



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Ingegneria Civile, Edile e Ambientale (ICEA)

SCUOLA DI DOTTORATO DI RICERCA IN:
SCIENZE DELL'INGEGNERIA CIVILE ED AMBIENTALE
XXX CICLO

**Numerical models for the large-scale simulation of fault and
fracture mechanics**

Direttore della Scuola: Ch.mo Prof. Stefano Lanzoni

Relatore: Prof. Massimiliano Ferronato

Correlatore: Prof. Carlo Janna

Dottorando: Andrea Franceschini

Abstract

Department of Civil, Architectural and Environmental Engineering

Doctor of Philosophy

Numerical models for the large-scale simulation of fault and fracture mechanics

Andrea Franceschini

The possible activation of pre-existing faults and the generation of new fractures in the subsurface may play a critical role in several fields of great social interest, such as the management and the exploitation of groundwater resources, especially in arid areas, the hydrocarbon recovery and storage, and the monitoring of the seismic activity in the Earth's crust. The sliding and/or opening of a fault can create preferential leakage paths for the pore fluid escape, causing a matter of great concern in the process of storing fluids and hydrocarbons underground. The most challenging effect connected to a fault activation is the possible earthquake triggering. Many earthquakes associated with the production and injection of fluids have been recently reported. Similar issues arise also in the development of unconventional hydrocarbon reservoirs, that has recently experienced a dramatic increase thanks to the deployment of the "fracking" technology, which is based on the massive generation of fractures through the injection of fluids at high pressures. The use of this technique in densely populated areas has raised a large scientific debate on the possible connected environmental risks. The over-exploitation of fresh aquifers in arid regions has caused the generation of significant ground fissures. In this thesis, a novel formulation based on the use of Lagrange multipliers has been developed for the stable and robust numerical modeling of fault mechanics. A fault or fracture is simulated as a pair of inner surfaces included in a 3D geological formation where Lagrange multipliers are used to prescribe the contact constraints. The standard variational formulation of the contact problem with Lagrange multipliers is modified to take into account the energy dissipated by the frictional work

along the activated fault portion. This term is computed by making use of the principle of maximum plastic dissipation, whose application defines the direction of the limiting shear stress vector. The novel approach has been verified against analytical solutions and applied in a number of real-world problems. In particular, we test the novel approach in four cases: (i) mechanics of two adjacent blocks, to investigate the numerical properties of the algorithm; (ii-iii) ground fractures due to groundwater withdrawal, with different geometries; (iv) fault reactivation in an underground reservoir subject to primary production and Underground Gas Storage cycles. The results are analyzed and commented. In the fourth case, the possible magnitude of the seismic events triggered by fault reactivation is computed, in order to evaluate whether underground human activities may generate seismicity.

The application of the fault model to large-scale problems gives rise to a set of sparse discrete systems of linearized equations with a generalized non-symmetric saddle point structure. The second part of this thesis is devoted to the development of efficient algorithms for the iterative solution of this kind of system. We focus on a preconditioning technique, denoted as “constraint preconditioning”, which exploits the native block structure of the Jacobian. The quality and performance of the preconditioner relies on two steps: (i) the preconditioning of the leading block and (ii) the Schur complement computation. In this work, novel preconditioning techniques for the leading block based on a multilevel framework are developed and tested. The main idea behind the multilevel preconditioner is to improve the quality of the factorized approximate inverses borrowing the scheme of incomplete factorizations, thus introducing some sequentiality in perfectly parallelizable algorithms. The proposed approach is robust, from a theoretical point of view, and very efficient in parallel environment. As to the latter point, i.e. the Schur complement computation, it can be done with the aid of different approximations. The main difference is whether the Jacobian is symmetrized or not. The computation can be founded on the FSAI approximation of the leading block inverse or on a physically-based block diagonal block algorithm. The Schur complement must be inverted, thus other possibilities come in. The approximate Schur complement can be inverted through FSAI, if symmetric, or an incomplete factorization, if non-symmetric, but it can also be solved exactly, thanks to a direct solver.

The performances of the proposed algorithms are finally investigated and discussed in a set of real-world numerical examples.

Sommario

Dipartimento di Ingegneria Civile, Edile e Ambientale

Dottorato di Ricerca

Numerical models for the large-scale simulation of fault and fracture mechanics

Andrea Franceschini

L'eventuale attivazione di faglie preesistenti e la generazione di nuove fratture nel sottosuolo può svolgere un ruolo cruciale in diversi campi di grande interesse sociale, come la gestione e lo sfruttamento delle risorse idriche sotterranee, in particolare in aree aride, l'estrazione e lo stoccaggio di idrocarburi e il monitoraggio dell'attività sismica nella crosta terrestre. Lo scorrimento e/o l'apertura di una faglia può creare percorsi di dispersione preferenziali per la fuoriuscita del fluido interstiziale causando gravi problemi nel processo di stoccaggio sotterraneo di fluidi e idrocarburi. L'effetto più problematico connesso ad un'attivazione di faglia è l'eventuale innesco di un terremoto. Recentemente, sono stati riportati molti terremoti associati all'estrazione e all'iniezione di fluidi. Problemi simili si verificano anche nella coltivazione di giacimenti di idrocarburi non convenzionali, che recentemente hanno registrato un drastico aumento grazie alla diffusione della tecnologia chiamata "fracking", che si basa sulla creazione di numerose fratture attraverso l'iniezione di fluidi ad alta pressione. L'uso di questa tecnica in aree densamente popolate ha sollevato un ampio dibattito scientifico sui possibili rischi ambientali ad essa connessi. L'eccessivo sfruttamento di acquiferi in regioni aride ha generato notevoli fratture nel terreno. In questa tesi, è stata sviluppata una nuova formulazione basata sui moltiplicatori di Lagrange, in grado di offrire una robusta modellazione numerica della meccanica delle faglie. Una faglia o una frattura viene simulata come una coppia di superfici interne inserite in una formazione geologica tridimensionale, in cui i moltiplicatori di Lagrange vengono utilizzati per imporre le condizioni di contatto. La classica formulazione variazionale del problema

di contatto con i moltiplicatori di Lagrange è modificata per includere l'energia dissipata dal lavoro delle forze d'attrito nella parte di faglia attivata. Questo quantità viene calcolata utilizzando il principio della massima dissipazione plastica, la cui applicazione definisce la direzione del vettore di tensione tangenziale in caso di scorrimento. Il nuovo approccio è stato validato grazie a soluzioni analitiche e utilizzato in diversi problemi derivanti da applicazioni reali. In particolare, il nuovo approccio è stato testato in quattro casi: (i) la meccanica di due blocchi adiacenti, per analizzare il comportamento dell'algoritmo da un punto di vista numerico; (ii-iii) fratture del sottosuolo causate da estrazione d'acqua subsuperficiale, con differenti geometrie; (iv) riattivazione di faglie in un giacimento soggetto a produzione primaria e cicli di stoccaggio di gas. I risultati sono stati analizzati e commentati. Nel quarto caso, si è determinata la possibile magnitudo relative ad eventi provocati dalla riattivazione delle faglie, per capire se le attività umane nel sottosuolo possano provocare fenomeni sismici.

L'applicazione del modello di faglia a problemi di grande scala genera un insieme di sistemi discreti e sparsi di equazioni linearizzate, con una struttura tipo punto sella non simmetrica generalizzata. La seconda parte di questa tesi è rivolta allo sviluppo di algoritmi efficienti per la soluzione iterativa di sistemi di questo tipo. In particolare, si utilizza la tecnica detta "constraint preconditioning", che sfrutta la suddivisione in blocchi dello jacobiano. La qualità del preconditionatore si basa su due fasi: (i) il preconditionamento del blocco principale e (ii) il calcolo del complemento di Schur. In questo lavoro, sono state sviluppate e testate nuove tecniche di preconditionamento per il blocco principale, basate su un framework multilevello. L'idea di base di questo approccio è il miglioramento delle inverse approssimate grazie allo schema delle fattorizzate incomplete, introducendo, quindi, una sequenzialità in algoritmi perfettamente paralleli. La tecnica presentata è robusta da un punto di vista teorico e molto efficiente in ambiente parallelo. Per quanto riguarda l'ultimo punto, ossia il calcolo del complemento di Schur, sono state proposte diverse approssimazioni. La differenza principale risiede nel considerare o meno lo jacobiano simmetrizzato. Il calcolo può essere basato sulla FSAI, per approssimare l'inversa del blocco principale, o su un approccio diagonale a blocchi, che è legato alle caratteristiche fisiche del problema. Il complemento di Schur deve essere invertito, quindi altre scelte entrano in gioco. Si può utilizzare la

FSAI, se la matrice è simmetrica, o una fattorizzata incompleta, se non simmetrica, ma è possibile anche risolvere esattamente il complemento di Schur, grazie ad un solutore diretto.

Infine, le performance degli algoritmi proposti sono analizzate e discusse su un insieme di problemi numerici derivanti da applicazioni reali.

Contents

Abstract	iii
Sommario	vii
List of Figures	xiii
List of Tables	xix
List of Algorithms	xxi
1 Introduction	1
1.1 Summary	6
2 Mathematical model of fault mechanics	9
2.1 Introduction	9
2.2 Linear momentum balance	10
2.2.1 Constitutive model	12
2.2.2 Pore pressure variation	14
2.2.3 Numerical model	19
2.3 Faults and fractures mechanics	20
2.3.1 Lagrange multiplier approach	21
2.3.2 Numerical model	25
2.4 Solution algorithms	29
2.4.1 Theoretical properties of the Jacobian matrix	30
2.5 Implementation details	33
2.5.1 Local stiffness matrices for tetrahedral and triangular discretiza- tions	35
2.6 Model validation	40
3 Numerical applications	43
3.1 Case study A	43
3.2 Case study B	50
3.3 Case study C	55
3.4 Case D	58
3.4.1 Faulted reservoirs	58
3.4.2 Seismic magnitude estimate	59
3.4.3 Reservoir setting	63
3.4.4 Numerical results and discussion	72

4	Efficient solvers for the linear problem	81
4.1	Introduction	81
4.2	Factorized sparse approximate inverse for SPD matrices	86
4.2.1	Static FSAI	89
4.2.2	Dynamic FSAI	90
4.3	Multilevel approaches for FSAI preconditioning	92
4.3.1	Block Tridiagonal FSAI	93
4.3.2	Domain Decomposition FSAI	95
4.3.3	Multilevel FSAI Low Rank preconditioner	97
4.3.4	Improving the MF performance with Low-Rank corrections . . .	110
4.4	Numerical performances for the SPD matrix	116
4.4.1	Block Tridiagonal FSAI and Domain Decomposition FSAI	116
4.4.2	Multilevel FSAI with Low Rank corrections	120
4.5	The Inexact Constraint Preconditioning framework	129
4.5.1	ICP description and algorithms	132
4.5.2	Schur complement approximations and inversion	133
4.6	Numerical results	135
5	Conclusions	141
	Bibliography	145

List of Figures

2.1	Conceptual scheme for the fault modelling.	22
2.2	Local reference frame on the fault surfaces.	22
2.3	Interface element.	35
2.4	Area associated to node m	37
2.5	Test case with analytical solution.	40
2.6	3D model for the test case with analytical solution.	41
2.7	Analytical test case: numerical and analytical solutions along the crack. .	42
3.1	Case study A.	44
3.2	Case study A: load history.	45
3.3	Case study A: deformed configuration for the load steps no. 8, 15 and 20. The displacement exaggeration factor is 15.	45
3.4	Case study A: normal stress on the crack surface for the load steps no. 8, 15 and 20.	46
3.5	Case study A: shear stress on the crack surface for the load steps no. 8, 15 and 20.	46
3.6	Case study A: σ_n , τ_s and Δu_z along the crack central axis for the load steps no. 8, 15, and 20.	47
3.7	Case study A: σ_n , τ_s and Δu_z vs load step at different elevation z	47
3.8	Case study A: convergence profile of the non-linear algorithm.	48
3.9	Case study A: convergence profile of the non-linear algorithm at step no. 8 (a), halving the load step size (b), and neglecting the numerical contribution of the integral over $\bar{\Gamma}$ in equation (2.70) for the full (c) and half-size load step (d).	49

3.10 Case study B. Sketch of the typical geological setting ground rupture in Queretaro City. The simulated configuration is highlighted by the red box.	50
3.11 Case study B.	50
3.12 Case study B: pressure maps in the aquifer after 3, 7 and 10 years of pumping (above) and along a vertical profile at the central well location (below). Pressure is in kPa.	51
3.13 Case study B: opening (left) and sliding (right) along the fracture.	52
3.14 Case study B: opening (above) and sliding (below) on the fissure at the end of the simulation.	53
3.15 Case study B: land subsidence along the x -direction.	54
3.16 Case study B: land subsidence map.	54
3.17 Case study B: distribution of sliding (light gray) and non sliding (dark gray) elements on the fissure at year 5 obtained with a penalty approach [Ferronato et al., 2008a; Janna et al., 2010b].	54
3.18 Case study C. Sketch of the typical geological setting ground rupture in Wuxi, China. The simulated configuration is highlighted by the red box.	55
3.19 Case study C: perspective views of the geological setting.	55
3.20 Case study C: 3D FE-IE grid. The materials are highlighted in various colors: rock in red, aquifers in blue, and aquitards in green.	56
3.21 Case study C: opening and sliding of the ground rupture after 10 years of water withdrawal.	57
3.22 Case study C: land subsidence after 10 years of water withdrawal.	57
3.23 Diagram description of hydrocarbon traps. Adapted from PetroWiki [2015].	58
3.24 Map of the Groningen gas field. Faults are black lines. The coastline is the blue line while the reservoir trace is the pink one. Adapter from Van Thienen-Visser et al. [2015].	59

3.25	Case D: (a) Seismic cross-section of the gas field. The main geological formations, the trace of the faults/thrusts and the GWC (gas water contact) location are shown. (b) Map of the top depth (m below msl) of the gas bearing pool PA. The two main faults (called "fault-1" and "fault-2") are indicated.	63
3.26	Case D: (a) Axonometric view and (b) plain view of the full geomechanical FE-IE model developed for the analyzed gas field. The vertical exaggeration in (a) is 5. The trace of the field (in red) and the faults (in green) crossing and confining northward the reservoir and the waterdrive are highlighted in (b). We designed two faults, that extend from -1200 to $-3000m$ depth, with a length of about $40km$ (regional fault, fault-1) and $5km$ (local fault, fault-2).	65
3.27	Case D: (a) Axonometric detail and (b) vertical southwest - northeast cross-section of the 3D FE-IE model of the gas field showing different portions of the geologic system. The colors are used to highlight the gas traps (FE in yellow), the two faults (IE in orange and red), the waterdrive (FE in light blue), the portion of the formation hydraulically disconnected from the reservoir north-east of the fault (light brown), the clayey interlayers (green), and the over- under- and sideburden (grey). The vertical exaggeration is 5. The trace of the vertical cross-section is shown in Figure 3.26b.	66
3.28	Case D: Average pore pressure behavior (bars) in 50 years.	67
3.29	Case D: (a) Pore pressure decline (bars) in level P_A at the end of the field production life and (b) pore pressure increase during a UGS injection cycle from April to November of the 45-th year since the start of the simulation. The traces of the mineralized pool and faults are highlighted by red and green lines, respectively.	67
3.30	Case D: Normal stress variations (bar) in level P_A along the west-east ($\Delta\sigma_x$), north-south ($\Delta\sigma_y$), and vertical ($\Delta\sigma_z$) directions after 12 years, i.e. at the end of the field production life (a, c, e), and during the extraction phase from one Autumn to the next Spring (b, d, f).	69

3.31	Case D: Normal stress variations (bar) in a vertical cross-section through the top of the gas field along the west-east ($\Delta\sigma_x$), north-south ($\Delta\sigma_y$), and vertical ($\Delta\sigma_z$) directions after 12 years, i.e. at the end of the field production life (a, c, e), and during the extraction phase from one Autumn to the next Spring (b, d, f).	71
3.32	Case D: (a) Direction of the maximum horizontal stress in the horizontal plane σ_{h2} . (b) Initial normal and tangential stress, and potential stress drop on the faults with scenario B.	72
3.33	Case D: Scenario B: change of the effective normal and (modulus of) shear stress on the faults at the end of the primary production ((a) and (c), respectively) and during an UGS withdrawal phase ((b) and (d), respectively).	73
3.34	Case D: Perspective view on the fault discretization with the cumulative slip during the primary production life of the reservoir, as computed by the IE solution in (a) scenario B and (b) scenario (D). Vertical exaggeration is 5.	75
3.35	Case D: Area A , and average (s_a) and maximal (s_{max}) yearly slip of the activated portion of the faults during the primary development of the reservoir for the four scenarios addressed by the study.	76
3.36	Case D: Average (bullet) and standard deviation (bar) of the magnitude M computed using the numerical results (equation (3.5)) and the Kanamori [1977] equation (3.3) during the UGS cycles. The M values for the four scenarios are plotted against the difference between the actual monthly p and the average pressure p_{avg} during this period. The dashed lines represent the linear regressions, colored according to the bullet/bar of the corresponding scenario.	79
4.1	Sparsity pattern of the original test case on the left and the reordered pattern employed by the BTFSAI preconditioner on the right.	94
4.2	Sparsity pattern of the original test case on the left and the reordered pattern employed by the DDFSAI preconditioner on the right.	96

4.3	Comparison between the eigenspectra of A , $\tilde{L}\tilde{L}^T$ and $(G^T G)^{-1}$ for the <code>bcsstk16</code> matrix from the University of Florida sparse matrix collection.	100
4.4	Schematic representation of the system of inequalities (4.63).	105
4.5	Number of iterations needed for convergence as a function of the number of blocks used, left. Wall clock time needed for computing the preconditioner and solving the linear system again as a function of the number of blocks used, right. These figures show the results obtained with DDFSAI and BTFSAI.	118
4.6	Convergence results for the matrices <code>Ecology2</code> , <code>Tmt-sym</code> , <code>Cylinder</code> and <code>Hook1498</code> , respectively.	119
4.7	Scalability profiles of DDFSAI and BTFSAI, respectively.	120
4.8	<code>bcsstk38</code> test case: eigenspectrum of A (left) and $(\tilde{S} - S)$ (right) for different Block FSAI F .	121
4.9	<code>bcsstk38</code> test case: eigenspectrum of E_K (left) and rightmost eigenvalues of the preconditioned matrices (4.67) and (4.59) (right) for $k_{G,max} = 30$ and $k_{F,max} = 60$. The upper bound (4.76) is also shown.	123
4.10	Strong scalability test for a regular 300^3 Laplacian.	128
4.11	Non-zero entries (in red) of B_2^T , A , and B_1 required in the approximate computation of $B_2^{T,(\lambda)} A^{-1} B_1^{(\lambda)}$. Adapted from Castelletto et al. [2016].	135
4.12	Strong scalability test for Case D matrix. Approaches 1 and 2 are used.	139

List of Tables

2.1	Algorithm for the control of the fault activation.	34
3.1	Case study B: Young modulus, E , Poisson ratio, ν , and permeability K of the different materials.	51
3.2	Case study C: Young modulus, E and Poisson ratio, ν of the different materials.	56
3.3	Case D: Geometric and petrophysical properties of the different pools of the gas field.	64
3.4	Case D: Geomechanical parameters for the transversely isotropic consti- tutive model.	68
3.5	Case D: Area (A), and average (s_a) and maximal (s_{max}) slip of the ac- tivated portion of the faults during UGS cycles for the four scenarios addressed by the study.	74
3.6	Case D: Average and standard deviation of the simulated magnitude M	78
4.1	BTFSAI - DDFSAI: Test matrices.	118
4.2	bcsstk38 test case: comparison between the conditioning numbers of (4.59) and (4.67) varying F with $\ E_K\ = 0$. λ and $\tilde{\lambda}$ denote the eigenval- ues of $M^{-1}AM^{-T}$ and $\tilde{M}^{-1}\tilde{A}\tilde{M}^{-T}$, respectively.	122
4.3	bcsstk38 test case: eigenvalue distribution of (4.59) and (4.67) varying F and G . The same notation as in Table 4.2 is used.	122
4.4	bcsstk38 test case: effect of Descending and Ascending Low-Rank Corrections, d_{lr} and a_{lr} , respectively, with $k_{G,max} = 30$ and $k_{F,max} = 60$. The target of Ascending Low-Rank correction is $\tilde{S}^{-1}$ and S^{-1} in the up- per and lower table, respectively.	124

4.5	Cube test case: MFLR performance varying n_l and ϵ_G ($\epsilon_F = 10^{-2}$, $dlr = alr = 0$).	125
4.6	Cube test case: MFLR performance varying n_l and ϵ_F ($\epsilon_G = 10^{-2}$, $dlr = alr = 0$).	126
4.7	Cube test case: MFLR performance varying either dlr or alr ($\epsilon_G = \epsilon_F = 10^{-2}$, $n_l = 10$).	126
4.8	Cube test case: MFLR performance varying both dlr and alr ($\epsilon_G = \epsilon_F = 10^{-2}$, $n_l = 10$).	127
4.9	MFLR: Test matrices.	127
4.10	Computational performance of the adaptive FSAI and MFLR preconditioners for the test matrices of Table 4.9.	127
4.11	FSAI set-up parameters.	137
4.12	Size and number of non-zeroes of the test matrices.	137
4.13	Best numerical results with approach 1, FSAI + FSAI.	137
4.14	Best numerical results with approach 2, FSAI + Direct.	138
4.15	Best numerical results with approach 3, FSAI + BDSchur.	138

List of Algorithms

1	CONSTRUCTION OF THE BTFSAI PRECONDITIONER	95
2	CALCULATION OF $\mathbf{u} = M^{-1}\mathbf{v}$	95
3	MULTILEVEL FACTORIZATION SET-UP	99
4	MULTILEVEL FACTORIZATION APPLICATION	99
5	MULTILEVEL FSAI SET-UP	103
6	MFLR SET-UP	115
7	MFLR APPLY	116
8	PRECONDITIONED BI-CGSTAB	130
9	ICP APPLICATION	133
10	ICP SPLIT APPLICATION (LEFT)	133
11	ICP SPLIT APPLICATION (RIGHT)	134

Chapter 1

Introduction

The possible activation of pre-existing faults and the generation of new fractures in the subsurface may play a critical role in several fields of great social interest, such as the management and the exploitation of groundwater resources, especially in arid areas, the hydrocarbon recovery and storage, and the monitoring of the seismic activity in the Earth's crust. The most known and studied effect related to the exploitation of subsurface resources, either groundwater or hydrocarbons, is land subsidence, that is still an important challenge in several world sites, e.g., [Galloway and Burbey \[2011\]](#); [Gambolati and Teatini \[2015\]](#). It has been observed that the effect of a deep fault activation on the surface deformation is often negligible [\[Ferronato et al., 2008a\]](#), but it can be much more significant from other points of view. For instance, the sliding and/or opening of a fault can create preferential leakage paths for the pore fluid escape, causing a matter of great concern in the process of storing fluids and hydrocarbons underground, such as CO_2 [\[Rutqvist and Tsang, 2002; Birkholzer and Zhou, 2009; Ferronato et al., 2010b; Cappa and Rutqvist, 2011a; Mazzoldi et al., 2012; Rinaldi and Rutqvist, 2013; Rutqvist et al., 2013a; Verdon et al., 2015\]](#) or natural gas [\[Teatini et al., 2011; Castelletto et al., 2013\]](#). Another important consequence is the possible damage, or even break, of the well casings, that generally brings significant economical and environmental losses [\[Fredrich et al., 2000; Castelletto et al., 2010; Eslamizadeh and Samanirad, 2010\]](#). However, the most challenging effect connected to a fault activation is the possible earthquake triggering. Earthquakes associated with the production and injection of fluids have been recently reported, e.g., [\[González et al., 2012; Brodsky and Lajoie, 2013; Lei et al., 2013; Figueiredo et al., 2015; Rutqvist et al., 2016\]](#), as well as with the waste water disposal, e.g., [\[Keranen et al., 2013; Van der Elst et al., 2013\]](#). Similar issues arise also in

the development of unconventional hydrocarbon reservoirs, i.e., the so-called oil and gas shales, and in the groundwater withdrawal from arid and semi-arid regions. The production from shale reservoirs has recently experienced a dramatic increase thanks to the deployment of the “fracking” technology, which is based on the massive generation of fractures due to the injection of dense fluids at high pressures. The use of this technique in densely populated areas has raised a large scientific debate on the possible connected environmental risks [Engelder, 2012; Warner et al., 2012; Cueto-Felgueroso and Juanes, 2013]. The over-exploitation of fresh aquifers in arid regions has caused the generation of significant ground fissures, that have been reported in many alluvial basins since the late '70s. Examples of such localized ruptures, which can be regarded as newly generated faults emerging to the ground surface, have been recorded in the South-Western part of the United States [Holzer et al., 1979; Carpenter, 2015], the Libyan desert [Rothenburg et al., 1995], China [Da-Yu and Li, 2000; Wang et al., 2009; Ye et al., 2015; Zhang et al., 2016], central Mexico [Carreón-Freyre et al., 2005; Teatini et al., 2015] and Central Iran [Eslamizadeh and Samanirad, 2010]. These problems are common also in geothermal reservoirs, where massive water injection has the potential to elevate pore pressures within porous and fractured formations and to reactivate faults, either seismically or aseismically [Gan and Elsworth, 2014].

Despite the growing interest in the prediction of the mechanics of faults, many issues still remain unsolved. A most important one is the capability to reliably simulate the mechanical behavior of a fault in a numerically stable way. The use of numerical models to predict the variation of the full stress tensor, the pore pressure and the related fault slip during the production and/or injection of subsurface fluids is of paramount importance to develop reliable computational tools for the efficient and safe management of underground resources. Currently, the simplest way to deal with faults in a geomechanical model consists of regarding them as a part of the continuous porous body characterized by a different rheological behavior [Rutqvist et al., 2008; Cappa and Rutqvist, 2011a,c]. This approach, however, has several limitations. The most significant ones concern the difficulty in representing in a reliable way the actual constitutive behavior of the fault and modeling the possible slip of the discontinuity surface. The latter aspect turns out to be particularly important when the model is

used to predict the expected magnitude of the earthquake triggered by the fault activation. In fact, recent studies suggest the use of simple mathematical relationships for correlating the maximum magnitude of a seismic event to the sliding fault area and the average slippage [Kanamori and Anderson, 1975; Mazzoldi et al., 2012]. Another numerical approach that makes easier the identification of the localized discontinuous displacement possibly arising along the fault is based on the introduction of interface frictional elements. Several formulations have been proposed starting from the original work by Goodman et al. [1968], e.g., Ferronato et al. [2008a]; Beer [1985]; Cescotto and Charlier [1993]; Juanes et al. [2002]; Aagaard et al. [2013]; Jha and Juanes [2014]. Most of the currently available numerical methods make use of the penalty approach to enforce the continuity of the displacement across the fault whenever the activation condition is not achieved. This method is largely used for the ease of implementation and the fact that it does not increase the number of unknowns of the problem, but the resulting stiffness matrix can be severely ill-conditioned, requiring the use of special solvers [Ferronato et al., 2008b; Janna et al., 2009; Ferronato et al., 2012]. Moreover, such an ill-conditioning can cause severe difficulties in the convergence of the Newton-like scheme used for the solution of the non-linear problem, with oscillatory behaviors between two non-equilibrated configurations.

An effective alternative to the penalty method relies on the use of Lagrange multipliers to enforce the contact constraints over the fault surfaces, e.g., Bertsekas [1982]. The main numerical drawback connected with this technique is the fact that the introduction of the multipliers increases the size of the discrete problem, that typically exhibits a saddle-point structure. The present thesis introduces a novel formulation for the stable numerical simulation of the fault and fracture mechanics with the aid of the Lagrange multipliers. The basic variational formulation is modified by adding the contribution of the frictional work along the activated fault portion, computed according to the “Principle of Maximum Plastic Dissipation” [Simo and Hughes, 2006; Wriggers, 2006] which prescribes that the friction energy must be as large as possible. This formulation results in a non-symmetric generalized saddle point Jacobian matrix that ensures a robust convergence of the non-linear problem.

In the second part of this work, the solution of the large sparse linear system of

equations arising from the fault discretization is addressed. The Lagrangian approach to impose the contact constraints along a fault produces a generalized non-symmetric saddle point linear system:

$$J = \begin{bmatrix} A & B_1 \\ B_2^T & 0 \end{bmatrix} \quad (1.1)$$

where the leading block A is symmetric and positive definite. To solve large and sparse linear system arising from 3D discretizations, iterative solvers based on Krylov subspaces are usually preferred [Benzi, 2002; Ferronato, 2012]. The key factor to ensure robustness and efficiency, however, is the choice of the preconditioner. In our case, the so-called constraint preconditioning [Benzi et al., 2005; Bergamaschi, 2012] is used. The performance of this method basically relies on two steps: (i) the preconditioning of the leading block and (ii) the Schur complement computation.

The leading block A is an SPD matrix arising from the discretization of the elasticity operator. There are several different approaches for building an efficient preconditioner for this problem, either physics-based or purely algebraic. Among the algebraic algorithms, the most popular categories are incomplete factorizations, multigrid methods, and sparse approximate inverses [Benzi, 2002; Ferronato, 2012], though different preconditioners can be composed to form new ones, e.g., blending domain decomposition approaches with incomplete factorizations or approximate inverses, or using nested Krylov methods [Janna et al., 2013; McInnes et al., 2014]. Compared to incomplete factorizations, there are some sparse approximate inverses that are generally more robust and more appropriate for parallel computational architectures, i.e. the Stabilized Approximate Inverse (SAINV) [Benzi et al., 2001]. In particular, the Factorized Sparse Approximate Inverse (FSAI) [Kolotilina and Yeremin, 1993] is an algebraic preconditioner for SPD problems that proves effective in a wide range of applications, especially in its dynamically adaptive variants [Janna and Ferronato, 2011; Janna et al., 2015b]. This algorithm provides a factorized approximation of A^{-1} :

$$A^{-1} \simeq G^T G \quad (1.2)$$

where G is a lower triangular matrix explicitly computed in the set-up phase so as to

resemble the inverse of the lower Cholesky factor of A in the sense of the Frobenius norm. One of the most attractive features of FSAI for the modern computers is its intrinsic high degree of parallelism. In fact, each row of G can be fully formed independently from the others, with the parallelization trivially accomplished by evenly subdividing the rows among threads and/or processes. The degree of parallelism is even redundant, as the number of rows is much larger than the usually available number of computing cores. The central idea presented in this work to improve the FSAI performance is to introduce some sequentiality in the FSAI computation, in order to use the information extracted from the earlier set-up stages for the remaining rows. This concept, which is the basis for incomplete factorizations, has been already introduced in the context of approximate inverses in [Chow and Saad \[1998\]](#); [Raghavan and Teranishi \[2010\]](#), and more recently in [Paludetto Magri et al. \[2017\]](#), where both a block tridiagonal and a domain decomposition approach have been used to improve the FSAI performance. In the present work, we formalize a general multilevel framework that can include all the former approaches.

One of the difficulties arising in such multilevel generalization of FSAI is related to the accuracy in the computation of the Schur complement at each level. If the earlier levels are not well approximated, the resulting Schur complement can be inaccurate, with a consequent degradation of the solver performance. This issue has been recently addressed in the context of multilevel incomplete factorizations with the aid of low-rank corrections [\[Xi et al., 2016\]](#). Low-rank compression algorithms are gaining an increasing attention especially in direct linear solution methods, e.g., [Amestoy et al. \[2015\]](#); [Hackbusch \[1999\]](#); [Wang et al. \[2013\]](#); [Xia \[2012, 2013\]](#); [Xia et al. \[2010\]](#); [Li et al. \[2016\]](#); [Benzi et al. \[2017\]](#), with the basic idea of taking advantage from data sparsity instead of structural sparsity. Within the present work, low-rank corrections are introduced for both enhancing the preconditioner quality at the earlier levels (Descending Low-Rank, DLR) and improving the accuracy in the Schur complement computation (Ascending Low-Rank, ALR).

The second step of the constraint preconditioner is the Schur complement computation. This can be done with the aid of approximations, the main difference being

the symmetrization of the Jacobian. The computation can be founded on the FSAI approximation of the leading block inverse or on a physically based block diagonal block algorithm. The Schur complement must be inverted, thus other possibilities come in. The approximate Schur complement can be inverted through FSAI, if symmetric, or an incomplete factorization, if non-symmetric, but it can also be solved exactly, thanks to a direct solver. All these approaches are implemented and tested on a set of real-world numerical examples. Results and efficiency are discussed.

1.1 Summary

The main body of the thesis is organized in Chapters as follows.

- **Chapter 1** introduces the work presented in this thesis.
- **Chapter 2** presents the mathematical model used for the fault simulation in the framework of the finite element method. First, the governing differential equations are presented and discretized, then the presence of a discontinuity is addressed discussing the possible strategies for its numerical treatment. The numerical model is fully developed and verified against analytical solutions.
- **Chapter 3** discusses some examples of real-world applications of the algorithm. We present a simple test case, where the expected theoretical properties are verified, two realistic test cases, dealing with ground fissure and aquifer over-exploitation, where a comparison with another approach is done, and a reservoir simulation, with the prediction of the possible induced seismicity.
- **Chapter 4** investigates the solution of the discrete algebraic problem, focusing on the preconditioning of symmetric positive definite matrices, i.e. the leading block of the generalized saddle point matrix arising from the presented algorithm. The multilevel FSAI algorithm is developed and compared with existing techniques. Then, the overall constraint preconditioner is presented and discussed with different approximations for the global Schur complement. Numerical results are shown to investigate the overall preconditioning performance.

- **Chapter 5** concludes the thesis by emphasizing the major outcomes and the possible future developments.

Chapter 2

Mathematical model of fault mechanics

2.1 Introduction

In porous media, displacements and strains are caused by changes in effective stress, that is carried by the soil skeleton. The relation between total stress, pore pressure and effective stress is the well known *Terzaghi's principle* for saturated porous media [Terzaghi, 1925; Bishop and Bjerrum, 1960; Verruijt, 2013].

