10	4.e-15	1.e+8	5	6	9
		1.e+10	7	6	6
	4.e-18	1.e+8	6	7	10
		1.e+10	7	7	7
	4.e-21	1.e+8	6	7	10
		1.e+10	7	6	7

In order to analyse the results in Table 4.5, we propose first a row-wise reading then a column-wise reading. The row-wise reading provides information on the influ-

ence of the mesh size, the material parameters being fixed. Two cases are observed. First when r=1.e+10 there is good independence with respect to the mesh size. The outer number of FGMRES iterations remains constant enough even though the size of the system is multiplied by 50. Secondly when r=1.e+8, the number of iterations increases in the bigger mesh, especially for (E=1.e+9, r=1.e+8). This behaviour started to appear for $\boldsymbol{P}_{GFg}$ and is confirmed for $\boldsymbol{P}_{Final}$.
The column-wise reading provides information on the influence of the material parameters, the mesh size being fixed. Some variation of the outer number of FGMRES iterations is observed, remaining between 5 and 49. The preconditioner shows an excellent robustness when E=2.5e+10 and E=5.e+10. Overall $\boldsymbol{P}_{Final}$ is robust except for the worst cases when (E=1.e+9, r=1.e+8).

Conclusion

The complex geometry of Cigeo gallery crossings requires 3D coupled Thermo-Hydro-Mechanical with a second gradient of dilation regularization models. Consequently, long and difficult numerical simulations have to be solved. Hence, the need of an optimal solver in order to reduce the time of solution. The approach taken in this manuscript is to study independently the THM and the second gradient of dilation systems.

For the THM system, a Block Jacobi and a Block Gauss-Seidel preconditioner sharing the same tailored sub-solvers (Krylov methods preconditioned by AMG preconditioners) are investigated. We assessed their robustness and weak and strong scalability on a simple yet representative test case. Both preconditioners show excellent mesh size independence, good robustness with respect to parameters variation and good scalability. Due to the large difference in the order of magnitude from the parameters, a scaling algorithm that efficiently re-balances the Jacobian was tested along side the preconditioners. However it is not necessary in our case since the preconditioners naturally handle the unbalance inside the system, even though special attention in the order of the unknowns of the Block Gauss-Seidel is needed.

For the second gradient of dilation system, an in-depth analysis of the system was done to prove the *inf-sup* condition in the continuous and in the discrete case. The proves led us to investigate four different preconditioners. Two Block Jacobi preconditioners, the first one incorporates Krylov methods preconditioned by AMG preconditioners as sub-solvers and the second one incorporates AMG preconditioners as sub-solvers. Two Block Gauss-Seidel preconditioners, the first one incorporates Krylov methods preconditioned by AMG preconditioners as sub-solvers and the second one incorporates AMG preconditioners as sub-solvers. The mesh size independence of the different preconditioners is tested on a simple yet representative test case. Overall, the preconditioners are robust in regards to the mesh size, with the Block Gauss-Seidel with Krylov methods preconditioned by AMG preconditioners as sub-solvers being extremely robust. Consequently, we assess it's robustness in regard to parameters variation where it shows overall good results. Nevertheless, for the two outlier values of r, 0 and 1.e+14, the preconditioner did not converge.

Finally, for the THM with a second gradient of dilation system a Block Gauss-Seidel preconditioner with Krylov methods preconditioned by AMG preconditioners as sub-solvers is investigated. The robustness of the solver is tested in on a simple yet representative test. Overall, the final preconditioner shows good mesh size robustness and a good robustness in regard to parameter variation.

With this thesis we have contributed to the efficient solution of THM problems in the industrial context at EDF. The developed preconditioners are now available for use in `code_aster`. Before they can be fully applied to industrial calculations, the results of this thesis must be completed and deepened. The two outlier values of r cases, 0 and 1.e+14, still need further consideration. Additionally, these results have been established in the linear regime and ultimately need to be applied to non-linear constitutive equations. All of these points have being investigated and encouraging results have already been obtained.

Appendices

Appendix A

THM extra results

A.1 Illustrative numerical results

A.1.1 Industrial problem

In this section, we apply the proposed model and the associated preconditioner to the long-term evolution of the rock surrounding a deep geological repository for radioactive waste. The problem is based on a repository for high-level, long-lived radioactive waste in the Callovo-Oxfordian clay and is fully presented in[50]. The goal of the simulation is to model the excavation of the disposal and the placement of packaged wastes modeled by a representative thermal flow: indeed, due to the remaining radioactivity, heat is emitted and its influence on the surrounding media needs to be evaluated.

The geometry and the associated dimensions of the repository are shown in Figure A.2. It consists of a main access gallery from which multiple storage cells branch off, into which packaged waste are placed. The repository is located at a depth of -560m. Dimensions taken into account are realistic but doesn't correspond to a real and precise architecture.

Given the the symmetry of the site, the geometry of the model consists of a section of 18m of the main gallery that crosses a single storage cell, as shown in Figure A.3. The domain is confined by a layer of argillite and we apply symmetry conditions on the lateral and upper parts of the domain. The material parameters of the clay and of the concrete are given in Table 3.3. The measured initial pore pressure at this depth is $p_0 = 5.6$ MPa and it varies linearly with depth according to $p(z) = \rho_f \cdot g \cdot (z - 560) + 5.610^6$. Similarly, the initial stress state is $\sigma_{xx} = -12.4$ MPa, $\sigma_{yy} = -16.1$ MPa, $\sigma_{zz} = -12.4$ MPa and it varies linearly with z to the surface. The initial temperature varies linearly with respect to z from 25°C at $z = -560$ to 22.7°C at $z = -483.75$. The initial porosity is 0.18 throughout the argillite.

The methodology to model the excavation follows the classical Convergence-Confinement method [43] (also called CV-CF). It begins with an initial state with no galleries,

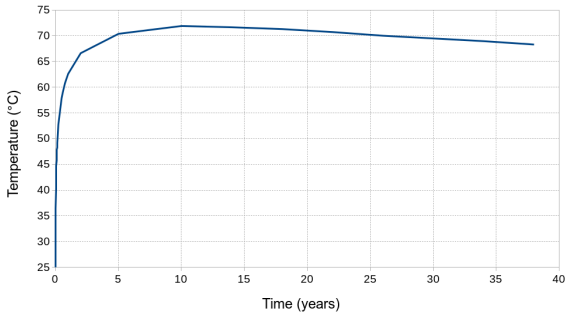


Figure A.1: Applied Temperature on the wall of the cell

where the domain is only submitted to gravity. In a first step, a quasi-instantaneous excavation of the gallery and cell is considered ; that is to say that the pore pressure and total radial stress at the galleries walls become zero in one second. In a second step, the concrete lining is introduced. In a third and final step, the radioactivity causes heat to be emitted around the cell in the course of time, and we therefore apply a representative temperature at the cell wall, according to the curve in Figure A.1. Fixed radial displacements are maintained at the cell wall. At the wall between the concrete lining and the gallery, the total radial stress is set to zero. The simulation is run until 40 years after the waste has been placed in the cell.

The evolution of pressure over time is shown in Figure A.4. A rapid and significant increase in hydraulic pressure of the hydraulic pressure is observed in the vicinity of the cell. This is due to the differential expansion caused by the thermal load since the thermal dilation of the water is bigger than that of the solid. After reaching a maximum, the pressure decreases steadily due to the diffusion of water and the decrease in temperature. The Figure A.5 shows that the main gallery undergoes a vertical collapse of about 1.3 cm as a result of its excavation, which is clearly present from the first year. Subsequently, due to the significant thermal expansion and the increase in fluid pressure, the upper part of the domain undergoes a strong upward shift. This leads to a complex and highly three-dimensional stress state as can be seen in Figure A.6. The zoom on the crossing of the gallery and the cell shows an intense traction zone where the signed Von Mises stress reaches 10 MPa. This is due to the fact that this section of the cell does not contain any waste. The complexity of the stress distribution fully justifies the fineness of the mesh.

A.1.2 Solver performance

The domain is discretized using 77 190 165 tetrahedral elements with quadratic shape function resulting in 341 292 114 DoF. The geometry and the mesh have been generated using the Salomé Platform 9.8 [83]. The domain has been splited in 2560 subdomains so that the simulation was run on 64 nodes of the Cronos cluster, each one running 40 MPI processes.

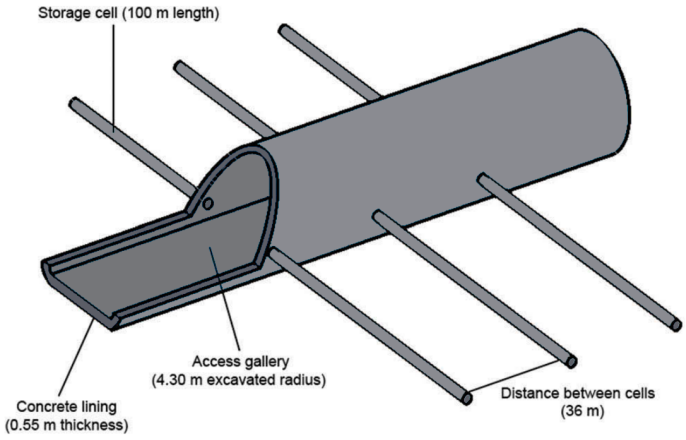


Figure A.2: Dimensions of the storage facility (courtesy from [50])

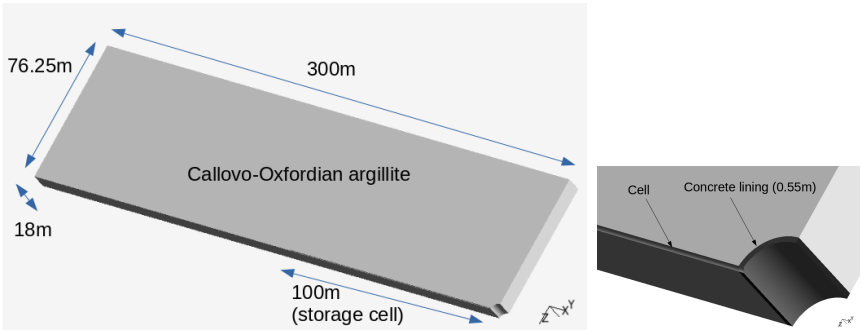


Figure A.3: Geometry of the site and zoom on the cell and concrete lining (courtesy from [50])

From a practical point of view, the 4 phases of the method consist in 4 non-linear simulations, each one being the initial state of the following. The last one deals with the effect of the waste on 40 years. The size of the time step follows the variation of the heat induced by the waste, shorter at the beginning when it varies strongly (say the order of one day) and longer when it stabilizes (say the order of one month). The non-linear convergence is easily reached in 2 or 3 iterations, exhibiting quadratic convergence. Each linear solve required less than 9 outer iterations and took roughly 35s so that the simulation was achieved in 2 hours.

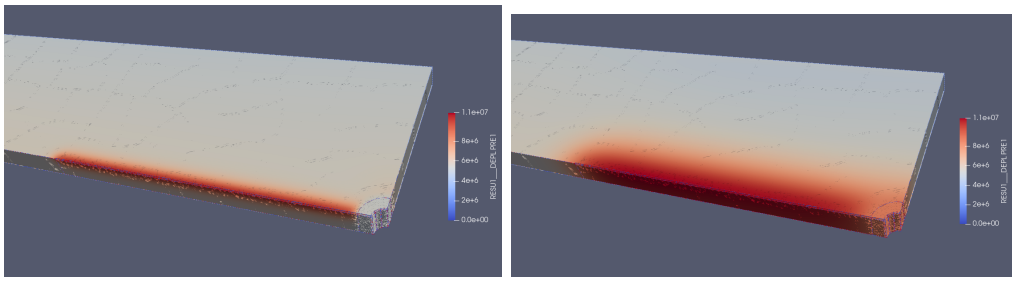


Figure A.4: Pressure distribution after 1 and 10 years

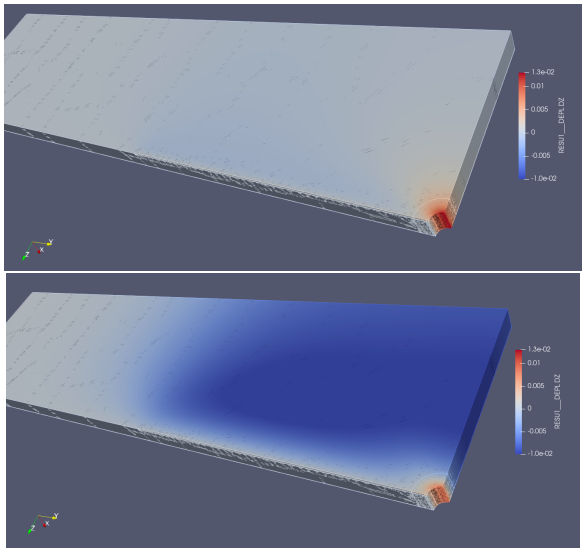


Figure A.5: Vertical displacement distribution after 1 and 10 years

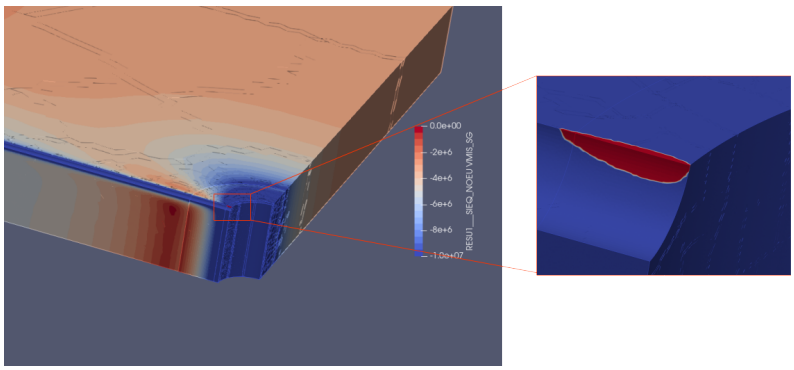


Figure A.6: Signed VonMises stress and zoom after 10 years

Appendix B

Resumé en français

B.1 Introduction

En 2006, la France a adopté une loi sur la gestion des déchets radioactifs qui prévoit une solution de stockage pour les déchets de haute activité et de moyenne activité à vie longue, c'est-à-dire les déchets les plus radioactifs. Cette solution est le projet Cigéo, un stockage géologique en profondeur situé en Meuse/Haute-Marne [3].

Le stockage des déchets se fera pendant plus de 100 ans et Cigéo sera agrandi en fonction des besoins en espace. Il sera ensuite fermé pour assurer le confinement des déchets sur de très longues périodes, sans intervention humaine. À l'heure actuelle, une installation de stockage est la seule solution pour les déchets radioactifs à très longue durée de vie. Toutefois, si une meilleure solution devait être trouvée à l'avenir, l'un des autres objectifs de Cigéo est de pouvoir y retourner pour récupérer les déchets radioactifs. Même si le projet est mené par l'Andra (Agence Nationale pour la gestion des Déchets RAdioactifs) qui est responsable de la mise en place de solutions de gestion sûres pour les déchets radioactifs français, la sécurité et le coût de la gestion des déchets restent de la responsabilité du producteur, c'est-à-dire en grande partie d'EDF. La sûreté du site repose en grande partie sur sa formation géologique. Environ 270 km de galeries et alvéoles vont être creusés à 500 m de profondeur et les travaux de cette thèse vont permettre d'évaluer l'impact de ces galeries et alvéoles. le travail de cette thèse va porter sur le dimensionnement des croisement de galeries.

Le problème de sécurité dans les plus de 300 galeries est l'infiltration d'eau apportant une contamination radiologique à la surface, une fois que le béton commence à être poreux et que des fissures se développent à proximité des galeries. L'excavation, le mouvement du sol et la dilatation des matériaux due à la chaleur des déchets radioactifs pourraient générer une telle dégradation. Ainsi, Cigéo est creusé dans une couche géologique argileuse stable à faible perméabilité et les déchets radioactifs sont traités avant leur élimination. Même si les risques sont fortement réduits, une modélisation précise est encore nécessaire pour le croisement de galeries. Compte tenu de la géométrie complexe, une modélisation thermo-hydro-mécanique avec une regularization par un second gradient de dilatation en 3D est nécessaire.

La mécanique traite la couche d'argile, du béton et des structures métalliques. Ils sont décrits comme des milieux poreux saturés d'un fluide monophasé faiblement compressible, ce qui est pris en charge par l'hydraulique. Enfin, la thermique est obligatoire pour modéliser la chaleur rayonnée par les déchets nucléaires. Puisque les géomateriaux ont un comportement mécanique complexe, des lois constitutives non linéaires sont nécessaires. Cependant, le couplage de ces lois constitutives avec les équations d'équilibre conduit à une perte d'unicité. La mécanique classique des milieux continus n'est pas adaptée à ce type de problème, en particulier dans le cas de la mécanique des matériaux. Notamment parce qu'elle ne caractérise pas les modèles

à un niveau niveau microscopique. Par conséquent, les fissures microscopiques que nous observons expérimentalement ne sont pas bien caractérisées, ce qui conduit à une dépendance pathologique de la simulation au maillage. Il est donc nécessaire d'introduire une longueur dans la modélisation des fissures afin de calculer des simulations physiquement correctes et indépendantes du maillage. Ceci est réalisé par un second gradient de dilatation régularisation. En raison de la taille du problème dans le temps (des milliers de pas de temps) et dans l'espace (des centaines de millions de degrés de liberté), le temps de résolution d'algorithmes inappropriés peut atteindre des mois. Cela n'est évidemment pas compatible avec les contraintes de temps des ingénieurs. Par conséquent, notre principale préoccupation est de réduire le temps de resolution en développant des solveurs optimaux

B.2 Thermo-Hydro-Mécanique

Système final

Le détail de chaque équation d'équilibre est dans la partie 3.1.1, on obtient le système final avec l'équation d'équilibre mécanique, l'équation de conservation de la masse et l'équation de conservation de l'énergie.

Soit Ω un domaine de dimension d, $1 \leq d \leq 3$, et t_f le temps final de la simulation. Le modèle THM décrit l'évolution de 3 inconnues primaires : le champ de déplacement vectoriel,$\underline{u}(x,t)$, le champ de pression du fluide, $p(x,t)$, le champ de température $T(x,t)$.

Le système couplé est composé de , $\forall x \in \Omega$ et $\forall t > 0 \in [0, t_f]$:

$$-\operatorname{div}(\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u})) + \nabla p + 3K_s\alpha_s\nabla T = \underline{f}^e \qquad \text{in } \Omega \times (0,t_f)$$

$$-\operatorname{div}(\rho_f\lambda_H\nabla p) + \rho_f(\operatorname{div}(\underline{\dot{u}}) + \frac{\varphi}{K_l}\dot{p} - \alpha_m 3\dot{T}) = 0 \qquad \text{in } \Omega \times (0,t_f)$$

$$\begin{aligned} -\operatorname{div}(\lambda_T\nabla T) - \operatorname{div}(\rho_f h_f \lambda_H \nabla p) \\ +\rho_f h_f(\operatorname{div}(\underline{\dot{u}}) + \frac{\varphi}{K_l}\dot{p} - \alpha_m 3\dot{T}) \\ +(3K_0\alpha_s\operatorname{div}(\underline{\dot{u}}) - 3\alpha_m\dot{p} - 9K_0\alpha_s^2\dot{T})T + C_\sigma^0\dot{T} = \Theta \qquad \text{in } \Omega \times (0,t_f) \end{aligned}$$

La frontière de Ω est notée $\partial\Omega$ et 6 partitions différentes sont nécessaires pour définir les conditions aux limites. Pour chaque inconnue primaire, nous pouvons définir des conditions aux limites de Dirichlet et de Neumann, par exemple le déplacement $\underline{u}$ et la contrainte $\underline{\underline{\sigma}}$, la pression p et le flux de fluide q, la température T et le flux

thermique Ψ.
Nous avons donc, respectivement, les conditions aux limites sur les inconnues de déplacement, sur les inconnues de pression et sur les inconnues de température telles que :

$$\begin{aligned}\partial\Omega = \partial\Omega^{\underline{u}} \cup \partial\Omega^{\underline{t}} \text{ with } \partial\Omega^{\underline{u}} \cap \partial\Omega^{\underline{t}} = \emptyset \\ \partial\Omega = \partial\Omega^{p} \cup \partial\Omega^{q} \text{ with } \partial\Omega^{p} \cap \partial\Omega^{q} = \emptyset \\ \partial\Omega = \partial\Omega^{T} \cup \partial\Omega^{\Psi} \text{ with } \partial\Omega^{T} \cap \partial\Omega^{\Psi} = \emptyset\end{aligned}$$

Les conditions aux limites sont données par :

$$\begin{aligned}\underline{\underline{\sigma}}(\underline{u}) \cdot \underline{n} &= \underline{t}^e && \text{on } \partial\Omega^{\underline{t}} \times (0, t_f) \\ -\lambda_H \nabla p \cdot \underline{n} &= q^e && \text{on } \partial\Omega^{q} \times (0, t_f) \\ -\lambda_T \nabla T \cdot \underline{n} &= \Psi^e && \text{on } \partial\Omega^{\Psi} \times (0, t_f) \\ \underline{u} &= \underline{u}^e && \text{on } \partial\Omega^{\underline{u}} \times (0, t_f) \\ p &= p^e && \text{on } \partial\Omega^{p} \times (0, t_f) \\ T &= T^e && \text{on } \partial\Omega^{T} \times (0, t_f) \\ \underline{u}(\,,0) &= \underline{u}_0 && \text{in } \Omega \\ p(\,,0) &= p_0 && \text{in } \Omega \\ T(\,,0) &= T_0 && \text{in } \Omega\end{aligned}$$

où $\underline{n}$ est la normale extérieure. En outre, les définitions des paramètres des matériaux sont données dans le tableau 3.1.

B.2.1 Préconditionnement

Comme nous l'avons vu au chapitre 2, les solveurs itératifs souffrent souvent d'un mauvais conditionnement du système linéaire et nécessitent un préconditionnement pour obtenir des performances satisfaisantes en termes de nombre d'itérations et de temps de simulation. Cela est particulièrement vrai pour les problèmes THM, qui sont en général mal conditionnés en raison des propriétés de chaque composant physique qui sont incluses via des paramètres dans le système linéaire.

Le système linéaire à résoudre a la structure suivante

$$\begin{bmatrix} \boldsymbol{J}_{\underline{u}\underline{u}} & \boldsymbol{J}_{\underline{u}p} & \boldsymbol{J}_{\underline{u}T} \\ \boldsymbol{J}_{p\underline{u}} & \boldsymbol{J}_{pp} & \boldsymbol{J}_{pT} \\ \boldsymbol{J}_{T\underline{u}} & \boldsymbol{J}_{Tp} & \boldsymbol{J}_{TT} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_{\underline{u}} \\ \boldsymbol{\delta}_{p} \\ \boldsymbol{\delta}_{T} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{R}_{\underline{u}} \\ \boldsymbol{R}_{p} \\ \boldsymbol{R}_{T} \end{bmatrix}. \tag{B.1}$$

Ce système est mal conditionné. De plus, il existe une différence significative dans l'ordre de grandeur de chaque paramètre (voir le tableau 3.3 qui se traduit par des

ordres de grandeur différents entre les 2-normes de chaque bloc basé sur la physique dans la matrice en (3.24)

$$\boldsymbol{S} := \begin{bmatrix} \left\|\boldsymbol{J}_{\underline{uu}}\right\|_2 & \left\|\boldsymbol{J}_{\underline{up}}\right\|_2 & \left\|\boldsymbol{J}_{uT}\right\|_2 \\ \left\|\boldsymbol{J}_{p\underline{u}}\right\|_2 & \left\|\boldsymbol{J}_{pp}\right\|_2 & \left\|\boldsymbol{J}_{pT}\right\|_2 \\ \left\|\boldsymbol{J}_{Tu}\right\|_2 & \left\|\boldsymbol{J}_{Tp}\right\|_2 & \left\|\boldsymbol{J}_{TT}\right\|_2 \end{bmatrix} \approx \begin{bmatrix} 1.\,\mathrm{e}{+13} & 1.\,\mathrm{e}{+01} & 1.\,\mathrm{e}{+06} \\ 1.\,\mathrm{e}{+04} & 1.\,\mathrm{e}{-08} & 1.\,\mathrm{e}{-02} \\ 1.\,\mathrm{e}{+08} & 1.\,\mathrm{e}{-03} & 1.\,\mathrm{e}{+05} \end{bmatrix}. \tag{B.2}$$

Nous avons une différence d'échelle maximale de 10^{21} entre le bloc de déplacement et le bloc de pression. Résoudre ce système naïvement pourrait conduire à des effets d'annulation dans la solution. Une mise à l'échelle préalable ou un préconditionnement de la matrice sont donc obligatoires.

Dans cette section, nous définissons un préconditionneur adapté à notre application et discutons de l'impact de la différence en ordres de grandeur entre les paramètres.

Définition des préconditionneurs par blocs

Notez que la matrice $\boldsymbol{J}$ est non symétrique. Nous avons donc besoin d'un solveur itératif pour les systèmes non-symétriques et choisissons la méthode *flexible GMRES* (FGMRES). Nous allons ensuite définir un préconditionneur qui peut être appliqué au système THM.

L'idée derrière le préconditionnement est de construire une matrice $\boldsymbol{P}$ qui est une assez bonne approximation de $\boldsymbol{J}$ mais qui est facilement inversible. Dans nos calculs, le préconditionnement est appliqué depuis la droite, ce qui signifie que nous résolvons le système

$$\boldsymbol{J}\boldsymbol{P}^{-1}y = r,$$

avec $y = \boldsymbol{P}x$. La solution x du système reste la même mais si $\boldsymbol{P}$ est une bonne approximation de $\boldsymbol{J}$, alors $\boldsymbol{J}\boldsymbol{P}^{-1}$ devient "plus proche" de l'identité et la méthode itérative convergera plus rapidement que pour le système non conditionné.

Le système linéaire de (3.24) a une structure en blocs, où chaque bloc diagonal correspond à l'un des trois modèles physiques. Il semble donc naturel de choisir un préconditionneur par blocs, comme cela a été par exemple décrit dans la référence [64]. Le préconditionneur le plus simple est probablement le préconditionneur de bloc de Jacobi donné par

$$\boldsymbol{P}_{Jac} = \begin{bmatrix} \boldsymbol{J}_{\underline{uu}} & 0 & 0 \\ 0 & \boldsymbol{J}_{pp} & 0 \\ 0 & 0 & \boldsymbol{J}_{TT} \end{bmatrix}.$$

L'application d'un préconditionneur du type Jacobi standard est simple, car il est trivial d'inverser une matrice diagonale. Pour un préconditionneur du type Jacobi par blocs, nous avons besoin des inverses des trois blocs physiques individuels, c'est-à-dire $\boldsymbol{J}_{\underline{uu}}^{-1}$, $\boldsymbol{J}_{pp}^{-1}$ et $\boldsymbol{J}_{TT}^{-1}$. Ceci est coûteux et nous cherchons donc une approximation de chaque bloc qui peut être plus facilement inversée. Avant de poursuivre la discussion, nous présentons notre deuxième et notre troisième choix de préconditionneur. Il s'agit des préconditionneurs de Gauss-Seidel à blocs inférieur et supérieur, désignés respectivement par $\boldsymbol{P}_{LGS}$ et $\boldsymbol{P}_{UGS}$, donnés par

$$\boldsymbol{P}_{LGS} = \begin{bmatrix} \boldsymbol{J}_{\underline{uu}} & 0 & 0 \\ \boldsymbol{J}_{p\underline{u}} & \boldsymbol{J}_{pp} & 0 \\ \boldsymbol{J}_{T\underline{u}} & \boldsymbol{J}_{Tp} & \boldsymbol{J}_{TT} \end{bmatrix}, \qquad \boldsymbol{P}_{UGS} = \begin{bmatrix} \boldsymbol{J}_{\underline{uu}} & \boldsymbol{J}_{\underline{u}p} & \boldsymbol{J}_{\underline{u}T} \\ 0 & \boldsymbol{J}_{pp} & \boldsymbol{J}_{pT} \\ 0 & 0 & \boldsymbol{J}_{TT} \end{bmatrix}$$

Bien que ces préconditionneurs utilisent les blocs triangulaires rectangulaires inférieurs (ou supérieurs) du système, lors de l'application de leur inverse, nous n'avons toujours besoin que de calculer les inverses des trois blocs diagonaux $\boldsymbol{J}_{\underline{uu}}^{-1}$, $\boldsymbol{J}_{pp}^{-1}$ et $\boldsymbol{J}_{TT}^{-1}$. Soit $\mathbf{I}$ la matrice identité de taille appropriée pour chaque bloc. Pour faciliter la notation, nous n'ajoutons pas la taille dans l'indice. L'inverse du préconditionneur de Gauss-Seidel inférieur est donné par

$$\boldsymbol{P}_{LGS}^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \boldsymbol{J}_{TT}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -\boldsymbol{J}_{T\underline{u}} & -\boldsymbol{J}_{Tp} & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & \boldsymbol{J}_{pp}^{-1} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ -\boldsymbol{J}_{p\underline{u}} & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \boldsymbol{J}_{\underline{uu}}^{-1} & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & I \end{bmatrix}$$

et l'inverse de $\boldsymbol{P}_{UGS}$ est donné par

$$\boldsymbol{P}_{UGS}^{-1} = \begin{bmatrix} \boldsymbol{J}_{\underline{uu}}^{-1} & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & -\boldsymbol{J}_{\underline{u}p} & -\boldsymbol{J}_{\underline{u}T} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & \boldsymbol{J}_{pp}^{-1} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & -\boldsymbol{J}_{pT} \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \boldsymbol{J}_{TT}^{-1} \end{bmatrix}$$

Comme mentionné ci-dessus, pour chacun de ces trois préconditionneurs, nous avons besoin, en théorie, des matrices inverses des blocs diagonaux. En pratique, ces inverses ne sont jamais calculées explicitement, car cela est trop coûteux.

Dans les expériences suivantes, nous utilisons une approche de préconditionnement imbriquée, où nous appliquons les préconditionneurs des blocs $\boldsymbol{P}_{Jac}, \boldsymbol{P}_{LGS}, \boldsymbol{P}_{UGS}$ en utilisant un certain nombre d'itérations de FGMRES préconditionnée par un cycle en V d'un solveur multigrille algébrique (AMG) pour chaque bloc $\boldsymbol{J}_{\underline{u}}, \boldsymbol{J}_{pp}, \boldsymbol{J}_{TT}$. Ceci offre la possibilité de contrôler la qualité de l'inverse pour chaque bloc en définissant une tolérance d'arrêt ou un nombre fixe d'itérations FGMRES. Dans cette stratégie, nous avons une interaction entre le nombre d'itérations externes de FGMRES sur le système de blocs et le nombre d'itérations internes de FGMRES appliquées à chaque

bloc diagonal. Une tolérance plus stricte pour les solveurs internes de FGMRES pourrait conduire à un plus petit nombre d'itérations externes de FGMRES et vice-versa. Ce choix est guidé par les expériences numériques décrites ci-après, où le compromis entre performance et robustesse était un objectif primordial.