Following the classical Voigt notation for stress and strain [Voigt, 1887], the total stress vector $\hat{\sigma}$ is $[\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z, \hat{\tau}_{xy}, \hat{\tau}_{yz}, \hat{\tau}_{xz}]^T$ and Terzaghi's principle can be written as:

$$\hat{\sigma} = \sigma - \alpha p \mathbf{i} \quad (2.1)$$

with σ the effective stress vector, i.e. $[\sigma_x, \sigma_y, \sigma_z, \tau_{xy}, \tau_{yz}, \tau_{xz}]^T$, α the Biot coefficient [Biot, 1941], p the pore fluid pressure, always positive, and $\mathbf{i} = [1, 1, 1, 0, 0, 0]^T$ the vectorial form of the Kronecker delta. Tensile stresses are taken as positive. The Biot coefficient is defined as:

$$\alpha = 1 - \frac{C_r}{C_b} \quad (2.2)$$

where C_r is the grain compressibility and C_b is the bulk compressibility of the porous medium. All the classical equilibrium relationships (mechanical, flow, chemical, etc.) are partial differential equations (PDE) holding true point by point. If the medium is continuum, all points have the same features and the equations are valid for any

infinitesimal volume. By distinction, with porous media the material is no longer continuous beyond a given dimension, but solid and liquid phases can be recognized with different behaviors. Thus, to solve the classical PDEs we need to define a Representative Elementary Volume (REV) [Bear, 2013; Verruijt, 2013]. From the REV up to the whole body, the material can be considered as a continuum with mean properties between solid and liquid phases and the PDEs must be valid REV by REV, instead of point by point. Conversely, from the REV to an infinitesimal volume, different phases are distinguishable and the problem should be solved on each phase. This topic is discussed, for example, on Chapter 3 of Vafai [2015].

2.2 Linear momentum balance

Focusing on static problems, i.e. avoiding inertial accelerations, the governing equations are Cauchy's balance of linear momentum, that, written in term of total stress [Zienkiewicz and Taylor, 2000], read:

$$\begin{aligned}\frac{\partial \hat{\sigma}_x}{\partial x} + \frac{\partial \hat{\tau}_{yx}}{\partial y} + \frac{\partial \hat{\tau}_{zx}}{\partial z} + F_x &= 0 \\ \frac{\partial \hat{\tau}_{xy}}{\partial x} + \frac{\partial \hat{\sigma}_y}{\partial y} + \frac{\partial \hat{\tau}_{zy}}{\partial z} + F_y &= 0 \\ \frac{\partial \hat{\tau}_{xz}}{\partial x} + \frac{\partial \hat{\tau}_{yz}}{\partial y} + \frac{\partial \hat{\sigma}_z}{\partial z} + F_z &= 0\end{aligned}\tag{2.3}$$

On the boundary of the continuum, the surface stress equilibrium is imposed by:

$$\begin{aligned}\hat{\sigma}_x n_x + \hat{\tau}_{yx} n_y + \hat{\tau}_{zx} n_z &= t_x \\ \hat{\tau}_{xy} n_x + \hat{\sigma}_y n_y + \hat{\tau}_{zy} n_z &= t_y \\ \hat{\tau}_{xz} n_x + \hat{\tau}_{yz} n_y + \hat{\sigma}_z n_z &= t_z\end{aligned}\tag{2.4}$$

In the remainder, we assume to work with a non-polar continuum [Hadjesfandiari and Dargush, 2011], i.e. the moment equilibrium (balance of angular momentum) yields a symmetric stress tensor ($\hat{\tau}_{xy} = \hat{\tau}_{yx}$, $\hat{\tau}_{yz} = \hat{\tau}_{zy}$ and $\hat{\tau}_{xz} = \hat{\tau}_{zx}$). This is the reason why the classical Voigt notation implies vectors with 6 components, instead of 9.

In the mechanics of porous media, it is common to solve equations (2.3) with respect to effective stress, which is the responsible of the soil skeleton deformation. By introducing Terzaghi's principle (2.1), equations (2.3) become:

$$\begin{aligned}\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} + F_x &= \alpha \frac{\partial p}{\partial x} \\ \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} + F_y &= \alpha \frac{\partial p}{\partial y} \\ \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \sigma_z}{\partial z} + F_z &= \alpha \frac{\partial p}{\partial z}\end{aligned}\quad (2.5)$$

In this work, all body forces F_x , F_y and F_z are assumed to be constant during the simulation process, so the structural problem can be solved in term of stress increments with respect to an equilibrated configuration. Assuming that the initial stress state σ_0 is already in equilibrium with the body forces, the unknowns in equations (2.5) are the stress changes caused by pressure gradients. Therefore, the governing system of equations is:

$$\begin{aligned}\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} &= \alpha \frac{\partial p}{\partial x} \\ \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} &= \alpha \frac{\partial p}{\partial y} \\ \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \sigma_z}{\partial z} &= \alpha \frac{\partial p}{\partial z}\end{aligned}\quad (2.6)$$

where now, and in the remainder of this work, σ are effective stress variations with respect to σ_0 .

A classical way to address the solution of (2.6) is using the *Virtual Work Principle* [Zienkiewicz and Taylor, 2000]. Equating the internal to the external virtual work we obtain:

$$\delta W_e = \delta W_i \Rightarrow \int_{\Omega} \delta \bar{\mathbf{u}}^T \mathbf{q} dV + \int_{\partial\Omega} \delta \bar{\mathbf{u}}^T \mathbf{t} dS = \int_{\Omega} \delta \boldsymbol{\varepsilon}^T \boldsymbol{\sigma}(\boldsymbol{\varepsilon}) dV \quad (2.7)$$

where $\bar{\mathbf{u}}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}(\boldsymbol{\varepsilon})$ represent the displacement, the strain and the effective stress vector, respectively. Furthermore, the δ symbol denotes the virtual variables, δW_e and δW_i are the external and internal virtual work, respectively, Ω is the continuum domain with $\partial\Omega$ its frontier, $\mathbf{q} = -\alpha \text{grad}(p) = -\alpha \nabla(p)$ and $\mathbf{t}$, as in (2.4), is the surface traction. The virtual work principle can be also prescribed on the total stress, providing

an equivalent formulation. The former approach (2.7) is also known as pore pressure gradient formulation, while the latter is the total stress formulation [Gambolati et al., 2001].

Under the hypothesis of small displacements, the strain ε is related to the displacements according to the following relationship:

$$\varepsilon = L\bar{\mathbf{u}} \Rightarrow \begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ \gamma_{xy} \\ \gamma_{yz} \\ \gamma_{xz} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x} & 0 & 0 \\ 0 & \frac{\partial}{\partial y} & 0 \\ 0 & 0 & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \\ 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} & 0 & \frac{\partial}{\partial x} \end{bmatrix} \begin{bmatrix} \bar{u} \\ \bar{v} \\ \bar{w} \end{bmatrix} \quad (2.8)$$

where $\bar{u}$, $\bar{v}$ and $\bar{w}$ are the displacements along the coordinate directions. Thus, the virtual work principle can be written as:

$$\int_{\Omega} \delta \bar{\mathbf{u}}^T \mathbf{q} dV + \int_{\partial\Omega} \delta \bar{\mathbf{u}}^T \mathbf{t} dS = \int_{\Omega} \delta \bar{\mathbf{u}}^T L^T \boldsymbol{\sigma}(\varepsilon) dV \quad (2.9)$$

2.2.1 Constitutive model

The equilibrium problem consists of 6 congruence PDEs (2.8) and of 3 equilibrium PDEs (2.6), but there are more than 9 unknowns. Therefore, 6 additional equations are needed in order to link the stress and the strain. These relationships are provided by the constitutive law. The material model used in this work is a hypo-elastic hardening model with hysteresis, as described in Janna et al. [2012a]. This means that the response to a load increase is different following loading and unloading stress paths.

In (2.7), effective stress $\boldsymbol{\sigma}(\varepsilon)$ depends on the strain according to the selected constitutive model, that can be written in the classical incremental form:

$$d\boldsymbol{\sigma} = D_t(\boldsymbol{\sigma}) d\varepsilon \quad (2.10)$$

with $D_t(\boldsymbol{\sigma})$ the tangent constitutive matrix.

For a general anisotropic behaviour, the tangent compliance matrix $C_t = D_t^{-1}$ linking strain and stress variations is:

$$d\boldsymbol{\varepsilon} = C_t(\boldsymbol{\sigma}) d\boldsymbol{\sigma} \Rightarrow \begin{bmatrix} d\varepsilon_x \\ d\varepsilon_y \\ d\varepsilon_z \\ d\gamma_{xy} \\ d\gamma_{yz} \\ d\gamma_{xz} \end{bmatrix} = \begin{bmatrix} \frac{1}{E_h} & -\frac{\nu_h}{E_h} & -\frac{\nu_v}{E_h} & 0 & 0 & 0 \\ -\frac{\nu_h}{E_h} & \frac{1}{E_h} & -\frac{\nu_v}{E_h} & 0 & 0 & 0 \\ -\frac{\nu_v}{E_h} & -\frac{\nu_v}{E_h} & \frac{1}{E_v} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{G_h} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{G_v} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{G_v} \end{bmatrix} \begin{bmatrix} d\sigma_x \\ d\sigma_y \\ d\sigma_z \\ d\tau_{xy} \\ d\tau_{yz} \\ d\tau_{xz} \end{bmatrix} \quad (2.11)$$

where the pedices h and v indicate the horizontal plane and the vertical direction, respectively. The matrix (2.11) depends on 5 material parameters. The usual set of such parameters is:

- E_v : Young modulus along the vertical direction;
- ν_v : Poisson ratio linking horizontal and vertical strain;
- $\beta = \frac{E_h}{E_v}$: ratio between the horizontal and vertical Young moduli;
- $\gamma = \frac{\nu_h}{\nu_v}$: ratio between the horizontal and vertical Poisson ratios;
- $\eta = \frac{G_h}{G_v}$: ratio between the horizontal and vertical shear moduli.

The horizontal shear modulus G_h is linked to E_h and ν_h according to the well-known relationship:

$$G_h = \frac{E_h}{2(1 + \nu_h)} \quad (2.12)$$

Now, let us consider a classical oedometric test. As the sample lateral deformation is forbidden, we have:

$$\begin{aligned} \varepsilon_v &= \varepsilon_z \\ \sigma_x &= \sigma_y \end{aligned} \quad (2.13)$$

The computation of the vertical compaction of a sample with prevented lateral deformation under axial load provides the theoretical expression of the vertical uniaxial compressibility C_m . Equation (2.11) in a cylindrical reference frame [Janna et al., 2012a]

reads:

$$\begin{bmatrix} d\varepsilon_r \\ d\varepsilon_\theta \\ d\varepsilon_z \\ d\gamma_{rz} \end{bmatrix} = \begin{bmatrix} \frac{1}{E_h} & -\frac{\nu_h}{E_h} & -\frac{\nu_v}{E_h} & 0 \\ -\frac{\nu_h}{E_h} & \frac{1}{E_h} & -\frac{\nu_v}{E_h} & 0 \\ -\frac{\nu_v}{E_h} & -\frac{\nu_v}{E_h} & \frac{1}{E_v} & 0 \\ 0 & 0 & 0 & \frac{1}{G_v} \end{bmatrix} \begin{bmatrix} d\sigma_r \\ d\sigma_\theta \\ d\sigma_z \\ d\tau_{rz} \end{bmatrix} \quad (2.14)$$

Setting:

$$d\varepsilon_r = 0, \quad d\varepsilon_\theta = 0 \quad \text{and} \quad d\gamma_{rz} = 0 \quad (2.15)$$

we obtain:

$$d\tau_{rz} = 0 \quad \text{and} \quad d\sigma_r = d\sigma_\theta = \frac{\nu_v}{1 - \nu_h} d\sigma_z \quad (2.16)$$

Introducing (2.16) in the definition of $d\varepsilon_z$, given by (2.14), provides:

$$\frac{d\varepsilon_z}{d\sigma_z} = \frac{1}{E_v} \left(1 - \frac{1}{\beta} \frac{2\nu_v^2}{1 - \gamma\nu_v} \right) \equiv C_m \quad (2.17)$$

Hence, equation (2.11) can be expressed as a function of C_m , which can be used as independent material parameter instead of E_v .

2.2.2 Pore pressure variation

In the discussion carried out so far, the strengths associated to a pore pressure variation are treated as external forcing functions. However, the pore fluid flow and deformation occur simultaneously in the porous medium and influence each other. The governing equations, originally developed by Biot [1941], are therefore “coupled” and should be solved together.

In a general framework, we can consider three-phase porous media, where the phases are:

- solid, denoted by s ;
- liquid, denoted by w ;
- gaseous, denoted by g .

We denote by n the porosity and S_r the saturation index of the material. For the sake of simplicity, we assume $S_r = 1$, i.e. fully saturated porous medium without gaseous

phase. Darcy's equation is:

$$\mathbf{v} = -K \nabla \Phi \quad (2.18)$$

where:

- $\mathbf{v}$ is the velocity vector;
- K is the hydraulic conductivity tensor;
- Φ is the hydraulic head, that is $\Phi = \int_{z_0}^z dz + \int_{p_0}^p \frac{dp}{\gamma_w(p)}$, where z_0 and p_0 are reference values for the vertical direction and the pressure, respectively.

Darcy's law is cast in terms of the relative velocity of fluid to grains [Verruijt, 1969]:

$$\mathbf{v} = n (\mathbf{v}_w - \mathbf{v}_s) \quad (2.19)$$

The mass balance equations for both phases are:

$$-\nabla \cdot [(1 - n) \gamma_s \mathbf{v}_s] = \frac{\partial}{\partial t} [(1 - n) \gamma_s] \quad (2.20)$$

$$-\nabla \cdot [n \gamma_w \mathbf{v}_w] = \frac{\partial}{\partial t} [n \gamma_w] \quad (2.21)$$

where γ_s and γ_w denote the specific weight of solid and liquid phases, respectively. In this context, γ_s is assumed to be constant while γ_w depends on p . From (2.21), we have:

$$-\gamma_w \nabla \cdot (n \mathbf{v}_w) - n \mathbf{v}_w \cdot \nabla \gamma_w = n \frac{\partial \gamma_w}{\partial t} + \gamma_w \frac{\partial n}{\partial t} \quad (2.22)$$

Equation (2.20) can be written as:

$$\frac{\partial n}{\partial t} = \nabla \cdot [(1 - n) \mathbf{v}_s] \quad (2.23)$$

Setting $n \mathbf{v}_w = \mathbf{v} + n \mathbf{v}_s$ from (2.19) and using (2.23), equation (2.22) reads:

$$-\nabla \cdot \mathbf{v} - \frac{\mathbf{v}}{\gamma_w} \cdot \nabla \gamma_w = \frac{n}{\gamma_w} \left(\frac{\partial \gamma_w}{\partial t} + \mathbf{v}_s \cdot \nabla \gamma_w \right) + \nabla \cdot \mathbf{v}_s \quad (2.24)$$

Introducing the total derivative:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{v}_s \cdot \nabla \quad (2.25)$$

equation (2.24) becomes:

$$-\nabla \cdot \mathbf{v} - \frac{\mathbf{v}}{\gamma_w} \cdot \nabla \gamma_w = \frac{n}{\gamma_w} \left(\frac{D\gamma_w}{Dt} \right) + \nabla \cdot \mathbf{v}_s \quad (2.26)$$

The following classical state equation for γ_w is used:

$$\gamma_w = \gamma_{w0} e^{\beta(p-p_0)} \quad (2.27)$$

where p_0 and γ_{w0} are reference values for pressure and specific weight, respectively, and β is the fluid compressibility. Developing the total derivative of γ_w with respect to time, we have:

$$\frac{D\gamma_w}{Dt} = \gamma_{w0} e^{\beta(p-p_0)} \beta \frac{Dp}{Dt} = \gamma_w \beta \frac{Dp}{Dt} = \gamma_w \beta \left(\frac{\partial p}{\partial t} + \mathbf{v}_s \cdot \nabla p \right) \quad (2.28)$$

The material derivative on the last term of (2.28) can be replaced by a partial derivative if the following inequality is satisfied:

$$\mathbf{v}_s \cdot \nabla p \ll \frac{\partial p}{\partial t} \quad (2.29)$$

With this assumption, equation (2.26) becomes:

$$-\nabla \cdot \mathbf{v} - \frac{\mathbf{v}}{\gamma_w} \cdot \nabla \gamma_w = n\beta \frac{\partial p}{\partial t} + \nabla \cdot \mathbf{v}_s \quad (2.30)$$

De Wiest [1966] has provided an expansion of the left-hand side of (2.22) for the homogeneous case taking into account the relationship between the hydraulic conductivity tensor K and γ_w (see [Gambolati and Freeze, 1973]). This approach leads to:

$$-\nabla \cdot \mathbf{v} - \frac{\mathbf{v}}{\gamma_w} \cdot \nabla \gamma_w = \nabla \cdot \left[K \left(\frac{\nabla p}{\gamma_w} + \nabla z \right) \right] + 2K\beta \frac{\partial p}{\partial z} + K \frac{\beta}{\gamma_w} \nabla p \cdot \nabla p \quad (2.31)$$

Assuming that the inequality:

$$2K\beta\frac{\partial p}{\partial z} + K\frac{\beta}{\gamma_w}\nabla p \cdot \nabla p \ll \nabla \cdot \left[K \left(\frac{\nabla p}{\gamma_w} + \nabla z \right) \right] \quad (2.32)$$

holds true, equation (2.31) simplifies into:

$$\nabla \cdot \left[K \left(\frac{\nabla p}{\gamma_w} + \nabla z \right) \right] = n\beta\frac{\partial p}{\partial t} + \nabla \cdot \mathbf{v}_s \quad (2.33)$$

The term $\nabla \cdot \mathbf{v}_s$ depends on the porous medium displacement, that on its turn is related to the pressure field. This interrelationship was firstly exposed by Biot [1941], then Verruijt [1969] provided a clear derivation. By definition, $\mathbf{v}_s$ collects the time-derivatives of the displacement field $\mathbf{u} = [u, v, w]^T$:

$$\nabla \cdot \mathbf{v}_s = \nabla \cdot \frac{\partial \mathbf{u}}{\partial t} = \frac{\partial}{\partial t} \nabla \cdot \mathbf{u} \quad (2.34)$$

The operator $\nabla \cdot$ is the sum of the space derivation with respect to coordinate directions, i.e.:

$$\nabla \cdot \mathbf{u} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = \varepsilon_x + \varepsilon_y + \varepsilon_z = \varepsilon_v \quad (2.35)$$

where ε_v is the volumetric strain. Finally, rearranging (2.33) with (2.35), we can write the continuity equation:

$$\nabla \cdot \left[K \left(\frac{\nabla p}{\gamma_w} + \nabla z \right) \right] = n\beta\frac{\partial p}{\partial t} + \frac{\partial \varepsilon_v}{\partial t} \quad (2.36)$$

Other slightly different formulations take into account also the Biot coefficient α [Gambolati et al., 2000]. Following Biot's formulation [Biot, 1941], as modified later by Van der Knaap [1959] and Geertsma [1966], the coupled model for an isotropic medium reads:

$$\nabla \cdot \left[K \left(\frac{\nabla p}{\gamma_w} + \nabla z \right) \right] = [n\beta + C_r(\alpha - n)]\frac{\partial p}{\partial t} + \alpha\frac{\partial \varepsilon_v}{\partial t} \quad (2.37)$$

where α and C_r are the Biot coefficient (see Eq. (2.2)) and the grain compressibility, respectively. Another formulation can be found in [Gambolati et al. \[1996\]](#):

$$\nabla \cdot \left[K \left(\frac{\nabla p}{\gamma_w} + \nabla z \right) \right] = [n\beta + C_r (1 - n)] \frac{\partial p}{\partial t} + \frac{\partial \varepsilon_v}{\partial t} \quad (2.38)$$

Finally, [Bai and Abousleiman \[1997\]](#) proposed:

$$\nabla \cdot \left[K \left(\frac{\nabla p}{\gamma_w} + \nabla z \right) \right] = [n\beta + C_r (1 - n)] \frac{\partial p}{\partial t} + \alpha \frac{\partial \varepsilon_v}{\partial t} \quad (2.39)$$

Under some circumstances, the equation (2.36) can be decoupled from the linear momentum balance. For instance, in oedometric conditions the volumetric strain equates the vertical strain, i.e. $\varepsilon_v = \varepsilon_z$, and ε_z is linked to the effective vertical stress by the vertical compressibility (2.17).

From Terzaghi's principle (2.1) with a constant total stress we also have $d\sigma_z = \alpha dp$, hence:

$$\frac{\partial \varepsilon_v}{\partial t} = \frac{\partial}{\partial t} (C_m \sigma_z) = \alpha C_m \frac{\partial p}{\partial t} \quad (2.40)$$

and equation (2.36) reads:

$$\nabla \cdot \left[K \left(\frac{\nabla p}{\gamma_w} + \nabla z \right) \right] = S_h \frac{\partial p}{\partial t} \quad (2.41)$$

where $S_h = n\beta + \alpha C_m$ is the elastic storage coefficient.

In general, however, a theoretical uncoupling is not possible. The solution of a fully coupled model is complex and, from a numerical point of view, quite ill-conditioned [[Ferronato et al., 2001, 2009](#)]. For the applications considered in this work, we assume a one-way coupled approach, i.e. the flow problem is solved first and then used in the structural one. The experience shows that the solution obtained in this way does not differ significantly from the one obtained with a fully coupled model, especially at large time and space scales [[Gambolati et al., 2000](#)].

[Segura et al. \[2011\]](#) demonstrated that in reservoir simulations the one-way coupled approach provides accurate results (compared with a fully-coupled approach) when the changes of the pore volume (i.e. porosity and rock compressibility), due to the

stress path within the reservoir, are properly accounted for by the fluid flow simulator. Consequently, in the geomechanical model only the external strength changes with time according to the time stepping of the flow model.

2.2.3 Numerical model

The equilibrium equations (2.7) or (2.9) are solved in the framework of the Finite Elements Method (FEM) [Zienkiewicz and Taylor, 2000].

The displacement $\bar{\mathbf{u}}(\mathbf{x})$ is approximated by the function $\mathbf{u}^h(\mathbf{x}) \in \mathcal{U}^h$, where $\mathcal{U}^h$ is the finite Hilbert function space generated by the piecewise polynomials ℓ_i , $i = 1, \dots, n_\Omega$, with n_Ω the number of FE nodes in Ω :

$$\mathbf{u}^h(\mathbf{x}) = \begin{bmatrix} u_x^h(\mathbf{x}) \\ u_y^h(\mathbf{x}) \\ u_z^h(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n_\Omega} \ell_i(\mathbf{x}) u_{x,i} \\ \sum_{i=1}^{n_\Omega} \ell_i(\mathbf{x}) u_{y,i} \\ \sum_{i=1}^{n_\Omega} \ell_i(\mathbf{x}) u_{z,i} \end{bmatrix} = N_u(\mathbf{x}) \mathbf{u} \quad (2.42)$$

where $\mathbf{u} \in \mathbb{R}^{3n_\Omega}$ is the array with the nodal values of displacements. From (2.42), we can compute the approximate strains, introducing the deformation matrix $B = LN_u$:

$$\boldsymbol{\varepsilon}^h = L\mathbf{u}^h = LN_u\mathbf{u} = B\mathbf{u} \quad (2.43)$$

The approximate effective stress $\boldsymbol{\sigma}^h$ depends on the strain according to the selected constitutive model, as already seen in (2.10) for the continuous case:

$$d\boldsymbol{\sigma}^h = D_t(\boldsymbol{\sigma}^h) d\boldsymbol{\varepsilon}^h \quad (2.44)$$

where now the tangent constitutive matrix $D_t(\boldsymbol{\sigma}^h)$ depends on the approximate effective stresses $\boldsymbol{\sigma}^h$.

Introducing all approximations above in (2.7), we obtain the numerical model:

$$\int_{\Omega} \delta \mathbf{u}^T N_u^T \mathbf{q} dV + \int_{\partial\Omega} \delta \mathbf{u}^T N_u^T \mathbf{t} dS = \int_{\Omega} \delta \mathbf{u}^T B^T \boldsymbol{\sigma}^h dV \quad (2.45)$$

As equation (2.45) must hold true for any virtual displacement compatible with the constraints, the final governing equation reads:

$$\int_{\Omega} N_u^T \mathbf{q} dV + \int_{\partial\Omega} N_u^T \mathbf{t} dS = \int_{\Omega} B^T \boldsymbol{\sigma}^h dV \quad (2.46)$$

The problem (2.46) is well-posed after the prescription of the boundary conditions, which are of Dirichlet kind on the displacements:

$$\mathbf{u} = \mathbf{u}_0 \quad \text{on} \quad \partial\Omega_D \quad (2.47)$$

and of Neumann kind on the stresses:

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_0 \quad \text{on} \quad \partial\Omega_N \quad (2.48)$$

Being (2.6) a system of elliptic differential equations, the boundary $\partial\Omega$ must be the union of two disjoint subsets $\partial\Omega_D$ and $\partial\Omega_N$, with Dirichlet boundary condition imposed on $\partial\Omega_D$ and an oblique derivative boundary condition, i.e. Neumann, imposed on $\partial\Omega_N$ [Taylor, 2011; Jovanović and Süli, 2013]. From a numerical point of view, the former condition is strongly imposed while the latter is weakly imposed.

2.3 Faults and fractures mechanics

Simulating the mechanics of faults and ground fissures is of paramount interest in a large number of engineering applications. In the field of physical modeling, they can be regarded as discontinuities within the field parameters of the porous medium. From a mathematical point of view, they are described by an internal boundary whose existence depends on the displacements and stresses acting on the continuum. On their turn, displacements and stresses depend on the domain and its boundaries, thus yielding a strongly non-linear problem where the domain boundary definition is itself part of the solution.

In literature, two methods are mainly used to deal with discontinuities in a continuous domain.

The first method, referred to as “penalty approach” [Zienkiewicz and Taylor, 2000; Bathe, 2006], consists of introducing very stiff springs between the faces of the rupture, in order to impose the condition of non-penetration. When the stress satisfies a certain failure criterion, they break and set the contact surfaces free to move. From a mathematical point of view, this method is not exact, because the elastic springs deform for any non-zero stress value. Furthermore, from a numerical point of view, this method may cause a severe ill-conditioning of the stiffness matrix [Ferronato et al., 2012], because of the introduction of the penalty coefficients. However, this approach is widely used, mainly for the ease of implementation. Some examples of application of the penalty approach, even in fields that differ from geomechanics, can be found in Khoei and Lewis [1999]; Onate and Rojek [2004]; Rutqvist and Tsang [2005]; Rutqvist et al. [2007]; Ferronato et al. [2008a]; Rinaldi and Rutqvist [2013]; Benkhira et al. [2016]; Burman and Ern [2016]; Sabetamal et al. [2016].

Alternatively, the constraint conditions can be imposed by using Lagrange multipliers [Zienkiewicz and Taylor, 2000; Bathe, 2006; Wriggers, 2006; Simo and Hughes, 2006], namely in an analytically exact way. The Lagrange multipliers, which physically represent the contact stresses, are additional unknowns, so that the problem is numerically enlarged. Moreover, the discrete system changes its nature and typically becomes indefinite with a 2×2 block structure. However, the increase of the computational cost is generally offset by a more robust convergence in the non-linear steps and a more stable numerical behavior.

Other methods, based on the previously discussed algorithms, have been also introduced. For example, Simo and Laursen [1992] and Zienkiewicz and Taylor [2000] present the augmented Lagrangian method, as a compromise between the two techniques. Nitsche’s method [Hansbo, 2005] is another intermediate approach between penalty and Lagrange techniques. Improvements and applications of Nitsche’s method can be found in Annavarapu et al. [2013, 2014].

2.3.1 Lagrange multiplier approach

In this work, we focus our attention on the Lagrange multiplier method. Conceptually, if we are minimizing the objective function $F(\mathbf{x})$ subject to the constraint $G(\mathbf{x}) = 0$,

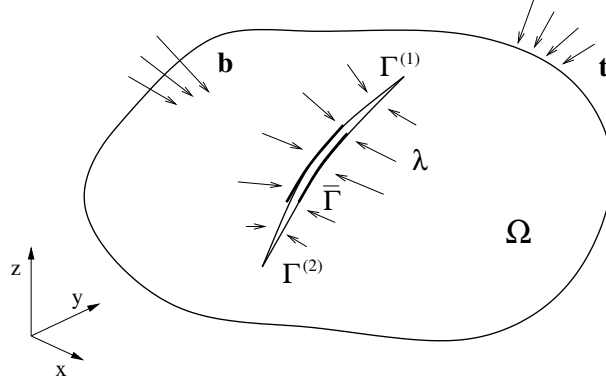


FIGURE 2.1: Conceptual scheme for the fault modelling.

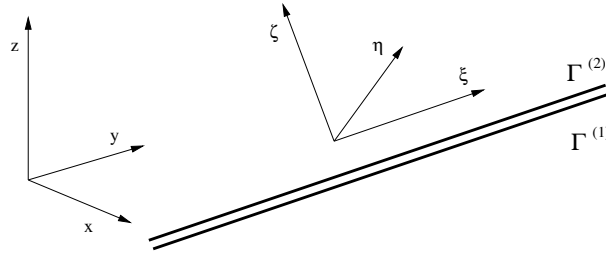


FIGURE 2.2: Local reference frame on the fault surfaces.

where $F : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\mathbf{G} : \mathbb{R}^m \rightarrow \mathbb{R}^m$, we can equivalently look for the free minimization of the function:

$$L(\mathbf{x}, \boldsymbol{\lambda}) = F(\mathbf{x}) - \boldsymbol{\lambda} \cdot \mathbf{G}(\mathbf{x}) \quad (2.49)$$

where $L : \mathbb{R}^{n+m} \rightarrow \mathbb{R}$ [Bertsekas, 1982]. In this sense, Lagrange multipliers are additional unknowns.

In our case, the functions F and G are the total potential energy of the system and the fault constraints, respectively.

Let us consider Fig. 2.1 as a reference, where $\Gamma^{(1)}$ and $\Gamma^{(2)}$ are the surfaces defining the inner discontinuity Γ . For every point $\mathbf{x}$ of Γ , we denote by σ_n and $\|\boldsymbol{\tau}\|_2$ the normal and tangential stress components. With $\mathbf{u}_r$ we indicate the relative displacements of two opposite points located on $\Gamma^{(1)}$ and $\Gamma^{(2)}$. With respect to the tangent plane, the normal direction is ζ , conventionally oriented from $\Gamma^{(1)}$ to $\Gamma^{(2)}$, while the two tangential directions ξ and η form an anti-clockwise tern (Fig. 2.2). At every point $\mathbf{x}$ we can define 3 Lagrange multipliers collected in the vector $\hat{\boldsymbol{\lambda}}$, which in a local reference frame is $\hat{\boldsymbol{\lambda}} = [\hat{\lambda}_\zeta, \hat{\boldsymbol{\lambda}}_s]^T = [\hat{\lambda}_\zeta, \hat{\lambda}_\xi, \hat{\lambda}_\eta]^T$, with $\sigma_n = \hat{\lambda}_\zeta$ and $\boldsymbol{\tau} = \hat{\boldsymbol{\lambda}}_s$.

The contact is subject to specific restrictions that are described by the governing

constitutive law, such as the non-compenetrability of solid bodies. Such a law can be very problem dependent, for a general overview see for instance [Wriggers and Zavarise \[2004\]](#); [Wriggers \[2006\]](#). We choose the Mohr - Coulomb failure criterion to describe the friction behavior on the fault plane [[Labuz and Zang, 2012](#)]. According to this criterion, the limit value for the modulus of the tangential stress is:

$$\tau_L = \begin{cases} c + \mu |\sigma_n| & \text{if } \sigma_n < 0 \\ 0 & \text{if } \sigma_n \geq 0 \end{cases} \quad (2.50)$$

where c is the cohesion and $\mu = \tan \varphi$ is the friction coefficient, with φ the friction angle.

In our case, the constraints on stresses imposed by the fault are:

1. continuity of displacements whenever the failure criterion is not violated:

$$\sigma_n \leq 0 \wedge \|\boldsymbol{\tau}\|_2 \leq \tau_L \longrightarrow \begin{cases} u_\zeta^{(1)} = u_\zeta^{(2)} \\ u_\xi^{(1)} = u_\xi^{(2)} \\ u_\eta^{(1)} = u_\eta^{(2)} \end{cases} \quad (2.51)$$

2. continuity of normal displacements and free slipping if the fault is compressed but the tangential stress exceeds the limiting value τ_L :

$$\sigma_n \leq 0 \wedge \|\boldsymbol{\tau}\|_2 > \tau_L \longrightarrow \begin{cases} u_\zeta^{(1)} = u_\zeta^{(2)} \\ \|\boldsymbol{\tau}\|_2 = \tau_L \end{cases} \quad (2.52)$$

3. free surface motion if the fault is subject to traction:

$$\sigma_n > 0 \longrightarrow \text{points are completely free} \quad (2.53)$$

where u_ζ , u_ξ and u_η are the components of $\bar{\mathbf{u}}$ at $\mathbf{x}$ along the local frame $\zeta - \xi - \eta$ (see [Fig. 2.2](#)). The only constraint that must hold true on the displacement is $\mathbf{u}_{r,n} \geq 0$, i.e. there is no compenetrability between opposite surfaces. This is the classical non-compenetrability condition for solid bodies.

In Fig. 2.1, Γ is the whole fault surface, while $\bar{\Gamma}$ is the activated part, i.e., where sliding and/or opening is recorded. The virtual work principle, as formulated in (2.9), has to be modified to account for the contribution given by the fault stresses and displacements, which are both unknown. The fault virtual work δW_f consists of two contributions:

- the term due to the virtual displacements and stresses along the portion of $\Gamma \setminus \bar{\Gamma}$, where the continuity of the solution is prescribed;
- the term due to the fault motion along the activated part $\bar{\Gamma}$.