Le choix d'un préconditionneur AMG pour chaque itération interne de FGMRES peut être expliqué comme suit. Les blocs diagonaux du système discrétisé (voir section 3.1.2) impliquent des opérateurs elliptiques et paraboliques non dégénérés, pour lesquels les préconditionneurs multigrilles à cycle en V sont particulièrement adaptés [60, 64]. Notez que dans notre cas particulier, nous pourrions tout aussi bien utiliser la méthode GMRES, puisqu'un cycle en V est un préconditionneur linéaire constant et ne nécessite donc pas de version flexible. Cela s'accompagnerait d'un léger gain de mémoire.

Le système de la THM présente une très mauvaise mise à l'échelle en raison de la présence de paramètres d'ordres de grandeur différents. Ce problème peut être résolu par l'utilisation d'un algorithme de mise à l'échelle dédié qui rééquilibre efficacement le jacobien. Ce n'est cependant pas obligatoire dans notre cas puisqu'il est montré que les préconditionneurs par blocs proposés peuvent gérer naturellement le déséquilibre des différents blocs (voir section 3.2.2). Dans le cas de la variante Block Gauss-Seidel, une attention particulière est nécessaire pour éliminer les inconnues dans un ordre bien choisi.

B.2.2 Performance du solveur

La robustesse et l'efficacité du solveur proposé sont cruciales pour les applications industrielles. Nous nous tournons donc vers la présentation des résultats d'un cas test illustratif, mettant à l'épreuve la robustesse du préconditionneur en faisant varier certains paramètres. L'efficacité parallèle est également évaluée par des tests de scalabilité faible et forte. La méthode est implémentée dans code_aster, le solveur éléments finis massivement parallèle et open source développé à EDF R&D [41].

Cas test

Le cas test doit être suffisamment simple pour que le maillage puisse être facilement raffiné mais suffisamment complexe pour ressembler au problème industriel considéré. À cette fin, un échantillon rectangulaire 3D est modélisé comme indiqué sur la figure 3.1, avec une longueur de 0,1 m suivant x, une hauteur de 0,1 m suivant y et une largeur de 0,05 m suivant z. Le maillage tétraédrique a été généré à l'aide de Gmsh 4.4.1.

Le déplacement a été fixé à 0 sur la surface inférieure ($y = 0$), une pression mécanique de 5 MPa a été appliquée sur la surface supérieure ($y = 0,1$) et une température de 80°C a été imposée sur toute la surface de l'échantillon.

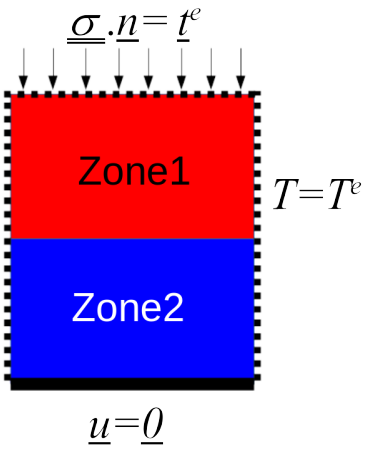


Figure B.1: Cas test

Selon l'évaluation considérée, l'échantillon est constitué d'un seul matériau ou de 2 matériaux différents. Pour les expériences de robustesse, il se compose uniquement d'argile, tandis que pour les expériences de mise à l'échelle, il se compose d'argile et de béton. Les 2 sous-domaines différents sont illustrés dans la figure 3.1. Nous soulignons que l'ordre de grandeur des paramètres des matériaux est d'une grande importance dans les applications industrielles. Les valeurs des paramètres des matériaux, affichées dans le tableau 3.3, sont représentatives d'un problème industriel typique de stockage des déchets géologiques [50].
Les essais ont été résolus avec code_aster en utilisant le cadre THM présenté dans la section ci-dessus, correspondant à un milieu THM isotrope saturé monophasique utilisant des éléments finis $P2$-$P1$-$P1$.

Robustesse

La robustesse des préconditionneurs est évaluée en faisant varier les valeurs du module de Young E, de la perméabilité intrinsèque k_{int} et de la conductivité thermique λ_T. Ces paramètres sont choisis car ils apparaissent respectivement dans chaque équation de bilan et y ont une influence majeure. Les tests sont réalisés à partir du cas d'essai de la figure 3.1 avec les deux zones 1 et 2 constituées d'argile, en utilisant les paramètres du matériau dans le tableau 3.3.
Les résultats sont compilés dans le tableau 3.4 pour P_J et le tableau 3.5 pour $\boldsymbol{P}_{LGS}$. Le nombre maximum d'itérations FGMRES externes pendant les itérations de Newton est affiché en premier, suivi du nombre total d'itérations de Newton entre parenthèses. La très grande plage de variation du maillage et de chaque paramètre (jusqu'à 6 ordres de grandeur) est soulignée.

Afin d'analyser les résultats des tableaux 3.4 et 3.5, nous proposons d'abord une lecture en ligne puis une lecture en colonne. La lecture en ligne nous montre l'influence

Table B.1: Paramètres du cas test

Symbol	Value	Unit
μ_l	10^{-3}	Pa.s
K_l	2.10^9	Pa
C_s	1000	J.kg^{-1}.K^{-1}
C_f	4180	J.kg^{-1}.K^{-1}
C_f^p	4180	J.kg^{-1}.K^{-1}
ρ_f	1000	kg.m^{-3}
λ_T	1.6	W.m^{-1}.K
T_0	273	K
p_{atm}	10^5	Pa
α_s	10^{-5}	K^{-1}
α_l	10^{-4}	K^{-1}
h_{f0}	$\frac{p_{atm}}{\rho_f}$	J.kg^{-1}
K_s	$\frac{E}{3(1-2\nu)}$	Pa
K_0	K_s	Pa
λ_H	K_{int}/μ_l	Pa^{-1}.m^2.s^{-1}
C_σ^0	$C_s\rho_s(1-\varphi)+C_l\rho_f\varphi$	J.K^{-1}.m^{-3}
ρ_s	$(\rho_m-\varphi\rho_f)/(1-\varphi)$	kg.kg^{-3}
α_m	$\varphi\alpha_l+(1-\varphi)\alpha_s$	K^{-1}

Clay		
Symbol	Value	Unit
E	6.10^9	Pa
ν	0.3	-
ρ_m	2410	kg.m^{-3}
K_{int}	4.10^{-21}	
φ	0.18	-
Concrete		
Symbol	Value	Unit
E	15.10^9	Pa
ν	0.2	-
ρ_m	2500	kg.m^{-3}
K_{int}	10^{-11}	
φ	0.2	-

du maillage, les paramètres du matériau étant fixes. Une excellente indépendance vis-à-vis de la taille de la maille est observée. Le nombre extérieur d'itérations FGMRES reste constant bien que la taille du système soit multipliée par 20, sauf pour P_J avec le jeu de paramètres (E=1.e+9, k_{int}=4.e-21, λ_T=2.3) où le nombre d'itérations passe de 70 à 44 mais semble se stabiliser en atteignant 40 dans la plus grande maille.
La lecture en colonne nous montre l'influence des paramètres matériels, le maillage étant fixe. On observe une variation modérée du nombre externe d'itérations FGMRES, qui reste le plus souvent inférieur à 12 sauf pour le "pire" jeu de paramètres (E=1.e+9, k_{int}=4.e-21), où il atteint jusqu'à 70 itérations pour P_J et 22 pour $\boldsymbol{P}_{LGS}$. Ce résultat particulier tend à montrer une meilleure robustesse de la variante Block Gauss-Seidel par rapport à la variante Jacobi, qui est analysée plus en détail dans la section suivante. Malgré cela, les deux préconditionneurs semblent être très robustes car ils atteignent la convergence à chaque exécution et l'augmentation des itérations de Krylov reste modérée par rapport aux grandes variations des paramètres.
Enfin, nous soulignons l'excellente robustesse en ce qui concerne les itérations de Newton, qui restent entre 2 et 4 pour chaque exécution.

Table B.2: Robustesse des paramètres du bloc Jacobi

P_J					
Parameters			ddl		
E	k_{int}	λ_T	10 000	60 000	200 000
1.e+9	4.e-15	4.e-01	6 (3)	6 (2)	6 (4)
		2.3e+00	5 (3)	5 (2)	4 (2)
	4.e-18	4.e-01	8 (3)	8 (3)	9 (3)
		2.3e+00	8 (3)	7 (3)	8 (3)
	4.e-21	4.e-01	48 (3)	46 (3)	50 (3)
		2.3e+00	70 (3)	44 (3)	40 (3)
2.5e+10	4.e-15	4.e-01	6 (2)	6 (2)	7 (3)
		2.3e+00	6 (2)	6 (2)	4 (2)
	4.e-18	4.e-01	7 (2)	6 (2)	8 (2)
		2.3e+00	6 (2)	6 (2)	7 (2)
	4.e-21	4.e-01	12 (2)	12 (2)	11 (2)
		2.3e+00	11 (2)	11 (2)	11 (2)
5.0e+10	4.e-15	4.e-01	6 (2)	6 (2)	7 (3)
		2.3e+00	6 (2)	6 (2)	5 (2)
	4.e-18	4.e-01	6 (2)	6 (2)	8 (2)
		2.3e+00	6 (2)	6 (2)	6 (2)
	4.e-21	4.e-01	10 (2)	11 (2)	10 (2)
		2.3e+00	10 (2)	9 (2)	9 (2)

Scalabilité parallèle

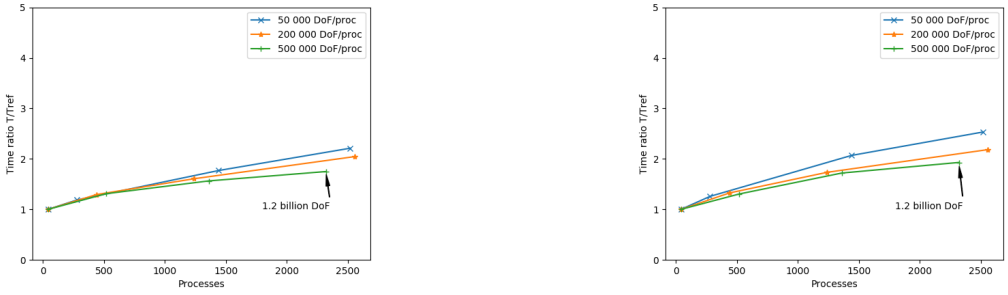
Une bonne scalabilité du préconditionneur proposé est essentielle pour maintenir le temps de résolution raisonnable lors du passage à des systèmes plus grands. Les tests de scalabilité faible et forte sont considérés en utilisant le cas bi-matériau de la figure 3.1 avec la zone 1 en argile et la zone 2 en béton. Des valeurs de paramètres réalistes ont été choisies dans le tableau 3.3. Les deux tests de scalabilité sont exécutés sur le cluster Cronos d'EDF. Il est composé de 1272 nœuds, équipés de 2 processeurs Xeon Platinum 8260 24C 2,4 GHz avec 24 cœurs chacun.

Un test de scalabilité faible consiste à fixer un nombre fixe de degrés de liberté (ddl) par processus et à augmenter la taille du problème en augmentant le nombre de processus. En d'autres termes, nous fixons la taille d'un sous-domaine et rendons le problème plus grand en augmentant le nombre total de sous-domaines. Notre objectif est d'étudier si l'algorithme de résolution nécessite le même temps de résolution que nous résolvions N ddl sur 1 processus ou 1000xN ddl sur 1000 processus. En cas de scalabilité faible parfaite, le temps devrait rester constant lorsque le nombre de processus augmente.

Comme on peut le voir sur la figure B.2, le nombre de ddl par processus est fixé à 50 000 (ligne bleue), 200 000 (ligne orange) et 500 000 (ligne verte) et le cas test est exécuté de 40 processus à 2500 processus. Le rapport entre le temps résolution et le

Table B.3: Robustesse des paramètres du bloc Gauss-Seidel

$\boldsymbol{P}_{LGS}$					
Parameters			ddl		
E	k_{int}	λ_T	10 000	60 000	200 000
1.e+9	4.e-15	4.e-01	4 (3)	5 (2)	5 (4)
		2.3e+00	4 (3)	4 (2)	3 (2)
	4.e-18	4.e-01	6 (3)	5 (3)	6 (3)
		2.3e+00	5 (3)	5 (3)	6 (3)
	4.e-21	4.e-01	21 (3)	22 (3)	20 (3)
		2.3e+00	18 (3)	20 (3)	18 (3)
2.5e+10	4.e-15	4.e-01	5 (2)	6 (2)	5 (3)
		2.3e+00	5 (2)	5 (2)	4 (2)
	4.e-18	4.e-01	5 (2)	6 (2)	6 (2)
		2.3e+00	5 (2)	5 (2)	6 (2)
	4.e-21	4.e-01	8 (2)	8 (2)	8 (2)
		2.3e+00	7 (2)	7 (2)	7 (2)
5.0e+10	4.e-15	4.e-01	5 (2)	5 (2)	5 (3)
		2.3e+00	5 (2)	5 (2)	4 (2)
	4.e-18	4.e-01	5 (2)	5 (2)	7 (2)
		2.3e+00	5 (2)	5 (2)	7 (2)
	4.e-21	4.e-01	7 (2)	7 (2)	7 (2)
		2.3e+00	6 (2)	7 (2)	6 (2)



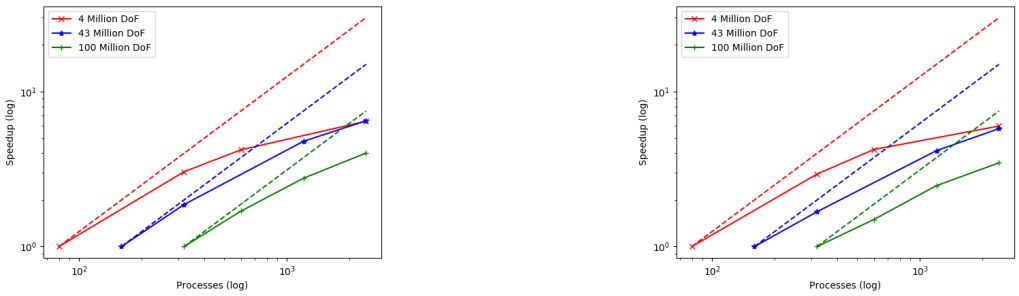
(a) $\boldsymbol{P}_J$ (b) $\boldsymbol{P}_{LGS}$

Figure B.2: Scalabilité faible

temps sur 40 processus est présenté. Pour un petit nombre de ddl par processus, il reste entre 1. et 2.5 pour $\boldsymbol{P}_{LGS}$ et entre 1. et 2.2 pour P_J. Alors que pour 500 000 ddl par processus, il reste entre 1. et 1,9 pour $\boldsymbol{P}_{LGS}$ et entre 1. et 1,7 pour P_J. Ce comportement sous-optimal pour les petits sous-domaines est souvent dû à la latence du réseau du cluster. Lorsque les sous-domaines sont grands et qu'il y a plus de travail par processus, le calcul domine le coût associé à la communication. Même si P_J évolue légèrement mieux, le temps de résolution est plus élevé qu'avec $\boldsymbol{P}_{LGS}$ en raison

du nombre plus élevé d'itérations qui varie entre 8 et 16 pour P_J et entre 7 et 11 pour $\boldsymbol{P}_{LGS}$. Nous soulignons qu'en utilisant $\boldsymbol{P}_{LGS}$ pour 500 000 ddl par processeur (ligne verte), la taille du système linéaire varie de 20 millions avec un temps de résolution de 465 secondes à plus de 1,2 milliard de ddl avec un temps de résolution de 891 secondes. La taille du problème est multipliée par 60 alors que le temps de résolution n'augmente que de 1,9. Il s'agit d'un très bon résultat de scalabilité puisque le cas test est plutôt complexe, notamment en raison de la variation des paramètres des matériaux entre l'argile et le béton.