As in case of opening, the normal and shear stresses $\sigma_n = \hat{\lambda}_n$ and $\tau = \hat{\lambda}_s$ are null, a non-zero contribution to the virtual work is provided only by the shear stress $\hat{\lambda}_s$ along $\bar{\Gamma}$. With the contribution δW_f , the virtual work principle (2.9) becomes:

$$\begin{aligned} \delta W_i + \delta W_f - \delta W_e = 0 \Rightarrow \\ \int_{\Omega} \delta \bar{\mathbf{u}}^T L^T \boldsymbol{\sigma}(\boldsymbol{\varepsilon}) d\Omega + \int_{\Gamma \setminus \bar{\Gamma}} \delta \mathbf{u}_r^T \hat{\boldsymbol{\lambda}} dS + \int_{\Gamma \setminus \bar{\Gamma}} \delta \hat{\boldsymbol{\lambda}}^T \mathbf{u}_r dS + \int_{\bar{\Gamma}} \delta \mathbf{u}_r^T \boldsymbol{\tau}_L dS + \\ - \int_{\Omega} \delta \bar{\mathbf{u}}^T \mathbf{q} d\Omega - \int_{\partial\Omega} \delta \bar{\mathbf{u}}^T \mathbf{t} d\partial\Omega = 0 \end{aligned} \quad (2.54)$$

where $\boldsymbol{\tau}_L$ is the stress limiting vector on the fault plane.

The Mohr-Coulomb failure criterion (2.50) provides the modulus τ_L of the limiting shear strength for the slippage inception, but gives no indication as to the direction of the limiting shear vector $\boldsymbol{\tau}_L$.

An idea could be to keep the direction of the projection on the tangent plane of the previously known Lagrange multiplier $\hat{\lambda}_s$. This idea is presented and developed in Jha and Juanes [2014], where the vector $\boldsymbol{\tau}_L$ is computed as:

$$\boldsymbol{\tau}_L = (1 - \delta_i) \hat{\lambda}_s \left(\frac{\|\boldsymbol{\tau}\| - \tau_L}{\tau_L} \right) \quad (2.55)$$

where δ_i is Kronecker's delta function ensuring that only the shear components of the Lagrange multiplier are perturbed. The main assumption of this technique, shared with other works, e.g. Aagaard et al. [2013], is that a local perturbation of the displacement is able to balance the part of stress exceeding its limiting value, according

to the failure criterion (2.50). Aagaard et al. [2013] consider only the DOF associated to the fault interface when computing the slip due to a perturbation of the Lagrange multipliers. Sometimes, the fault slip extends across the entire domain and other algorithms must be employed. In Aagaard et al. [2013], for example, a bilinear search in logarithmic space is used to minimize a suitable objective function.

A different approach, with no “locality” assumptions, is based on the *Principle of Maximum Plastic Dissipation* [Simo and Hughes, 2006; Wriggers, 2006] which is a general law governing any dissipative deformation process. In this problem, such a principle can be formulated as follows. Consider the set $\mathcal{T} \subset \mathbb{R}^2$ of vectors $\boldsymbol{\tau}(\mathbf{x})$ lying in the fault plane and such that $\|\boldsymbol{\tau}(\mathbf{x})\|_2 = \tau_L$. When a relative displacement $\mathbf{u}_r$ between the fault surfaces occurs, the shear vector $\boldsymbol{\tau}_L \in \mathcal{T}$, acting at the point $\mathbf{x}$, is the one that maximizes the work W_f of the friction stress, i.e., the dissipated plastic energy. As W_f reads:

$$W_f(\mathbf{x}) = [\boldsymbol{\tau}(\mathbf{x})]^T \mathbf{u}_r(\mathbf{x}), \quad \forall \boldsymbol{\tau}(\mathbf{x}) \in \mathcal{T} \quad (2.56)$$

the vector $\boldsymbol{\tau}_L$ satisfying the principle must have the same direction as $\mathbf{u}_r$, i.e.:

$$\boldsymbol{\tau}_L(\mathbf{x}) = \tau_L \frac{\mathbf{u}_r}{\|\mathbf{u}_r\|_2} \quad (2.57)$$

2.3.2 Numerical model

There are different ways to treat numerically a frictional discontinuity surface that can develop unrecoverable plastic deformations with energy dissipation. Basically, they can be classified in two big groups:

- approximating a fault by a 3D thin volume with properties different from the surrounding. At a numerical level, in the context of the Finite Element Method, this approach implies a finer discretization of the neighborhood of the fault to simulate a layer of different material. See, among others, Rutqvist et al. [2008]; Cappa and Rutqvist [2011b,a];
- introduction of a discontinuity surface, which behaves as an internal boundary and can be active or not depending on the stress state.

The first approach has several limitations. The most significant concern the difficulty in representing in a reliable way the actual constitutive behavior of the fault and modeling accurately the possible slip of the discontinuity surface. The latter aspect turns out to be particularly important, for instance, when the model is used to predict the expected magnitude of the earthquake triggered by the fault activation. In fact, seismological studies suggest the use of simple mathematical relationships for correlating the maximum magnitude of a seismic event to the sliding fault area and the average slippage [Kanamori and Anderson, 1975; Hanks and Kanamori, 1979; Mazzoldi et al., 2012]. For the straightforward application of such empirical relationships the accurate knowledge of the fault slippage is of paramount importance, as it will be shown later.

The other numerical approach, used in this work, makes the identification of the localized discontinuous displacements, possibly arising along the fault, much more straightforward. This technique is based on the introduction of interface frictional elements. Several formulations have been proposed starting from the original work by Goodman et al. [1968], e.g., Beer [1985]; Cescotto and Charlier [1993]; Juanes et al. [2002]; Ferronato et al. [2008a]; Aagaard et al. [2013]; Jha and Juanes [2014].

Within this framework and similarly to the mathematical formulation developed for the continuous domain, the fault stress, i.e. Lagrange multipliers $\hat{\lambda}(\mathbf{x})$, is approximated by $\lambda^h(\mathbf{x}) \in \mathcal{L}^h$, with $\mathcal{L}^h$ a finite Hilbert function space generally different from $\mathcal{U}^h$. $\mathcal{L}^h$ is generated by the set of piecewise polynomials ϕ_j , $j = 1, \dots, n_\Gamma$, with n_Γ the number of FE nodes in $\Gamma^{(1)}$ and $\Gamma^{(2)}$ (see Fig. 2.1):

$$\lambda^h(\mathbf{x}) = \begin{bmatrix} \lambda_x^h(\mathbf{x}) \\ \lambda_y^h(\mathbf{x}) \\ \lambda_z^h(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n_\Gamma} \phi_j(\mathbf{x}) \lambda_{x,j} \\ \sum_{j=1}^{n_\Gamma} \phi_j(\mathbf{x}) \lambda_{y,j} \\ \sum_{j=1}^{n_\Gamma} \phi_j(\mathbf{x}) \lambda_{z,j} \end{bmatrix} = N_\lambda(\mathbf{x}) \lambda_g \quad (2.58)$$

As already mentioned, the fault stresses are more conveniently expressed in a local reference frame. For each element that discretizes the surfaces $\Gamma^{(1)}$ and $\Gamma^{(2)}$, the local reference frame $\xi - \eta - \zeta$ (see Fig. 2.2) is defined such that the ζ axis identifies the normal direction to the contact surfaces, conventionally oriented from $\Gamma^{(1)}$ to $\Gamma^{(2)}$, and the transversal directions ξ and η are defined so as to constitute an anti-clockwise tern with ζ . Introducing the rotation matrix R that transforms the local nodal stresses λ into

the global vector λ_g we have:

$$\lambda^h(\mathbf{x}) = N_\lambda(\mathbf{x}) R \lambda \quad (2.59)$$

The relative displacement $\mathbf{u}_r$ appearing in equation (2.54) is conventionally defined as $\mathbf{u}_r = \mathbf{u}^{(2)} - \mathbf{u}^{(1)}$. Working in finite Hilbert spaces, the approximate relative displacement $\mathbf{u}_r^h$ becomes:

$$\mathbf{u}_r^h = \mathbf{u}^{h,(2)} - \mathbf{u}^{h,(1)} \quad (2.60)$$

where $\mathbf{u}^{h,(1)}$ and $\mathbf{u}^{h,(2)}$ are the restriction of $\mathbf{u}^h$ over $\Gamma^{(1)}$ and $\Gamma^{(2)}$, respectively. Using matrices $S^{(1)}$ and $S^{(2)}$ that map the nodal displacements on the fault surfaces from the global nodal displacement vector $\mathbf{u}$, we obtain:

$$\mathbf{u}_r^h = N_u S^{(2)} \mathbf{u} - N_u S^{(1)} \mathbf{u} = N_u S \mathbf{u} \quad (2.61)$$

where $S = S^{(2)} - S^{(1)}$.

The discrete shear stress τ_L^h along $\bar{\Gamma}$ depends on τ_L and $\mathbf{u}_r^h$ according to equation (2.57), which was defined from the principle of maximum plastic dissipation. Using the discrete function λ^h defined in equation (2.58), τ_L^h reads:

$$\tau_L^h = c - \tan \varphi \mathbf{n}^T N_\lambda R \lambda \quad (2.62)$$

where $\mathbf{n}$ is the normal vector to Γ , conventionally oriented from $\Gamma^{(1)}$ and $\Gamma^{(2)}$. Introducing equations (2.61) and (2.62) into (2.57), with $\mathbf{u}_r^h$ that replaces $\mathbf{u}_r$, provides the expression of τ_L^h :

$$\tau_L^h = (c - \tan \varphi \mathbf{n}^T N_\lambda R \lambda) \frac{\mathbf{u}_r^h}{\|\mathbf{u}_r^h\|_2} = (c - \tan \varphi \mathbf{n}^T N_\lambda R \lambda) \frac{N_u S \mathbf{u}}{\sqrt{\mathbf{u}^T S^T N_u^T N_u S \mathbf{u}}} \quad (2.63)$$

Setting $H = S^T N_u^T N_u S$, the expression (2.63) simplifies into:

$$\tau_L^h = (c - \tan \varphi \mathbf{n}^T N_\lambda R \lambda) \frac{N_u S \mathbf{u}}{\|\mathbf{u}\|_H} \quad (2.64)$$

where $\|\mathbf{u}\|_H$ is the H -energy norm of $\mathbf{u}$:

$$\|\mathbf{u}\|_H = \sqrt{\mathbf{u}^T H \mathbf{u}} \quad (2.65)$$

Notice that H is positive semidefinite because S has not full rank, hence $\|\mathbf{u}\|_H$ is a semi-norm that can be zero even if $\mathbf{u} \neq 0$.

Thus, the contribution to the virtual work principle given by the introduction of a discontinuity, after the approximation in the finite Hilbert function space, becomes:

$$\begin{aligned} \delta W_f = & \int_{\Gamma \setminus \bar{\Gamma}} \delta \mathbf{u}^T S^T N_u^T N_\lambda R \boldsymbol{\lambda} dS + \int_{\Gamma \setminus \bar{\Gamma}} \delta \boldsymbol{\lambda}^T R^T N_\lambda^T N_u S \mathbf{u} dS + \\ & + \int_{\bar{\Gamma}} \delta \mathbf{u}^T (c - \tan \varphi \mathbf{n}^T N_\lambda R \boldsymbol{\lambda}) \frac{H \mathbf{u}}{\|\mathbf{u}\|_H} dS \end{aligned} \quad (2.66)$$

The numerical model (2.45) for a continuous porous medium is updated as follows:

$$\begin{aligned} \delta \mathbf{u}^T \int_{\Omega} N_u^T \mathbf{q} dV + \delta \mathbf{u}^T \int_{\partial \Omega} N_u^T \mathbf{t} dS = & \delta \mathbf{u}^T \int_{\Omega} B^T \boldsymbol{\sigma}^h dV + \\ & + \delta \mathbf{u}^T \int_{\Gamma \setminus \bar{\Gamma}} S^T N_u^T N_\lambda R \boldsymbol{\lambda} dS + \delta \boldsymbol{\lambda}^T \int_{\Gamma \setminus \bar{\Gamma}} R^T N_\lambda^T N_u S \mathbf{u} dS + \\ & + \delta \mathbf{u}^T \int_{\bar{\Gamma}} (c - \tan \varphi \mathbf{n}^T N_\lambda R \boldsymbol{\lambda}) \frac{H \mathbf{u}}{\|\mathbf{u}\|_H} dS \end{aligned} \quad (2.67)$$

The equation (2.67) must hold true for any virtual displacement $\delta \mathbf{u}$ and virtual stress $\delta \boldsymbol{\lambda}$. Setting:

$$C = \int_{\Gamma \setminus \bar{\Gamma}} S^T N_u^T N_\lambda R dS \quad (2.68)$$

$$\mathbf{f} = \int_{\Omega} N_u^T \mathbf{q} dV + \int_{\partial \Omega} N_u^T \mathbf{t} dS \quad (2.69)$$

equation (2.67) can be recast as a non-linear system in the unknowns $\mathbf{u}$ and $\boldsymbol{\lambda}$:

$$\begin{cases} \int_{\Omega} B^T \boldsymbol{\sigma}^h dV + C \boldsymbol{\lambda} + \delta \mathbf{u}^T \int_{\bar{\Gamma}} (c - \tan \varphi \mathbf{n}^T N_\lambda R \boldsymbol{\lambda}) \frac{H \mathbf{u}}{\|\mathbf{u}\|_H} dS - \mathbf{f} = 0 \\ C^T \mathbf{u} = 0 \end{cases} \quad (2.70)$$

i.e.:

$$\begin{cases} \mathbf{F}_1(\mathbf{u}, \boldsymbol{\lambda}) = 0 \\ \mathbf{F}_2(\mathbf{u}, \boldsymbol{\lambda}) = 0 \end{cases} \quad (2.71)$$

2.4 Solution algorithms

The numerical solution of the system of equations (2.71) is obtained by using a classical Newton-Raphson scheme. The Jacobian matrix $J(\mathbf{u}, \boldsymbol{\lambda})$ can be written in a 2×2 block form:

$$J = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix} \quad (2.72)$$

where:

$$J_{11} = \frac{\partial \mathbf{F}_1}{\partial \mathbf{u}} = \int_{\Omega} B^T D_t B dV + \int_{\Gamma} (c - \tan \varphi \mathbf{n}^T N_{\lambda} R \boldsymbol{\lambda}) \frac{\|\mathbf{u}\|_H^2 H - (H\mathbf{u})(H\mathbf{u})^T}{\|\mathbf{u}\|_H^3} dS = K(\mathbf{u}) + E(\mathbf{u}, \boldsymbol{\lambda}) \quad (2.73)$$

$$J_{12} = \frac{\partial \mathbf{F}_1}{\partial \boldsymbol{\lambda}} = C - \int_{\Gamma} \tan \varphi \frac{H\mathbf{u}}{\|\mathbf{u}\|_H} \mathbf{n}^T N_{\lambda} R dS = C - F(\mathbf{u}) \quad (2.74)$$

$$J_{21} = \frac{\partial \mathbf{F}_2}{\partial \mathbf{u}} = C^T \quad (2.75)$$

$$J_{22} = \frac{\partial \mathbf{F}_2}{\partial \boldsymbol{\lambda}} = 0 \quad (2.76)$$

The solution $(\mathbf{u}, \boldsymbol{\lambda})$ is obtained by starting from an initial guess $(\mathbf{u}^{(0)}, \boldsymbol{\lambda}^{(0)})$ with the following iterative scheme:

$$\begin{bmatrix} \mathbf{u}^{(k+1)} \\ \boldsymbol{\lambda}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}^{(k)} \\ \boldsymbol{\lambda}^{(k)} \end{bmatrix} + \begin{bmatrix} \Delta \mathbf{u} \\ \Delta \boldsymbol{\lambda} \end{bmatrix} \quad (2.77)$$

The corrections $\Delta \mathbf{u}$ and $\Delta \boldsymbol{\lambda}$ are computed by solving a linear system with the Jacobian matrix $J^{(k)} = J(\mathbf{u}^{(k)}, \boldsymbol{\lambda}^{(k)})$:

$$\begin{bmatrix} K(\mathbf{u}^{(k)}) + E(\mathbf{u}^{(k)}, \boldsymbol{\lambda}^{(k)}) & C - F(\mathbf{u}^{(k)}) \\ C^T & 0 \end{bmatrix} \begin{bmatrix} \Delta \mathbf{u} \\ \Delta \boldsymbol{\lambda} \end{bmatrix} = - \begin{bmatrix} \mathbf{F}_1(\mathbf{u}^{(k)}, \boldsymbol{\lambda}^{(k)}) \\ \mathbf{F}_2(\mathbf{u}^{(k)}, \boldsymbol{\lambda}^{(k)}) \end{bmatrix} \quad (2.78)$$

The Jacobian $J^{(k)}$ is a non-symmetric generalized saddle-point matrix [Benzi et al.,

2005], so the linearized step (2.78) has to be solved by a non-symmetric solver. A detailed discussion on the linear solver choice will be developed in Chapter 4.

The convergence of the Newton-Raphson procedure (2.77) is achieved when some vector norm falls below a prescribed tolerance. Possible choices are:

- the norm of the displacement increment, i.e., $r_u = \|\Delta \mathbf{u}\|$;
- the norm of the unbalanced forces, i.e., $r_f = \left\| \mathbf{F}_1 \left(\mathbf{u}^{(k)}, \boldsymbol{\lambda}^{(k)} \right) \right\|$;
- the deformation energy, i.e. $r_e = \left| \Delta \mathbf{u}^T \mathbf{F}_1 \left(\mathbf{u}^{(k)}, \boldsymbol{\lambda}^{(k)} \right) + \Delta \boldsymbol{\lambda}^T \mathbf{F}_2 \left(\mathbf{u}^{(k)}, \boldsymbol{\lambda}^{(k)} \right) \right|$.

The norms above are usually normalized by their initial value. Different values for the exit tolerance have to be set in order to guarantee the same solution accuracy when using different vector norms. The norm of the unbalanced forces r_f is the most demanding one, while the energy norm r_e is the weakest, thus an exit tolerance set to 10^{-10} for r_f may correspond to 10^{-20} or less for r_e .

2.4.1 Theoretical properties of the Jacobian matrix

The matrix $K(\mathbf{u})$ resulting by the application of the Finite Element method to the linear momentum balance is symmetric if the tangent constitutive matrix D_t is so. This is guaranteed in the framework of elasto-plasticity if the plastic flow rule is associative and the kinematic hardening is symmetric [Zienkiewicz and Taylor, 2000]. In this work, we use a symmetric tangent constitutive matrix D_t . Moreover, $K(\mathbf{u})$ is also positive definite, as most of the matrices resulting from FEM discretizations of poromechanics problems.

It is well known that the solution of a generalized saddle point system can be obtained in a more efficient and robust way if the leading block is positive definite, that is $K(\mathbf{u}) + E(\mathbf{u}, \boldsymbol{\lambda})$ in our case (see (2.73)). For a better understanding, see, e.g. Liesen [2006]; Bergamaschi [2012] for the definite case and Gould and Simoncini [2009] for the indefinite case. Here we prove the following result:

Proposition 2.4.1 *The matrix $E(\mathbf{u}, \boldsymbol{\lambda})$ is symmetric positive semidefinite.*

Proof. By definition [Saad, 2003], a matrix A is positive semidefinite, if:

$$\mathbf{v}^T A \mathbf{v} \geq 0, \quad \forall \mathbf{v} \in \mathbb{R}^n, \mathbf{v} \neq 0 \quad (2.79)$$

With matrix $E(\mathbf{u}, \boldsymbol{\lambda})$, (2.79) becomes:

$$\mathbf{v}^T \left[\int_{\bar{\Gamma}} \tau_L \frac{\|\mathbf{u}\|_H^2 H - (H\mathbf{u})(H\mathbf{u})^T}{\|\mathbf{u}\|_H^3} dS \right] \mathbf{v} \geq 0 \quad (2.80)$$

where $\tau_L = c - \tan \varphi \mathbf{n}^T N_\lambda R \boldsymbol{\lambda} > 0$ by definition. Being $\|\mathbf{u}\|_H^3 \geq 0$, it is sufficient to prove that:

$$\int_{\bar{\Gamma}} \mathbf{v}^T \left[\|\mathbf{u}\|_H^2 H - (H\mathbf{u})(H\mathbf{u})^T \right] \mathbf{v} dS \geq 0 \quad (2.81)$$

If the integrand is greater than or equal to zero, than (2.81) holds. We have:

$$(\mathbf{u}^T H \mathbf{u})(\mathbf{v}^T H \mathbf{v}) - (\mathbf{v}^T H \mathbf{u})(\mathbf{u}^T H \mathbf{v}) \geq 0 \quad (2.82)$$

By definition, $H = S^T N_u^T N_u S$ can be written as $H = LL^T$, where $L = S^T N_u^T$. Defining $\mathbf{u}_L = L^T \mathbf{u}$ and $\mathbf{v}_L = L^T \mathbf{v}$, we can simplify (2.82) into:

$$\begin{aligned} \langle \mathbf{u}, \mathbf{u} \rangle_H \langle \mathbf{v}, \mathbf{v} \rangle_H - \langle \mathbf{u}, \mathbf{v} \rangle_H^2 &\geq 0 \\ (\mathbf{u}_L^T \mathbf{u}_L)(\mathbf{v}_L^T \mathbf{v}_L) - (\mathbf{v}_L^T \mathbf{u}_L)^2 &\geq 0 \end{aligned} \quad (2.83)$$

The scalar product of $\mathbf{x}$ and $\mathbf{y}$ can be also written as:

$$\mathbf{x}^T \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos \theta \quad (2.84)$$

where θ is the angle between the two vectors. Thus, (2.83) becomes:

$$\left(\|\mathbf{u}_L\|_2^2 \right) \left(\|\mathbf{v}_L\|_2^2 \right) - (\|\mathbf{u}_L\|_2 \|\mathbf{v}_L\|_2 \cos \theta)^2 = \|\mathbf{u}_L\|_2^2 \|\mathbf{v}_L\|_2^2 (1 - \cos^2 \theta) \geq 0 \quad (2.85)$$

It is straightforward to show that $\mathbf{u}$ lies in the null-space of $E(\mathbf{u}, \boldsymbol{\lambda})$. Indeed:

$$\left[\int_{\bar{\Gamma}} \tau_L \frac{\|\mathbf{u}\|_H^2 H - (H\mathbf{u})(H\mathbf{u})^T}{\|\mathbf{u}\|_H^3} dS \right] \mathbf{u} = \int_{\bar{\Gamma}} \tau_L \frac{\|\mathbf{u}\|_H^2 H \mathbf{u} - (\mathbf{u}^T H \mathbf{u}) H \mathbf{u}}{\|\mathbf{u}\|_H^3} dS = 0 \quad (2.86)$$

This means that matrix $E(\mathbf{u}, \boldsymbol{\lambda})$ always has at least one null eigenvalue, and so it is positive semidefinite. $\square$

In the generalized saddle-point matrix J , the contributions $E(\mathbf{u}, \boldsymbol{\lambda})$ and $F(\mathbf{u}, \boldsymbol{\lambda})$ appear at the inception of the fault activation. It is therefore interesting to investigate the behavior of these two matrices when $\|\mathbf{u}_r\|_2 \rightarrow 0$ and $\|\mathbf{u}\|_H \rightarrow 0$. Setting $\mathbf{v} = H\mathbf{u}$, we have two cases:

I) $\|\mathbf{v}\|_2 \rightarrow 0$ but $\|\mathbf{u}\|_2 \neq 0$: this means that, with $\mathbf{v} = \alpha \hat{\mathbf{v}}$ and $\alpha \rightarrow 0$, we can write:

$$E(\mathbf{u}, \boldsymbol{\lambda}) = \int_{\bar{\Gamma}} \tau_L \frac{\|\mathbf{u}\|_H^2 H - (H\mathbf{u})(H\mathbf{u})^T}{\|\mathbf{u}\|_H^3} dS = \int_{\bar{\Gamma}} \tau_L \frac{\alpha \mathbf{u}^T \hat{\mathbf{v}} H - \alpha^2 \hat{\mathbf{v}} \hat{\mathbf{v}}^T}{\alpha^{3/2} (\mathbf{u}^T \hat{\mathbf{v}})^{3/2}} dS \quad (2.87)$$

Hence:

$$\lim_{\|\mathbf{u}\|_2 \rightarrow 0} \|E(\mathbf{u}, \boldsymbol{\lambda})\| = \lim_{\alpha \rightarrow 0} \left\| \int_{\bar{\Gamma}} \tau_L \frac{\alpha \mathbf{u}^T \hat{\mathbf{v}} H - \alpha^2 \hat{\mathbf{v}} \hat{\mathbf{v}}^T}{\alpha^{3/2} (\mathbf{u}^T \hat{\mathbf{v}})^{3/2}} dS \right\| = \infty \quad (2.88)$$

By distinction, for matrix $F(\mathbf{u}, \boldsymbol{\lambda})$ we have:

$$F(\mathbf{u}, \boldsymbol{\lambda}) = \int_{\bar{\Gamma}} \tan \varphi \frac{H\mathbf{u}}{\|\mathbf{u}\|_H} \mathbf{n}^T N_\lambda R dS = \int_{\bar{\Gamma}} \tan \varphi \frac{\alpha \hat{\mathbf{v}}}{\sqrt{\alpha} \sqrt{\mathbf{u}^T \hat{\mathbf{v}}}} \mathbf{n}^T N_\lambda R dS \quad (2.89)$$

Hence:

$$\lim_{\|\mathbf{u}\|_2 \rightarrow 0} \|F(\mathbf{u}, \boldsymbol{\lambda})\| = \lim_{\alpha \rightarrow 0} \left\| \int_{\bar{\Gamma}} \tan \varphi \frac{\alpha \hat{\mathbf{v}}}{\sqrt{\alpha} \sqrt{\mathbf{u}^T \hat{\mathbf{v}}}} \mathbf{n}^T N_\lambda R dS \right\| = 0 \quad (2.90)$$

So, F does not diverge and can be neglected for $\alpha \rightarrow 0$.

II) $\|\mathbf{v}\|_2 \rightarrow 0$ because $\|\mathbf{u}\|_2 \rightarrow 0$: as before, setting $\mathbf{u} = \alpha \hat{\mathbf{u}}$ with $\alpha \rightarrow 0$, we can write:

$$\begin{aligned} E(\mathbf{u}, \boldsymbol{\lambda}) &= \int_{\bar{\Gamma}} \tau_L \frac{\|\mathbf{u}\|_H^2 H - (H\mathbf{u})(H\mathbf{u})^T}{\|\mathbf{u}\|_H^3} dS = \\ &= \int_{\bar{\Gamma}} \tau_L \frac{\alpha^2 \|\hat{\mathbf{u}}\|_H^2 H - \alpha^2 (H\hat{\mathbf{u}})(H\hat{\mathbf{u}})^T}{\alpha^3 \|\hat{\mathbf{u}}\|_H^3} dS \end{aligned} \quad (2.91)$$

Again, any norm of matrix E tends to infinity. For $F(\mathbf{u}, \boldsymbol{\lambda})$, we have:

$$\begin{aligned} F(\mathbf{u}, \boldsymbol{\lambda}) &= \int_{\bar{\Gamma}} \tan \varphi \frac{H\mathbf{u}}{\|\mathbf{u}\|_H} \mathbf{n}^T N_\lambda R dS = \int_{\bar{\Gamma}} \tan \varphi \frac{\alpha H \hat{\mathbf{u}}}{\alpha \|\hat{\mathbf{u}}\|_H} \mathbf{n}^T N_\lambda R dS = \\ &= \int_{\bar{\Gamma}} \tan \varphi \frac{H \hat{\mathbf{u}}}{\|\hat{\mathbf{u}}\|_H} \mathbf{n}^T N_\lambda R dS \end{aligned} \quad (2.92)$$

Hence, matrix F does not diverge, but it cannot be neglected and its behavior depends on the direction of vector $\mathbf{u}$, i.e. $\hat{\mathbf{u}}$.

2.5 Implementation details

The main non-linearity of the proposed model stems from the activation of a fault portion. A major difficulty with this formulation is that $\bar{\Gamma}$ is itself unknown and has to be computed during the non-linear iterations. In the present approach, three operating modes are possible for a pair of corresponding nodes on $\Gamma^{(1)}$ and $\Gamma^{(2)}$.

1. Full opening: the nodes are not in contact and a free relative displacement is allowed. The Lagrange multipliers are known and equal to zero.
2. Full closure: the fault is compressed and the Mohr-Coulomb criterion (2.50) is satisfied. The Lagrange multipliers are unknown and such that no relative movement is allowed between two corresponding nodes.
3. Slipping: the fault is compressed, but the limiting shear strength is not able to prevent sliding. The Lagrange multiplier λ_ζ acting on the normal direction is unknown, while the components lying on the contact surface are computed using equation (2.57).

Hence, according to the operating mode, we have a different number of unknowns to compute. The algorithm summarizing the sequence of steps that control the fault activation is provided in Table 2.1. Lines in italic are comments. For each node pair j lying on Γ , the relative displacement $\mathbf{u}_{r,j}$ of $\Gamma^{(2)}$ with respect to $\Gamma^{(1)}$ is computed. If either the normal component of $\mathbf{u}_{r,j}$ or the normal stress $\lambda_{\zeta,j}$ obtained from the last linearized step is positive, then the contact surfaces open at the node pair j and the Lagrange multipliers are zero. At the following solution of system (2.78), the nodes of the pair j on $\Gamma^{(1)}$ and $\Gamma^{(2)}$ can move freely. Because of rounding errors the check on the opening condition is relaxed by using the tolerances ε_u and ε_n on the normal displacement and stress, respectively. If the opening check is not satisfied, the current tangential stress τ and the limiting value τ_L are computed. If τ does not exceed the limiting value τ_L , the node pair j is closed and the related contributions to C and C^T

TABLE 2.1: Algorithm for the control of the fault activation.

1.	FOR $j = 1, \dots, n_\Gamma$
2.	Compute the relative displacement $\mathbf{u}_{r,j}$
3.	IF $(\mathbf{n}_j^T \mathbf{u}_{r,j} > \varepsilon_u) \vee (\lambda_{\zeta,j} \geq \varepsilon_n)$ THEN
4.	Full opening mode: Lagrange multipliers are zero
5.	$\lambda_{\zeta,j} = \lambda_{\xi,j} = \lambda_{\eta,j} = 0$
6.	ELSE
7.	Full closure or Slipping mode: check the tangential stress
8.	$\tau = \sqrt{\lambda_{\xi,j}^2 + \lambda_{\eta,j}^2}$
9.	$\tau_L = c - \lambda_{\zeta,j} \tan \varphi$
10.	IF $(\tau < \tau_L + \varepsilon_t)$ THEN
12.	Full closure mode: Lagrange multipliers are unknown
13.	Compute the contribution of node j to C and C^T
14.	ELSE
15.	Slipping mode: only the Lagrange multiplier $\lambda_{\zeta,j}$ is unknown
16.	Compute the contribution of node j to C and C^T related to $\lambda_{\zeta,j}$
17.	IF $(\ \mathbf{u}_{r,j}\ _2 > \varepsilon_u)$ THEN
18.	Compute the contribution to E and $-F$
19.	END IF
20.	END IF
21.	END IF
22.	END FOR

are computed. The Lagrange multipliers at j are unknown and so as to ensure the displacement continuity through $\Gamma^{(1)}$ and $\Gamma^{(2)}$. Otherwise, if τ exceeds the limiting value τ_L the nodes of the pair j can slide. The stress components on the surface plane $\lambda_{\xi,j}$ and $\lambda_{\eta,j}$ are not considered as unknowns in the solution of system (2.78) and are computed according to equation (2.57). By distinction, $\lambda_{\zeta,j}$ is still unknown and its contribution to C and C^T must be accounted for. Again, because of the rounding errors it is advisable to relax the check on τ_L by introducing another tolerance ε_t . In the slipping mode, the local contributions to E and $-F$ have to be computed as well. As shown in equations (2.73) and (2.74), the matrices E and $-F$ depend on $\|\mathbf{u}\|_H$ that locally coincides with $\|\mathbf{u}_{r,j}\|_2$ (see (2.61) and (2.65)). If the current shear stress exceeds the limiting value τ_L but the sliding is close to zero, the coefficients of E and $-F$ can become numerically unstable. Actually, if the limiting shear stress is exceeded with no sliding there's no friction work on $\bar{\Gamma}$ and the contributions E and $-F$ should be zero. This is why these terms are computed only if $\|\mathbf{u}_{r,j}\|_2$ is "large enough", i.e., is numerically larger than the tolerance ε_u .

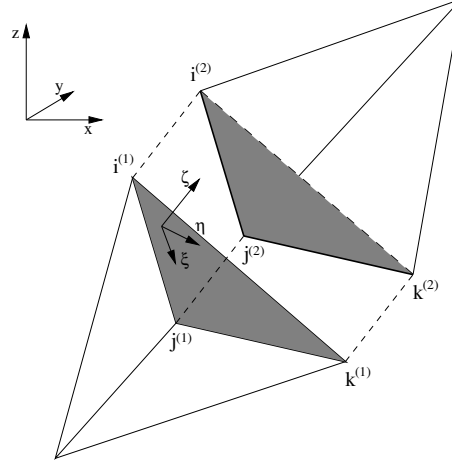


FIGURE 2.3: Interface element.

2.5.1 Local stiffness matrices for tetrahedral and triangular discretizations

The specific contributions arising from a 3D tetrahedral discretization of the porous body Ω are derived. The fault contact surfaces $\Gamma^{(1)}$ and $\Gamma^{(2)}$ are therefore discretized by triangular elements. For the sake of brevity, we denote as Interface Element (IE) a pair of opposite triangular elements located on the contact surfaces (see Fig. 2.3). The global contributions C , E and F are computed in the classical FE framework by assembling the local contributions $C^{(e)}$, $E^{(e)}$ and $F^{(e)}$ over each IE.

With this choice, we use piecewise linear interpolation functions for the displacements. As to the Lagrange multipliers, we use piecewise constant basis functions to be consistent with the stress computed in the finite elements that discretize the continuous porous body. Quite obviously, other choices are possible, but more stable results are typically obtained using for the Lagrange multipliers a lower-order approximation than for the displacements. This behaviour is similar to what happens in consolidation problems, where Galerkin finite element solutions often exhibit pressure oscillations, especially with equal orders of interpolation for displacement and pressure [Murad and Loula, 1994; Wan, 2002].