Passons au test de scalabilité forte, qui consiste à fixer la taille du problème et à augmenter le nombre de processus. Le but est de résoudre le système plus rapidement en ajoutant des ressources. Par exemple, si nous résolvons un système d'une taille donnée à l'aide de N processus, en utilisant $N\ fois M$ processus, le temps de résolution devrait être réduit de M. En cas de forte scalabilité parfaite, le temps de résolution diminue proportionnellement à l'augmentation du nombre de processus.



(a) $\boldsymbol{P}_J$ (b) $\boldsymbol{P}_{LGS}$

Figure B.3: Scalabilité forte

Les tests de scalabilité forte sont présentés dans la figure B.3. La taille du problème est fixée à 4 millions ddl (ligne rouge), 43 millions ddl (ligne bleue) et 100 millions ddl (ligne verte). Le nombre de processus passe de 80 à 2 400. La ligne pointillée représente la scalabilité forte idéale. Le speed-up par rapport au nombre de processus est présentée. Dans tous les cas, lorsque l'on augmente le nombre de processus de 80 à 320 avec 4 millions de ddl, la scalabilité reste satisfaisante avec une efficacité de 76% pour P_J et de 73% pour $\boldsymbol{P}_{LGS}$. Ensuite, l'efficacité commence à décliner jusqu'à atteindre 2 400 processus et se situe à 20% pour P_J et pour $\boldsymbol{P}_{LGS}$. Pour le cas de 100 millions de ddl, de 320 à 600 processus, l'efficacité pour P_J est de 90% et pour $\boldsymbol{P}_{LGS}$ de 79%. En augmentant les processus à 2400, l'efficacité pour P_J est de 53% et pour $\boldsymbol{P}_{LGS}$ 46%. De même que pour la scalabilité faible, la différence d'efficacité entre chaque problème est souvent due à la latence du réseau du cluster, car à 4 millions de ddl, il y a moins de travail par processus. P_J évolue légèrement mieux que $\boldsymbol{P}_{LGS}$ mais $\boldsymbol{P}_{LGS}$ reste plus rapide pour tous les cas testés. Compte tenu de la complexité du cas

test et du fait que nous avons commencé à 80 processus afin de pouvoir résoudre des problèmes représentatifs en termes de taille, la forte évolutivité du préconditionneur proposé est satisfaisante.

B.3 Second gradient de dilatation

B.3.1 Application à la mécanique

Commençons maintenant par faire une analyse approfondie du second gradient de dilatation appliquée à la mécanique. D'après les travaux de Fernandes sur le second gradient de dilatation, le modèle suivant décrit l'évolution du champ de déplacement vectoriel $\underline{u}(x)$, du champ de deformation volumique microscopique scalaire $\chi(x)$ et de λ en tant que multiplicateurs de Lagrange, $\forall x \in \Omega$ [37]. Une régularisation par Lagrangien augmenté peut être utilisée à travers la constante de pénalisation $r > 0$. L'utilisation d'un Lagrangien augmenté est optionnelle : pour $r = 0$ la régularisation est seulement imposée par les multiplicateurs de Lagrange.
Les équations d'équilibre sont :

$$-\operatorname{div}(\underline{\underline{\sigma}}(\underline{u})) + \nabla\lambda - r\nabla(\operatorname{div}(\underline{u})) + r\nabla\chi = 0 \qquad \text{in } \Omega \qquad \text{(B.3)}$$

$$\lambda - \operatorname{div}(\underline{\underline{S}}(\chi)) - r\operatorname{div}(\underline{u}) + r\chi = 0 \qquad \text{in } \Omega \qquad \text{(B.4)}$$

$$\operatorname{div}(\underline{u}) - \chi = 0 \qquad \text{in } \Omega \qquad \text{(B.5)}$$

$$\underline{u} = 0 \qquad \text{on } \Gamma_1 \qquad \text{(B.6)}$$

$$\frac{\partial \underline{u}}{\partial \underline{n}} = 0 \qquad \text{on } \Gamma_2 \qquad \text{(B.7)}$$

$$\underline{\underline{S}}(\chi)\,\underline{n} = 0 \qquad \text{on } \partial\Omega \qquad \text{(B.8)}$$

Où $\underline{\underline{\sigma}}(\underline{u}) = 2\mu\underline{\underline{\varepsilon}}(\underline{u}) + \lambda_m \operatorname{div}(\underline{u})I$. $\underline{\underline{S}}(\chi) = 3a^1\nabla\chi$ où a^1 est un paramètre matériel donné et ≥ 0. (B.3) et (B.4) sont des équations d'équilibre de la contrainte mécanique $\underline{\underline{\sigma}}$ et des doubles contraintes de dilatation microscopiques $\underline{\underline{S}}$; (B.5) est une équation de contrainte qui force l'égalité des deformations volumique microscopique et macroscopiques. La limite de Ω est notée $\partial\Omega$ et 2 partitions différentes sont nécessaires pour définir les conditions aux limites. Les deux partitions différentes sont définies comme suit : $\partial\Omega = \Gamma_1 \cup \Gamma_2$ avec $\Gamma_1 \cap \Gamma_2 = \emptyset$.

Remark 5. *Pour des raisons de simplicité, nous avons choisi que les doubles forces microscopiques externes soient égales à 0, ce qui donne $\underline{\underline{S}}(\chi)\,\underline{n} = 0$ à la frontière. Etant donné que nos calculs visent à caractériser les bandes de localisation à un niveau macroscopique, il est prouvé que l'effet de cette condition aux limites est limité. [44].*

Formulation variationnelle
Nous définissons l'espace de Sobolev :

$$H^1_{\Gamma_1}(\Omega) = \{v \in H^1(\Omega) | v = 0 \text{ sur } \Gamma_1\}$$

Nous introduisons également la norme :

$$\|v\|_{H^1_{\Gamma_1}\times H^1} = (\|v\|^2_{H^1_{\Gamma_1}} + \|v\|^2_{H^1})^{\frac{1}{2}}$$

Soit $\underline{u}, \underline{v} \in (H^1_{\Gamma_1}(\Omega))^d$, $\chi, \psi \in H^1(\Omega)$ et $\lambda, \tau \in L^2(\Omega)$. En considérant les espaces de Sobolev appropriés définis ci-dessus et par intégration par parties, nous obtenons la formulation faible suivante:
Trouver $(\underline{u}, \chi, \lambda) \in (H^1_{\Gamma_1}(\Omega))^d \times H^1(\Omega) \times L^2(\Omega)$ tel que pour tout $(\underline{v}, \psi, \tau)$ $dans (H^1_{\Gamma_1}(\Omega))^d \times H^1(\Omega) \times L^2(\Omega)$, on a

$$\int_\Omega \Big(\underline{\underline{\sigma}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v}) - \lambda \ \mathrm{div}(\underline{v}) + r \ \mathrm{div}(\underline{u}) \ \mathrm{div}(\underline{v}) - r\,\chi \ \mathrm{div}(\underline{v}) \Big)\, dx = 0$$

$$\int_\Omega \Big(\underline{\underline{S}}(\chi)\, \nabla\psi + \lambda\, \psi + r\, \chi\, \psi - r\, \psi \ \mathrm{div}(\underline{u}) \Big)\, dx = 0$$

$$\int_\Omega \big(\tau \ \mathrm{div}(\underline{u}) - \tau\, \chi \big)\, dx = 0$$

où $r \geq 0$ est le terme de pénalisation.

Remark 6. *Une autre façon d'écrire la formulation variationnelle est*

$$\int_\Omega (\underline{\underline{\sigma}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v}) + \underline{\underline{S}}(\chi)\nabla\psi - \lambda(\ div(\underline{v}) - \psi) + \tau(\ div(\underline{u}) - \chi) + r(\ div(\underline{u}) - \chi)(\ div(\underline{v}) - \psi) dx = 0$$

C'est ainsi qu'elle est écrite dans la documentation de `code_aster`.

En suivant la théorie de Brezzi décrite dans la sous-section 2.2.1, une façon de s'assurer du caractère bien posé du problème est de prouver le théorème 2.2.1.
Pour cela, nous définissons les formes bilinéaires

$$\begin{aligned} a\big((\underline{u}, \chi), (\underline{v}, \psi)\big) = {} & \underline{\underline{\sigma}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v}) + r \ \mathrm{div}(\underline{u}) \ \mathrm{div}(\underline{v}) - r\,\chi \ \mathrm{div}(\underline{v}) \\ & + \underline{\underline{S}}(\chi)\, \nabla\psi + r\, \chi\, \psi - r\, \psi \ \mathrm{div}(\underline{u}) \end{aligned} \tag{B.9}$$

et

$$b\big((\underline{u}, \chi), \tau\big) = \tau \ \mathrm{div}(\underline{u}) - \tau\, \chi \tag{B.10}$$

Nous désignons par A et B, respectivement, les opérateurs linéaires continus associés à la forme bilinéaire $a(.,.)$ et $b(.,.)$ comme dans 2.2.1. Le noyau de B est noté $KerB$.

Le système peut alors être écrit comme suit

$$a\big((\underline{u}, \chi), (\underline{v}, \psi)\big) - b\big((\underline{v}, \psi), \lambda\big) = 0 \tag{B.11}$$

$$b\big((\underline{u}, \chi), \tau\big) = 0 \tag{B.12}$$

Nous voyons clairement qu'il s'agit d'un système point-selle antisymétrique. Par conséquent, le théorème 2.2.1 s'applique. En nous inspirant de ce théorème, nous définissons un nouveau théorème B.3.1 pour le problème B.11.

Theorem B.3.1. *Soient* $V := (H^1_{\Gamma_1})^d \times H^1$ *et* $Q := L^2$.
Il existe une constante $\alpha > 0$ *, telle que*

$$a\left((\underline{u},\chi),(\underline{u},\chi)\right) \geq \alpha \left\|(\underline{u},\chi)\right\|_V^2, \ (\underline{u},\chi) \in KerB \tag{B.13}$$

et il existe une constante $\beta > 0$*, telle que*

$$\inf_{\lambda\in Q}\sup_{(\underline{u},\chi)\in V} \frac{b((\underline{u},\chi),\lambda)}{\left\|(\underline{u},\chi)\right\|_V \|\lambda\|_Q} \geq \beta. \tag{B.14}$$

Nous pouvons conclure qu'il existe une solution unique $(\underline{u},\chi,\lambda) \in V \times Q$ *pour le système.*

La preuve du Theoreme B.3.1 se trouve en section 4.2.

Discrétisation

Pour la discrétisation spaciale, nous utilisons la méthode des éléments finis avec des éléments Taylor-Hood $P2$-$P1$-$P1$ en 3D et $P2$-$P1$-$P0$ en 2D. Cela se traduit par l'utilisation de fonctions quadratiques continues par morceaux pour approximer le déplacement et de fonctions linéaires continues par morceaux pour approximer la dilatation microscopique et les multiplicateurs de Lagrange en 3D et de fonctions constantes continues par morceaux pour les multiplicateurs de Lagrange en 2D. Dans [37] et [44], les éléments $P1$ et $P2$ ont également été testés pour les multiplicateurs de Lagrange en 2D. Néanmoins, $P0$ reste le meilleur par rapport au temps, avec qualité de solution égale. Pour vérifier la stabilité de la discrétisation, la condition *inf-sup* doit également être remplie pour le problème discret.

Soit $H^1_{\Gamma_{1h}}(\Omega)$ l'espace discret d'éléments finis $P2$ tel que $H^1_{\Gamma_{1h}}(\Omega) \subset H^1_{\Gamma_1}(\Omega)$ de dimension N_u avec $h > 0$ un paramètre qui se réfère à la taille de la maille. $H^1_h(\Omega)$ est l'espace discret d'éléments finis $P1$ tel que $H^1_h(\Omega) \subset H^1(\Omega)$ de dimension N_χ. $L^2_h(\Omega)$ est l'espace d'éléments finis discrets $P1$ tel que $L^2_h(\Omega) \subset L^2(\Omega)$ de dimension N_λ.

Soit $\{\phi_{\underline{v}_j}\}_{j=1}^{N_u}$ une base pour l'espace d'éléments finis $(H^1_{\Gamma_{1h}}(\Omega))^d$, pour tout $v_h \in (H^1_{\Gamma_{1h}}(\Omega))^d$ nous avons

$$\underline{v}_h = \sum_{j=1}^{N_u} v_j \, \phi_{\underline{v}_j}$$

Soit $\{\phi_{\psi_j}\}_{j=1}^{N_\chi}$ une base pour l'espace d'éléments finis $H^1_h(\Omega)$, pour tout $\psi_h \in H^1_h(\Omega)$ nous avons

$$\psi_h = \sum_{j=1}^{N_\chi} \psi_j \, \phi_{\psi_j}$$

Soit $\{\phi_{\tau_j}\}_{j=1}^{N_\lambda}$ une base pour l'espace d'éléments finis $L_h^2(\Omega)$, pour tout $\tau_h \in L_h^2(\Omega)$ nous avons

$$\tau_h = \sum_{j=1}^{N_\lambda} \tau_j \, \phi_{\tau_j}$$

Nous considérons la restriction des formes bilinéaires a et b sur les espaces discrets. De plus, nous considérons la restriction des opérateurs A et B que nous désignons par A_h et B_h.

Le problème discret consiste a trouver $(\underline{u}_h, \chi_h, \lambda_h) \in (H^1_{\Gamma_1})^d_h \times H^1_h \times L^2_h$ tel que pour tout $(\underline{v}_h, \psi_h, \tau_h) \in (H^1_{\Gamma_1})^d_h \times H^1_h \times L^2_h$, nous avons

$$a\left((\underline{u}_h, \chi_h), (\underline{v}_h, \psi_h)\right) - b\left((\underline{v}_h, \psi_h), \lambda_h\right) = 0 \tag{B.15}$$

$$b\left((\underline{u}_h, \chi_h), \tau_h\right) = 0 \tag{B.16}$$

Le système linéaire est alors

$$\begin{bmatrix} \boldsymbol{A}_{\underline{u}\underline{u}} & \boldsymbol{A}_{\underline{u}\chi} & \boldsymbol{A}_{u\lambda} \\ \boldsymbol{A}_{\chi\underline{u}} & \boldsymbol{A}_{\chi\chi} & \boldsymbol{A}_{\chi\lambda} \\ \boldsymbol{A}_{\lambda\underline{u}} & \boldsymbol{A}_{\lambda\chi} & \boldsymbol{A}_{\lambda\lambda} \end{bmatrix} \begin{bmatrix} \underline{\boldsymbol{u}}_h \\ \boldsymbol{\chi}_h \\ \boldsymbol{\lambda}_h \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_{\underline{u}} \\ \boldsymbol{b}_\chi \\ \boldsymbol{b}_\lambda \end{bmatrix} \tag{B.17}$$

Le caractère bien posé du problème discret ne découle pas automatiquement du cas continu. Pour cela, nous pouvons utiliser le même raisonnement que pour le cas continu mais avec les espaces discrets.