The local unknowns defined at the elemental level are the relative displacement $\mathbf{u}_r^{(e)}$ and the contact stress $\boldsymbol{\lambda}^{(e)}$. Recalling the global definitions (2.60)-(2.61) and with

reference to the element in Fig. 2.3, $\mathbf{u}_r^{(e)}$ reads:

$$\mathbf{u}_r^{(e)} = \begin{bmatrix} \mathbf{u}_{r,i} \\ \mathbf{u}_{r,j} \\ \mathbf{u}_{r,k} \end{bmatrix} = \begin{bmatrix} -I_3 & I_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & -I_3 & I_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -I_3 & I_3 \end{bmatrix} \begin{bmatrix} \mathbf{u}_i^{(1)} \\ \mathbf{u}_i^{(2)} \\ \mathbf{u}_j^{(1)} \\ \mathbf{u}_j^{(2)} \\ \mathbf{u}_k^{(1)} \\ \mathbf{u}_k^{(2)} \end{bmatrix} = S^{(e)} \mathbf{u}^{(e)} \quad (2.93)$$

where $\mathbf{u}_{r,m}$ is the vector of the relative displacements along x , y and z between the opposite nodes $m^{(1)}$ and $m^{(2)}$, and I_n denotes the rank- n identity matrix. The matrix $N_u^{(e)}(\mathbf{x})$ of the local basis functions such that $\mathbf{u}_r^{h,(e)} = N_u^{(e)}(\mathbf{x}) \mathbf{u}_r^{(e)}$ is:

$$N_u^{(e)}(\mathbf{x}) = \begin{bmatrix} \psi_i^{(e)}(\mathbf{x}) I_3 & \psi_j^{(e)}(\mathbf{x}) I_3 & \psi_k^{(e)}(\mathbf{x}) I_3 \end{bmatrix} \quad (2.94)$$

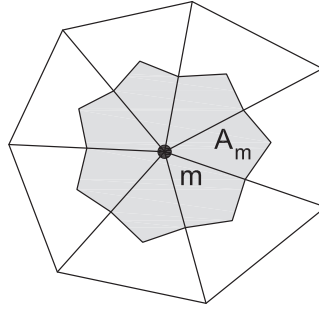
The function $\psi_m^{(e)}(\mathbf{x})$ is the restriction on the IE e of the basis functions $\ell_{m^{(1)}}(\mathbf{x})$ and $\ell_{m^{(2)}}(\mathbf{x})$ introduced in equation (2.42). Similarly, recalling the global definition (2.59) and with reference to Fig. 2.3, $\boldsymbol{\lambda}^{(e)}$ reads:

$$\boldsymbol{\lambda}^{(e)} = \begin{bmatrix} \lambda_i \\ \lambda_j \\ \lambda_k \end{bmatrix} \quad (2.95)$$

where $\boldsymbol{\lambda}_m$ is the vector of the nodal strengths along the local ζ , ξ and η directions acting between the nodes $m^{(1)}$ and $m^{(2)}$. Using the rotation matrix $R^{(e)}$, the matrix $N_\lambda^{(e)}(\mathbf{x})$ of the local basis functions such that $\boldsymbol{\lambda}^{h,(e)} = N_\lambda^{(e)}(\mathbf{x}) R^{(e)} \boldsymbol{\lambda}^{(e)}$ is:

$$N_\lambda^{(e)}(\mathbf{x}) = \begin{bmatrix} \phi_i^{(e)}(\mathbf{x}) I_3 & \phi_j^{(e)}(\mathbf{x}) I_3 & \phi_k^{(e)}(\mathbf{x}) I_3 \end{bmatrix} \quad (2.96)$$

with $\phi_m^{(e)}(\mathbf{x})$ representing the restriction on the IE e of the basis functions $\phi_{m^{(1)}}(\mathbf{x})$ and $\phi_{m^{(2)}}(\mathbf{x})$ of equation (2.58). In particular, $\phi_m(\mathbf{x})$ is the piecewise constant function

FIGURE 2.4: Area associated to node m .

defined as follows:

$$\phi_m(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A_m \\ 0 & \text{otherwise} \end{cases} \quad (2.97)$$

In equation (2.97) A_m is the area associated to the node m (see Fig. 2.4), i.e.:

$$A_m = \sum_e \frac{A^{(e)}}{3} \quad (2.98)$$

By using equations (2.93), (2.94) and (2.96), and recalling the global definition (2.68), the 18×9 local matrix $C^{(e)}$ can be easily obtained as:

$$C^{(e)} = \int_{A^{(e)}} S^{(e),T} N_u^{(e),T} N_\lambda^{(e)} R^{(e)} dS = S^{(e),T} \bar{H}^{(e)} R^{(e)} \quad (2.99)$$

where:

$$\bar{H}^{(e)} = \int_{A^{(e)}} N_u^{(e),T} N_\lambda^{(e)} dS = \frac{A^{(e)}}{108} \begin{bmatrix} 22I_3 & 7I_3 & 7I_3 \\ 7I_3 & 22I_3 & 7I_3 \\ 7I_3 & 7I_3 & 22I_3 \end{bmatrix} \quad (2.100)$$

The matrix $\bar{H}^{(e)}$ can be conveniently lumped, obtaining:

$$\tilde{H}^{(e)} = \frac{A^{(e)}}{3} I_9 \quad (2.101)$$

The use of $\tilde{H}^{(e)}$ instead of $\bar{H}^{(e)}$ in the computation of $C^{(e)}$ is equivalent to the assumption $N_u^{(e)} = N_\lambda^{(e)}$, i.e., using the same piecewise constant basis functions for both displacements and stresses on the IEs.

The local contribution $E^{(e)}$ can be derived from the global definition of $E(\mathbf{u}, \boldsymbol{\lambda})$ provided in equation (2.73). The first factor in the integral, i.e.:

$$\tau_L^{(e)} = \left(c - \tan \varphi \mathbf{n}^{(e),T} N_\lambda^{(e)} R^{(e)} \boldsymbol{\lambda}^{(e)} \right) \quad (2.102)$$

is actually the value of the limiting shear strength $\tau_L^{(e)}$ computed at the elemental level:

$$\tau_L^{(e)} = \phi_i^{(e)}(\mathbf{x}) \tau_{L,i} + \phi_j^{(e)}(\mathbf{x}) \tau_{L,j} + \phi_k^{(e)}(\mathbf{x}) \tau_{L,k} \quad (2.103)$$

where $\tau_{L,m}$ is the limiting shear strength value at the nodes $m^{(1)}$ and $m^{(2)}$. The H -energy norm of $\mathbf{u}^{(e)}$ reads:

$$\|\mathbf{u}^{(e)}\|_H = \sqrt{\mathbf{u}^{(e),T} S^{(e),T} N_u^{(e),T} N_u^{(e)} S^{(e)} \mathbf{u}^{(e)}} = \sqrt{\mathbf{u}_r^{h,(e),T} \mathbf{u}_r^{h,(e)}} = \|\mathbf{u}_r^{h,(e)}\|_2 \quad (2.104)$$

Using equations (2.103) and (2.104), $E^{(e)}$ is obtained as:

$$E^{(e)} = S^{(e),T} \left[\int_{A^{(e)}} \tau_L^{(e)} \frac{\|\mathbf{u}_r^{h,(e)}\|_2^2 N_u^{(e),T} N_u^{(e)} - N_u^{(e),T} \mathbf{u}_r^{h,(e)} \mathbf{u}_r^{h,(e),T} N_u^{(e)}}{\|\mathbf{u}_r^{h,(e)}\|_2^3} dS \right] S^{(e)} \quad (2.105)$$

The integral in equation (2.105) has to be evaluated numerically. However, the computation of $E^{(e)}$ becomes straightforward using the assumption that leads to the lumped matrix $\tilde{H}^{(e)}$ of equation (2.101), i.e., $N_u^{(e)} = N_\lambda^{(e)}$ on the IEs. In this case, equation (2.105) reads:

$$E^{(e)} = \frac{A^{(e)}}{3} S^{(e),T} \begin{bmatrix} \tau_{L,i} \frac{Q_i}{\|\mathbf{u}_{r,i}\|_2^3} & 0 & 0 \\ 0 & \tau_{L,j} \frac{Q_j}{\|\mathbf{u}_{r,j}\|_2^3} & 0 \\ 0 & 0 & \tau_{L,k} \frac{Q_k}{\|\mathbf{u}_{r,k}\|_2^3} \end{bmatrix} S^{(e)} \quad (2.106)$$

where Q_m is the 3×3 matrix:

$$Q_m = \|\mathbf{u}_{r,m}\|_2^2 I_3 - \mathbf{u}_{r,m} \mathbf{u}_{r,m}^T \quad (2.107)$$

According to (2.85), we know that matrix $E^{(e)}$ is positive semidefinite.

Finally, the contribution $F^{(e)}$ is computed by writing at the elemental level the global definition of $F(\mathbf{u})$ in equation (2.74). The matrix $N_{n,\lambda}^{(e)} = \mathbf{n}^{(e),T} N_\lambda^{(e)} R^{(e)}$ selects from the local vector $\lambda^{(e)}$ the components $\lambda_{\zeta,m}$ of the normal strength for each pair of nodes $m = i, j, k$ of the IE e . Hence, $N_{n,\lambda}^{(e)}$ is actually the first row of $N_\lambda^{(e)}$ in equation (2.96):

$$N_{n,\lambda}^{(e)}(\mathbf{x}) = \begin{bmatrix} \phi_i^{(e)}(\mathbf{x}) & 0 & 0 & \phi_j^{(e)}(\mathbf{x}) & 0 & 0 & \phi_k^{(e)}(\mathbf{x}) & 0 & 0 \end{bmatrix} \quad (2.108)$$

Recalling equation (2.104) and using the definition (2.108), $F^{(e)}$ can be computed as:

$$F^{(e)} = S^{(e),T} \int_{A^{(e)}} \tan \varphi \frac{N_u^{(e),T} N_u^{(e)} \mathbf{u}_r^{(e)}}{\|\mathbf{u}_r^{h,(e)}\|_2} N_{n,\lambda}^{(e)} dS \quad (2.109)$$

Similarly to $E^{(e)}$, the integral in equation (2.109) has to be evaluated numerically. However, the assumption $N_u^{(e)} = N_\lambda^{(e)}$ on the IEs allows for a straightforward computation of $F^{(e)}$. Approximating φ at the elemental level as:

$$\varphi^{(e)} = \phi_i^{(e)}(\mathbf{x}) \varphi_i + \phi_j^{(e)}(\mathbf{x}) \varphi_j + \phi_k^{(e)}(\mathbf{x}) \varphi_k \quad (2.110)$$

where φ_m is the friction angle at the pair of nodes $m^{(1)}$ and $m^{(1)}$, $F^{(e)}$ reads:

$$F^{(e)} = \frac{A^{(e)}}{3} S^{(e),T} \begin{bmatrix} \tan \varphi_i \frac{\mathbf{u}_{r,i}}{\|\mathbf{u}_{r,i}\|_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \tan \varphi_j \frac{\mathbf{u}_{r,j}}{\|\mathbf{u}_{r,j}\|_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \tan \varphi_k \frac{\mathbf{u}_{r,k}}{\|\mathbf{u}_{r,k}\|_2} & 0 & 0 \end{bmatrix} \quad (2.111)$$

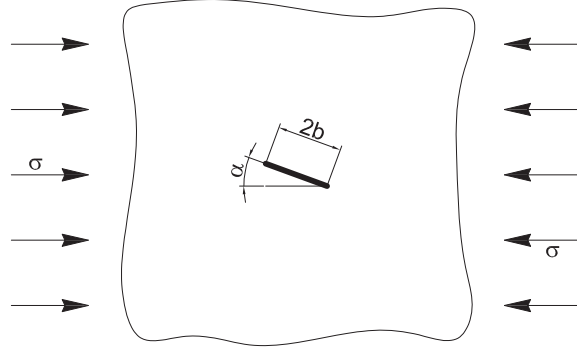


FIGURE 2.5: Test case with analytical solution.

2.6 Model validation

The proposed model is validated by using a test case that has an analytical solution. The problem, available in [Phan et al. \[2003\]](#), concerns a single crack under compression in plane stress conditions.

A single crack in an unbounded domain with plane stress conditions and subject to a compressive external remote stress σ is studied. The material is linear elastic with Young modulus $E = 70GPa$ and Poisson ratio $\nu = 0.2$. The crack length and inclination angle are $2b = 10m$ and $\alpha = 20^\circ$, respectively (Fig. 2.5). Assuming $\sigma = 200MPa$ with zero cohesion and friction angle $\varphi = 30^\circ$, the crack surfaces slide with the following analytical results for the normal stress σ_n and slip Δu_t [[Phan et al., 2003](#)]:

$$\sigma_n = -\sigma \sin^2 \alpha \quad (2.112)$$

$$\Delta u_t = \frac{4(1 - \nu^2) \sigma \sin \alpha (\cos \alpha - \sin \alpha \tan \varphi)}{E} \sqrt{b^2 - (\eta - b)^2} \quad (2.113)$$

where $0 \leq \eta \leq 2b$ is the linear length coordinate along the crack.

We want to reproduce the 2D plane stress condition of [Phan et al. \[2003\]](#) with our 3D model. To this aim, a cubic volume with side $L = 500m$, i.e., 50 times the crack size, is considered, with the boundary conditions shown in Fig. 2.6. The plane stress

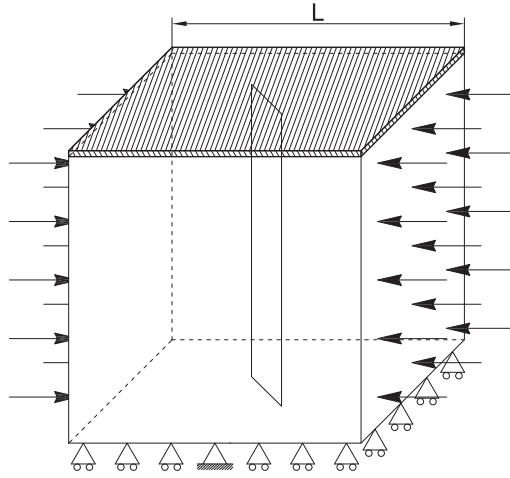


FIGURE 2.6: 3D model for the test case with analytical solution.

condition is obtained on the upper face (hatched in Fig. 2.6). The FE grid consists of 1,933,830 elements and 347,183 nodes, while the crack is discretized by 2,895 nodes. Fig. 2.7 provides a comparison between the numerical and analytical solutions along the crack. The results show quite a good agreement, especially on consideration of the approximate representation of the plane stress state condition with our full 3D model.

The numerical outcome differs from the analytical solution (2.112) and (2.113) by about 2% and less than 1% in terms of σ_n and Δu_t , respectively. Finally, notice that the larger differences for σ_n at the cracks ends are due to the grid size.

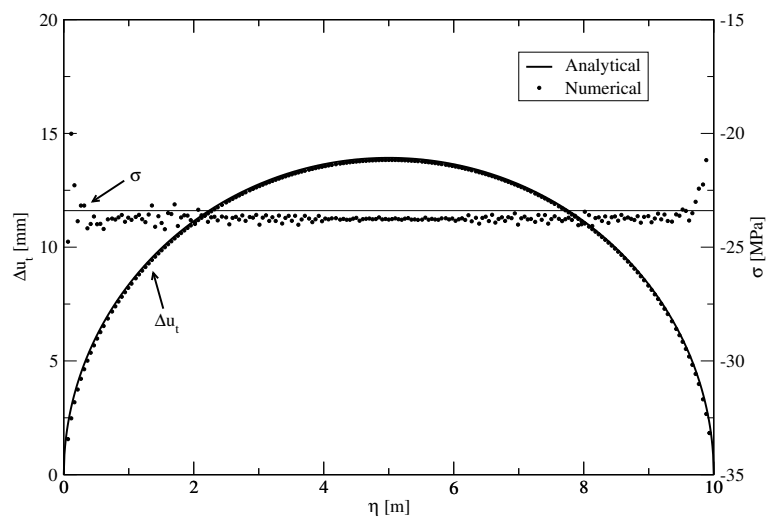


FIGURE 2.7: Analytical test case: numerical and analytical solutions along the crack.

Chapter 3

Numerical applications

Four test cases are presented to analyze the effectiveness and accuracy of the fault model.

1. Case study A: two adjacent blocks under compression and shear. This example is used to investigate the numerical convergence of the non-linear algorithm, together with its robustness and stability.
2. Case study B: ground fracture due to groundwater withdrawal in arid regions. This is a realistic engineering problem, used to validate the proposed model in a large-size application.
3. Case study C: ground fracture due to groundwater withdrawal. The driving mechanism is different from Case study B.
4. Case D: fault reactivation in an underground reservoir subjected to primary production and UGS (Underground Gas Storage) cycles. The prediction of fault motion is used to estimate the possible earthquake magnitude triggered by the UGS activities.

3.1 Case study A

The problem is shown in Fig. 3.1. A prismatic elastic body is divided in two blocks by a vertical crack. The prism has a rectangular basis $5 \times 10m$ wide and is 15-m high. Boundary conditions and loads are prescribed as shown in Fig. 3.1. The FE grid consists of 36,000 elements and 7,749 nodes, of which 588 are located on the crack. The linear elastic material is characterized by $E = 2GPa$ and $\nu = 0.25$. An initial compressive

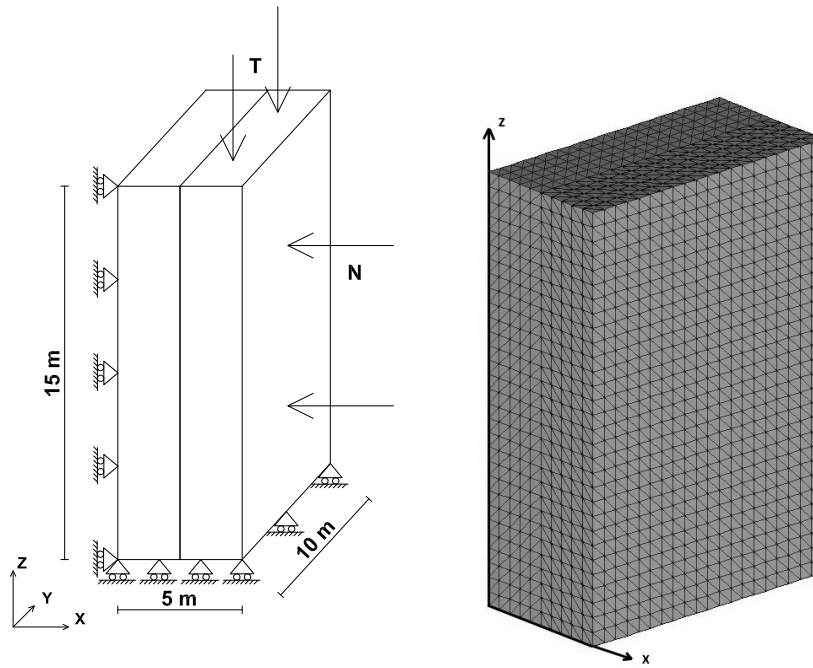


FIGURE 3.1: Case study A.

stress state $\sigma_0 = \sigma_n = 1kPa$ along the x -direction is assumed. The crack has zero cohesion and friction angle $\varphi = 30^\circ$.

The normal and shear loads, N and T , respectively, vary during the simulation as shown in Fig. 3.2. At first, the prism, hence the crack, is compressed along the x -direction by an increasing load up to $N = 750kN$, corresponding to a distributed pressure equal to $5kPa$. Then, the shear force T begins to act, up to $T = 450kN$, i.e., a uniform vertical pressure on the upper face of the rightmost block equal to $18kPa$ is applied. In the second part of the simulation, first N and then T are removed. In this way the crack is expected to start slipping as T increases, with the sliding portion later increasing more rapidly as N decreases. Finally, the reduction of T produces a rebound of the rightmost block. As the initial compressive state σ_0 is still acting, the induced shear stress turns to be smaller than the limiting value τ_L , for small values of T , and the crack closes again. As a consequence, the body does not return to the initial configuration, although at the end of the simulation the loads vanish.

The main numerical results are provided in Fig. 3.3 through Fig. 3.7. Fig. 3.3 shows the deformed configuration of the elastic body at some steps of interest, namely the load steps no. 8, 15 and 20, while Fig. 3.4 and 3.5 gives the normal and shear stress, σ_n

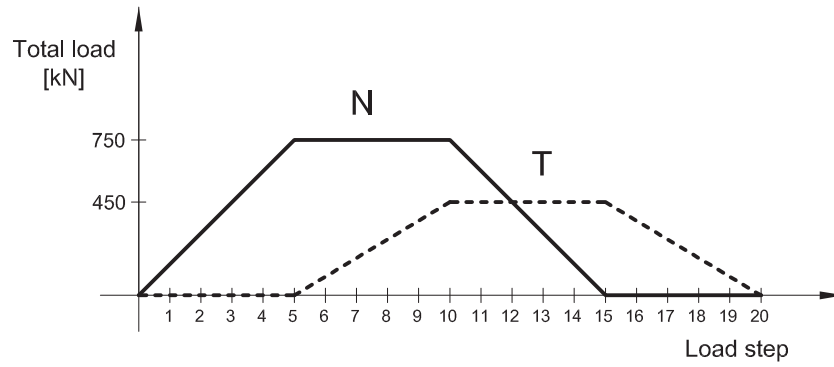


FIGURE 3.2: Case study A: load history.

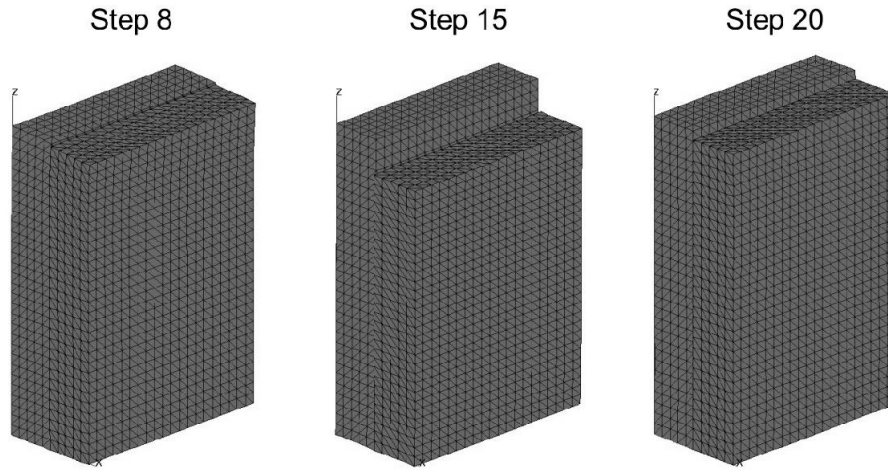


FIGURE 3.3: Case study A: deformed configuration for the load steps no. 8, 15 and 20. The displacement exaggeration factor is 15.

and τ_s , respectively, at the same instants. Finally, fig. 3.6 and 3.7 provide the behavior of σ_n , τ_s and the relative slip Δu_z in the vertical direction along the crack central axis and during the simulation at some points of interest at different elevation z .

At the step 8, the crack starts sliding in the upper part (see Fig. 3.6). Fig. 3.4 and 3.6 show also a σ_n reduction in this region, which is induced by the increase of T . This yields a decrease of the limiting stress τ_L , hence the inception of the sliding condition. At the step 15, the crack is fully sliding. The τ_s reduction is due to the absence of N that reduces σ_n , hence τ_L . In this situation, the rightmost volume of the elastic body is almost independent of the leftmost part. At the end of the simulation there are no external loads. The body does not return to the original configuration because of the model non-linearity, i.e., the energy dissipation generated by the fault friction.

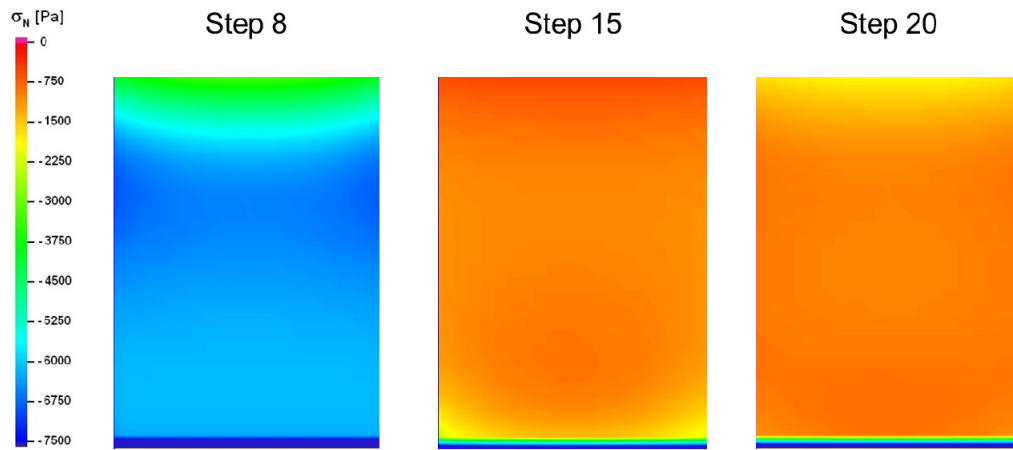


FIGURE 3.4: Case study A: normal stress on the crack surface for the load steps no. 8, 15 and 20.

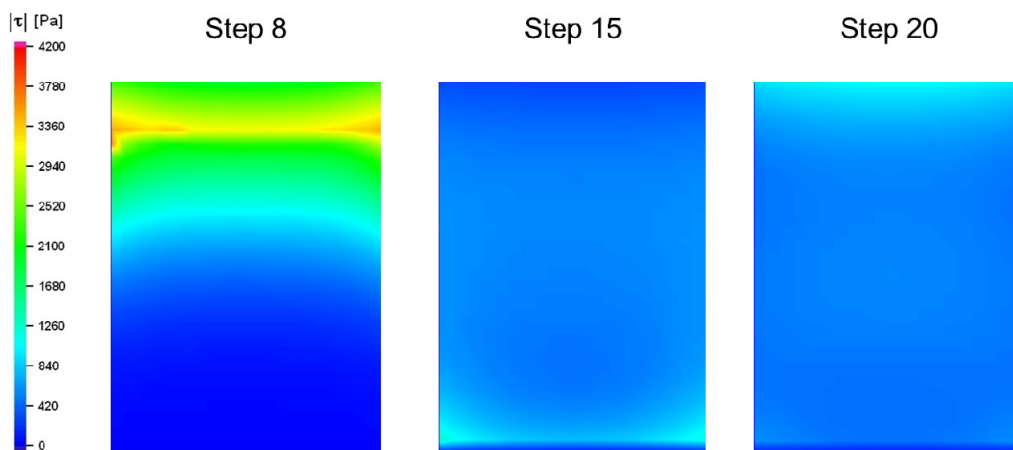


FIGURE 3.5: Case study A: shear stress on the crack surface for the load steps no. 8, 15 and 20.

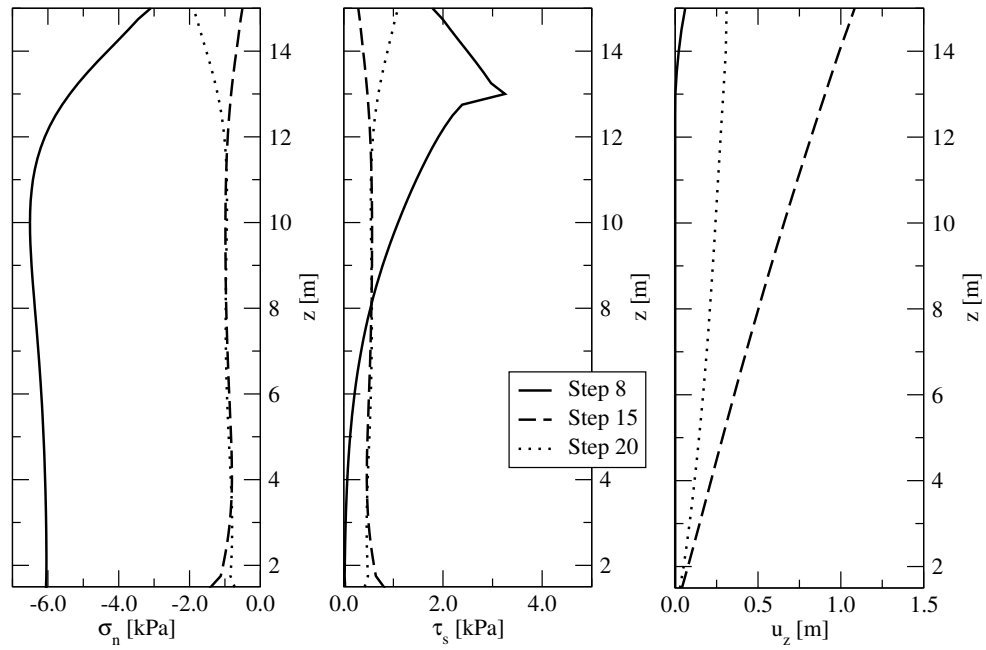


FIGURE 3.6: Case study A: σ_n , τ_s and Δu_z along the crack central axis for the load steps no. 8, 15, and 20.

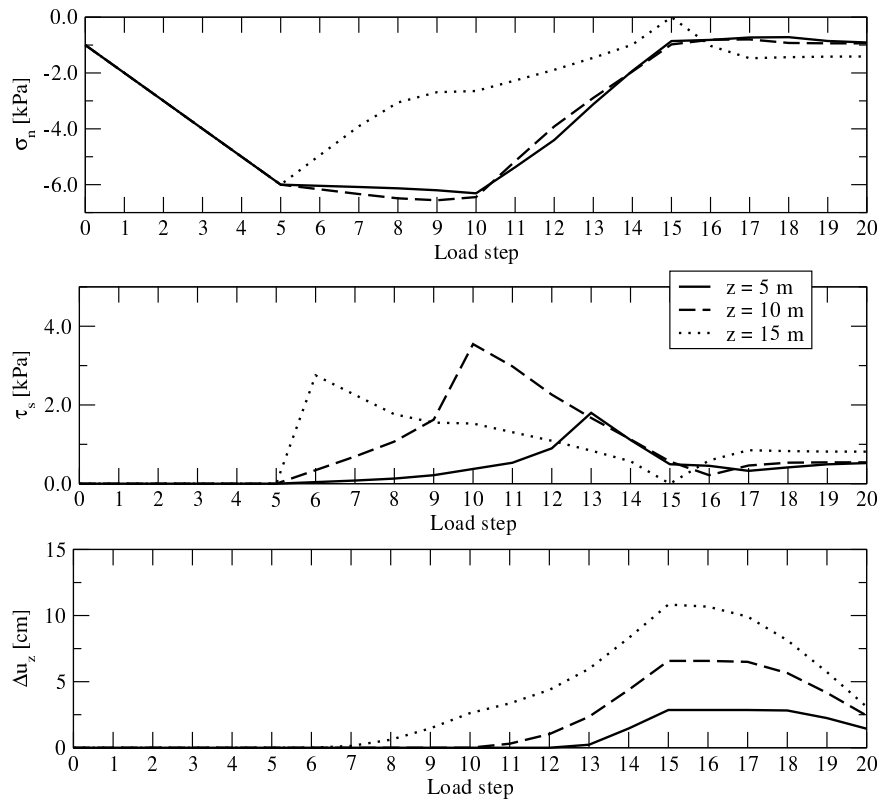


FIGURE 3.7: Case study A: σ_n , τ_s and Δu_z vs load step at different elevation z .

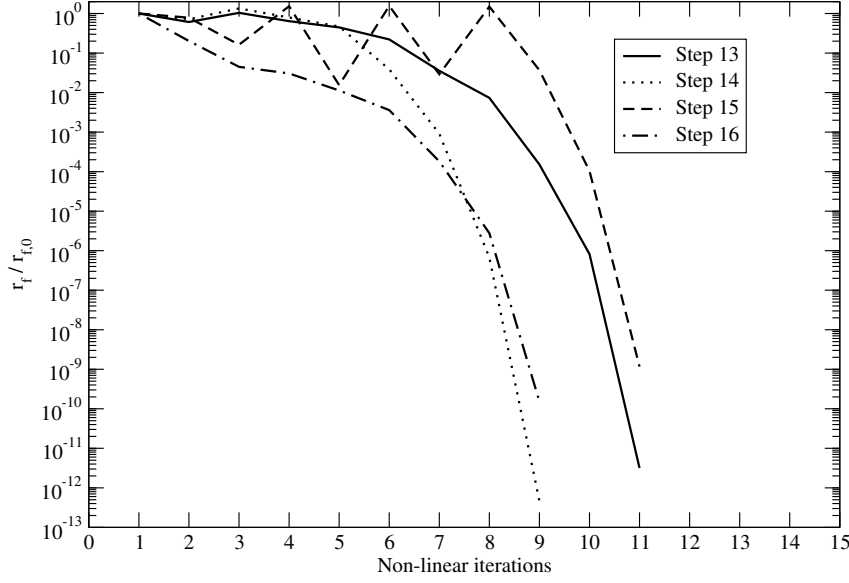


FIGURE 3.8: Case study A: convergence profile of the non-linear algorithm.

Fig. 3.8 shows four representative convergence profiles of the non-linear algorithm. The selected exit criterion is based on the relative 2-norm of the unbalanced forces, i.e.:

$$\frac{r_f}{r_{f,0}} = \frac{\left\| \mathbf{F}_1 \left(\mathbf{u}^{(k)}, \boldsymbol{\lambda}^{(k)} \right) \right\|_2}{\left\| \mathbf{F}_1 \left(\mathbf{u}^{(0)}, \boldsymbol{\lambda}^{(0)} \right) \right\|_2} \quad (3.1)$$

From a numerical point of view, the steps from 13 to 16 are the most challenging ones because the crack is fully sliding. Nevertheless, the non-linear algorithm exhibits a quadratic convergence at any step.

The novelty of the proposed formulation for the fault simulation relies on the contribution arising from the principle of maximum plastic dissipation, i.e., the integral over $\bar{\Gamma}$ in equation (2.70) and the related matrices E and $-F$ in the computation of the Jacobian. The numerical importance of such a contribution is emphasized in Fig. 3.9, where the convergence profiles of the non-linear algorithm at the inception of sliding (step no. 8) are shown for two different sizes of the load step. When a portion of the crack starts sliding, $\bar{\Gamma} \neq \emptyset$ and neglecting the corresponding integral yields a stagnation of the unbalanced forces, possibly far from the equilibrated configuration, independently of the load step size.