Theorem B.3.2. *Soit $V_h := (H^1_{\Gamma_{1h}})^d(\Omega) \times H^1_h(\Omega)$, $Q_h := L^2_h(\Omega)$, $a(.,.)$ est définis comme dans* (B.9) *et $b(.,.)$ est définis comme dans* (B.10).
Il existe une constante $\alpha > 0$, telle que

$$a\left((\underline{u}_h, \chi_h), (\underline{u}_h, \chi_h)\right) \geq \alpha \left\|(\underline{u}_h, \chi_h)\right\|_V^2, \ (\underline{u}_h, \chi_h) \in Ker B_h \tag{B.18}$$

et il existe une constante $\beta > 0$, telle que

$$\inf_{\lambda_h \in Q_h} \sup_{(\underline{u}_h, \chi_h) \in V_h} \frac{b((\underline{u}_h, \chi_h), \lambda_h)}{\left\|(\underline{u}_h, \chi_h)\right\|_V \|\lambda_h\|_Q} \geq \beta. \tag{B.19}$$

Nous pouvons conclure qu'il existe une solution unique $(\underline{u}_h, \chi_h, \lambda_h) \in V_h \times Q_h$ pour le système (B.11) *et que la discrétisation est stable.*

La preuve du Theoreme B.3.2 se trouve en section 4.2.

Preconditioneur

Afin de construire un préconditionneur pour le problème de mécanique avec une régularization pas second gradient de dilatation, nous suivons l'approche développée par Mardal et Winther, expliquée dans le premier chapitre[64]. La technique consiste à trouver un préconditionneur approprié pour le problème continu et à appliquer une discrétisation stable par éléments finis afin d'identifier un préconditionneur correct pour le problème discret. Comme nous suivons la théorie développée dans [64], il est plus facile d'introduire les mêmes notations, de la même manière que dans le premier chapitre.

Tout d'abord, nous proposons un préconditionneur pour le problème continu. Pour ce faire, nous commençons par introduire l'opérateur des coefficients correspondant à la mécanique avec une régularization par second gradient de dilatation. Puisque les preuves d'existence et unicité ont été faites avec $r = 0$, nous commençons par ce cas.

$$\mathcal{A}_0 = \begin{pmatrix} -\Delta & 0 & \operatorname{grad} \\ 0 & -\Delta & I \\ \operatorname{div} & -I & 0 \end{pmatrix}.$$

En conséquence des deux conditions de Brezzi, $\mathcal{A}_0$ est un isomorphisme qui mappe $H^1_{\Gamma_1} \times H^1 \times L^2$ sur $H^{-1} \times H^{-1} \times L^2$ où H^{-1} est le dual de H^1. Le choix canonique du préconditionneur est alors l'opérateur diagonal par bloc

$$\mathcal{B} = \begin{pmatrix} (-\Delta)^{-1} & 0 & 0 \\ 0 & (-\Delta)^{-1} & 0 \\ 0 & 0 & I \end{pmatrix}$$

qui mappe l'espace $H^{-1} \times H^{-1} \times L^2$ sur $H^1_{\Gamma_1} \times H^1 \times L^2$. Sauf que le second gradient de dilatation dépend du paramètre de pénalisation r. Par conséquent, une approche dépendant du paramètre pourrait être plus adaptée. Pour ce faire, nous introduisons l'opérateur des coefficients correspondant au problème où $r \in [0, 1.e + 14]$.

$$\mathcal{A}_r = \begin{pmatrix} -\Delta - r \operatorname{grad}\operatorname{div} & r \operatorname{grad} & \operatorname{grad} \\ r \operatorname{div} & -\Delta - rI & I \\ \operatorname{div} & -I & 0 \end{pmatrix}.$$

Nous pouvons maintenant proposer un préconditionneur dépendant de r pour le problème continu $\mathcal{A}_r$ où chaque opérateur peut être remplacé par des opérateurs spectralement équivalents.

$$\mathcal{B}_r = \begin{pmatrix} (-\Delta - r \operatorname{grad}\operatorname{div})^{-1} & 0 & 0 \\ 0 & (-\Delta - r\, I)^{-1} & 0 \\ 0 & 0 & I^{-1} \end{pmatrix}$$

$\mathcal{B}_r$ est un isomorphisme qui mappe $H^{-1} \times H^{-1} \times L^2$ sur $H^1_{\Gamma_1} \times H^1 \times L^2$. Ceci est dû au fait que les opérateurs $\operatorname{grad}\operatorname{div} : H(\operatorname{div}) \to H(\operatorname{div})^*$ et $I : L^2 \to L^2$ mappent des espaces plus faibles que l'opérateur $\Delta : H^1 \to H^{-1}$ tel que $H^1 \subset H(\operatorname{div}) \subset L^2$. De plus, puisque l'opérateur Δ ne dépend pas de r, il reste présent dans $-\Delta - r \operatorname{grad}\operatorname{div}$ et $-\Delta - r\, I$ pour $r \in [0, 1.\text{e+}14]$. Il en résulte un mapping isomorphe de $\mathcal{B}_r$ de $H^{-1} \times H^{-1} \times L^2$ sur $H^1_{\Gamma_1} \times H^1 \times L^2$ pour $r \in [0, 1.\text{e+}14]$.

Nous considérons maintenant une discrétisation $P2$-$P1$-$P1$, avec les espaces $V_h \times Q_h$ définis ci-dessus. La discrétisation est stable dans la norme de $H^1_{\Gamma_1} \times H^1 \times L^2$. Le préconditionneur discret proposé est alors

$$\mathcal{B}_{r,h} = \begin{pmatrix} (-\Delta - r \operatorname{grad}\operatorname{div})_h^{-1} & 0 & 0 \\ 0 & (-\Delta - r\, I)_h^{-1} & 0 \\ 0 & 0 & I_h^{-1} \end{pmatrix}$$

Nous pouvons maintenant présenter les préconditionneurs que nous utilisons dans nos calculs, basés sur $\mathcal{B}_{r,h}$. À l'intérieur des préconditionneurs, nous remplaçons chaque opérateur discret $(-\Delta - r \operatorname{grad}\operatorname{div})_h^{-1}$, $(-\Delta - r\, I)_h^{-1}$ et I_h^{-1} par des préconditionneurs correspondants efficaces. Nous commençons par choisir des opérateurs spectralement équivalents pour $(-\Delta - r \operatorname{grad}\operatorname{div})_h$, $(-\Delta - r\, I)_h$ et I_h pour ensuite choisir une méthode d'inversion appropriée pour chaque opérateur.

Premièrement, $(-\Delta - r \operatorname{grad}\operatorname{div})_h$ et $(-\Delta - r\, I)_h$ sont représentés dans le préconditionneur par $\boldsymbol{A}_{\underline{uu}}$ et $\boldsymbol{A}_{\chi\chi}$ respectivement. $\boldsymbol{A}_{\underline{uu}}$ et $\boldsymbol{A}_{\chi\chi}$ sont essentiellement équivalents aux opérateurs $(-\Delta - r \operatorname{grad}\operatorname{div})_h$ et $(-\Delta - r\, I)_h$ avec l'ajout de paramètres physiques. Ici, I_h est l'opérateur de Riez qui mappe L^2 sur L^2 et est représenté par la matrice de masse discrète et non par la matrice identité. Nous avons constaté empiriquement que la mise à l'échelle de la matrice masse discrète par $-\frac{1}{r}$ ajoute de la robustesse au préconditionneur. Enfin, puisque les inverses ne sont jamais calculées explicitement, nous avons choisi des méthodes pour les approximer.

L'une des méthodes pour inverser les blocs, passe par un cycle en V d'un solveur multigrille algébrique (AMG) comme préconditionneur. Dans le chapitre 3, nous avons vu que cette méthode est particulièrement adaptée aux opérateurs elliptiques et paraboliques non dégénérés, ce qui est le cas ici avec les faibles valeurs de r. Pour $\boldsymbol{A}_{\underline{uu}}^{-1}$ et $\boldsymbol{A}_{\chi\chi}^{-1}$ nous avons également testé l'ajout d'un accélérateur, le résultat est l'approximation de l'inverse calculée à travers une méthode FGMRES avec un préconditionneur AMG à un cycle en V. Pour chaque approximation nous avons choisi une certaine tolérance pour FGMRES, cela permet une meilleure qualité de l'approximation de l'inverse de chaque bloc et nous pouvons avoir une diminution du nombre d'itérations externes. En ce qui concerne la méthode itérative externe, étant donné que le system du second gradient de dilatation est antisymétrique et que le préconditionneur peut varier à chaque itération lorsque nous ajoutons FGMRES

interne, nous avons besoin d'une version flexible du solveur itératif, c'est pour cela nous avons choisi FGMRES.

Quatre préconditionneurs sont testés, deux diagonaux $\boldsymbol{P}_{JH}$ et $\boldsymbol{P}_{JFg}$, où J représente un préconditionneur de type Block Jacobi,

$$\boldsymbol{P}_{JH}^{-1} = \begin{bmatrix} \tilde{\boldsymbol{A}}_{\underline{u}\underline{u}}^{-1} & 0 & 0 \\ 0 & \tilde{\boldsymbol{A}}_{\chi\chi}^{-1} & 0 \\ 0 & 0 & \tilde{\boldsymbol{M}}_{\lambda}^{-1} \end{bmatrix}, \qquad \boldsymbol{P}_{JFg}^{-1} = \begin{bmatrix} \hat{\boldsymbol{A}}_{\underline{u}\underline{u}}^{-1} & 0 & 0 \\ 0 & \hat{\boldsymbol{A}}_{\chi\chi}^{-1} & 0 \\ 0 & 0 & \tilde{\boldsymbol{M}}_{\lambda}^{-1} \end{bmatrix}.$$

où le symbole $\tilde{}$ représente l'approximation de l'inverse par un cycle en V AMG et le symbole $\hat{}$ par une méthode FGMRES avec un préconditionneur AMG à un cycle en V où le critère d'arrêt est la tolérance relative fixée à 10^{-3}. Nous testons également deux triangles inférieurs $\boldsymbol{P}_{GH}$ et $\boldsymbol{P}_{GFg}$, où G représente un préconditionneur de type Block Gauss-Seidel,

$$\boldsymbol{P}_{GH}^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \tilde{\boldsymbol{M}}_{\lambda}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -\boldsymbol{A}_{\lambda\underline{u}} & -\boldsymbol{A}_{\lambda\chi} & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & \tilde{\boldsymbol{A}}_{\chi\chi}^{-1} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ \boldsymbol{A}_{\chi\underline{u}} & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{A}}^{-1} & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & I \end{bmatrix}$$

$$\boldsymbol{P}_{GFg}^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \tilde{\boldsymbol{M}}_{\lambda}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -\boldsymbol{A}_{\lambda\underline{u}} & -\boldsymbol{A}_{\lambda\chi} & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & \hat{\boldsymbol{A}}_{\chi\chi}^{-1} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ \boldsymbol{A}_{\chi\underline{u}} & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{A}}_{\underline{u}\underline{u}}^{-1} & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & 0 & I \end{bmatrix}$$

Résultats numériques

La robustesse et l'efficacité du solveur proposé sont cruciales pour les applications industrielles. Nous nous tournons donc vers la présentation des résultats d'un cas test illustratif, mettant à l'épreuve la robustesse du préconditionneur en faisant varier certains paramètres. La méthode est implémentée dans Firedrake [77]

Cas test

Le cas de test doit être suffisamment simple pour que le maillage puisse être facilement raffiné, mais suffisamment complexe pour ressembler au problème industriel considéré.

À cette fin, un échantillon rectangulaire 3D est modélisé comme indiqué sur la figure B.4, avec une longueur de 0,5 m dans la direction x, une hauteur de 0,5 m dans la direction y et une largeur de 1 m dans la direction z. Le maillage tétraédrique a été généré à l'aide de la fonction BoxMesh de Firedrake. Le déplacement a été fixé à 0 sur la surface inférieure ($y = 0$), une pression mécanique de 5 MPa a été appliquée sur la surface supérieure ($y = 0, 1$).

Figure B.4: Test case

L'échantillon est constitué uniquement d'argile. Nous soulignons que la valeur des paramètres du matériau est d'une grande importance dans les applications industrielles. Elles sont affichées dans le tableau B.4, et sont représentatives d'un problème industriel typique de stockage des déchets radioactifs [44].

Les essais ont été résolus avec Firedrake en utilisant le cadre de la mécanique avec une régularisation par second gradient de dilatation présenté dans la section précédente, où des éléments finis $P2$-$P1$-$P1$ sont utilisés.

Table B.4: Paramètres du cas de test

Clay		
Symbol	Value	Unit
E	6.10^9	Pa
ν	0.3	-
a_1	500	-
r	10^{10}	-

Robustesse
La robustesse de $\boldsymbol{P}_{GFg}$ est évaluée en faisant varier les valeurs du module de Young E et du paramètre de pénalisation r. Ces paramètres sont choisis car ils apparaissent dans les deux premières équations d'équilibre et peuvent avoir de grandes variations. Les tests sont effectués à l'aide du cas test de la figure B.4 constitué d'argile avec les paramètres du tableau B.4.

Les résultats sont compilés dans le tableau B.5. Le nombre d'itérations externes de FGMRES est affiché. La très grande plage de discrétisation du maillage et de chaque paramètre (jusqu'à 14 ordres de grandeur) est soulignée. Nous avons choisi cette grande variation de r afin de tester toutes les valeurs évaluées par Fernandes dans [36] où $r \in [1.e+8, 1.e+14]$ et par Gantier dans [44] où $r \in 0, 1.e+10, 4.e+11$. Lorsque $r > 0$, la valeur de r est choisie par rapport à E afin d'évaluer correctement l'influence du terme de pénalité ce qui explique les grandes valeurs.

Table B.5: Robustesse des paramètres de Gauss-Seidel du bloc inférieur

Parameters		ddl		
E	r	16 197	117 637	895 749
1.e+9	0	-	-	-
	1.e+8	6	8	13
	1.e+10	25	25	25
	4.e+11	96	106	108
	1.e+14	-	-	-
2.5e+10	0	-	-	-
	1.e+8	5	6	10
	1.e+10	7	7	7
	4.e+11	30	30	30
	1.e+14	-	-	-
5.0e+10	0	-	-	-
	1.e+8	4	5	9
	1.e+10	6	6	6
	4.e+11	22	22	22
	1.e+14	-	-	-

Afin d'analyser les résultats du Tableau B.5, nous proposons d'abord une lecture en ligne puis une lecture en colonne.

La lecture en ligne renseigne sur l'influence du maillage, les paramètres matériau étant fixes. Une bonne indépendance par rapport au maillage est observée. Le nombre extérieur d'itérations FGMRES reste assez constant bien que la taille du système soit multipliée par 500, avec une légère exception pour r=1.e+8 où le nombre d'itérations a tendance à augmenter pour les plus grand maillages. La lecture en colonne renseigne sur l'influence des paramètres matériau, le maillage étant fixe. Nous commençons par les cas où le préconditionneur a convergé. On observe une variation du nombre extérieur d'itérations FGMRES, qui reste compris entre 4 et 30, sauf pour le "pire" jeu de paramètres (E=1.e+9, r=4.e+11), où il atteint jusqu'à 108 itérations. Le préconditionneur tend à montrer une meilleure robustesse lorsque r est plus proche en ordre de grandeur de E. Ceci est clairement mis en évidence par le fait que pour les

deux valeurs extrêmes de r, 0 et 1.e+14, le préconditionneur n'a pas convergé. Ceci est dû à deux raisons différentes. Premièrement, puisque l'opérateur pour le bloc de déplacement est $-\nabla - r\,\mathrm{grad}\,\mathrm{div}$ lorsque r=1.e+14, grad div devient plus grand. Il en résulte que l'opérateur n'est plus elliptique et que le préconditionneur AMG n'est plus adapté [64]. Deuxièmement, dans la preuve continue, lorsque r=0 la forme bilinéaire $a(.,.)$ n'est plus coercitive sur l'espace entier mais seulement sur le noyau de B. La perte de coercivité de la forme bilinéaire $a(.,.)$ est due à la perte de coercivité de la déformation volumique microscopique χ, où la partie correspondant à la matrice masse disparaît lorsque r=0.

B.3.2 Application à la thermo-hydro-mécanique linéaire.

Nous testons maintenant les préconditionneurs sur l'ensemble du problème, c'est-à-dire la Thermo-Hydro-Mécanique avec une régularisation par second gradient de dilatation. L'enthalpie spécifique du fluide, h_f, a été négligée afin de simplifier l'équation de conservation de l'énergie. De plus, la température non dépendante du temps T à l'intérieur de la chaleur non convective, Q', est remplacée par la température de référence T_0. En conséquence de ces fortes simplifications, le problème devient linéaire. Néanmoins, la stratégie de préconditionnement n'est pas compromise dans le cas général.

Soit Ω un domaine de dimension d, $1 \leq d \leq 3$, et t_f le temps final de la simulation. Le système est , $\forall x \in \Omega$ et $\forall t > 0 \in [0, t_f]$,

$$
\begin{aligned}
-\operatorname{div}(\underline{\underline{\underline{\underline{A}}}} : \underline{\underline{\varepsilon}}(\underline{u})) + \nabla p + 3K_s\alpha_s\nabla T + \nabla\lambda - r\nabla(\operatorname{div}(\underline{u})) + r\nabla\chi &= \underline{f}^e \quad \text{in } \Omega\times(0,t_f)\\
-\operatorname{div}(\rho_f\lambda_H\nabla p) + \rho_f(\operatorname{div}(\dot{\underline{u}}) + \frac{\varphi}{K_l}\dot{p} - \alpha_m 3\dot{T}) &= 0 \quad \text{in } \Omega\times(0,t_f)\\
-\operatorname{div}(\lambda_T\nabla T)&\\
+(3K_0\alpha_s\operatorname{div}(\dot{\underline{u}}) - 3\alpha_m\dot{p} - 9K_0\alpha_s^2\dot{T})T_0 + C_\sigma^0\dot{T} &= \Theta \quad \text{in } \Omega\times(0,t_f)\\
\lambda - \operatorname{div}(\underline{\underline{S}}(\chi)) - r\operatorname{div}(\underline{u}) + r\chi &= 0 \quad \text{in } \Omega\times(0,t_f)\\
\operatorname{div}(\underline{u}) - \chi &= 0 \quad \text{in } \Omega\times(0,t_f)
\end{aligned}
$$

La frontière de Ω est notée $\partial\Omega$ et six partitions différentes sont nécessaires pour définir les conditions aux limites. Nous appliquons des conditions aux limites de Dirichlet pour le déplacement, la pression et la température. Les conditions aux limites de Neumann pour le déplacement $\underline{u}$ suivent la contrainte $\underline{\underline{\sigma}}$, celles pour la pression $\boldsymbol{P}$ suivent le flux du fluide q et celles pour la température T suivent le flux thermique Ψ.

Nous avons donc, respectivement, les conditions aux limites sur les inconnues de déplacement, sur les inconnues de pression et sur les inconnues de température telles que :

$$\partial\Omega = \partial\Omega^{\underline{u}} \cup \partial\Omega^{\underline{t}} \text{ with } \partial\Omega^{\underline{u}} \cap \partial\Omega^{\underline{t}} = \emptyset$$
$$\partial\Omega = \partial\Omega^{p} \cup \partial\Omega^{q} \text{ with } \partial\Omega^{p} \cap \partial\Omega^{q} = \emptyset$$
$$\partial\Omega = \partial\Omega^{T} \cup \partial\Omega^{\Psi} \text{ with } \partial\Omega^{T} \cap \partial\Omega^{\Psi} = \emptyset$$

Les conditions aux limites sont données par :

$$\begin{aligned}
\underline{\underline{\sigma}}(\underline{u}) \cdot \underline{n} &= \underline{t}^e && \text{on } \partial\Omega^{\underline{t}} \times (0, t_f) \\
\underline{\underline{S}}(\chi) \cdot \underline{n} &= 0 && \text{on } \partial\Omega \times (0, t_f) \\
-\lambda_H \nabla p \cdot \underline{n} &= q^e && \text{on } \partial\Omega^{q} \times (0, t_f) \\
-\lambda_T \nabla T \cdot \underline{n} &= \Psi^e && \text{on } \partial\Omega^{\Psi} \times (0, t_f) \\
\underline{u} &= \underline{u}^e && \text{on } \partial\Omega^{\underline{u}} \times (0, t_f) \\
p &= p^e && \text{on } \partial\Omega^{p} \times (0, t_f) \\
T &= T^e && \text{on } \partial\Omega^{T} \times (0, t_f) \\
\underline{u}(\,, 0) &= \underline{u}_0 && \text{in } \Omega \\
p(\,, 0) &= p_0 && \text{in } \Omega \\
T(\,, 0) &= T_0 && \text{in } \Omega
\end{aligned}$$

où $\underline{n}$ est la normale à la frontière.

Pour la discrétisation du problème, on suit les mêmes étapes que pour la THM. Un Euler implicite est appliqué pour la discrétisation temporelle et des éléments finis $P2$-$P1$-$P1$-$P1$-$P1$ sont considérés pour la discrétisation spatiale. Le système linéaire à résoudre a la structure suivante :

$$\begin{bmatrix}
\boldsymbol{A}_{\underline{u}\underline{u}} & \boldsymbol{A}_{\underline{u}\chi} & \boldsymbol{A}_{\underline{u}\lambda} & \boldsymbol{A}_{\underline{u}p} & \boldsymbol{A}_{\underline{u}T} \\
\boldsymbol{A}_{\chi\underline{u}} & \boldsymbol{A}_{\chi\chi} & \boldsymbol{A}_{\chi\lambda} & 0 & 0 \\
\boldsymbol{A}_{\lambda\underline{u}} & \boldsymbol{A}_{\lambda\chi} & \boldsymbol{A}_{\lambda\lambda} & 0 & 0 \\
\boldsymbol{A}_{p\underline{u}} & 0 & 0 & \boldsymbol{A}_{pp} & \boldsymbol{A}_{pT} \\
\boldsymbol{A}_{T\underline{u}} & 0 & 0 & \boldsymbol{A}_{Tp} & \boldsymbol{A}_{TT}
\end{bmatrix}
\begin{bmatrix}
\underline{\boldsymbol{u}} \\ \boldsymbol{\chi} \\ \boldsymbol{\lambda} \\ \boldsymbol{p} \\ \boldsymbol{T}
\end{bmatrix}^{n+1}
=
\begin{bmatrix}
\boldsymbol{b}_{\underline{u}} \\ \boldsymbol{b}_{\chi} \\ \boldsymbol{b}_{\lambda} \\ \boldsymbol{b}_{p} \\ \boldsymbol{b}_{T}
\end{bmatrix} \tag{B.20}$$

Préconditionneur

La raison pour laquelle nous avons décidé de séparer la THM du second gradient de dilatation est qu'il n'y a pas de couplage entre (p, T) et (χ, λ). Cela nous a permis de traiter les difficultés de chaque sous-problème individuellement afin de proposer un préconditionneur approprié. Puisqu'il n'y a pas de couplage entre les deux sous-problèmes, le préconditionneur final est une fusion entre les deux sous-conditionneurs

et doit conserver les qualités positives de chacun. Le meilleur préconditionneur pour le problème THM est $\boldsymbol{P}_{LGS}$ et pour le second gradient de dilatation $\boldsymbol{P}_{Gfg}$. Le préconditionneur final est

$$\boldsymbol{P}_{Final} = \begin{bmatrix} \boldsymbol{A}_{\underline{u}\underline{u}} & 0 & 0 & 0 & 0 \\ \boldsymbol{A}_{\chi\underline{u}} & \boldsymbol{A}_{\chi\chi} & 0 & 0 & 0 \\ \boldsymbol{A}_{\lambda\underline{u}} & \boldsymbol{A}_{\lambda\chi} & \boldsymbol{M}_{\lambda} & 0 & 0 \\ \boldsymbol{A}_{p\underline{u}} & 0 & 0 & \boldsymbol{A}_{pp} & 0 \\ \boldsymbol{A}_{Tu} & 0 & 0 & \boldsymbol{A}_{Tp} & \boldsymbol{A}_{TT} \end{bmatrix}.$$

$\boldsymbol{P}_{Final}$ est un préconditionneur de type Bloc Gauss-Seidel inférieur. Pour l'appliquer, il suffit d'approximer les inverses des blocs diagonaux. On utilise la même approximation que dans $\boldsymbol{P}_{LGS}$ et $\boldsymbol{P}_{Gfg}$. Cela se traduit par les inverses de $\boldsymbol{A}_{\underline{u}\underline{u}}^{-1}$, $\boldsymbol{A}_{\chi\chi}^{-1}$, $\boldsymbol{A}_{pp}^{-1}$, et $\boldsymbol{A}_{TT}^{-1}$ approximé par une méthode FGMRES avec un seul cycle en V d'AMG. Le critère d'arrêt de chaque FGMRES interne est de 10 itérations pour $\boldsymbol{A}_{\underline{u}\underline{u}}^{-1}$, trois itérations pour $\boldsymbol{A}_{pp}^{-1}$, trois itérations pour $\boldsymbol{A}_{TT}^{-1}$ et une tolérance relative de 10^{-3} pour $\boldsymbol{A}_{\chi}^{-1}$. L'inverse de $\boldsymbol{M}_{\lambda}^{-1}$ est approximé par un cycle en V d'AMG.

Résultats numériques

Cas test

Nous utilisons le même test que dans la section ci-dessus, Figure B.4, où nous ajoutons des conditions limites de Dirichlet à la température. Il s'agit d'un échantillon

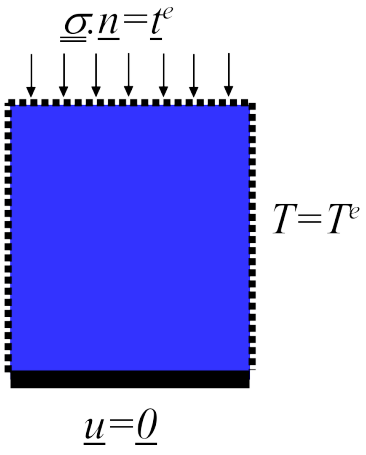


Figure B.5: Cas test

rectangulaire 3D modélisé comme indiqué sur la figure 4.6, avec une longueur de 0,5 m dans la direction x, une hauteur de 0,5 m dans la direction y et une largeur de

1 m dans la direction z. Le maillage tétraédrique a été généré à l'aide de la fonction BoxMesh de firedrake. Le déplacement a été fixé à 0 sur la surface inférieure ($y = 0$), une pression mécanique de 5 MPa a été appliquée sur la surface supérieure ($y = 0.1$) et une température de 80°C est imposée sur toute la surface de l'échantillon.

L'échantillon est constitué uniquement d'argile. Nous soulignons que la valeur des paramètres matériau est d'une grande importance dans les applications industrielles. Ils sont affichés dans le tableau B.4 et dans le tableau B.1, et sont représentatifs d'un problème industriel typique de stockage des déchets radioactifs [44].

Les essais ont été résolus avec firedrake en utilisant le modèle de Thermo-Hydro-Mécanique avec une régularisation par second gradient de dilatation présenté ci-dessus, où les éléments finis $P2$-$P1$-$P1$-$P1$-$P1$ sont utilisés.

Robustesse

Les performances de chaque sous-préconditionneur ayant déjà été testées, nous testons directement, pour le problème final, la robustesse de $\boldsymbol{P}_{Final}$ en faisant varier les valeurs du module de Young E, de la perméabilité intrinsèque K_{int} et du paramètre de pénalisation r. Comme le préconditionner du second gradient de dilatation n'est plus adapté lorsque $r = 0$ et lorsque $r = 1.e + 14$, nous testons le préconditionneur final avec $r \in [1.e + 8, 1.e + 10]$. Les résultats sont compilés dans le Tableau B.6. Le nombre maximal d'itérations externes de FGMRES est affiché.

Afin d'analyser les résultats du Tableau B.6, nous proposons d'abord une lecture en ligne puis une lecture en colonne. La lecture en ligne nous renseigne sur l'influence du maillage, les paramètres matériau étant fixes. Deux cas sont observés. Tout d'abord lorsque r=1.e+10 il y a une bonne indépendance par rapport au maillage. Le nombre d'itérations externes d'FGMRES reste assez constant même si la taille du système est multipliée par 50. Ensuite, lorsque r=1.e+8, le nombre d'itérations augmente dans les maillages plus grands, en particulier pour (E=1.e+9, r=1.e+8). Ce comportement commence à apparaître pour $\boldsymbol{P}_{GFg}$ et se confirme pour $\boldsymbol{P}_{Final}$. La lecture en colonne renseigne sur l'influence des paramètres matériau, le maillage étant fixe. On observe une certaine variation du nombre d'itérations externes d'FGMRES, qui reste compris entre 5 et 49. Le préconditionneur montre une excellente robustesse lorsque E=2.5e+10 et E=5.e+10. Globalement, $\boldsymbol{P}_{Final}$ est robuste, sauf dans les pires cas lorsque (E=1.e+9, r=1.e+8).

B.4 Conclusion

La géométrie complexe des croisements de galeries Cigeo nécessite de modèles de thermo-hydro-mécanique en 3D avec une régularisation par second gradient de di-

Table B.6: Robustesse des paramètres du bloc inférieur de Gauss-Seidel

Parameters			ddl		
E	K_{int}	r	17 655	127 463	967 623
1.e+9	4.e-15	1.e+8	16	18	24
		1.e+10	22	23	23
	4.e-18	1.e+8	21	29	49
		1.e+10	10	14	17
	4.e-21	1.e+8	19	24	37
		1.e+10	9	9	10
2.5e+10	4.e-15	1.e+8	6	7	10
		1.e+10	8	7	7
	4.e-18	1.e+8	8	9	12
		1.e+10	7	7	7
	4.e-21	1.e+8	8	8	12
		1.e+10	7	7	7
5.0e+10	4.e-15	1.e+8	5	6	9
		1.e+10	7	6	6
	4.e-18	1.e+8	6	7	10
		1.e+10	7	7	7
	4.e-21	1.e+8	6	7	10
		1.e+10	7	6	7

latation. Par conséquent, des simulations numériques longues et difficiles doivent être résolues. D'où la nécessité d'un solveur optimal afin de réduire le temps de résolution. L'approche adoptée dans ce manuscrit consiste à étudier indépendamment le système THM et le système du second gradient de dilatation.