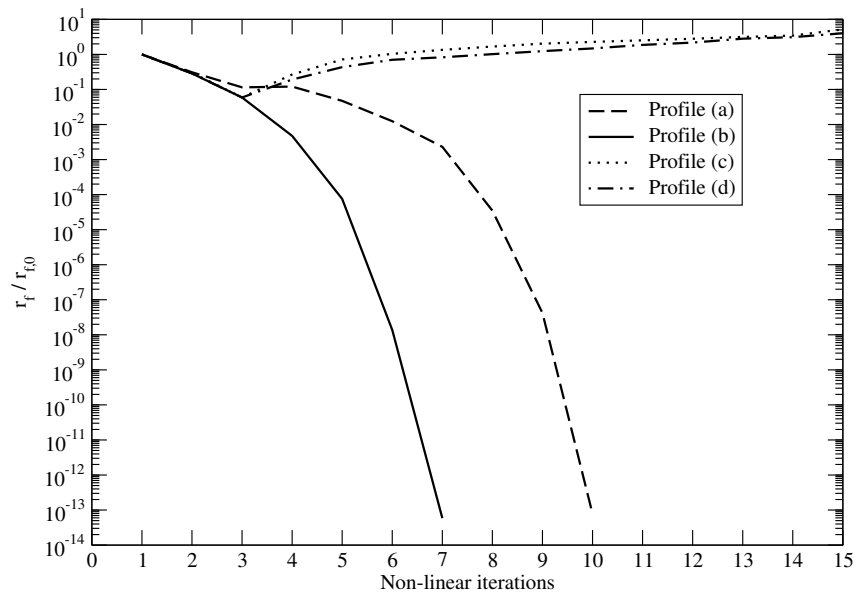


FIGURE 3.9: Case study A: convergence profile of the non-linear algorithm at step no. 8 (a), halving the load step size (b), and neglecting the numerical contribution of the integral over $\bar{\Gamma}$ in equation (2.70) for the full (c) and half-size load step (d).

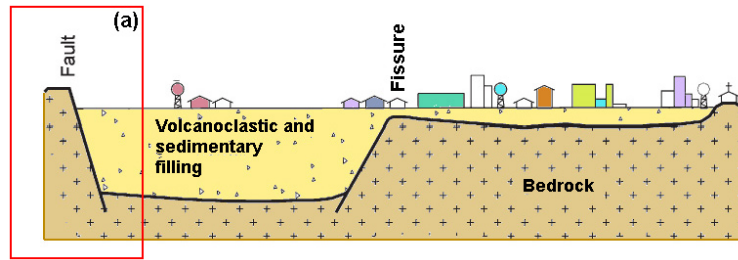


FIGURE 3.10: Case study B. Sketch of the typical geological setting ground rupture in Queretaro City. The simulated configuration is highlighted by the red box.

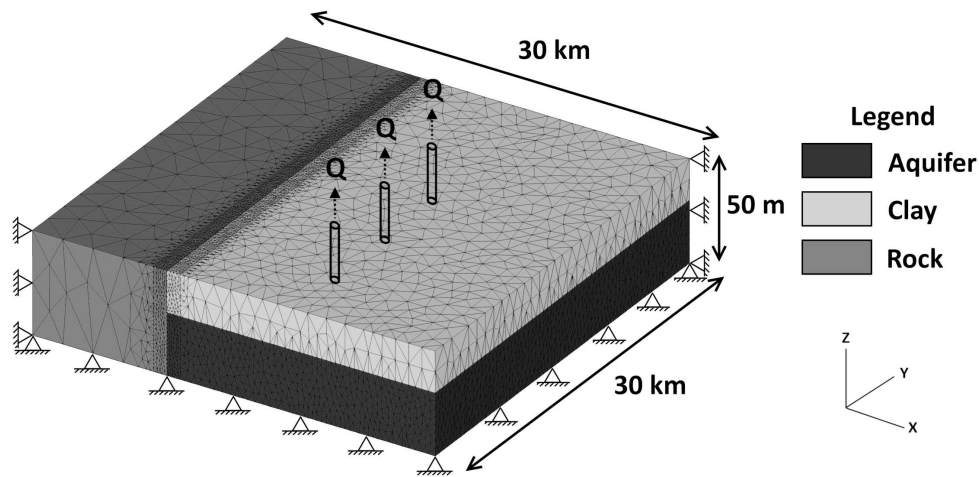


FIGURE 3.11: Case study B.

3.2 Case study B

The numerical model is experimented with in a realistic test case simulating the possible generation and evolution of a ground fracture in a 3D geological setting. The case where a rigid basement outcrops and abruptly bounds an alluvial sequence is investigated. For example, such a situation often occurs in the urban areas located in the lacustrine or fluvio-lacustrine depressions of the central sector of the Mexican volcanic belt, such as Morelia, Celaya, and Queretaro [Carreón-Freyre et al. \[2005\]](#). A sketch of the hydrogeological setting of the test case is shown in Fig. 3.10 (after [Franceschini et al. \[2015\]](#)).

The problem domain is sketched in Fig. 3.11. A 30-m thick aquifer is overlain by a 20-m thick clay unit and bounded on a side by an outcropping rock formation that intercepts the sedimentary sequence vertically. The hydro-geomechanical properties of the different materials are summarized in Table 3.1. For the sake of simplicity, a linear

TABLE 3.1: Case study B: Young modulus, E , Poisson ratio, ν , and permeability K of the different materials.

Material	E [MPa]	ν	K [m/s]
Aquifer	15	0.25	10^{-4}
Clay	15	0.25	10^{-10}
Rock	4900	0.25	0

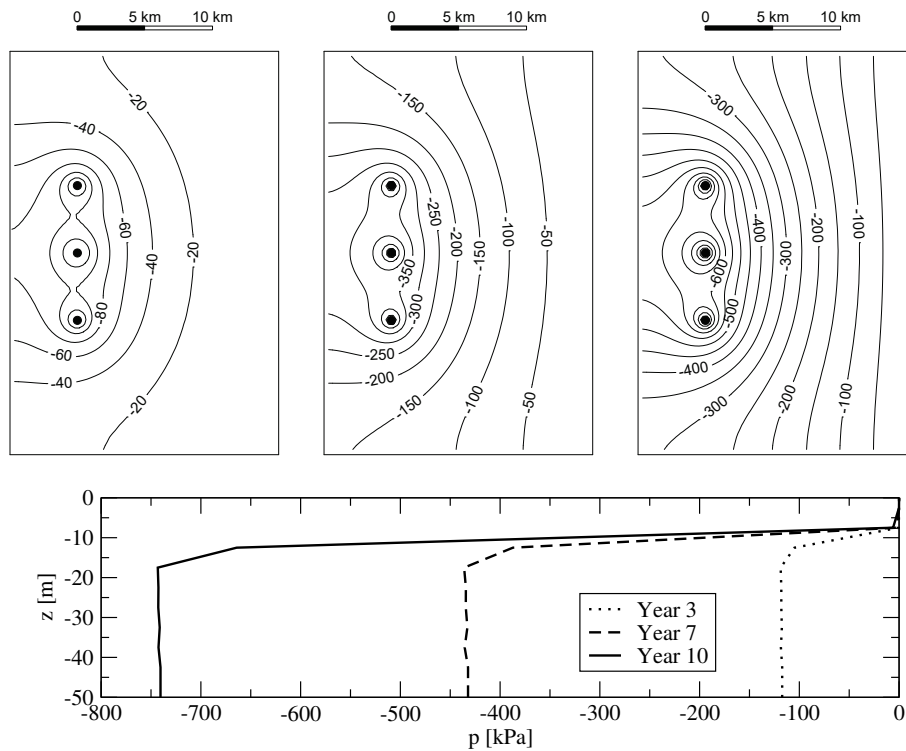


FIGURE 3.12: Case study B: pressure maps in the aquifer after 3, 7 and 10 years of pumping (above) and along a vertical profile at the central well location (below). Pressure is in kPa.

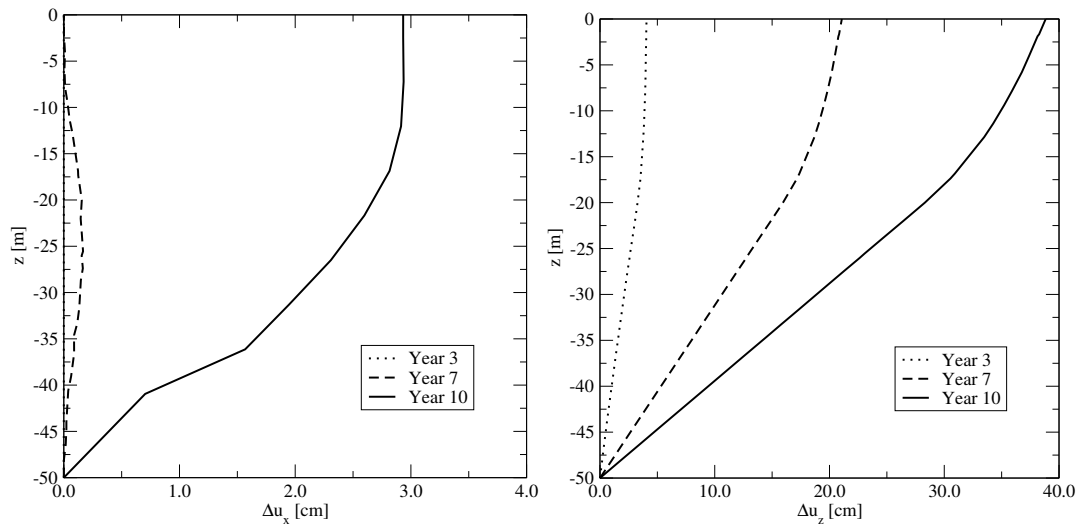


FIGURE 3.13: Case study B: opening (left) and sliding (right) along the fracture.

elastic constitutive behavior is assumed. The groundwater is withdrawn for a 10-year period by three wells located at a 5-km distance from the interface between the rock and the alluvial formation (see Fig. 3.11). The pore pressure variation induced in the sandy and clay units (see Fig. 3.12) is used as an external source of strength in the geomechanical model.

The interface between the rock and alluvial formation is modelled as a vertical fault with zero cohesion and friction angle $\varphi = 30^\circ$. The lithostatic load provides the initial stress condition. The overall FE mesh consists of 297,971 elements with 57,050 nodes, of which 4,009 are located on the fault.

The groundwater withdrawal causes both a vertical and lateral contraction of the aquifer, and partially of the overlying clay layer. By distinction, the rock formation is rigid and impervious, so that a fracture can develop at the interface with the alluvial deposits. According to the pressure evolution in time, the fracture starts to slide vertically after 3 years from the inception of pumping. At year 7, the lateral contraction of the aquifer causes an opening that enlarges to the whole contact surface at the end of simulation. Fig. 3.13 and 3.14 show that the opening Δu_x and the vertical sliding Δu_z are about 3 and 40 cm on the ground surface, respectively, and that the numerical solution appears to be quite smooth on the contact surface. The generation of the ground fracture has also a significant impact on the land motion. Fig. 3.15 provides the land subsidence, i.e., the vertical movement of the ground surface, along the x direction,

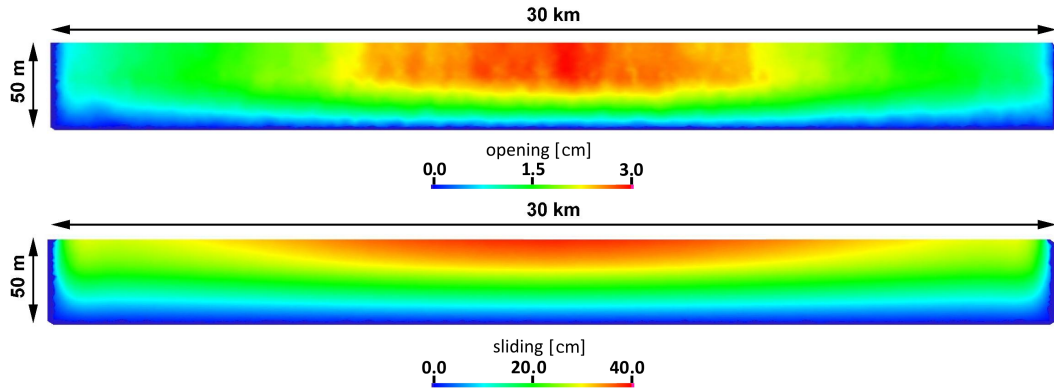


FIGURE 3.14: Case study B: opening (above) and sliding (below) on the fissure at the end of the simulation.

showing that the fissure slip has the same order as the maximum settlement. Finally, Fig. 3.16 represents the subsidence map.

Case study B has been already presented and discussed in Janna et al. [2010b], where a penalty approach with the explicit algorithms developed in Ferronato et al. [2008a] was used. The formulation developed in this work provides a much more stable solution along the fissure, avoiding the numerical oscillations that may typically occur using a penalty approach. For the sake of the comparison, Fig. 3.17 provides the distribution on the fracture plane of the sliding elements at year 5 obtained with the penalty approach. At this year the fissure is actually completely sliding. The mean slippage of the fracture has been computed as the integral average of the relative vertical movements of the nodes at the interface between the rock and the alluvial sediments, resulting in 4.0cm and 3.8cm with the penalty and the Lagrangian formulation, respectively, i.e., on the average the two approaches give approximately the same result. However, locally the outcome of the two models can be quite different.

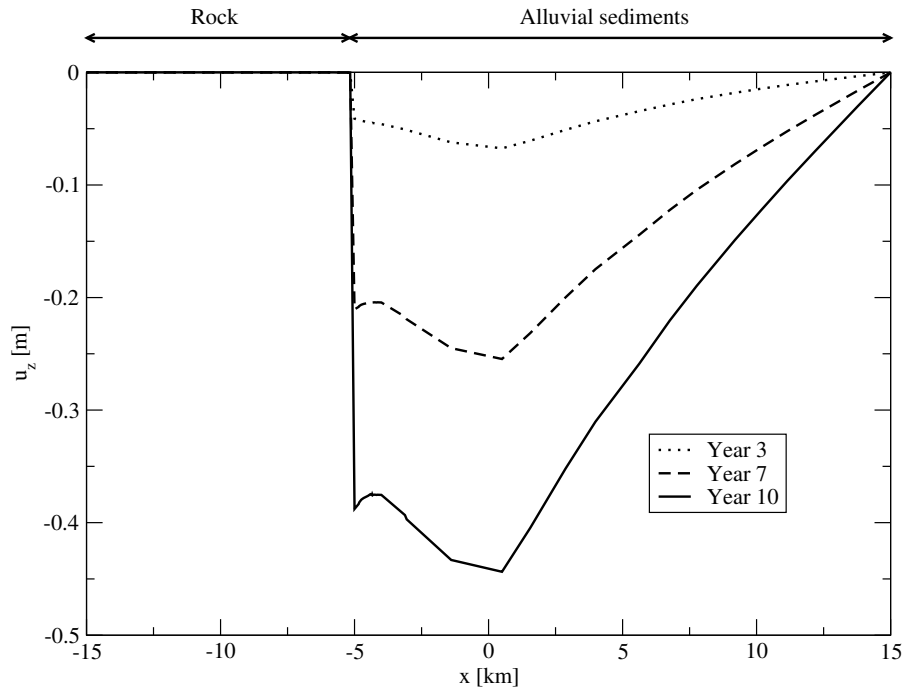


FIGURE 3.15: Case study B: land subsidence along the x -direction.

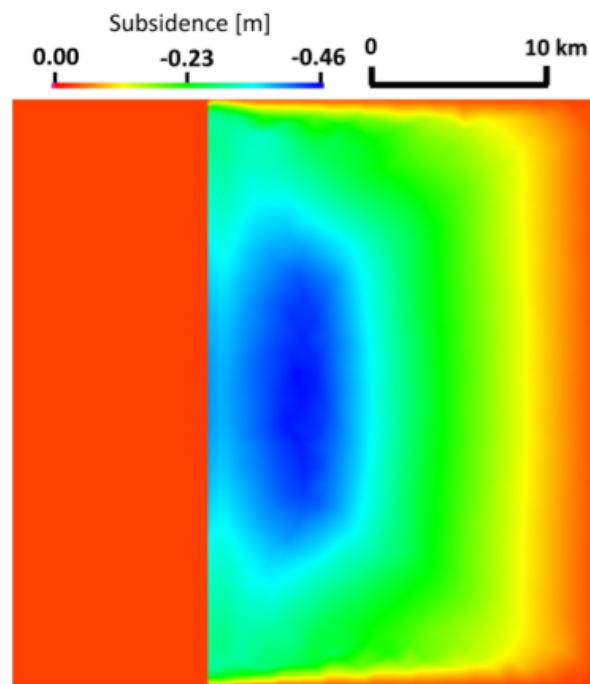


FIGURE 3.16: Case study B: land subsidence map.

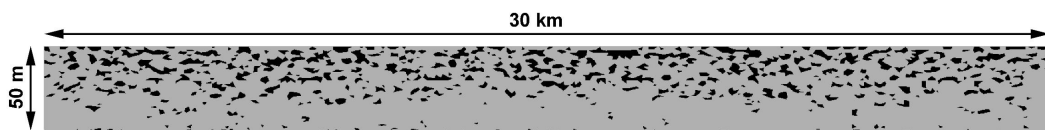


FIGURE 3.17: Case study B: distribution of sliding (light gray) and non-sliding (dark gray) elements on the fissure at year 5 obtained with a penalty approach [Ferronato et al., 2008a; Janna et al., 2010b].

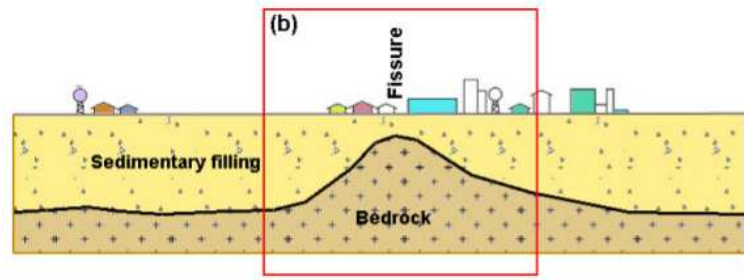


FIGURE 3.18: Case study C. Sketch of the typical geological setting ground rupture in Wuxi, China. The simulated configuration is highlighted by the red box.

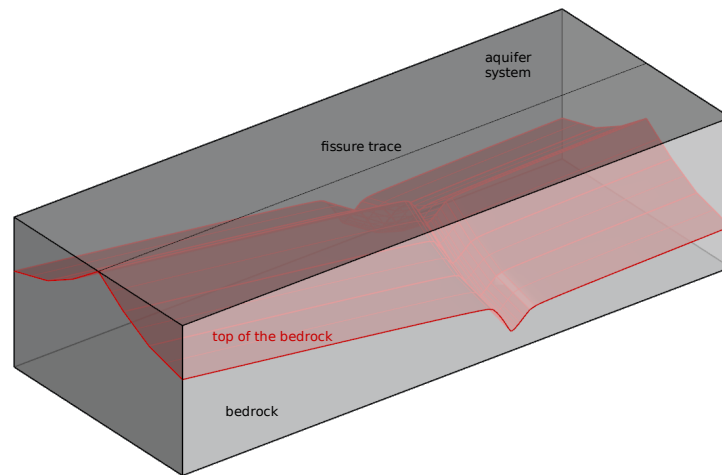


FIGURE 3.19: Case study C: perspective views of the geological setting.

3.3 Case study C

In China, earth fissuring related to groundwater pumping occurred since the 1970s. To cope with the rapid economic development, a large groundwater over-exploitation occurred in the Wuxi area, Jiangsu Province, yielding large land subsidence [Shi et al., 2007] and fissure development [Wang et al., 2009]. Earth fissures cause serious damages, including cracking of buildings and failure in underground pipelines, with huge economic losses. The geologic setting, which is characterized by an undulating shaped, is a relatively shallow rocky paleobasement covered by Quaternary compressible sedimentary deposits from the Yangtze River, that strongly enhances the risk of fissure development. A sketch of the hydrogeological setting of the test case is shown in Fig. 3.18 (after Franceschini et al. [2015]).

The geological information has been used to develop the static model. A rock ridge,

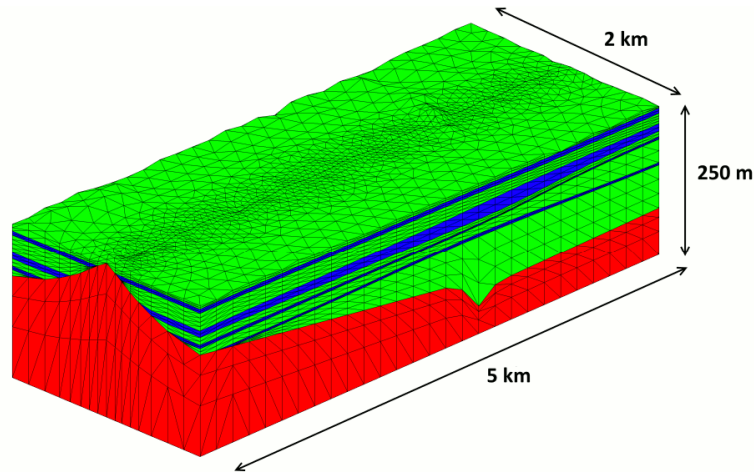


FIGURE 3.20: Case study C: 3D FE-IE grid. The materials are highlighted in various colors: rock in red, aquifers in blue, and aquitards in green.

TABLE 3.2: Case study C: Young modulus, E and Poisson ratio, ν of the different materials.

Material	E [MPa]	ν
Aquifer	12	0.30
Clay	8	0.30
Rock	1200	0.30

characterized by a depth between few meters and about 100m from the ground surface, is buried below the Quaternary sedimentary sequence (see Fig. 3.19). The large land subsidence measured in the area between the 1980s and 2000s has been accompanied by the development of a fissure in correspondence of the tip of the ridge. The geometry of the main sandy and clayey layers forming the multi-aquifer system has been derived from a number of boreholes drilled down to the bedrock. The target of the study is to simulate the development and propagation of the ground rupture, using a realistic history of groundwater extraction.

The domain extends $2\text{km} \times 5\text{km}$ in the horizontal plane and from the land surface down to 250m depth in the vertical direction. A refined mesh consisting of 23,303 nodes and 121,942 tetrahedral elements is used to accurately represent the actual litho-stratigraphic configuration (see Fig. 3.20). A number of 1,876 IEs are introduced into the 3D FE grid along the trace of the ridge tip, extending from the land surface to the bedrock top. The IEs are characterized by a friction angle $\varphi = 30^\circ$ and zero cohesion. The geomechanical properties of the bedrock, aquifers, and aquitards are provided in Table 3.2.

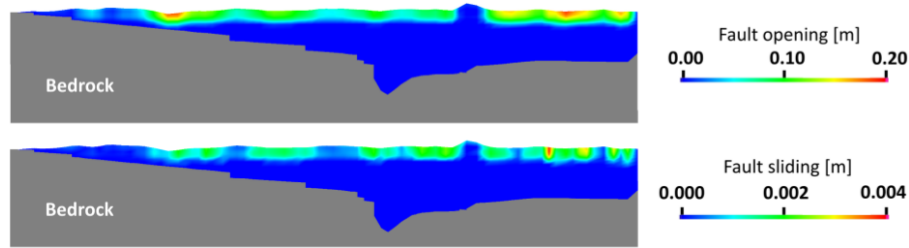


FIGURE 3.21: Case study C: opening and sliding of the ground rupture after 10 years of water withdrawal.

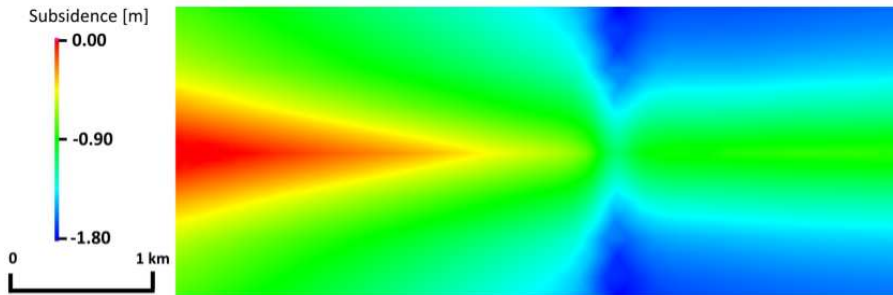


FIGURE 3.22: Case study C: land subsidence after 10 years of water withdrawal.

A uniform pressure decline is prescribed in the sandy layers. The drawdown increases linearly from 0 to 20m over 10 years. To account for a likely delay of pressure propagation in the clay layers, one half of the pressure change (i.e., 10m at the end of the simulation period) is prescribed in the aquitards. No pressure change propagates in the bedrock. The numerical results confirm the development of an earth fissure along the ridge tip. Due to the quasi-symmetric geological setting, the rupture is characterized by a certain opening and a negligible sliding. At the end of the simulation period the maximum opening corresponds to about 20 cm (see Fig. 3.21). Fig. 3.22 provides the corresponding land subsidence distribution. The basement depth of burial significantly affects both the rupture opening and land subsidence.

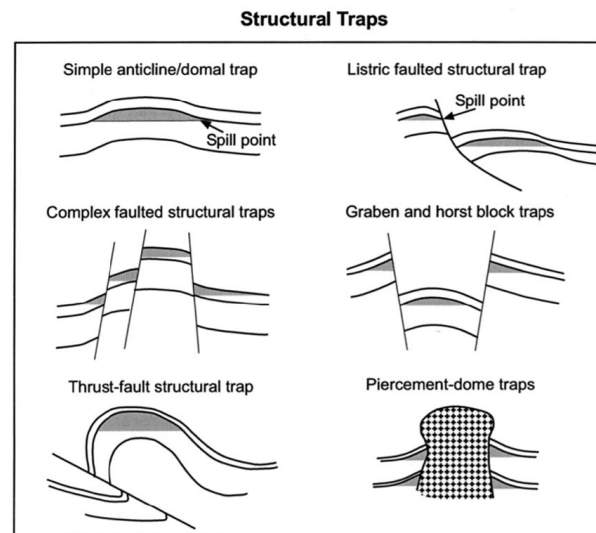


FIGURE 3.23: Diagram description of hydrocarbon traps. Adapted from [PetroWiki \[2015\]](#).

3.4 Case D

The oil/gas production or injection in faulted reservoirs can induce/trigger seismicity due to fault reactivation. In this case, we are interested in the prevision of the possible human caused seismic events and in the magnitude estimation. In particular, our aim is to analyze the beginning of the breaking on the faults.

3.4.1 Faulted reservoirs

An accurate knowledge of the geology of an underground reservoir is essential for a reliable forecast of its development, production, and management.

The structural traps giving rise to hydrocarbon reservoirs are usually formed by structural uplift and differential compaction. Typical traps are structural domes and doubly plunging anticlines (see Fig. 3.23). The bulk of the world's oil is found in these four-way-closure traps [Demaïson and Huizinga, 1991], which were the first type to be exploited by surface mapping. Many major oil fields in the world were discovered by using surface mapping to locate domal structures.

A more complex method of forming a structural trap is by faulting and structural uplift (see Fig. 3.23). Faulted structures can vary from a simple faulted anticline to complex faulting around piercement structures and domal uplifts. Faulted structures are very common and form some of the most complex known reservoirs. Types of

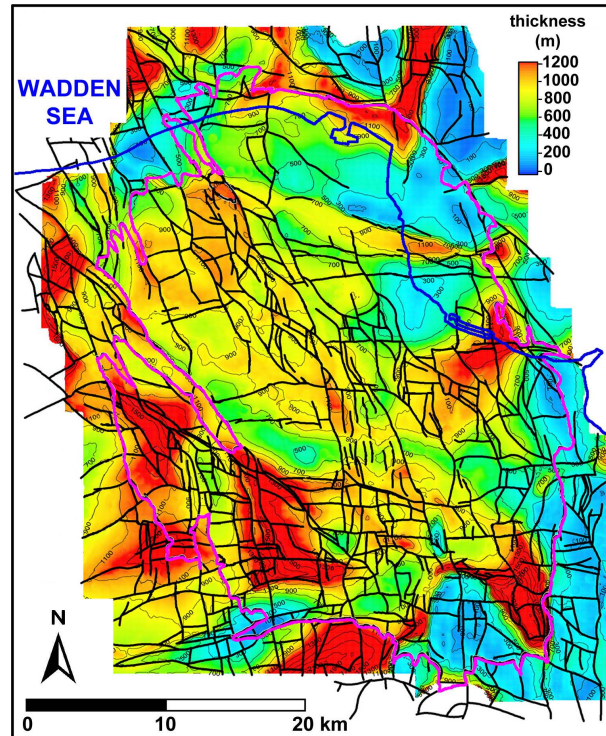


FIGURE 3.24: Map of the Groningen gas field. Faults are black lines. The coastline is the blue line while the reservoir trace is the pink one. Adapter from [Van Thienen-Visser et al. \[2015\]](#).

faults include normal, listric, reverse, and thrust, which are related to the stress fields generated during the structural movement. Closure is obtained by the uplift of sediments juxtaposed to the piercement dome. Faulted reservoirs commonly have a bottom seal formed by the lower contact of sand with shale. The bottom seal, along with the oil/water contact within the sand body, forms the base of the reservoir [[PetroWiki, 2015](#)].

A famous example of highly faulted reservoir is the Groningen gas field, in the north of The Netherlands [[Van Eck et al., 2006](#)]. Figure 3.24 represents this gas field and its faults [[Van Thienen-Visser et al., 2015](#)].

3.4.2 Seismic magnitude estimate

In their pioneering paper on the anthropogenic land subsidence over the Goose Creek (TX) oil field, [Pratt and Johnson \[1926\]](#) wrote that “the movements were accompanied by slight earthquakes which shook the houses, displaced dishes, spilled water, and disturbed the inhabitants generally”. This is perhaps the first mention of a possible

relationship between oil production and seismic activity. Since then, the correlation between subsurface fluid removal/injection and resulting earthquakes has been postulated a number of times, and particularly so in the second half of the past century as the amount of deep fluid production/injection has greatly increased [Healy et al., 1968; Yerkes and Castle, 1976; Raleigh et al., 1976; Segall, 1985, 1989; Pennington et al., 1986; Ahmad and Smith, 1988; Simpson et al., 1988; Davis and Pennington, 1989; Nicholson and Wesson, 1990, 1992; McGarr, 1991; Zoback and Zinke, 2002; Ake et al., 2005; Dahm et al., 2007; Majer and Peterson, 2007; Lei et al., 2013; Kim, 2013; Zbinden et al., 2017]. Davies et al. [2013] compiled a list of 198 published examples of earthquakes with magnitudes $M \geq 1$ that have occurred since 1929 and are suspected to be of anthropogenic origin. According to authors, the potential causes and magnitudes may be correlated to: (a) mining ($1.6 \leq M \leq 5.6$); (b) oil and gas field depletion ($1.0 \leq M \leq 7.3$); (c) water injection for secondary oil recovery ($1.9 \leq M \leq 5.1$); (d) reservoir impoundment ($2.0 \leq M \leq 7.9$); (e) waste disposal ($2.0 \leq M \leq 5.3$); (f) academic research boreholes investigating induced seismicity and stress ($2.8 \leq M \leq 3.1$); (g) solution mining ($1.0 \leq M \leq 5.2$); (h) geothermal operations ($1.0 \leq M \leq 4.6$); and (i) hydraulic fracturing for recovery of gas and oil from low-permeability sedimentary rocks ($1.0 \leq M \leq 3.8$).

According to McGarr [1991], “net extraction of oil reduces slightly the average density of the upper crust, causing an isostatic imbalance. The lower crust deforms in response to this imbalance, thus increasing the load on the seismogenic layer, which fails seismically to thicken the crust so as to restore static equilibrium”. As a major consequence “oil production induces an upward force on the upper crust that stimulates compensatory aseismic deformation below the seismogenic layer that eventually results in an earthquake sequence” [McGarr, 1991]. Focusing much more precisely on the actual physical process that may induce an earthquake, Ellsworth [2013] argues that the mechanism responsible for inducing seismicity “appears to be the well-understood process of weakening a pre-existing fault” by changing the fault loading conditions. In essence: “increasing the shear stress, reducing the normal stress and/or elevating

the pore pressure can bring the fault to failure triggering the nucleation of an earthquake” [Ellsworth, 2013]. Similarly Lei et al. [2013] in an investigation on injection-induced seismicity around a natural gas reservoir in Zingong, southwestern Sichuan Basin, China, conclude that “the induced seismicity in the region resulted from the reactivation of pre-existing faults. Injected fluids diffuse outward along pre-existing faults weakening the faults and leading to their activation. ... Prospecting for faults in the formation is thus a key in hazard assessment of induced seismicity”.

The classic seismological literature provides different empirical models to connect the activation of a fault to the possible magnitude of the generated earthquake. There are seismological models based on probabilistic approaches, see e.g. Bourne et al. [2014]; Dahm et al. [2015]; Roche and Van der Baan [2015]. In particular, Bourne et al. [2014] develop a seismological model for earthquakes induced by volume changes in subsurface reservoirs. Their approach is based on the works of Kostrov [1974] and McGarr [1976] that link the total strain to the summed seismic moment in an earthquake catalog, using a Monte Carlo method to estimate the probability distribution of the total seismic moment associated with a given reservoir compaction. Dahm et al. [2015] develop a probabilistic discrimination scheme for the problem of conventional production of hydrocarbon fields and depletion-induced and depletion-triggered earthquakes. The method is based on physical-statistical seismicity models, with a 3D boundary element method adapted for the nucleus of strain approach which is integrated with the Gutenberg-Richter distribution between the magnitude and total number of earthquakes in any given region and time period. It addresses the uncertainty in earthquake location and other input parameters and distinguishes between natural, human-triggered, and human-induced earthquakes. Roche and Van der Baan [2015] analyze the likelihood of a tensile and shear failure due to 1-D variations in local stresses and rock strengths caused by pore pressure changes and layering at the Carthage Cotton Valley gas reservoir, Texas. A Mohr-Coulomb and Griffith’s failure criterion was used to compute the likelihood of shear and tensile failure, respectively, as a function of depth with ad hoc estimate of the tensile strength and cohesion of the local lithologies.

In this application, we use another approach, i.e. a deterministically-based analysis that makes use of the model developed in previous chapter. The prediction of the

seismic moment (M_0) of an earthquake can be written as a function of the fault rupture by the relation [Hanks and Kanamori, 1979]:

$$M_0 = G A s_a \quad (3.2)$$

where G is the shear modulus of the rock incorporating the reactivated fault, A is the area of the fault-surface rupture, and s_a is the average slip of the fault surface during a single event. The seismic moment (3.2) is converted into the moment magnitude M , used to measure the strength of the seismic event. The relationship between M and M_0 is defined as [Kanamori and Anderson, 1975; Kanamori, 1977]:

$$M = \frac{2}{3} (\log_{10} M_0 - 9.1) \quad (3.3)$$

with M_0 expressed in $[Nm]$. According to Hanks and Kanamori [1979], the radiated energy of the seismic event is related to M_0 as follows:

$$E_s = \frac{\Delta\tau}{2G} M_0 \quad (3.4)$$

where $\Delta\tau$ is the earthquake stress drop [Bailey and Ben-Zion, 2009], i.e. the difference between initial and final stresses $\Delta\tau = \tau_0 - \tau_f$.

Unlike the most recent dynamic models of earthquake developed for the simulation of major seismic events of natural origin [Bizzarri, 2010; Kato, 2016; Nielsen et al., 2016] the simplified “quasi-static” approach has long been used to estimate the maximum magnitude M caused by anthropogenic activities, such as fluid production and injection/storage from/into the surface. Its use is fully warranted by the likely negligible inertia of the system when small (e.g. centimetric) slip and small areal extent characterize the fault activation.

Since our model is heterogeneous, equation (3.2) must be written in an integral form. As G and s are functions of time t and space $\mathbf{x}$, equation (3.2) can be expressed as:

$$M_0(t) = \int_{A^*} G(\mathbf{x}, t) s(\mathbf{x}, t) dS \quad (3.5)$$

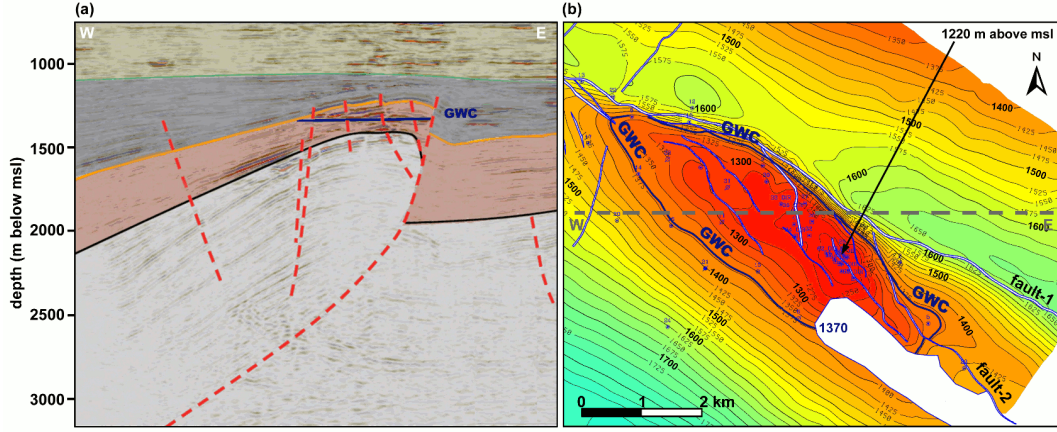


FIGURE 3.25: Case D: (a) Seismic cross-section of the gas field. The main geological formations, the trace of the faults/thrusts and the GWC (gas water contact) location are shown. (b) Map of the top depth (m below msl) of the gas bearing pool PA. The two main faults (called “fault-1” and “fault-2”) are indicated.

where s is the fault slip and G is the current local shear modulus of the host rock in the surrounding of dS . Notice that the integration domain A^* in equation (3.5) is the area where a new activation occurs at time t . This is consistent with the assumption that the fault portion already activated slips plastically with energy dissipation (aseismic movements), thus does not contribute to the seismic energy release [Rutqvist et al., 2013b].

Taking advantage of the discrete solution both in time and in space, equation (3.5) can be numerically computed as:

$$M_0(T_C) = \sum_{i=1}^{n_{IE}} G_i s_i A_i \quad (3.6)$$

where G_i is the arithmetic average of the shear modulus of the two FEs sharing the i -th IE, s_i and A_i its slippage and area, respectively, n_{IE} is the total number of newly activated IEs and T_C is the current simulation time.

3.4.3 Reservoir setting

The reservoir, made up of sandy turbidites, is seated at an average depth of 1300 m below mean sea level (msl) with a main orientation from north-west to south-east and is within a series of thrusts. It represents the north-westernmost part of an elongated

TABLE 3.3: Case D: Geometric and petrophysical properties of the different pools of the gas field.

Parameter	Pool P_A	Pool P_B	Pool P_C	Pool P_D
Initial gas pressure P_i (bars)	150.5	151.4	151.8	152.5
Hydraulic conductivity K (mD)	133-160	163-257	195.1	200
Net/Gross pay p_{ng} (%)	0.757	0.854	0.789	0.676
Porosity ϕ (%)	24	24	23	23
Water saturation S_w (%)	17.8	12.4	12.7	14.7
Average thickness (m)	15	14	18	10
Gas-water contact depth (m below msl)	1370	1370	1370	1370
Average aquifer conductivity (mD)	219	126	224	96

anticline bound by north-eastern reversed faults (Figure 3.25a). The gas traps are of a mixed stratigraphic - structural type combining anticlines, faults, and pinch-outs. The caprock is represented by a silty-clayey facies up to 200-m thick (Figure 3.25b). It is supposed that the field has been developed for gas production for 12 years, while the UGS activities start 10 years after the end of the primary production.

The reservoir consists of a few main sandy layers, 10 to 20-m thick and hydraulically disconnected. Four levels, namely P_A , P_B , P_C , and P_D , which account for a cumulative 100-m thickness, have been converted to UGS activity. Pools P_A and P_B are the most important ones. Table 3.3 summarizes the main features of the reservoir levels. The original pore pressure p_i prior to field development was about 150 bars.

The reservoir is confined to the north-east by two main reversed faults (Figure 3.25). Because of their positions, they may be both partially reactivated by production and UGS activities. The discontinuities, which do not completely cross the caprock and end a few dozens of meters above the reservoir top, are almost vertical at least in the depth range spanned by the mineralized formation. A large and active lateral aquifer (waterdrive) is connected to the reservoir south-westward.

The numerical model is centered upon the reservoir with a $50 \times 50 \text{ km}$ areal size and 10 km depth (from the ground surface to the rigid basement). Fig. 3.26 shows a perspective view of the FE-IE model with the trace of the two faults of interest, namely the regional fault bounding the reservoir in the north/north-east boundary (called “fault-1” in the sequel) and a local thrust crossing the easternmost part of the reservoir (called “fault-2”). The tetrahedral FEs discretizing the domain are 2,154,316 for a total of 1,142,655 degrees of freedom. Zoomed axonometric and cross-section views of the

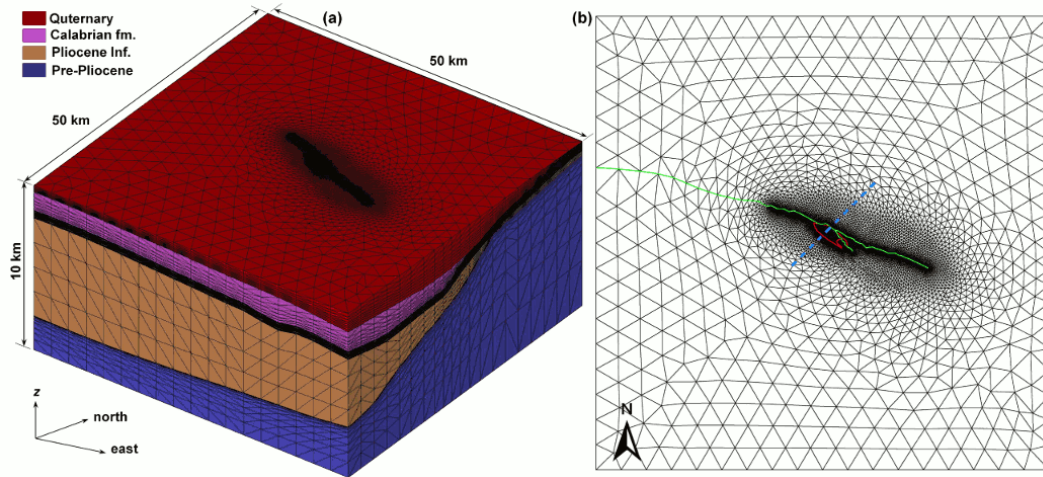


FIGURE 3.26: Case D: (a) Axonometric view and (b) plain view of the full geomechanical FE-IE model developed for the analyzed gas field. The vertical exaggeration in (a) is 5. The trace of the field (in red) and the faults (in green) crossing and confining northward the reservoir and the waterdrive are highlighted in (b). We designed two faults, that extend from -1200 to $-3000m$ depth, with a length of about $40km$ (regional fault, fault-1) and $5km$ (local fault, fault-2).

reservoir and the nearby faults as represented in the model are provided in Figure 3.27.

We use 26, 274 IEs altogether to address the geomechanical fault/thrust behavior.

A history-matched fluid-dynamic reservoir model is available, providing for the pressure behaviour in space and time during both the primary production and the UGS stage, spanning an overall period of 50 years. As an example, the behavior of the average pore pressure p in level P_A is provided in Figure 3.28. The maximum 120 bar decline occurred in the 12-th year. Gas storage started at year 22, with a yearly excursion of about 50 bars at the most. The original pore pressure p_i is never exceeded during the UGS operations (see Fig. 3.28).

The faults preclude any hydraulic connection of the field with the waterdrive in the north eastern direction, while the aquifer is active south-westward. As can be seen from Figure 3.29a, the pore pressure decline Δp at the end of the field production life propagates several km within the waterdrive because of the relatively high hydraulic conductivity (Table 3.3). The porous volume involved by the pressure variation during the UGS cycles is much smaller restricted because of both the reduced seasonal change and the limited duration in time (Figure 3.29b). In both situations, the faults exert an important influence on the Δp distribution.

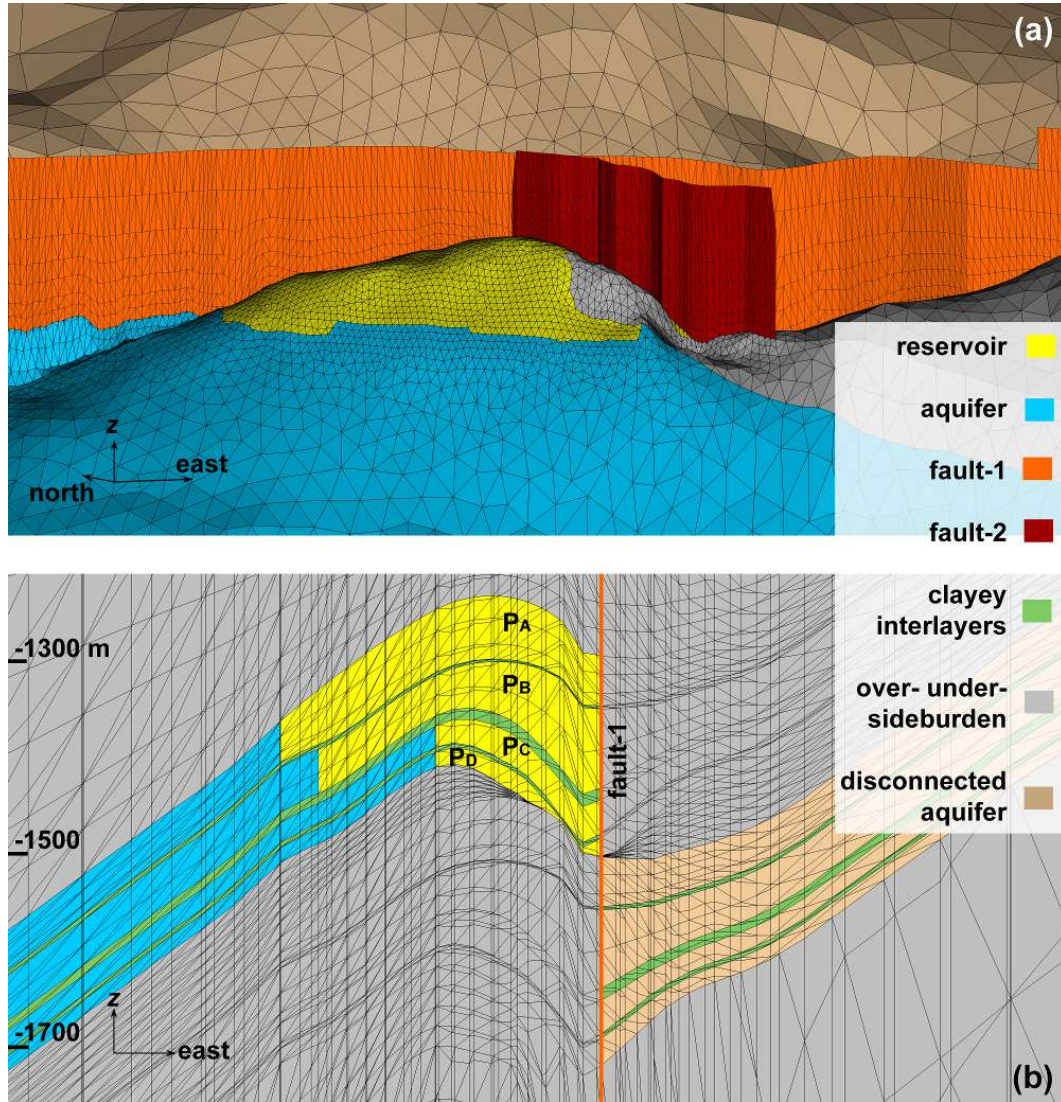


FIGURE 3.27: Case D: (a) Axonometric detail and (b) vertical southwest - northeast cross-section of the 3D FE-IE model of the gas field showing different portions of the geologic system. The colors are used to highlight the gas traps (FE in yellow), the two faults (IE in orange and red), the waterdrive (FE in light blue), the portion of the formation hydraulically disconnected from the reservoir north-east of the fault (light brown), the clayey interlayers (green), and the over- under- and sideburden (grey). The vertical exaggeration is 5. The trace of the vertical cross-section is shown in Figure 3.26b.

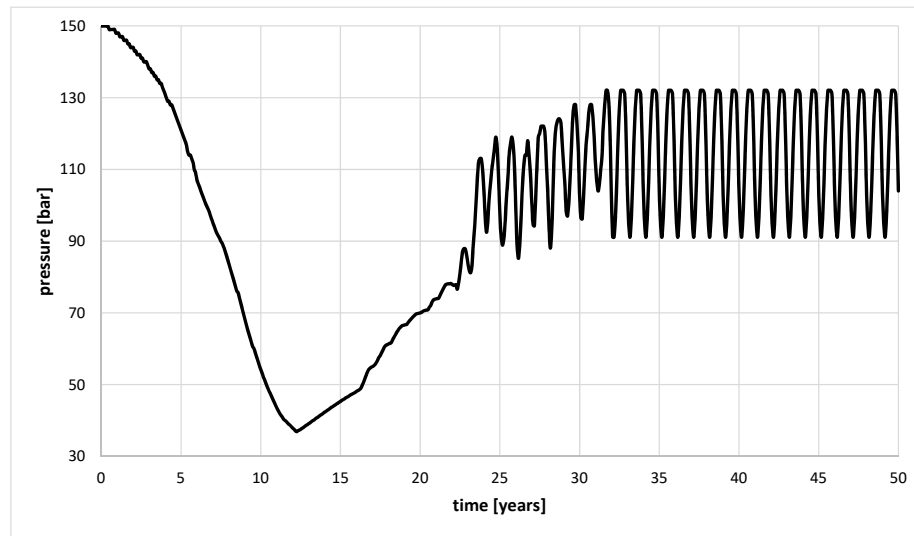


FIGURE 3.28: Case D: Average pore pressure behavior (bars) in 50 years.

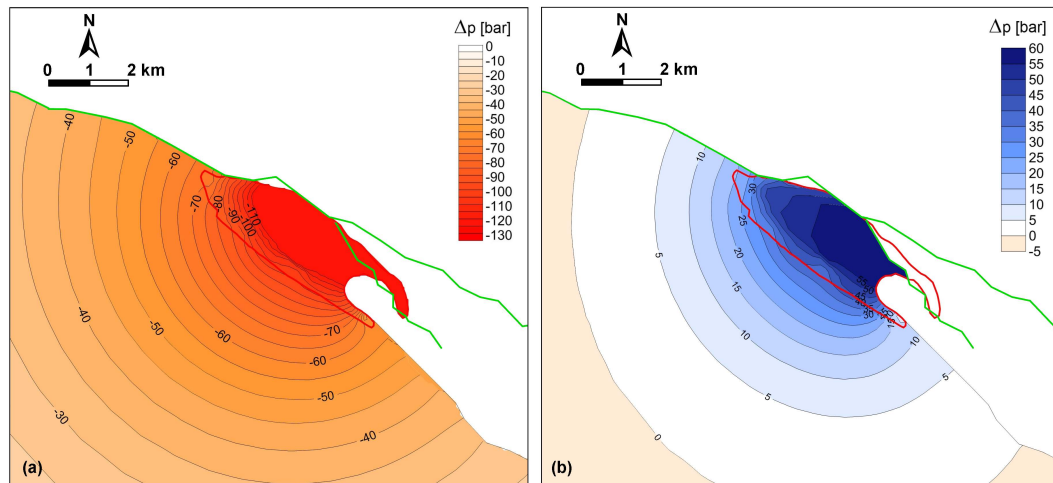


FIGURE 3.29: Case D: (a) Pore pressure decline (bars) in level P_A at the end of the field production life and (b) pore pressure increase during a UGS injection cycle from April to November of the 45-th year since the start of the simulation. The traces of the mineralized pool and faults are highlighted by red and green lines, respectively.

TABLE 3.4: Case D: Geomechanical parameters for the transversely isotropic constitutive model.

β	ν_h	ν_z	γ	s
3.0	0.15	0.25	1.0	5.0

Reliably predicting the geomechanical response of reservoirs mostly depends on the representativeness of the rock constitutive law. In this study, we use the hysteretic hypo-plastic law developed by Baù et al. [2002]; Janna et al. [2012a]; Ferronato et al. [2013]. The constitutive equation for the oedometric compressibility C_m originally advanced by Baù et al. [2002] in virgin loading conditions reads:

$$C_M = 0.01369\sigma_z^{-1.1347} \quad (3.7)$$

with C_M in $[\text{bar}^{-1}]$ and σ_z in $[\text{bar}]$, here taken as positive in compressive conditions. In unloading/reloading conditions C_m of equation (3.7) is reduced by a factor s between $3.5 \div 5.0$ over the range of interest [Ferronato et al., 2013]. The parameters of the transversely isotropic model are listed in Table 3.4.

One of the biggest uncertainties for the prediction of the possible fault reactivation is the stress regime in the horizontal directions. Regional reverse faults may reveal a general compressive stress regime, with the largest horizontal stress σ_{h2} approximately orthogonal to the discontinuity trace and the smallest σ_{h1} orthogonal to σ_{h2} . The direction of the maximal horizontal stress is estimated 18° North. Different scenarios have been investigated to account for the uncertainty in the initial stress regime. In particular, it has been assumed $\sigma_{h1} = K\sigma_z$ with K the confinement factor:

$$K = \frac{\nu_z}{1 - \nu_h} \quad (3.8)$$

The other horizontal stress σ_{h2} is obtained from one of the following three equations:

$$(A) \sigma_{h2} = \sigma_{h1}, \quad (B-C) \sigma_{h2} = \frac{\sigma_z + \sigma_{h1}}{2} \quad \text{and} \quad (D) \sigma_{h2} = \sigma_z \quad (3.9)$$

Concerning the Mohr-Coulomb failure criterion to assess the fracture stability, we used $c = 0 \div 10$ bars and $\varphi = 30^\circ$, i.e. a typical value suggested in the literature for

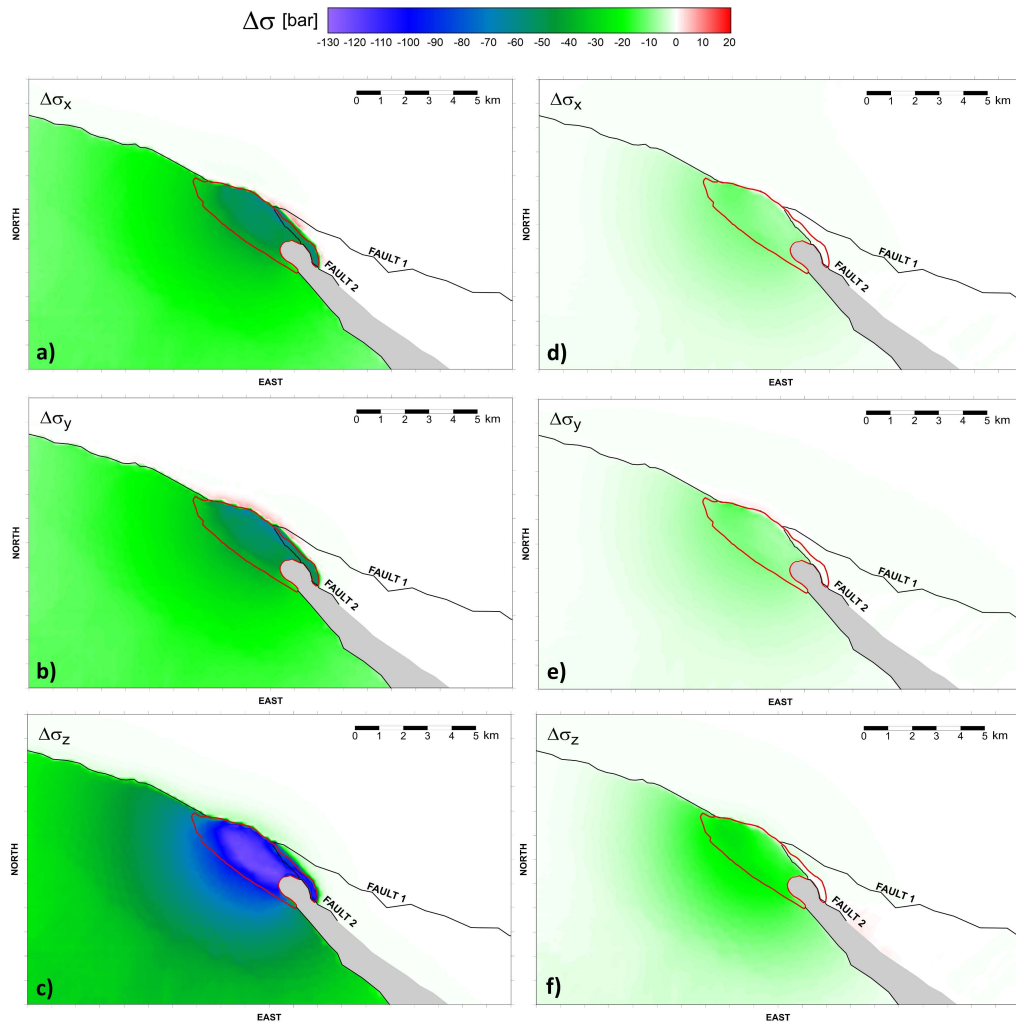


FIGURE 3.30: Case D: Normal stress variations (bar) in level P_A along the west-east ($\Delta\sigma_x$), north-south ($\Delta\sigma_y$), and vertical ($\Delta\sigma_z$) directions after 12 years, i.e. at the end of the field production life (a, c, e), and during the extraction phase from one Autumn to the next Spring (b, d, f).

sandstone [Fjær et al., 2008].

Figure 3.30 shows the normal horizontal (i.e. $\Delta\sigma_x$ and $\Delta\sigma_y$) and vertical (i.e. $\Delta\sigma_z$) stress variation within the P_A level at the end of the field production life, i.e. the most stressed reservoir conditions. Within the reservoir the variation $\Delta\sigma_z$ is approximately equal to the fluid pressure change, while $\Delta\sigma_x$ and $\Delta\sigma_y$ undergo a smaller almost identical change (about 40% of the vertical component). Notice that the use of a transversely isotropic material with $\beta = 3$ significantly impacts on the stress path in the production layers, with the generation of a large deviatoric stress ($\Delta\sigma_z - \Delta\sigma_h$).

Figure 3.30 also shows the propagation of the stress change in the geologic unit surrounding the reservoir. The porous volume experiencing a significant stress variation coincides with the volume of the aquifer hydraulically connected to the reservoir where the pore pressure variation propagates up to some 10 km south-west of the reservoir at the end of the primary production life. During UGS activities the volume affected by the fluctuating stress variation outside the reservoir is much smaller. This is so because of the relatively short UGS semi-cycle during which the reservoir pore pressure change only partially propagates within the waterdrive. Stress variations propagate negligibly to the north-eastward beyond the faults, less than 4 and 2 bars for $\Delta\sigma_z$ and $\Delta\sigma_x, \Delta\sigma_y$, respectively. The portion of fault-1 affected by a significant stress change (say larger than 10 bars) extends for about 12 km and 2 km to the west of the reservoir boundary after 12 years and during one of the UGS cycles, respectively. Fault-2, which compartmentalizes the field, is loaded by a significant stress change for its entire length.

Finally, Figure 3.31 shows the same results as Figure 3.30 along a south-north vertical cross section through the reservoir. The figure emphasizes the fact that the stress change mainly remains confined within the geologic unit where the reservoir and the waterdrive are located. Only a relatively small percentage of the $\Delta\sigma_x, \Delta\sigma_y$ and $\Delta\sigma_z$ as predicted in the reservoir extends upward in the caprock and below in the underburden. In correspondence of the anticline crest zone, a 10% to 20% of the reservoir stress change extends in the 100-m thick caprock and underburden bounding the produced layers. Moving further up- or downward, the stress change almost vanishes within the next 100 m. Note that the stress change has the opposite sign relative to that experienced by the reservoir, e.g. traction when the reservoir is compacting.

Four scenarios are addressed to account for the uncertainties on the fault parameters and stress condition, and to investigate their impact on the possible fault activation and magnitude of the induced seismicity:

- Scenario A: $c = 10$ bars, $\varphi = 30^\circ$ and $\sigma_{h2} = \sigma_z$ (equation (3.9)-A).
- Scenario B: $c = 10$ bars, $\varphi = 30^\circ$ and $\sigma_{h2} = \frac{\sigma_z + \sigma_{h1}}{2}$ (equation (3.9)-B).
- Scenario C: $c = 0$ bars, $\varphi = 30^\circ$ and $\sigma_{h2} = \frac{\sigma_z + \sigma_{h1}}{2}$ (equation (3.9)-C).
- Scenario D: $c = 0$ bars, $\varphi = 30^\circ$ and $\sigma_{h2} = \sigma_{h1}$ (equation (3.9)-D).

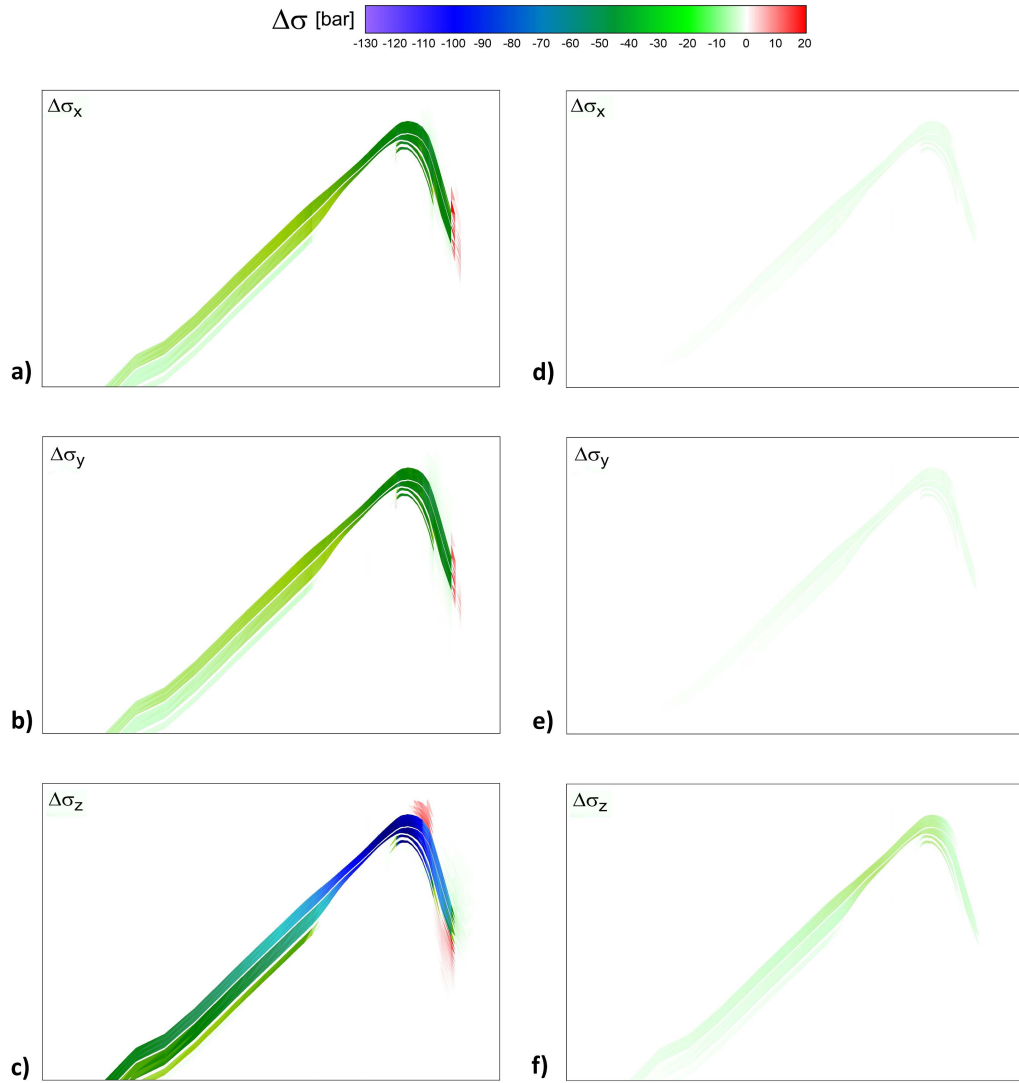


FIGURE 3.31: Case D: Normal stress variations (bar) in a vertical cross-section through the top of the gas field along the west-east ($\Delta\sigma_x$), north-south ($\Delta\sigma_y$), and vertical ($\Delta\sigma_z$) directions after 12 years, i.e. at the end of the field production life (a, c, e), and during the extraction phase from one Autumn to the next Spring (b, d, f).

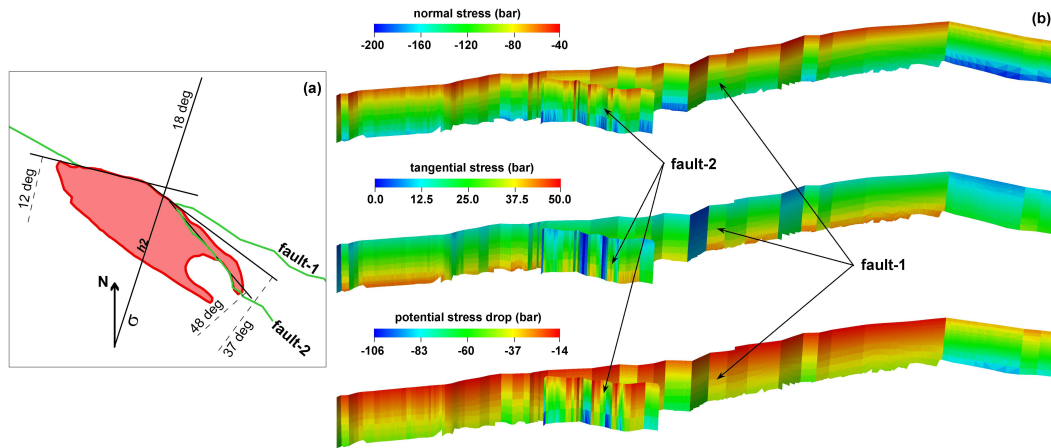


FIGURE 3.32: Case D: (a) Direction of the maximum horizontal stress in the horizontal plane σ_{h2} . (b) Initial normal and tangential stress, and potential stress drop on the faults with scenario B.

Among these possibilities, the least conservative scenario should be the A, with the maximum value for the horizontal stress, while the most probable is the scenario B. The most conservative is the scenario D, with no cohesion and low horizontal stress. No scenario has been addressed to test the model sensitivity to φ based on [King et al. \[1994\]](#) who suggest that an exact quantification of fault activation is rather insensitive to the friction angle over the range $25^\circ \div 35^\circ$.

3.4.4 Numerical results and discussion

The four scenarios are discussed in the following order: firstly the most probable one (B), then the most conservative (D) and finally the other two cases (A and C).