Pour le système THM, un préconditionneur Block Jacobi et un préconditionneur Block Gauss-Seidel partageant les mêmes sous-solveurs adaptés (méthodes de Krylov préconditionnées par des préconditionneurs AMG) sont étudiés. Nous avons évalué leur robustesse et leur scalabilité faible et forte sur un cas test simple mais représentatif. Les deux préconditionneurs montrent une excellente indépendance vis-à-vis du maillage, une bonne robustesse vis-à-vis de la variation des paramètres et une bonne scalabilité. En raison de la grande différence dans l'ordre de grandeur des paramètres, un algorithme de mise à l'échelle qui rééquilibre efficacement le Jacobien a été testé en même temps que les préconditionneurs. Cependant, cela n'est pas nécessaire dans notre cas puisque les préconditionneurs gèrent naturellement le déséquilibre à l'intérieur du système, même si une attention particulière dans l'ordre des inconnues du Block Gauss-Seidel est nécessaire.

Pour le système du second gradient de dilatation, une analyse approfondie du système a été effectuée pour prouver la condition *inf-sup* dans le cas continu et dans le cas discret. Les preuves nous ont conduit à étudier quatre préconditionneurs différents. Deux préconditionneurs Block Jacobi, le premier incorpore des méthodes de Krylov préconditionnées par des préconditionneurs AMG comme sous-solveurs et le second incorpore des préconditionneurs AMG comme sous-solveurs. Deux préconditionneurs Block Gauss-Seidel, le premier incorpore des méthodes de Krylov préconditionnées par des préconditionneurs AMG comme sous-solveurs et le second incorpore des préconditionneurs AMG comme sous-solveurs. L'indépendance du maillages des différents préconditionneurs est testée sur un cas test simple mais représentatif. Dans l'ensemble, les préconditionneurs sont robustes par rapport au maillage, la méthode Block Gauss-Seidel avec des méthodes de Krylov préconditionnée par des préconditionneurs AMG comme sous-solveurs étant extrêmement robuste. Par conséquent, nous évaluons sa robustesse en ce qui concerne la variation des paramètres où il montre de bons résultats dans l'ensemble. Néanmoins, pour les deux valeurs extrêmes de r, 0 et 1.e+14, le préconditionneur n'a pas convergé.

Enfin, pour la THM avec une régularisation par second gradient de dilatation, un préconditionneur Bloc Gauss-Seidel avec des méthodes de Krylov préconditionnées par des préconditionneurs AMG comme sous-solveurs est étudié. La robustesse du solveur est testée sur un cas test simple mais représentatif. Dans l'ensemble, le préconditionneur final montre une bonne robustesse au niveau du maillage et une bonne robustesse par rapport à la variation des paramètres.

Avec cette thèse, nous avons contribué à la résolution efficace des problèmes THM dans le contexte industriel d'EDF. Les préconditionneurs développés sont maintenant disponibles pour une utilisation dans `code_aster`. Avant de pouvoir être pleinement appliqués à des calculs industriels, les résultats de cette thèse doivent être complétés et approfondis. Les deux cas de valeurs extrêmes de r, 0 et 1.e+14, doivent encore être examinés. De plus, ces résultats ont été établis dans le régime linéaire et doivent finalement être appliqués aux équations constitutives non linéaires. Tous ces points ont été étudiés et des résultats encourageants ont déjà été obtenus.

Bibliography

[1] James Adler et al. "Robust Block Preconditioners for Biot's Model". In: (May 2017).

[2] P. R. Amestoy et al. "Hybrid scheduling for the parallel solution of linear systems". In: *Parallel Computing* 32.2 (2006), pp. 136–156.

[3] *Andra.fr*. 2020. URL: www.andra.fr.

[4] F. Armero and J. C. Simo. "A new unconditionally stable fractional step method for non-linear coupled thermomechanical problems". In: *International Journal for Numerical Methods in Engineering* 35.4 (1992), pp. 737–766. DOI: https://doi.org/10.1002/nme.1620350408. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nme.1620350408. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.1620350408.

[5] Douglas Arnold, Richard Falk, and Ragnar Winther. "Multigrid In H(div) and H(curl)". In: *Numerische Mathematik* 85 (Jan. 2000). DOI: 10.1007/s002110000137.

[6] Michele Benzi. "A Sparse Approximate Inverse Preconditioner For The Conjugate Gradient Method". In: *SIAM J Sci Comput* 17 (July 1996). DOI: 10.1137/S1064827594271421.

[7] Luca Bergamaschi, Massimiliano Ferronato, and G. Gambolati. "Mixed Constraint Preconditioners for the iterative solution of FE coupled consolidation equations". In: *Journal of Computational Physics* (Dec. 2008), pp. 9885–9897. DOI: 10.1016/j.jcp.2008.08.002.

[8] Luca Bergamaschi and Ángeles Martínez. "RMCP: Relaxed Mixed Constraint Preconditioners for saddle point linear systems arising in geomechanics". In: *Computer Methods in Applied Mechanics and Engineering* 221-222 (2012), pp. 54–62. ISSN: 0045-7825. DOI: https://doi.org/10.1016/j.cma.2012.02.004. URL: http://www.sciencedirect.com/science/article/pii/S0045782512000461.

[9] Silvia Bertoluzza et al. "Boundary conditions involving pressure for the Stokes problem and applications in computational hemodynamics". In: *Computer Methods in Applied Mechanics and Engineering* (May 2017). DOI: 10.1016/j.cma.2017.04.024. URL: https://hal.archives-ouvertes.fr/hal-01420651.

[10] M. A. Biot and D. G. Willis. "The Elastic Coefficients of the Theory of Consolidation". In: *Journal of Applied Mechanics* 24.4 (June 1957), pp. 594–601. ISSN: 0021-8936. DOI: 10.1115/1.4011606. eprint: https://asmedigitalcollection.asme.org/appliedmechanics/article-pdf/24/4/594/6750516/594_1.pdf. URL: https://doi.org/10.1115/1.4011606.

[11] Maurice A. Biot. "General Theory of Three-Dimensional Consolidation". In: *Journal of Applied Physics* 12.2 (1941), pp. 155–164. DOI: 10.1063/1.1712886.

[12] Daniele Boffi, Franco Brezzi, and Michel Fortin. *Mixed Finite Element Methods and Applications*. en. Google-Books-ID: mRhAAAAAQBAJ. Springer Science & Business Media, July 2013. ISBN: 9783642365195.

[13] Dietrich Braess. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, 2007.

[14] A. Brandt, Steve McCormick, and John Ruge. "Algebraic multigrid (AMG) for sparse matrix equations". In: (Jan. 1984).

[15] William Briggs, Van Henson, and Steve McCormick. *A Multigrid Tutorial, 2nd Edition*. Jan. 2000. ISBN: 978-0-89871-462-3.

[16] Hans-Joachim Bungartz et al. "preCICE – A fully parallel library for multi-physics surface coupling". In: *Computers and Fluids* 141 (2016). Advances in Fluid-Structure Interaction, pp. 250–258. ISSN: 0045-7930. DOI: https://doi.org/10.1016/j.compfluid.2016.04.003. URL: http://www.sciencedirect.com/science/article/pii/S0045793016300974.

[17] Xiao-Chuan Cai. "Multiplicative Schwarz Methods for Parabolic Problems". In: *SIAM Journal on Scientific Computing* 15.3 (1994), pp. 587–603. DOI: 10.1137/0915039. eprint: https://doi.org/10.1137/0915039. URL: https://doi.org/10.1137/0915039.

[18] *Parallel Scalable Unstructured CPR-Type Linear Solver for Reservoir Simulation*. Vol. All Days. SPE Annual Technical Conference and Exhibition. SPE-96809-MS. Oct. 2005. DOI: 10.2118/96809-MS. eprint: https://onepetro.org/SPEATCE/proceedings-pdf/05ATCE/All-05ATCE/SPE-96809-MS/1839702/spe-96809-ms.pdf. URL: https://doi.org/10.2118/96809-MS.

[19] Nicola Castelletto, Joshua A. White, and Massimiliano Ferronato. "Scalable algorithms for three-field mixed finite element coupled poromechanics". en. In: *Journal of Computational Physics* 327 (Dec. 2016), pp. 894–918. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2016.09.063. URL: http://www.sciencedirect.com/science/article/pii/S0021999116304843 (visited on 05/28/2020).

[20] R. Chambon, D. Caillerie, and N. El Hassan. "One-dimensional localisation studied with a second grade model". In: *European Journal of Mechanics - A/Solids* 17.4 (1998), pp. 637–656. ISSN: 0997-7538. DOI: https://doi.org/10.1016/S0997-7538(99)80026-6. URL: https://www.sciencedirect.com/science/article/pii/S0997753899800266.

[21] René Chambon, Denis Caillerie, and Takashi Matsushima. "Plastic continuum with microstructure, local second gradient theories for geomaterials: Localization studies". In: *International Journal of Solids and Structures* 38 (Nov. 2001), pp. 8503–8527. DOI: 10.1016/S0020-7683(01)00057-9.

[22] René Chambon, Denis Caillerie, and Claudio Tamagnini. "A strain gradient plasticity theory for finite strain". In: *Computer Methods in Applied Mechanics and Engineering - COMPUT METHOD APPL MECH ENG* 193 (July 2004), pp. 2797–2826. DOI: 10.1016/j.cma.2003.10.016.

[23] Shuangshuang Chen et al. *Robust block preconditioners for poroelasticity*. Jan. 2020.

[24] E. et F. Cosserat Cosserat. "Théorie des corps déformables". In: *A. Hermann et fils* (1909).

[25] O. Coussy. "Revisiting the constitutive equations of unsaturated porous solids using a Lagrangian saturation concept". In: *International Journal for Numerical and Analytical Methods in Geomechanics* 31.15 (2007), pp. 1675–1694. DOI: https://doi.org/10.1002/nag.613. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nag.613. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nag.613.

[26] O. Coussy. "Thermodynamics". In: *Poromechanics*. John Wiley & Sons, Ltd, 2003. Chap. 3, pp. 37–70. ISBN: 9780470092712. DOI: https://doi.org/10.1002/0470092718.ch3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470092718.ch3. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470092718.ch3.

[27] O. Coussy. "Thermoporoelasticity". In: *Poromechanics*. John Wiley & Sons, Ltd, 2003. Chap. 4, pp. 71–112. ISBN: 9780470092712. DOI: https://doi.org/10.1002/0470092718.ch4. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470092718.ch4. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470092718.ch4.

[28] Olivier Coussy. *Mechanics and Physics of Porous Solids*. June 2010. DOI: 10.1002/9780470710388.ch5.

[29] Matteo Cusini et al. "Constrained pressure residual multiscale (CPR-MS) method for fully implicit simulation of multiphase flow in porous media". In: *Journal of Computational Physics* 299 (2015), pp. 472–486. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2015.07.019. URL: https://www.sciencedirect.com/science/article/pii/S0021999115004647.

[30] G. Dhondt. *The Finite Element Method for Three-dimensional Thermomechanical Applications*. Chichester, England: Wiley, 2004.

[31] Howard Elman, David Silvester, and Andrew Wathen. "Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics". In: (Jan. 2006).

[32] Howard C. Elman. "Multigrid and Krylov Subspace Methods for the Discrete Stokes Equations". en. In: *International Journal for Numerical Methods in Fluids* 22.8 (1996), pp. 755–770. ISSN: 1097-0363. DOI: 10.1002/(SICI)1097-0363(19960430)22:8<755::AID-FLD377>3.0.CO;2-1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0363%2819960430%2922%3A8%3C755%3A%3AAID-FLD377%3E3.0.CO%3B2-1 (visited on 05/25/2020).

[33] Alexandre Ern and Sébastien Meunier. "A posteriori error analysis of Euler-Galerkin approximations to coupled elliptic-parabolic problems". en. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 43.2 (Mar. 2009), pp. 353–375. ISSN: 0764-583X, 1290-3841. DOI: 10.1051/m2an:2008048. URL: https://www.esaim-m2an.org/articles/m2an/abs/2009/02/m2an0756/m2an0756.html (visited on 06/11/2020).

[34] Robert Falgout, Jim Jones, and Ulrike Yang. "The Design and Implementation of hypre, a Library of Parallel High Performance Preconditioners". In: vol. 51. Jan. 2006, pp. 267–294. DOI: 10.1007/3-540-31619-1_8.

[35] R. Fernandes, Clément Chavant, and René Chambon. "A simplified second gradient model for dilatant materials: Theory and numerical implementation". In: *International Journal of Solids and Structures - INT J SOLIDS STRUCT* 45 (Oct. 2008), pp. 5289–5307. DOI: 10.1016/j.ijsolstr.2008.05.032.

[36] R. Fernandes, Clément Chavant, and René Chambon. "A simplified second gradient model for dilatant materials: Theory and numerical implementation". In: *International Journal of Solids and Structures - INT J SOLIDS STRUCT* 45 (Oct. 2008), pp. 5289–5307. DOI: 10.1016/j.ijsolstr.2008.05.032.

[37] Roméo FERNANDES. "Modélisation numérique objective des problèmes couplés hydromécaniques dans le cas des géomatériaux". PhD thesis. Université Joseph Fourier - Grenoble I, 2009.

[38] Massimiliano Ferronato, Nicola Castelletto, and Giuseppe Gambolati. "A fully coupled 3-D mixed finite element model of Biot consolidation". In: *Journal of Computational Physics* 229.12 (2010), pp. 4813–4830. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2010.03.018. URL: https://www.sciencedirect.com/science/article/pii/S0021999110001282.

[39] Samuel Forest and Eric Lorentz. "Localization phenomena and regularization methods". In: *Local approach to fracture*. Ed. by J. Besson. Ecole d'été "Mécanique de l'endommagement et approche locale de la rupture" (MEALOR), juillet 2004. Les presses de l'ecole des mines de paris, 2004, pp. 311–371. URL: https://hal.archives-ouvertes.fr/hal-00164479.

[40] Michel Fortin, R Glowinski, and Trans-inter-scientia (Firm). *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*. Translation of: Méthodes de Lagrangien augmenté. OCLC: 9556803. Amsterdam; New York; New York, N.Y.: North-Holland Pub. Co. ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co., 1983. ISBN: 9780444866806.

[41] Electricité de France. *Finite element code_aster, Analysis of Structures and Thermomechanics for Studies and Research, Year = 1989–2022*. Open source on www.code-aster.org.

[42] Roland W. Freund and Noël M. Nachtigal. "Software for simplified Lanczos and QMR algorithms". In: *Applied Numerical Mathematics* 19.3 (Dec. 1995), pp. 319–341. DOI: 10.1016/0168-9274(95)00089-5.

[43] Manuel de la Fuente et al. "Applicability of the Convergence-Confinement Method to Full-Face Excavation of Circular Tunnels with Stiff Support System". In: *Rock Mechanics and Rock Engineering* 52 (2019), pp. 2361–2376.

[44] Maxime Gantier. "Modélisation numérique robuste et fiable de la fissuration des roches et des interfaces". PhD thesis. Nov. 2021.

[45] F. J. Gaspar et al. "A systematic comparison of coupled and distributive smoothing in multigrid for the poroelasticity system". In: *Numerical Linear Algebra with Applications* 11.2-3 (2004), pp. 93–113. DOI: 10.1002/nla.372. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nla.372. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nla.372.

[46] Francisco J. Gaspar and Carmen Rodrigo. "On the fixed-stress split scheme as smoother in multigrid methods for coupling flow and geomechanics". In: *Computer Methods in Applied Mechanics and Engineering* 326 (2017), pp. 526–540. ISSN: 0045-7825. DOI: https://doi.org/10.1016/j.cma.2017.08.025. URL: http://www.sciencedirect.com/science/article/pii/S0045782517306047.