Scenario B. Figure 3.34a shows the cumulative slip of fault-1 and fault-2 over the primary production phase (12 years). Notice that the location of the more significant displacements depends not only on the shear stress change (Figure 3.33c) but also on the potential stress drop (Figure 3.32b), which is an indication of how far is the fault from the rupture. Figure 3.35 gives the yearly evolution of the total activated area (A), the average (s_a) and maximal (s_{max}) slip on the two faults during the primary production life of the reservoir. The outcome reveals that the largest local slip might amount to 5.4 cm with an average value s_a on the order of 0.6 cm. Notice in Figure 3.35a that the fault remains stuck over the first 3 years. The activated area increases to the maximum value of $106m^2$, as computed by the FE-IE model at the end of the

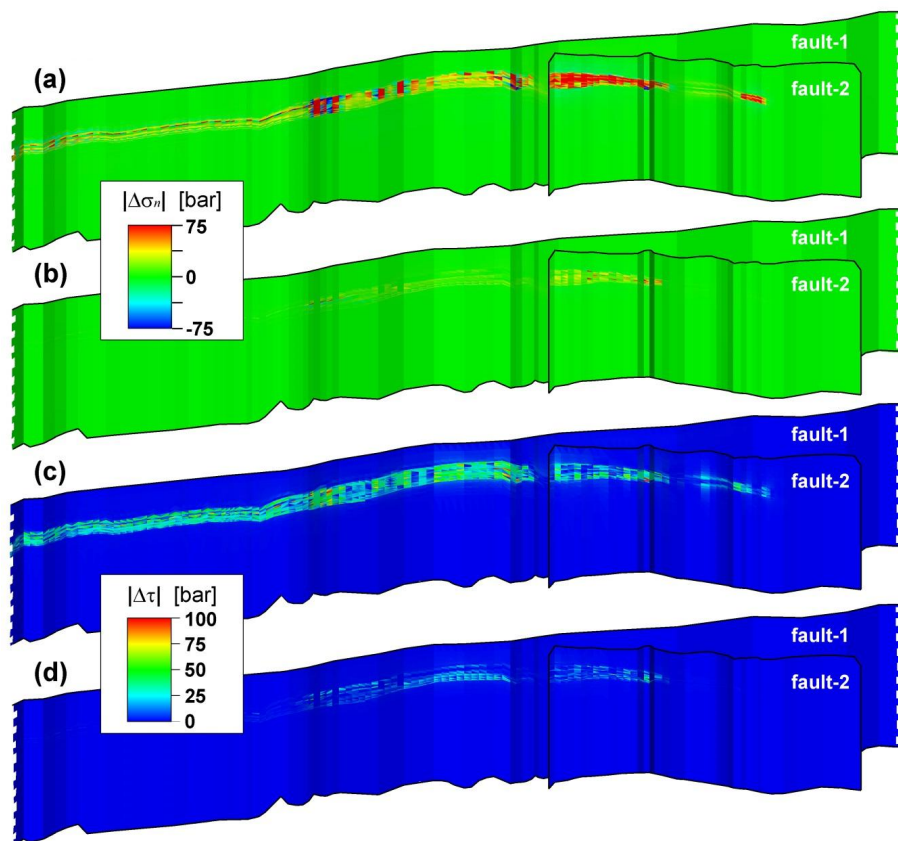


FIGURE 3.33: Case D: Scenario B: change of the effective normal and (modulus of) shear stress on the faults at the end of the primary production ((a) and (c), respectively) and during an UGS withdrawal phase ((b) and (d), respectively).

TABLE 3.5: Case D: Area (A), and average (s_a) and maximal (s_{max}) slip of the activated portion of the faults during UGS cycles for the four scenarios addressed by the study.

Scenario	Area A (m^2)		Average slip s_a (m)		Maximum slip s_{max} (m)	
	mean	st. dev	mean	st. dev	mean	st. dev
A	$8.58 \cdot 10^2$	$2.19 \cdot 10^3$	$2.99 \cdot 10^{-6}$	$7.14 \cdot 10^{-6}$	$5.86 \cdot 10^{-6}$	$1.59 \cdot 10^{-5}$
B	$2.79 \cdot 10^3$	$3.45 \cdot 10^3$	$3.42 \cdot 10^{-5}$	$4.32 \cdot 10^{-5}$	$4.34 \cdot 10^{-5}$	$6.14 \cdot 10^{-5}$
C	$1.51 \cdot 10^4$	$5.47 \cdot 10^3$	$4.76 \cdot 10^{-5}$	$2.35 \cdot 10^{-5}$	$1.29 \cdot 10^{-4}$	$2.92 \cdot 10^{-5}$
D	$8.20 \cdot 10^4$	$3.24 \cdot 10^4$	$8.08 \cdot 10^{-5}$	$4.89 \cdot 10^{-5}$	$2.32 \cdot 10^{-4}$	$1.85 \cdot 10^{-4}$

production period. The average horizontal length ΔL_a and vertical size ΔZ_a of the fault where activation occurred after 12 years to 6000 m and 75 m for fault-1, and 2600 m and 175 m for fault-2, respectively. As expected from the model outcome in term of stress variation, the fault may activate only at the reservoir depth. The activation started in the shallower part of the reservoir, where the majority of the production wells are located, and propagated laterally along the gas-bearing sandy layers. The fault portions in the basement and caprock remain inactive due to the limited stress change caused by the withdrawal operations. It is interesting to notice that, during the primary production, the maximum yearly increment of A and s_a is expected to occur between 8 and 10 years after the beginning of the production when the largest increase of the effective stress has been experienced by the reservoir.

A similar analysis has been carried out for the UGS development. The IE outcomes in terms of A , s_a , and s_{max} are summarized in Table 3.5. As the pressure range changes from cycle to cycle (3.28) and recalling that the computation in this phase has been carried out with a monthly time step, we have elected to provide the mean and the standard deviation computed using the maximum monthly values within each UGS cycle. This month generally corresponds to April, i.e. at the end of the production period when the minimum pressure is reached. Notice that the values shown in Table 3.5 are about 2 orders of magnitude smaller than those presented in Figure 3.35, i.e. the faults practically do not activate in this scenario. This is accounted for by the limited pore pressure variation confined for the most part within the field, hence impacting negligibly on the possible movement of the faults.

Scenario D. Figure 3.34 shows the cumulative slip of fault-1 and fault-2 over the primary production phase. In this scenario there is a significant increase of both A , s_{max}

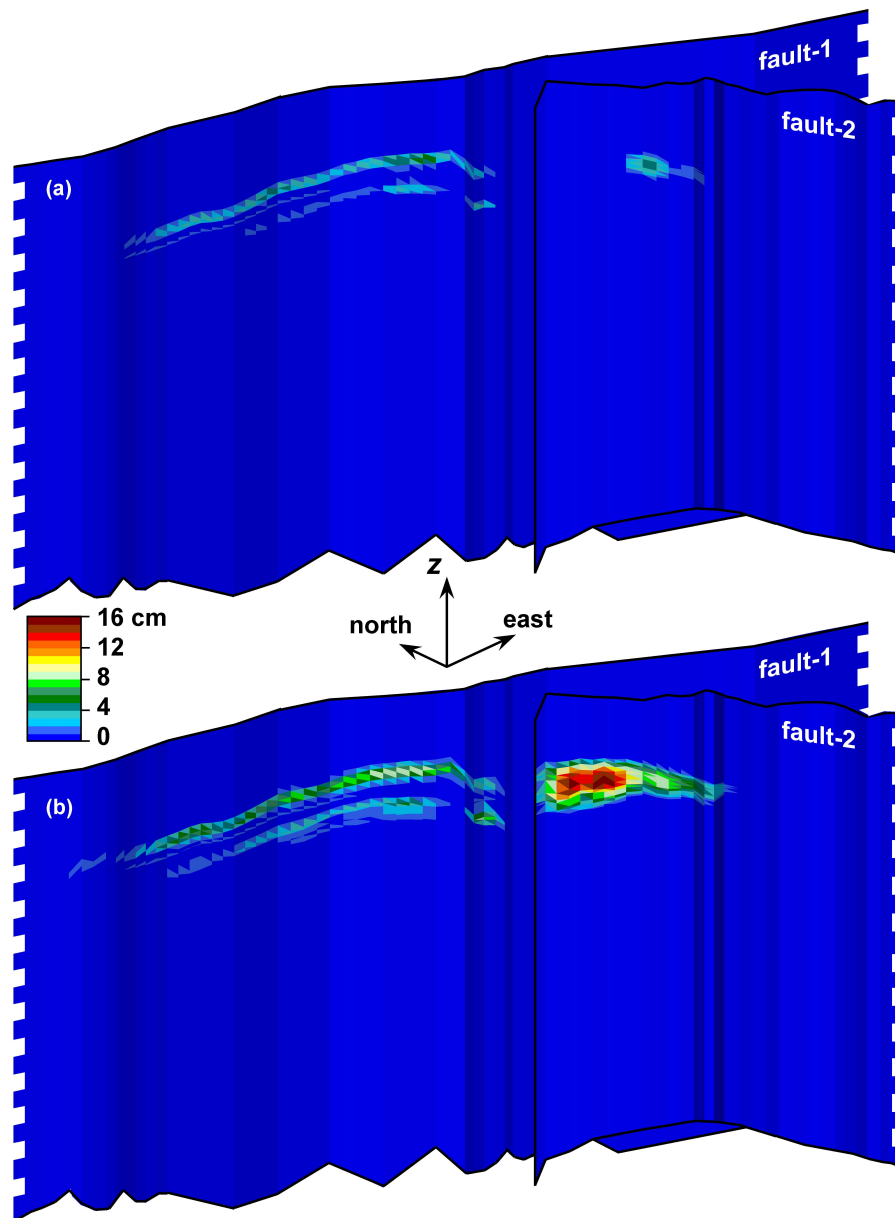


FIGURE 3.34: Case D: Perspective view on the fault discretization with the cumulative slip during the primary production life of the reservoir, as computed by the IE solution in (a) scenario B and (b) scenario (D). Vertical exaggeration is 5.

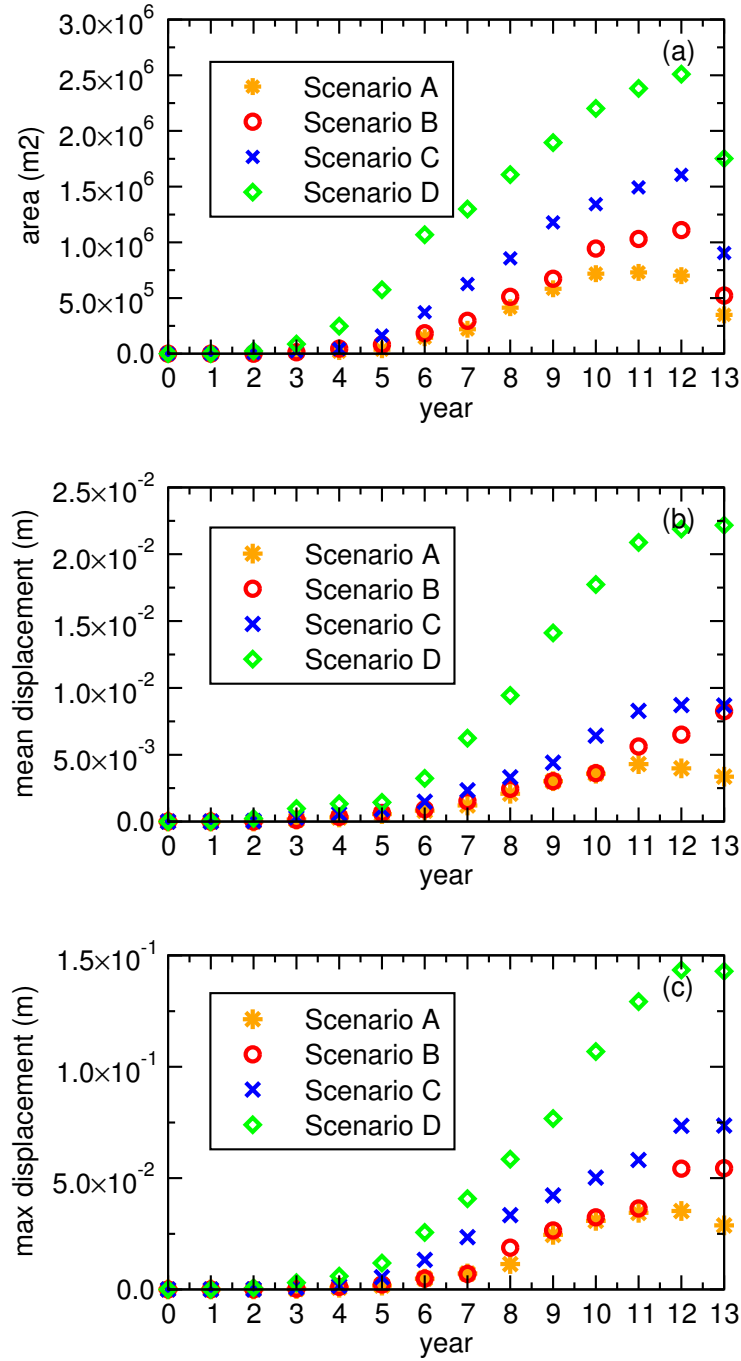


FIGURE 3.35: Case D: Area A , and average (s_a) and maximal (s_{max}) yearly slip of the activated portion of the faults during the primary development of the reservoir for the four scenarios addressed by the study.

(14.3 cm) and s_a (2.2 cm) during both the primary production and UGS procedures (Figures 3.35). As expected, a normal stress regime enhances the possibility of fault reactivation. On summary scenario D, with $c = 0$ bars and a normal stress state, is the most conservative and should be viewed as a likely upper limit. In this scenario ΔL_a and Δz_a , after 12 years, amount to 7000 m and 280 m for fault-1, and 2900 m and 190 m for fault-2, respectively. Similarly, during UGS cycles the largest A and slip values are predicted between April and May, i.e. at the end of extraction. As in Scenario A, during UGS the possibility of a fault activation is much reduced with respect to the primary production phase (Table 3.5), with the maximum displacements peaking 0.2 mm.

Scenario A and scenario C. As expected scenario A is the least conservative. The portion of the activated faults and the average and maximum sliding are the smallest one among the investigated scenarios (Figure 3.35 and Table 3.5). The difference with the outcome of scenario B is rather limited. Scenario C provides intermediate results between B and D (Figure 3.35 and Table 3.5). A comparison among the results obtained from the four scenarios highlights that the outcome from scenario C is generally closer to the one of B, suggesting that the assumption on the stress regime plays a more important role than that on cohesion.

Now, using the results of the activated area and slip, we can estimate the possible magnitude M of the triggered seismic events. It must be kept in mind that the predicted M values should be regarded as conservative as the volume V_a is assumed to activate entirely and instantaneously during the reference year/month. As mentioned above, the most conservative scenario is D, whose results are to be regarded as an unlikely upper bound. The extreme M value turns out to be equal to be 2.0 for the fault at the end of the production life and 0.6 ± 0.1 during the UGS operations. Negative magnitude values have been predicted for all the time steps during the UGS activities with scenarios A, B, and C (except for March, in scenario C). Table 3.6 and Figure 3.36 point out that M is correlated to the monthly amount and the sign of pressure change Δp . More specifically:

- Table 3.6 suggests that the extreme M values occur during the months experiencing the largest Δp , i.e. March (pressure decrease) and June/July (pressure

TABLE 3.6: Case D: Average and standard deviation of the simulated magnitude M .

Month	M ($\pm$ st. dev.)				Δp_{avg} (bar)	Δp (bar) (min/max)
	Scenario A	Scenario B	Scenario C	Scenario D		
January	-	-0.5 ($\pm$ 0.2)	-0.6 ($\pm$ 0.5)	-0.2 ($\pm$ 0.4)	-5.4	-11.1/-2.1
February	-0.5 ($\pm$ 0.3)	-0.6 ($\pm$ 0.3)	-0.4 ($\pm$ 0.3)	0.0 ($\pm$ 0.5)	-10.1	-19.7/-3.6
March	-0.5 ($\pm$ 0.2)	-0.3 ($\pm$ 0.3)	0.2 ($\pm$ 0.1)	0.6 ($\pm$ 0.1)	-10.6	-15.9/-3.6
April	-	-	-0.5 ($\pm$ 0.3)	-	-2.3	-5.6/-1.3
May	-	-0.5 ($\pm$ 0.3)	-	-0.4 ($\pm$ 0.3)	3.9	1.4/11.8
June	-1.3 ($\pm$ 0.0)	-0.5 ($\pm$ 0.1)	-0.4 ($\pm$ 0.1)	-0.7 ($\pm$ 0.5)	6.8	1.9/11.3
July	-1.0 ($\pm$ 0.4)	-0.6 ($\pm$ 0.2)	-0.5 ($\pm$ 0.2)	-0.2 ($\pm$ 0.3)	6.7	2.4/8.6
August	-1.0 ($\pm$ 0.4)	-0.7 ($\pm$ 0.2)	-0.6 ($\pm$ 0.6)	-0.4 ($\pm$ 0.4)	5.9	0.9/8.6
September	-1.0 ($\pm$ 0.3)	-0.7 ($\pm$ 0.1)	-0.7 ($\pm$ 0.5)	-0.6 ($\pm$ 0.5)	4.6	1.8/6.5
October	-0.9 ($\pm$ 0.1)	-1.0 ($\pm$ 0.1)	-0.6 ($\pm$ 0.1)	-0.9 ($\pm$ 0.5)	2.6	0.3/5.4
November	-	-	-1.1 ($\pm$ 0.1)	-0.4 ($\pm$ 0.6)	0.6	-0.5/1.3
December	-	-0.9 ($\pm$ 0.1)	-	-0.8 ($\pm$ 0.2)	-2.5	-5.7/0.6

increase); faults are not reactivated at the phase transition in April/May and November/December when Δp is at its minimum value;

- Figure 3.36 shows that M exhibits lower values as the pressure rises during gas injection, i.e. the magnitude during injection (negative values of $p - p_{avg}$ in Figure 3.36, with p the actual monthly pressure and p_{avg} the average pressure during the UGS period is smaller than during withdrawal (positive $p - p_{avg}$).

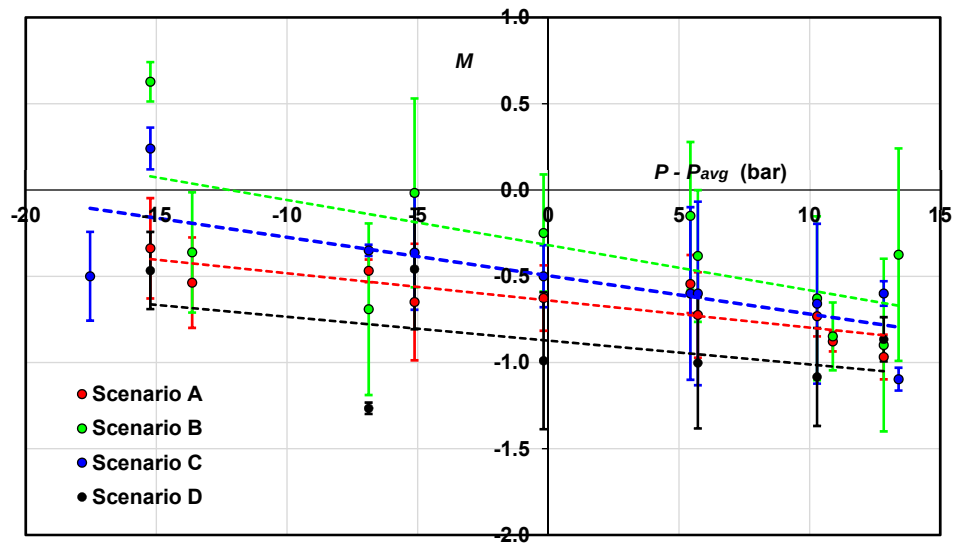


FIGURE 3.36: Case D: Average (bullet) and standard deviation (bar) of the magnitude M computed using the numerical results (equation (3.5)) and the Kanamori [1977] equation (3.3) during the UGS cycles. The M values for the four scenarios are plotted against the difference between the actual monthly p and the average pressure p_{avg} during this period. The dashed lines represent the linear regressions, colored according to the bullet/bar of the corresponding scenario.

Chapter 4

Efficient solvers for the linear problem

4.1 Introduction

A task of paramount importance in computational science and engineering is the solution of large sparse linear systems:

$$A\mathbf{x} = \mathbf{b} \tag{4.1}$$

where $A \in \mathbb{R}^{n \times n}$ is a non-singular matrix and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$. This kind of linear system arises not only from the discretization of partial differential equations (PDEs) of elliptic and parabolic type, but also from applications not governed by PDEs. The latter include the design and computer analysis of circuits, power system networks, chemical engineering processes, economics models, and queueing systems [[Benzi, 2002](#)].

Roughly speaking, to solve the linear system (4.1) there are two main methods: direct and iterative solvers. This traditional classification of solution methods is actually an oversimplification. First, the boundaries between the two classes of methods have become increasingly blurred; a certain number of ideas and techniques has been transferred from the area of sparse direct solvers (in the form of preconditioners) to the iterative field, with the result that iterative methods are becoming more and more reliable. Second, the field of iterative methods comprises a bewildering variety of techniques, ranging from truly iterative methods, such as the classical Jacobi, Gauss-Seidel and SOR iteration, to Krylov subspace methods, which theoretically converge in a finite number of steps in exact arithmetic, to multilevel methods, which combine both

techniques [Benzi, 2002].

In the present work, we use iterative solvers based on Krylov subspace methods. Such techniques are particularly appropriate for large-scale problems run on parallel architecture because of the ease of parallelization and the limited memory requirements. There are different algorithms according to the properties of the system matrix. The most popular methods are:

1. Conjugate Gradient (CG). The Conjugate Gradient method derives its name from the fact that it generates a sequence of A -conjugate vectors and that the residuals are mutually orthogonal. CG is effective when the coefficient matrix A is symmetric positive definite and is based on a short-term recurrence, thus allowing for a limited memory requirement [Hestenes and Stiefel, 1952].
2. Generalized Minimal Residual (GMRES). The Generalized Minimal Residual method computes a sequence of orthogonal vectors combining them through a least-squares solve and update. However, unlike CG, full GMRES is based on a long-term recurrence, requiring a large amount of storage. For this reason restarted and truncated versions are typically used. This method is used for general non-symmetric matrices [Saad and Schultz, 1986].
3. Bi-Conjugate Gradient Stabilized (Bi-CGstab). The Bi-Conjugate Gradient Stabilized method computes two sets of bi- A -conjugate vectors by orthogonalizing the residual to the Krylov subspace generated by the transposed of A . In particular, Bi-CGstab is a transpose-free stabilized version where the 2-norm of the residual vector is smoothed through a local steepest-descent step [Van der Vorst, 1992].

Other methods are: Minimum Residual (MINRES) [Paige and Saunders, 1975], Conjugate Gradient on the Normal Equations (CGNE and CGNR) [Craig, 1955], Bi-Conjugate Gradient (BiCG) [Lanczos, 1952; Fletcher, 1976], Quasi Minimal Residual (QMR) [Freund and Nachtigal, 1991], Conjugate Gradient Squared (CGS) [Sonneveld, 1989] and Chebyshev Iteration [Golub and Varga, 1961]. To complete this short list, we mention also a quite recently developed method, i.e. the Induced Dimension Reduction (IDR(s)) method [Sonneveld and Van Gijzen, 2008].

All Krylov subspace methods need the help of preconditioning to improve convergence [Benzi, 2002; Ferronato, 2012]. The term *preconditioning* refers to transforming the system (4.1) into another system with more favorable properties for the iterative solution. A *preconditioner* is the operator whose effect is such a transformation. Generally speaking, preconditioning attempts to improve the spectral properties of the coefficient matrix by clustering the eigenvalues around a non-zero value [Benzi, 2002].

The preconditioners can be classified into two main categories: algebraic or physically-based. In the former case, the algorithm needs to know the coefficient matrix only, independently of the problem it arises from. In the latter case, the method is built up for the specific problem. In general, the physically-based preconditioners can be more efficient but also less general. Algebraic preconditioners are usually more robust and more appropriate for general-purpose codes. There exists also the “Gray-box” technique, that is a third route for building an efficient preconditioner, with some features of both the physics-based and the purely algebraic approaches (see, e.g. [Carpentieri et al., 2005]).

Most known algebraic preconditioning techniques are [Ferronato, 2012]:

- Incomplete Factorizations: the basic idea of incomplete factorizations is to approximate the triangular factors L and U in a cost-effective way. Typically, if L and U are computed so as to satisfy $A = LU$, for example, by a Gaussian elimination procedure, fill-in takes place and L and U are much less sparse than A . The approximations $\tilde{L}$ and $\tilde{U}$ of L and U , respectively, can be obtained by simply discarding a number of fill-in entries according to some rules, e.g. Saad [1994]. This kind of computation is inherently sequential, nevertheless some parallelism can be gained through level-scheduling and other approaches, e.g. Anderson and Saad [1989]; Hysom and Pothen [1999]; Gupta [2017]. The issues concerning the existence and numerical stability can undermine the robustness of incomplete factorizations. Further developments based on incomplete factorizations are the multilevel ILUs preconditioners [Bollhöfer and Saad, 2006] and the recursive multilevel approaches, i.e. ARMS [Saad and Suchomel, 2002]. The latter one is provided also in a distributed version (pARMS, [Li et al., 2003]).

- **Approximate Inverses:** the idea is to provide an efficient sparse approximation of the inverse of A that can be applied to a vector simply by a matrix-vector product. This algorithm makes approximate inverses much more attractive for parallel computations. For example, an approximate inverse M^{-1} can be computed by minimizing the Frobenius norm $\|I - AM^{-1}\|_F$ over a set of matrices with a prescribed sparsity pattern [Frederickson, 1975; Grote and Huckle, 1997]. Another approach based on the bi-conjugation process, e.g. a two-sided Gram-Schmidt algorithm, is AINV [Benzi et al., 1996]. When matrix A is SPD, however, there is no guarantee that M^{-1} is SPD as well. For SPD matrices factorized sparse approximate inverses can be used, such as FSAI (Factorized Sparse Approximate Inverse) [Kolotilina and Yeregin, 1993];
- **Algebraic Multigrid Methods:** these methods can be seen as an acceleration of stationary iterative methods based on local relaxation, such as Gauss-Seidel and Jacobi methods. For linear systems arising from finite element discretization of elliptic boundary value problems, local relaxation methods were observed to converge very fast on the high frequency part of the solution. The low frequency part of the solution corresponds to a relatively smoother part of the function that can be well-approximated on a coarser grid. The main idea behind such multigrid methods is to project the error obtained after applying a few iterations of local relaxation methods onto a coarser grid. The projected error equations have two properties. First, the resulting system has a smaller size. Second, part of the low frequency error on a finer grid can be more easily corrected on the coarse grid. By repeating such a process and moving to coarser grids, a multilevel iterative process is obtained. For an overview of AMG methods see Stüben and Trottenberg [1982]; McBryan et al. [1991]; Trottenberg et al. [2000]; Xu and Zikatanov [2016].

Different preconditioners can be composed to form new ones, e.g., blending domain decomposition approaches [Dolean et al., 2015] with incomplete factorizations or approximate inverses, or using nested Krylov methods [Chan and Mathew, 1994; Janna et al., 2010a, 2013; McInnes et al., 2014].

In conclusion, to solve the linear system arising from the fault problem, we want to use the Bi-CGstab algorithm preconditioned with a technique based on the approximate inverses.

An important and computationally demanding task in the proposed model is the repeated solution of the linearized system (2.78). The matrix J is non-symmetric and has a generalized saddle-point structure. It can be written as:

$$J = \begin{bmatrix} K + E & C - F \\ C^T & 0 \end{bmatrix} = \begin{bmatrix} A & B_1 \\ B_2^T & 0 \end{bmatrix} \quad (4.2)$$

where $A = K + E$ is an SPD block (see proposition 2.4.1) and $B_1 \neq B_2$. In the remainder of this chapter, we will use the form defined in (4.2), assuming that B_1 and B_2 have full rank.

The solution to a saddle point problem frequently arises in several engineering applications, e.g. computational fluid dynamics, constrained and weighted least squares estimation, constrained optimization, etc. [Benzi et al., 2005]. An effective approach to solve a linear system with matrix (4.2) is to use a Krylov subspace method, with the so-called constraint preconditioning. The constraint preconditioning [Keller et al., 2000; Benzi et al., 2005; Bai et al., 2009; Bergamaschi, 2012] is based on the exploitation of the native 2×2 block structure of the original saddle point matrix. This approach has been extensively used in the solution of several problems, e.g. arising from mixed finite element formulations of elliptic partial differential equations [Bank et al., 1989; Tong and Sameh, 1998; Ewing et al., 1990; Perugia and Simoncini, 2000; Rozložník and Simoncini, 2002; Axelsson and Neytcheva, 2003; Mihajlović and Silvester, 2004], saddle point “KKT” systems in optimization [Dyn and Ferguson, 1983; Freund and Nachtigal, 1995; Shi, 1995; Luksan and Vlcek, 1998; Keller et al., 2000; Gould et al., 2001; Bergamaschi et al., 2004] or coupled Biot’s consolidation [Bergamaschi et al., 2007, 2008; Ferronato et al., 2009, 2010a; Janna et al., 2012b] and Navier-Stokes equations [Silvester et al., 2001; Elman et al., 2002].

In this work, we consider a constraint preconditioner M_C^{-1} with the following structure:

$$M_C^{-1} = \begin{bmatrix} \tilde{A} & B_1 \\ B_2^T & 0 \end{bmatrix}^{-1} \quad (4.3)$$

where $\tilde{A}$ is an approximation of the original leading block A (see Eq. (4.2)). The key point for the M_C^{-1} effectiveness is the choice of $\tilde{A}$ and the approximation of S , i.e. the Schur complement.

In the remainder, we discuss the possibilities to approximate the inverse of A , i.e. the creation of $\tilde{A}$, remembering that the leading block is an SPD matrix. We focus on preconditioners based on the factorized approximate inverse approach. Then, different ways to approximate the Schur complement computation are proposed and, finally, the numerical results are presented and discussed.

4.2 Factorized sparse approximate inverse for SPD matrices

For approximating A^{-1} , we focus on the FSAI preconditioner for SPD matrices. The use of a factorized approximate inverse preconditioner allows for the explicit computation of the Schur complement through matrix-matrix products. In fact, setting $A^{-1} = G^T G$, matrix (4.3) can be factorized as follows:

$$M_C = \begin{bmatrix} G^{-1} & 0 \\ H_2 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \tilde{S} \end{bmatrix} \begin{bmatrix} G^{-T} & H_1 \\ 0 & I \end{bmatrix} = \begin{bmatrix} G^{-1}G^{-T} & G^{-1}H_1 \\ H_2G^{-T} & \tilde{S} + H_2H_1 \end{bmatrix} \quad (4.4)$$

Equating (4.4) to (4.3), we can write the expressions for the matrices H_1 , H_2 and $\tilde{S}$:

$$H_1 = GB_1, \quad H_2 = B_2G^T \quad \text{and} \quad \tilde{S} = -H_2H_1 \quad (4.5)$$

A recent and efficient variant of FSAI is the adaptive version, denoted by *aFSAI*, advanced by Huckle [2003]; Janna et al. [2015b]. The basic idea consists of building a sparse approximation G of the inverse of the exact lower Cholesky factor L of A by

minimizing the following Frobenius norm:

$$\|I - GL\|_F \rightarrow \min \quad (4.6)$$

over the set of matrices having some prescribed lower triangular non-zero pattern. The idea, originally developed by [Kolotilina and Yereimin, 1993], has been extended following [Janna and Ferronato, 2011] so as to select dynamically and optimally the lower triangular non-zero pattern of G . The advantages of such an approach are manyfold:

1. the algorithm building and applying G is very robust, as it does not suffer from the typical numerical instabilities characterizing incomplete factorizations;
2. the computation of G can be done independently row-by-row, thus being perfectly parallelizable;
3. the convergence ensured by adaptive FSAI is typically very fast with SPD structural stiffness matrices.

In this work, we will refer to the aFSAI implementation available in the FSAIPACK software package (Algorithm 3 in Janna et al. [2015b]). Formally, the application of this Algorithm is denoted by the following function call:

$$G = \text{aFSAI}(A, k_{max}, s, \tau, \epsilon) \quad (4.7)$$

where G is such that $A^{-1} \simeq G^T G$ and $(k_{max}, s, \tau, \epsilon)$ is the set of user-specified parameters required by the aFSAI function:

- k_{max} : integer value denoting the maximum number of steps of the inner adaptive procedure for the dynamic construction of G non-zero pattern;
- s : integer value denoting the number of entries added to the G non-zero pattern at every step of the adaptive procedure;
- τ : real tolerance below which a new entry in G is dropped;
- ϵ : real tolerance to exit the adaptive procedure before achieving k_{max} steps.

To understand the meaning of the user-specified parameters we introduce in detail both the static and the adaptive FSAI algorithm. FSAI is computed by minimizing the Frobenius norm (4.6) over all $G \in \mathcal{W}_S$, where $\mathcal{W}_S$ is the set of matrices sharing the same lower triangular nonzero pattern S . Differentiating (4.6) with respect to G entries, $[G]_{ij}$, and setting to zero, we have:

$$[GA]_{ij} = [L^T]_{ij} \quad \forall (i, j) \in S \quad (4.8)$$

Since L^T is upper triangular while S is a lower triangular pattern, the matrix equation (4.8) can be rewritten as:

$$[GA]_{ij} = \begin{cases} 0 & i \neq j, (i, j) \in S \\ l_{ii} & i = j \end{cases} \quad (4.9)$$

L is unknown, so G cannot be computed from (4.9). Instead, it is possible to compute $\tilde{G}$ such that:

$$[\tilde{G}A]_{ij} = \delta_{ij} \quad \forall (i, j) \in S \quad (4.10)$$

where δ_{ij} is Kronecker's delta. The FSAI factor G is finally computed by scaling $\tilde{G}$ with a diagonal matrix D :

$$G = D\tilde{G} \quad (4.11)$$

The choice of D is not unique. It can be shown that if the diagonal entries of $P = GAG^T$ are unitary, the Kaporin number of the preconditioned matrix P :

$$\kappa = \frac{\frac{1}{n} \text{tr}(GAG^T)}{\det(GAG^T)^{\frac{1}{n}}} \quad (4.12)$$

is minimum over all matrices belonging to $\mathcal{W}_S$ [Kaporin, 1990, 1994]. The Kaporin number is an alternative measurement of the conditioning for symmetric positive definite matrices. According to the definition of Kaporin [1994], the estimated number of PCG iterations k_H needed to reduce of ε times the residual norm $\|r_k\|_2$ is:

$$k_H \leq \left\lceil n \log_2 \kappa + \log_2 \frac{1}{\varepsilon} \right\rceil \quad (4.13)$$

A unitary diagonal on P can be enforced by setting:

$$D = \left[\text{diag} \left(\tilde{G} \right) \right]^{-1/2} \quad (4.14)$$

One of the main difficulties in the computation of a good approximation G of L^{-1} is given by the selection of $\mathcal{S}$. The optimal a priori choice of an appropriate nonzero pattern is a difficult task to perform. This task has been addressed either statically [Huckle, 1999; Chow, 2000] or dynamically [Huckle, 2003; Janna and Ferronato, 2011].