[47] Gabriel N. Gatica, Antonio Márquez, and Salim Meddahi. "Analysis of the Coupling of Primal and Dual-Mixed Finite Element Methods for a Two-Dimensional Fluid-Solid Interaction Problem". In: *SIAM Journal on Numerical Analysis* 45.5 (2007), pp. 2072–2097. DOI: 10.1137/060660370. eprint: https://doi.org/10.1137/060660370. URL: https://doi.org/10.1137/060660370.

[48] Michael Gee et al. "ML 5.0 Smoothed Aggregation User's Guide". In: (Jan. 2006).

[49] P. Germain. "The Method of Virtual Power in Continuum Mechanics. Part 2: Microstructure". In: *SIAM Journal on Applied Mathematics* 25.3 (1973), pp. 556–575. ISSN: 00361399. URL: http://www.jstor.org/stable/2100123 (visited on 07/05/2022).

[50] Richard Giot et al. "A transversely isotropic thermo-poroelastic model for claystone: parameter identification and application to a 3D underground structure". In: *Geomechanics and Geoengineering* 13.4 (Mar. 2018), pp. 246–263. DOI: 10.1080/17486025.2018.1445874. URL: https://hal.archives-ouvertes.fr/hal-02017965.

[51] Anne Greenbaum. *Iterative Methods for Solving Linear Systems*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Jan. 1997. ISBN: 9780898713961. DOI: 10.1137/1.9781611970937. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611970937 (visited on 06/18/2020).

[52] Sebastian Gries et al. "Preconditioning for Efficiently Applying Algebraic Multigrid in Fully Implicit Reservoir Simulations". In: *SPE Journal* 19.04 (Jan. 2014), pp. 726–736. ISSN: 1086-055X. DOI: 10.2118/163608-PA. eprint: https://onepetro.org/SJ/article-pdf/19/04/726/2099424/spe-163608-pa.pdf. URL: https://doi.org/10.2118/163608-PA.

[53] Joachim Berdal Haga, Harald Osnes, and Hans Petter Langtangen. "A parallel block preconditioner for large-scale poroelasticity with highly heterogeneous material parameters". In: *Computational Geosciences* 16.3 (Mar. 2012), pp. 723–734. DOI: 10.1007/s10596-012-9284-4.

[54] Joachim Berdal Haga, Harald Osnes, and Hans Petter Langtangen. "Efficient block preconditioners for the coupled equations of pressure and deformation in highly discontinuous media". In: *International Journal for Numerical and Analytical Methods in Geomechanics* 35.13 (2011), pp. 1466–1482. DOI: 10.1002/nag.973.

[55] Gwendal Jouan, Panagiotis Kotronis, and Frédéric Collin. "Using a second gradient model to simulate the behaviour of concrete structural elements". In: *Finite Elements in Analysis and Design* 90 (2014), pp. 50–60. ISSN: 0168-874X. DOI: https://doi.org/10.1016/j.finel.2014.06.002. URL: http://www.sciencedirect.com/science/article/pii/S0168874X14001085.

[56] Carsten Keller, Nicholas I. M. Gould, and Andrew J. Wathen. "Constraint Preconditioning for Indefinite Linear Systems". In: *SIAM J. Matrix Analysis Applications* (2000). DOI: 10.1137/S0895479899351805.

[57] Axel Klawonn. "Block-Triangular Preconditioners for Saddle Point Problems with a Penalty Term". In: *SIAM Journal on Scientific Computing* 19.1 (Jan. 1998), pp. 172–184. ISSN: 1064-8275. DOI: 10.1137/S1064827596303624. URL: https://epubs.siam.org/doi/abs/10.1137/S1064827596303624 (visited on 05/25/2020).

[58] Philip A. Knight, Daniel Ruiz, and Bora Uçar. "A Symmetry Preserving Algorithm for Matrix Scaling". In: *SIAM Journal on Matrix Analysis and Applications* 35.3 (2014), pp. 931–955. DOI: 10.1137/110825753. URL: https://doi.org/10.1137/110825753.

[59] Carola Kruse et al. "Application of an iterative Golub-Kahan algorithm to structural mechanics problems with multi-point constraints". In: *Advanced Modeling and Simulation in Engineering Sciences* 7.45 (2020). DOI: 10.1186/s40323-020-00181-2.

[60] Jeonghun Lee, Kent-Andre Mardal, and Ragnar Winther. "Parameter-Robust Discretization and Preconditioning of Biot's Consolidation Model". In: *SIAM Journal on Scientific Computing* 39 (Jan. 2017), A1–A24. DOI: 10.1137/15M1029473.

[61] Jörg Liesen and Zdeněk Strakoš. *Krylov subspace methods. Principles and analysis*. English. MSC2010: 65F10 = Iterative numerical methods for linear systems MSC2010: 65-02 = Research exposition (monographs, survey articles) pertaining to numerical analysis. Oxford: Oxford University Press, 2013. ISBN: 9780199655410. URL: https://zbmath.org/?q=an%3A1263.65034 (visited on 06/18/2020).

[62] P. Luo et al. "Multigrid method for nonlinear poroelasticity equations". en. In: *Computing and Visualization in Science* 17.5 (Oct. 2015), pp. 255–265. ISSN: 1433-0369. DOI: 10.1007/s00791-016-0260-8. URL: https://doi.org/10.1007/s00791-016-0260-8 (visited on 05/18/2020).

[63] P. Luo et al. "On an Uzawa smoother in multigrid for poroelasticity equations". In: *Numerical Linear Algebra with Applications* 24.1 (2017). e2074 nla.2074, e2074. DOI: 10.1002/nla.2074. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nla.2074. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nla.2074.

[64] Kent-Andre Mardal and Ragnar Winther. "Preconditioning discretizations of systems of partial differential equations". In: *Numerical Linear Algebra with Applications* 18 (Jan. 2011), pp. 1–40. DOI: 10.1002/nla.716.

[65] B. Markert, Y. Heider, and W. Ehlers. "Comparison of monolithic and splitting solution schemes for dynamic porous media problems". In: *International Journal for Numerical Methods in Engineering* 82.11 (2010), pp. 1341–1383. DOI: 10.1002/nme.2789. URL: http://doi.org/10.1002/nme.2789.

[66] Takashi Matsushima, René Chambon, and Denis Caillerie. "Large strain finite element analysis of a local second gradient model: Application to localization". In: *International Journal for Numerical Methods in Engineering* 54 (June 2002), pp. 499–521. DOI: 10.1002/nme.433.

[67] Gérard Meurant. "A Multilevel AINV Preconditioner". en. In: *Numerical Algorithms* 29.1 (Mar. 2002), pp. 107–129. ISSN: 1572-9265. DOI: 10.1023/A:1014816109110. URL: https://doi.org/10.1023/A:1014816109110 (visited on 06/01/2020).

[68] R.D. Mindlin. "Second gradient of strain and surface-tension in linear elasticity". In: *International Journal of Solids and Structures* 1.4 (1965), pp. 417–438. ISSN: 0020-7683. DOI: https://doi.org/10.1016/0020-7683(65)90006-5. URL: https://www.sciencedirect.com/science/article/pii/0020768365900065.

[69] Malcolm Murphy, Gene Golub, and Andrew Wathen. "A Note on Preconditioning for Indefinite Linear Systems". In: *Siam Journal on Scientific Computing* 21 (June 2000). DOI: 10.1137/S1064827599355153.

[70] D. B. OLIVEIRA, S. S. PENNA, and R. L. S. PITANGUEIRA. "Elastoplastic constitutive modeling for concrete: a theoretical and computational approach". en. In: *Revista IBRACON de Estruturas e Materiais* (Feb. 2020), pp. 171–182. ISSN: 1983-4195.

[71] Maxim A. Olshanskii and Eugene E. Tyrtyshnikov. *Iterative Methods for Linear Systems: Theory and Applications*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2014.

[72] C. C. Paige and M. A. Saunders. "Solution of Sparse Indefinite Systems of Linear Equations". In: *SIAM Journal on Numerical Analysis* 12.4 (Sept. 1975), pp. 617–629. ISSN: 0036-1429. DOI: 10.1137/0712047. URL: https://epubs.siam.org/doi/10.1137/0712047 (visited on 05/25/2020).

[73] R. H. J. PEERLINGS et al. "GRADIENT ENHANCED DAMAGE FOR QUASI-BRITTLE MATERIALS". In: *International Journal for Numerical Methods in Engineering* 39.19 (1996), pp. 3391–3403. DOI: https://doi.org/10.1002/(SICI)1097-0207(19961015)39:19<3391::AID-NME7>3.0.CO;2-D. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-0207%2819961015%2939%3A19%3C3391%3A%3AAID-NME7%3E3.0.CO%3B2-D. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0207%2819961015%2939%3A19%3C3391%3A%3AAID-NME7%3E3.0.CO%3B2-D.

[74] R.H.J. Peerlings et al. "A critical comparison of nonlocal and gradient-enhanced softening continua". In: *International Journal of Solids and Structures* 38.44 (2001), pp. 7723–7746. ISSN: 0020-7683. DOI: https://doi.org/10.1016/S0020-7683(01)00087-7. URL: https://www.sciencedirect.com/science/article/pii/S0020768301000877.

[75] G. Pijaudier-Cabot and Zdenek Bazant. "Nonlocal Damage Theory". In: *Journal of Engineering Mechanics-asce - J ENG MECH-ASCE* 113 (Oct. 1987). DOI: 10.1061/(ASCE)0733-9399(1987)113:10(1512).

[76] Catherine Powell and David Silvester. "Black-Box Preconditioning for Mixed Formulation of Self-Adjoint Elliptic PDEs". en. In: *Challenges in Scientific Computing - CISC 2002*. Ed. by Eberhard Bänsch. Lecture Notes in Computational Science and Engineering. Berlin, Heidelberg: Springer, 2003, pp. 268–285. ISBN: 9783642190148. DOI: 10.1007/978-3-642-19014-8_13.

[77] Florian Rathgeber et al. "Firedrake: automating the finite element method by composing abstractions". In: *CoRR* abs/1501.01809 (2015). arXiv: 1501.01809. URL: http://arxiv.org/abs/1501.01809.

[78] K. H. Roscoe. "The Influence of Strains in Soil Mechanics". In: *Géotechnique* 20.2 (1970), pp. 129–170. DOI: 10.1680/geot.1970.20.2.129. eprint: https://doi.org/10.1680/geot.1970.20.2.129. URL: https://doi.org/10.1680/geot.1970.20.2.129.

[79] Miroslav Rozložník. "Iterative Solution of Saddle-Point Problems". en. In: ed. by Miroslav Rozložník. Nečas Center Series. Cham: Springer International Publishing, 2018, pp. 49–68. ISBN: 9783030014315. DOI: 10.1007/978-3-030-01431-5_6. URL: https://doi.org/10.1007/978-3-030-01431-5_6 (visited on 06/12/2020).

[80] Y. Saad. *Iterative Methods for Sparse Linear Systems*. 2nd. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003. ISBN: 0898715342.

[81] Youcef Saad and Martin H. Schultz. "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems". In: *SIAM Journal on Scientific and Statistical Computing* 7.3 (July 1986), pp. 856–869. ISSN: 0196-5204. DOI: 10.1137/0907058. URL: https://epubs.siam.org/doi/10.1137/0907058 (visited on 10/07/2020).

[82] Nikolay Sakharnykh. "High-Performance Geometric Multi-Grid with GPU Acceleration". URL: https://devblogs.nvidia.com/high-performance-geometric-multi-grid-gpu-acceleration/.

[83] *Salomé Platform*. https://www.salome-platform.org/. 2022.

[84] A. Settari and D. Walters. "Advances in Coupled Geomechanical and Reservoir Modeling With Applications to Reservoir Compaction". In: *SPE Journal - SPE J* 6 (Sept. 2001), pp. 334–342. DOI: 10.2118/74142-PA.

[85] John Y. Shu, Wayne E. King, and Norman A. Fleck. "Finite elements for materials with strain gradient effects". In: *International Journal for Numerical Methods in Engineering* 44.3 (1999), pp. 373–391. DOI: https://doi.org/10.1002/(SICI)1097-0207(19990130)44:3<373::AID-NME508>3.0.CO;2-7.

[86] Gerard Sleijpen, Henk Van der Vorst, and J. Modersitzki. "Differences In The Effects Of Rounding Errors In Krylov Solvers For Symmetric Indefinite Linear Systems". In: *SIAM Journal on Matrix Analysis and Applications* 22 (Feb. 2000). DOI: 10.1137/S0895479897323087.

[87] Barry F. Smith, Petter E. Bjørstad, and William D. Gropp. *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. English (US). United Kingdom: Cambridge University Press, 1996. ISBN: 9780521602860.

[88] *Scalable Multistage Linear Solver for Coupled Systems of Multisegment Wells and Unstructured Reservoir Models*. Vol. All Days. SPE Reservoir Simulation Conference. SPE-119175-MS. Feb. 2009. DOI: 10.2118/119175-MS. eprint: https://onepetro.org/spersc/proceedings-pdf/09RSS/All-09RSS/SPE-119175-MS/1784642/spe-119175-ms.pdf. URL: https://doi.org/10.2118/119175-MS.

[89] S. P Vanka. "Block-implicit multigrid solution of Navier-Stokes equations in primitive variables". en. In: *Journal of Computational Physics* 65.1 (July 1986), pp. 138–158. ISSN: 0021-9991. DOI: 10.1016/0021-9991(86)90008-2. URL: http://www.sciencedirect.com/science/article/pii/0021999186900082 (visited on 06/01/2020).

[90] I. Vardoulakis. "Shear band inclination and shear modulus of sand in biaxial tests". In: *International Journal for Numerical and Analytical Methods in Geomechanics* 4.2 (1980), pp. 103–119. DOI: https://doi.org/10.1002/nag.1610040202. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nag.1610040202. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nag.1610040202.

[91] Andrew Wathen, Bernd Fischer, and David Silvester. "The convergence rate of the minimal residual method for the Stokes problem". In: *Numerische Mathematik* 71.1 (Aug. 1995), pp. 121–134. DOI: 10.1007/s002110050138.

[92] Joshua White and Ronaldo Borja. "Block-preconditioned Newton–Krylov solvers for fully coupled flow and geomechanics". In: *Computational Geosciences* 15 (Sept. 2011), pp. 647–659. DOI: 10.1007/s10596-011-9233-7.

[93] Joshua A. White, Nicola Castelletto, and Hamdi A. Tchelepi. "Block-partitioned solvers for coupled poromechanics: A unified framework". In: *Computer Methods in Applied Mechanics and Engineering* 303 (2016), pp. 55–74. ISSN: 0045-7825. DOI: https://doi.org/10.1016/j.cma.2016.01.008. URL: http://www.sciencedirect.com/science/article/pii/S0045782516000104.

[94] R. Wienands et al. "An Efficient Multigrid Solver based on Distributive Smoothing for Poroelasticity Equations". In: *Computing* 73.1 (Mar. 2004), pp. 99–119. ISSN: 1572-9265. DOI: 10.1007/s00607-004-0078-y.

[95] Walter Zulehner. "Analysis of iterative methods for saddle point problems: A unified approach". In: *Math. Comput.* 71 (Jan. 2002), pp. 479–505. DOI: 10.1090/S0025-5718-01-01324-2.