4.2.1 Static FSAI

Assume that the nonzero pattern $\mathcal{S}$ is given. Denote by $\mathcal{P}_i$ the set of column indices belonging to the i -th row of $\mathcal{S}$:

$$\mathcal{P}_i = \{j : (i, j) \in \mathcal{S}\} \quad (4.15)$$

and denote $m_i = |\mathcal{P}_i|$. Let $A[\mathcal{P}_i, \mathcal{P}_i]$ be the submatrix of A collecting the entries a_{kl} such that $k, l \in \mathcal{P}_i$. Now, equation (4.10) can be rewritten as:

$$A[\mathcal{P}_i, \mathcal{P}_i] \tilde{\mathbf{g}}_i = \mathbf{e}_{m_i}, \quad i \in \{1, 2, \dots, n\} \quad (4.16)$$

where $\tilde{\mathbf{g}}_i$ is the dense vectors containing the nonzero coefficients of the i -th row of $\tilde{G}$ and $\mathbf{e}_{m_i}$ is the m_i -th vector of the canonical basis of $\mathbb{R}^{m_i}$. The row $\mathbf{g}_i$ of G is obtained by scaling $\tilde{\mathbf{g}}_i$ with the square root of its last entry:

$$\mathbf{g}_i = \frac{\tilde{\mathbf{g}}_i}{\sqrt{\tilde{g}_{i, m_i}}} \quad (4.17)$$

Traditionally, for the static pattern generation the nonzero pattern of small powers of A is exploited. The number of entries in $\mathcal{S}$ grows so quickly with the power exponent of A , so that only the second or third power can be used in real world applications. On the other hand, in order to include in $\mathcal{S}$ entries also belonging to higher powers of A prefiltration of A may be recommended [Chow, 2000, 2001].

The computational cost is dominated by the solution of dense linear systems (4.16). Denoting by $\overline{m}_G$ the average number of non-zero entries per row, the computational cost is proportional to the third power of $\overline{m}_G$:

$$c \propto n \cdot \overline{m}_G^3 = n \cdot \overline{m}_A^3 \mu_G^3 \quad (4.18)$$

where $\overline{m}_A$ is the average number of non-zero entries per row of the original matrix and μ_G is the preconditioner density, defined as the ratio between the number of nonzeros in G and that in A :

$$\mu_G = \frac{\text{nnz}(G)}{\text{nnz}(A)} \quad (4.19)$$

4.2.2 Dynamic FSAI

Adaptive procedures have been developed in order to dynamically generate S during the FSAI computation, with the aim of selecting in each row the “best” nonzero positions [Grote and Huckle, 1997; Huckle, 2003; Janna and Ferronato, 2011; Janna et al., 2015b].

Assume that an initial guess G_0 , complying with equations (4.10) and (4.11), is given and let S_0 be its nonzero pattern. A natural initial guess is $G_0 = \text{diag}(A)^{-1}$, with S_0 the diagonal pattern [Janna et al., 2015b]. Now, we write $G_0 = \tilde{D}_0 \tilde{G}_0$, with $\tilde{D}_0 = \text{diag}(G_0)$. Janna and Ferronato [2011] proved that the entries of $\tilde{G}_0$ can be directly computed by solving the dense linear systems:

$$A \begin{bmatrix} \overline{\mathcal{P}}_i^0 & \overline{\mathcal{P}}_i^0 \end{bmatrix} \tilde{\mathbf{g}}_{0,i} = -A \begin{bmatrix} \overline{\mathcal{P}}_i^0 & i \end{bmatrix}, \quad i \in \{1, 2, \dots, n\} \quad (4.20)$$

where $\overline{\mathcal{P}}_i^0 = \{j : (i, j) \in S_0\}$, $\overline{\mathcal{P}}_i^0 = \overline{\mathcal{P}}_i^0 \setminus i$ and $\tilde{\mathbf{g}}_{0,i}$ is the dense vector collecting the nonzero off-diagonal terms in the i -th row of $\tilde{G}_0$. Now, we consider the Kaporin conditioning number κ of the preconditioned matrix $G_0 A G_0^T$:

$$\kappa = \frac{\frac{1}{n} \text{tr}(G_0 A G_0^T)}{\det(G_0 A G_0^T)^{\frac{1}{n}}} = \frac{\frac{1}{n} \text{tr}(\tilde{D}_0 \tilde{G}_0 A \tilde{G}_0^T \tilde{D}_0)}{\det(\tilde{D}_0 \tilde{G}_0 A \tilde{G}_0^T \tilde{D}_0)^{\frac{1}{n}}} \quad (4.21)$$

Recalling that $G_0 A G_0^T$ has unitary diagonal entries and $\det(\tilde{G}_0) = 1$, it follows that:

$$\kappa = \frac{1}{\det(A)^{\frac{1}{n}}} \frac{1}{\det(\tilde{D}_0)^{\frac{2}{n}}} = \frac{\det\left[\text{diag}\left(\tilde{G}_0 A \tilde{G}_0^T\right)\right]^{\frac{1}{n}}}{\det(A)^{\frac{1}{n}}} \quad (4.22)$$

Denoting by $\tilde{G}_0[i, :]$ the i -th row of $\tilde{G}_0$, the numerator of Eq. (4.22) reads:

$$\det\left[\text{diag}\left(\tilde{G}_0 A \tilde{G}_0^T\right)\right] = \prod_{i=1}^n \tilde{G}_0[i, :] A \tilde{G}_0[i, :]^T = \prod_{i=1}^n \psi_{0,i} \quad (4.23)$$

with $\psi_{0,i} = \tilde{G}_0[i, :] A \tilde{G}_0[i, :]^T$. Thus, from (4.22) and (4.23), we can write:

$$\kappa^n = \frac{1}{\det(A)} \prod_{i=1}^n \psi_{0,i} \quad (4.24)$$

The target is to find an augmented pattern $\mathcal{S}_1$ allowing for a reduction of κ^n , defined in (4.24). The idea is to compute the gradient of κ^n with respect to $\tilde{G}_0$ and add in $\mathcal{S}_1$ the positions corresponding to its largest components. Notice that each $\psi_{0,i}$ in (4.24) depends on the i -th row of $\tilde{G}_0$ only, so the adaptive pattern generation can be efficiently carried out in parallel according to a row-wise scheme. Denoting by $\nabla\psi_{0,i}$ the gradient of $\psi_{0,i}$ with respect to the entries of $\tilde{G}_0[i, :]$, its components are:

$$\frac{\partial\psi_{0,i}}{\partial\tilde{g}_{0,ij}} = 2 \left(\sum_{r=1}^{i-1} a_{jr} \tilde{g}_{0,ir} + a_{ji} \right), \quad \forall j = 1, \dots, i-1 \quad (4.25)$$

According to (4.25), we can compute the gradient of the i -th contribution to the Kaporin conditioning number.

The nonzero pattern $\overline{\mathcal{P}}_i^0$ is enlarged by adding the s positions corresponding to the largest entries of $\nabla\psi_{0,i}$, to obtain $\overline{\mathcal{P}}_i^1$. This procedure can be repeated k times, until the achievement of a given criterion on the maximum number of entries per row or on the Kaporin conditioning number relative reduction. To reduce the computational cost of each iteration, it is also possible to enforce some dropping by neglecting entries of G smaller than the tolerance τ .

As to the computational cost, each iteration requires the same operations as the static FSAI algorithm plus the computation of $\nabla\psi_{k,i}$. Denoting by $m_{k,i}$ the number of

nonzeroes in the i -th row of $\tilde{G}_k$, the cost for computing $\nabla\psi_{k,i}$ is on the average proportional to $\overline{m}_A \cdot m_{k,i}$. The costs for gathering and solving the dense linear subsystems are proportional to $m_{k,i}^2$ and $m_{k,i}^3$, respectively. After k_{iter} iterations of this procedure starting from $\tilde{G}_0 = I$ without any dropping, the cost c_i for computing the i -th row is:

$$\begin{aligned} c_i &\propto \sum_{k=1}^{k_{max}} \left[\overline{m}_A \cdot sk + (sk)^2 + (sk)^3 \right] = \\ &= (\alpha_0 \overline{m}_A + \alpha_1) k_{max} + (\alpha_0 \overline{m}_A + \alpha_2) k_{max}^2 + \alpha_3 k_{max}^3 + \alpha_4 k_{max}^4 \end{aligned} \quad (4.26)$$

where $\alpha_i, i = 0, \dots, 4$ are five scalar coefficients depending upon s , i.e. the number of entries at each step of the adaptive procedure. For k_{max} large enough, that is approximately $k_{max} \geq \sqrt{\overline{m}_A}$, the overall cost of the adaptive procedure is proportional to the fourth power of the preconditioner density [Janna et al., 2015b]:

$$c = \sum_{i=1}^n c_i \propto n \cdot k_{max}^4 = n \cdot \overline{m}_A^4 \mu_G^4 \quad (4.27)$$

Therefore, the set-up cost of aFSAI grows very rapidly when the preconditioner density increases.

4.3 Multilevel approaches for FSAI preconditioning

In this section, we describe three new approaches to build up a factorized preconditioner suitable for SPD matrices. These methods are based on FSAI, but they introduce some kind of sequentiality which can be employed if A has a multilevel structure. It is well known that incomplete factorizations perform better than approximate inverses as preconditioner for Krylov methods in the sense that they yield a faster convergence with the same density. As the former are inherently sequential algorithms, they cannot exploit the growing parallelism of modern architectures. By distinction, FSAI is embarrassingly parallel. Thus, an idea is to mix the features of both methods by adding some sequentiality to aFSAI and making it more similar to an incomplete factorization. Similar ideas can be found for instance in Chow and Saad [1998]; Raghavan and Teranishi [2010].

We can distinguish three multilevel approaches:

1. Block Tridiagonal FSAI (BTFSAI);
2. Domain Decomposition FSAI (DDFSAI);
3. Multilevel FSAI with Low Rank corrections (MFLR).

4.3.1 Block Tridiagonal FSAI

With the BTFSAI approach, the matrix A is first reordered by the reverse Cuthill-McKee algorithm [Cuthill and McKee, 1969], with the aim to reduce its bandwidth, then it is divided into a block tridiagonal structure according to a given number of blocks. A block LDU decomposition of this new matrix is calculated, where the inverses of the diagonal blocks are approximated explicitly by the adaptive FSAI algorithm. With this approach, the global matrix is subdivided into small blocks, which could be approximated with less effort. The sequentiality of this method increases with the number of blocks, while parallelism can be exploited at the block level by the use of aFSAI.

Calling A_i the SPD diagonal submatrix and B_i the upper diagonal submatrix of the i -th block, the SPD matrix A can be written as (see Figure 4.1):

$$A = \begin{bmatrix} A_1 & B_1 & & & \\ B_1^T & A_2 & B_2 & & \\ & \ddots & \ddots & \ddots & \\ & & B_{n-2}^T & A_{n-1} & B_{n-1} \\ & & & B_{n-1}^T & A_n \end{bmatrix} \quad (4.28)$$

Computing the block LDU decomposition of the reordered matrix (4.28), we have:

$$A = \begin{bmatrix} I_1 & & & & \\ B_1^T S_1^{-1} & I_2 & & & \\ & \ddots & \ddots & & \\ & & B_{n-1}^T S_{n-1}^{-1} & I_n \end{bmatrix} \begin{bmatrix} S_1 & & & & \\ & S_2 & & & \\ & & \ddots & & \\ & & & S_n \end{bmatrix} \begin{bmatrix} I_1 & S_1^{-1} B_1 & & & \\ & I_2 & S_2^{-1} B_2 & & \\ & & \ddots & \ddots & \\ & & & I_n \end{bmatrix} \quad (4.29)$$

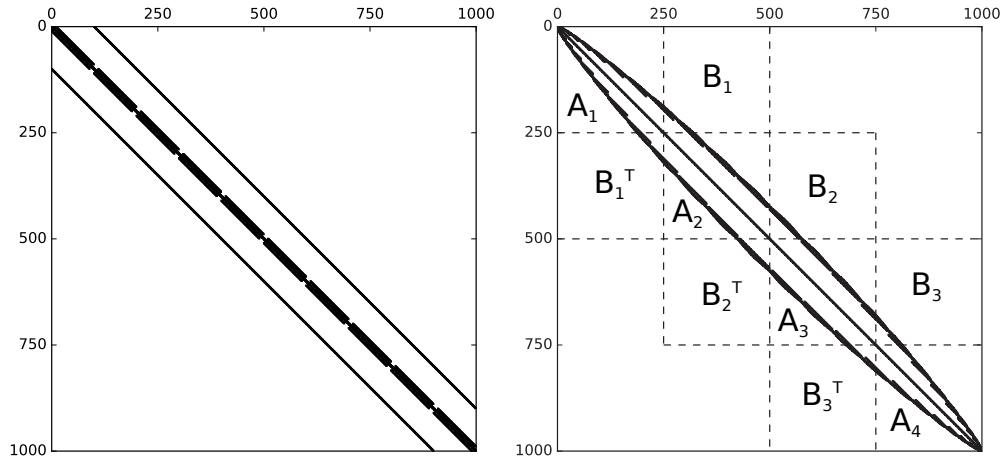


FIGURE 4.1: Sparsity pattern of the original test case on the left and the reordered pattern employed by the BTFSAI preconditioner on the right.

where S_i is the i -th Schur complement, defined as:

$$S_1 = A_1, \quad S_i = A_i - B_{i-1}^T S_{i-1}^{-1} B_{i-1}, \quad i = 2, \dots, n \quad (4.30)$$

According to (4.30), the inverse of S_{i-1} is needed to compute S_i . We can replace S_{i-1}^{-1} with its FSAI approximation $F_i^T F_i$. Denoting by H_i the matrix $F_i B_i$, the Schur complement (4.30) can be calculated recursively as:

$$S_i = A_i - H_{i-1}^T H_{i-1}, \quad i = 2, \dots, n \quad (4.31)$$

The BTFSAI preconditioner M^{-1} of the matrix A reads:

$$M^{-1} = \begin{bmatrix} F_1^T & -F_1^T H_1 F_2^T & & & \\ & F_2^T & -F_2^T H_2 F_3^T & & \\ & & \ddots & \ddots & \\ & & & F_n^T & \end{bmatrix} \begin{bmatrix} F_1 & & & & \\ -F_2 H_1^T F_1 & F_2 & & & \\ & \ddots & \ddots & & \\ & & -F_n H_n^T F_{n-1} & F_n & \end{bmatrix} \quad (4.32)$$

Usually no dropping is necessary to store efficiently H_i and S_i , because F_i is very sparse and the matrix-matrix products $F_i B_i$ and $H_i^T H_i$ can be thoroughly computed.

The procedure to build the BTFSAI preconditioner is summarized in the Algorithm 1 and its application to an arbitrary vector $\mathbf{v}$ is presented in the Algorithm 2.

Algorithm 1 CONSTRUCTION OF THE BTFSAI PRECONDITIONER

```

1: procedure CPTBTFSAI( $A, n$ )
2:   Compute  $A_1^{-1} \simeq F_1^T F_1$ , the approximate inverse of  $A_1$ 
3:   for  $i = 2, \dots, n$  do
4:     Compute  $H_{i-1} = F_{i-1} B_{i-1}$ 
5:     Compute  $S_i = H_{i-1}^H H_{i-1}$ 
6:     Compute  $S_i^{-1} \simeq F_i^T F_i$ , the approximate inverse of the Schur complement
7:   end for
8: end procedure

```

Algorithm 2 CALCULATION OF $\mathbf{u} = M^{-1}\mathbf{v}$

```

1: procedure APPLYBTFSAI( $F_i, H_i, \mathbf{v}$ )
2:    $\mathbf{w}_1 = F \mathbf{v}_1$ 
3:   for  $i = 2, \dots, n$  do
4:      $\mathbf{w}_i = F_i (\mathbf{v}_i - H_{i-1}^T \mathbf{w}_{i-1})$ 
5:   end for
6:    $\mathbf{u}_n = F_n^T \mathbf{w}_n$ 
7:   for  $i = n-1, \dots, 1$  do
8:      $\mathbf{u}_i = F_i^T (\mathbf{w}_i - H_i \mathbf{u}_{i+1})$ 
9:   end for
10:  return  $\mathbf{u}$ 
11: end procedure

```

The main drawback of this approach is the need for computing $(n-1)$ Schur complements along with their FSAI approximations. Although these matrices should be SPD from a theoretical point of view, the use of approximate inverses for their computation may cause the loss of this property. Indeed, we can have indefinite Schur complements if the approximate inverses are not accurate enough. In this case, the construction of the preconditioner should be restarted allowing for a larger number of nonzero entries per row [Paludetto Magri et al., 2017].

4.3.2 Domain Decomposition FSAI

Domain decomposition refers to a number of techniques which solve a problem defined on a large domain Ω via the solution of smaller problems defined over subdomains $\Omega_k, k = 1, \dots, n$. Each problem can be solved independently and the solutions can be gathered to build the outcome of the original problem. For more details, see Chan and Mathew [1994]; Dolean et al. [2015].

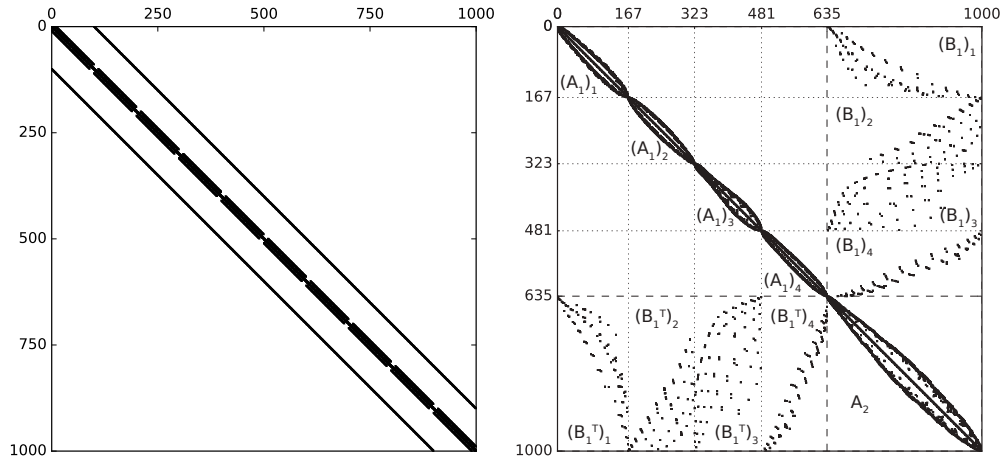


FIGURE 4.2: Sparsity pattern of the original test case on the left and the reordered pattern employed by the DDFSAI preconditioner on the right.

In the domain decomposition FSAI preconditioner (DDFSAI) the graph of the original matrix is reordered according to the k -way partition algorithm implemented in the *METIS* software library [Karypis and Kumar, 1998]. After this, each independent block of the matrix is reordered according to the reverse Cuthill-McKee algorithm [Cuthill and McKee, 1969] in order to reduce its bandwidth.

Let A^* be the reordered matrix (see Figure 4.2):

$$A^* = \begin{bmatrix} A_1 & B_1 \\ B_1^T & A_2 \end{bmatrix} = \begin{bmatrix} (A_1)_1 & & & (B_1)_1 \\ & (A_1)_2 & & (B_1)_2 \\ & & \ddots & \vdots \\ & & & (A_1)_n & (B_1)_n \\ (B_1^T)_1 & (B_1^T)_2 & \cdots & (B_1^T)_n & A_2 \end{bmatrix} \quad (4.33)$$

where the submatrices $(A_1)_i$ represent the communications between the internal nodes of the i -th subdomain for $i = 1, 2, \dots, n$, A_2 the communications between interface nodes and $(B_1)_i$ the coupling between the interface and internal nodes belonging to the i -th subdomain.

The block LDU decomposition of (4.33) is given by:

$$A^* = \begin{bmatrix} I_1 & \\ B_1^T A_1^{-1} & I_2 \end{bmatrix} \begin{bmatrix} A_1 & \\ & S_2 \end{bmatrix} \begin{bmatrix} I_1 & A_1^{-1} B_1 \\ & I_2 \end{bmatrix} \quad (4.34)$$

thus, the exact inverse of A^* can be written as:

$$(A^*)^{-1} = \begin{bmatrix} I_1 & -A_1^{-1}B_1 \\ & I_2 \end{bmatrix} \begin{bmatrix} A_1^{-1} \\ S_2^{-1} \end{bmatrix} \begin{bmatrix} I_1 \\ -B_1^T A_1^{-1} & I_2 \end{bmatrix} \quad (4.35)$$

The computation of the inverses A_1^{-1} and S_2^{-1} is approximated by the adaptive FSAI algorithm, giving rise to the following preconditioner for A^* :

$$M^{-1} = \begin{bmatrix} I_1 & -F_1^T F_1 B_1 \\ & I_2 \end{bmatrix} \begin{bmatrix} F_1^T F_1 \\ F_2^T F_2 \end{bmatrix} \begin{bmatrix} I_1 \\ -B_1^T F_1^T F_1 & I_2 \end{bmatrix} \quad (4.36)$$

Defining, as before, $H_1 = F_1 B_1$, the expression (4.36) can be rewritten as:

$$M^{-1} = \begin{bmatrix} F_1^T & -F_1^T H_1 F_2^T \\ & F_2^T \end{bmatrix} \begin{bmatrix} F_1 \\ -F_2 H_1^T F_1 & F_2 \end{bmatrix} \quad (4.37)$$

The algorithm for constructing the DDFSAI preconditioner is almost the same as Algorithm 1 restricted to two blocks, with the only difference consisting in the initial ordering of the matrix. The same argument applies also to the preconditioner application that can be summarized according to the Algorithm 2 restricted to two blocks only.

In this algorithm, only one Schur complement is computed, so it is less likely that the matrix S_2 becomes indefinite and the preconditioner calculation encounters a breakdown.

4.3.3 Multilevel FSAI Low Rank preconditioner

In this section, we present a new approach, that is free from the possibility to encounter a breakdown due to the indefiniteness of the approximate Schur complement.

Consider a standard multilevel approach applied to the SPD matrix A for a total number n_l of levels. The matrix A_l obtained at any level $l \in [0, n_l - 1]$ is the Schur complement of the previous level. Following for instance [Janna et al., 2009], we can

partition A_l in four blocks and perform the factorization:

$$A_l = \begin{bmatrix} K & B \\ B^T & C \end{bmatrix} = \begin{bmatrix} L_K & \\ B^T L_K^{-T} & I \end{bmatrix} \begin{bmatrix} I & \\ & S \end{bmatrix} \begin{bmatrix} L_K^T & L_K^{-1} B \\ & I \end{bmatrix} \quad (4.38)$$

where L_K is the exact Cholesky factor of block K , $K \in \mathbb{R}^{n_1 \times n_1}$, and $S = C - B^T K^{-1} B$, $S \in \mathbb{R}^{n_2 \times n_2}$, is the Schur complement of A_l with respect to the partition n_1, n_2 . Our goal is to get a preconditioner, so we can replace each block at the right-hand side of (4.38) with an approximation:

$$\tilde{L} \simeq L_K, \quad \tilde{H} \simeq L_K^{-1} B, \quad \tilde{S} \simeq S \quad (4.39)$$

A popular choice is to use an incomplete factorization as the main kernel for the approximations in (4.39), i.e., setting $\tilde{L}$ equal to the Incomplete Cholesky (IC) factor of K as in Saad and Suchomel [2002]; Li et al. [2003]; Janna et al. [2009]; Bu et al. [2016]. In the same framework, the use of FSAI as the main kernel is straightforward. After computing G such that $G^T G \simeq K^{-1}$, the approximations in (4.39) become:

$$\tilde{L} \simeq G^{-1}, \quad \tilde{H} \simeq GB, \quad \tilde{S} \simeq C - \tilde{H}^T \tilde{H} \quad (4.40)$$

If the matrix A has a block tridiagonal structure, the choice (4.40) brings to the already presented BTFSAI algorithm.

The recursive computation of (4.38) with the approximations (4.39) provides the general framework for a multilevel preconditioner M of A , made by the list of factors $M_l, l \in [0, n_l]$ (Algorithm 3). The application phase of this preconditioner is reported in Algorithm 4.

As already observed, this straightforward implementation of a multilevel FSAI preconditioner is very prone to breakdowns. In fact, the Schur complement approximation $S_i = A_i - H_{i-1}^T H_{i-1}$ is computed as the difference between two SPD matrices and can be indefinite. Such an experience is quite common even in relatively well-conditioned problems. The reason for this resides in the poor approximation of the leftmost eigenvalues usually obtained by FSAI. Fig. 4.3 compares the eigenspectra of the `bcsstk16`

Algorithm 3 MULTILEVEL FACTORIZATION SET-UP

```

1: procedure ML_SETUP( $n_l, A$ )
2:   Set  $A_0 = A$ 
3:   for  $l = 0, \dots, n_l - 2$  do
4:     Partition  $A_l$  as  $\begin{bmatrix} K & B \\ B^T & C \end{bmatrix}$ 
5:     Compute  $\tilde{L}$  such that  $\tilde{L}\tilde{L}^T \simeq K$ 
6:     Compute  $\tilde{H}$  such that  $\tilde{H} \simeq L_K^{-1}B$ 
7:     Compute  $\tilde{S}$  such that  $\tilde{S} \simeq C - B^T K^{-1}B$ 
8:     Form  $M_l = \begin{bmatrix} \tilde{L} & 0 \\ \tilde{H}^T & I \end{bmatrix}$ 
9:     Set  $A_{l+1} = \tilde{S}$ 
10:   end for
11:   Compute  $M_{n_l-1}$  such that  $M_{n_l-1}M_{n_l-1}^T \simeq A_{n_l-1}$ 
12:   Set  $M = \{M_0, M_1, \dots, M_{n_l-1}\}$ 
13: end procedure

```

Algorithm 4 MULTILEVEL FACTORIZATION APPLICATION

```

1: procedure ML_APPLY( $n_l, M, \mathbf{v}$ )
2:   Set  $lev_{start} = 1$ 
3:   for  $l = 0, \dots, n_l - 2$  do
4:     Solve  $M_l \mathbf{z} = \mathbf{v}$ 
5:     Set  $lev_{end} = lev_{start} + n_1 - 1$ 
6:     Form  $\mathbf{w} (lev_{start} : lev_{end}) = \mathbf{v} (1 : n_1)$ 
7:     Set  $\mathbf{v} = \mathbf{v} (n_1 + 1 : n)$ 
8:     Update  $lev_{start} = lev_{end} + 1$ 
9:   end for
10:  Solve  $(M_{n_l-1}M_{n_l-1}^T) \mathbf{z} = \mathbf{v}$ 
11:  Form  $\mathbf{w} (lev_{start} : n) = \mathbf{z}$ 
12:  Set  $lev_{end} = lev_{start} - 1$ 
13:  for  $l = n_l - 2, \dots, 0$  do
14:    Set  $lev_{start} = lev_{end} - n_1 + 1$ 
15:    Retrieve  $\mathbf{w}_l = \mathbf{w} (lev_{start} : n)$ 
16:    Solve  $M_l^T \mathbf{z} = \mathbf{w}_l$ 
17:    Form  $\mathbf{w} (lev_{start} : n) = \mathbf{w}_l$ 
18:    Update  $lev_{end} = lev_{start} - 1$ 
19:  end for
20:  return  $\mathbf{w}$ 
21: end procedure

```

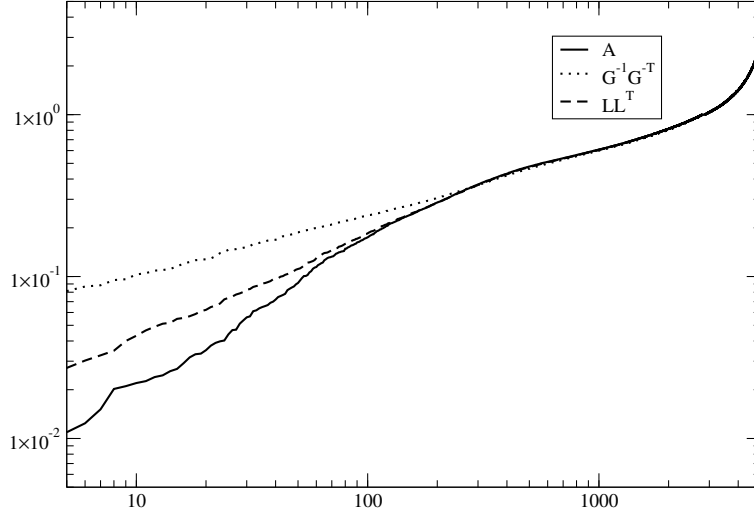


FIGURE 4.3: Comparison between the eigenspectra of A , $\tilde{L}\tilde{L}^T$ and $(G^TG)^{-1}$ for the bcsstk16 matrix from the University of Florida sparse matrix collection.

matrix (structural problem, $n = 4,884$ with 290,378 non-zero entries) from the University of Florida sparse matrix collection [Davis and Hu, 2011] with its approximations by IC and FSAI, $\tilde{L}\tilde{L}^T$ and $(G^TG)^{-1}$, respectively, having approximatively the same number of non-zeros. Both IC and FSAI are not able to capture the smallest eigenvalues of a matrix, though IC is generally better. The smallest eigenvalues of K are the largest of K^{-1} , and therefore control the most significant entries of $B^TK^{-1}B$. As a consequence, H computed as in (4.31) often fails in approximating accurately the largest entries of $B^TK^{-1}B$, leading to the appearance of negative eigenvalues in S and causing the breakdown of the procedure. Although some stabilization techniques can help increase the IC-based multilevel preconditioner robustness, e.g., [Janna et al., 2009; Scott and Tůma, 2014], they do not give satisfactory results with FSAI.

The robustness of the multilevel FSAI (MF) preconditioner can be ensured by computing $\tilde{S}$ so as to be SPD independently of G . To this aim, we can use the following result.

Theorem 4.3.1 *Let $A \in \mathbb{R}^{n \times n}$ be an SPD (2×2) -block matrix:*

$$A = \begin{bmatrix} K & B \\ B^T & C \end{bmatrix} \quad (4.41)$$

with $K \in \mathbb{R}^{n_1 \times n_1}$, $B \in \mathbb{R}^{n_1 \times n_2}$ and $C \in \mathbb{R}^{n_2 \times n_2}$, and $V \in \mathbb{R}^{n \times n_2}$ and $D \in \mathbb{R}^{n_2 \times n}$ the 2-block rectangular matrices:

$$V = \begin{bmatrix} F^T \\ I \end{bmatrix}, \quad D = \begin{bmatrix} 0 & Z \end{bmatrix} \quad (4.42)$$

with $F \in \mathbb{R}^{n_2 \times n_1}$ and $Z \in \mathbb{R}^{n_2 \times n_2}$, such that the Frobenius norm $\|D - V^T L\|_F$ is minimum for any Z , L being the lower Cholesky factor of A . Then, $S = V^T A V$ is the Schur complement of A with respect to the partition (n_1, n_2) .

Proof. Recalling equation (4.38), the lower Cholesky factor of A reads:

$$L = \begin{bmatrix} L_K & 0 \\ B^T L_K^{-T} & L_S \end{bmatrix} \quad (4.43)$$

where L_S is the lower Cholesky factor of the Schur complement of A , i.e., $L_S L_S^T = C - B^T K^{-1} B$. The matrix $(D - V^T L)$ is therefore:

$$D - V^T L = \begin{bmatrix} F L_K + B^T L_K^{-1} & Z - L_S \end{bmatrix} \quad (4.44)$$

whose Frobenius norm is minimum for any Z if $F L_K + B^T L_K^{-1} = 0$, i.e.:

$$F = -B^T K^{-1} \quad (4.45)$$

The matrix $S = V^T A V$ reads:

$$S = F K F^T + B^T F^T + B F + C \quad (4.46)$$

Introducing equation (4.45) into (4.46) provides:

$$V^T A V = C - B^T K^{-1} B \quad (4.47)$$

□

Remark. The matrix F of the previous Theorem 4.3.1 is generally dense. If equation (4.45) is enforced only for the entries located in a prescribed set of positions $\tilde{\mathcal{S}} \subset \mathcal{S} =$

$\{(i, j) : 1 \leq i \leq n_1, 1 \leq j \leq n_2\}$, the sparsity of F can be retained at a workable level. This definition for F coincides with the Block FSAI preconditioner introduced in Janna et al. [2010a] and Janna and Ferronato [2011], where the non-zero pattern $\tilde{S}$ is defined either statically or dynamically during the computation of F . Using a sparse F into equation (4.46) produces an approximation $\tilde{S}$ of the exact Schur complement S of A .

Corollary 4.3.2 *The Schur complement approximation $\tilde{S}$ computed with equation (4.46) and a sparse Block FSAI F is SPD.*

Proof. The expression of $\tilde{S}$ can be easily rearranged by adding and subtracting the term $B^T K^{-1} B$:

$$\tilde{S} = C - B^T K^{-1} B + (F + B^T K^{-1}) K (K^{-1} B + F^T) = S + W^T K W \quad (4.48)$$

The result immediately follows by noting that $W^T K W$ is SPD. $\square$

Based on these results, the MF preconditioner is built as follows. The zero-level preconditioner M_0^{-1} is made by two factors:

$$M_0^{-1} = P_2 P_1 \quad (4.49)$$

An explicit approximation of K^{-1} is computed as $G^T G$ using an adaptive FSAI procedure [Janna et al., 2015b] and introduced in P_1 :

$$P_1 = \begin{bmatrix} G & 0 \\ 0 & I \end{bmatrix} \quad (4.50)$$

Then, the preconditioned matrix $P_1 A P_1^T$ is computed:

$$P_1 A P_1^T = \begin{bmatrix} G K G^T & G B \\ B^T G^T & C \end{bmatrix} \quad (4.51)$$

and, to decouple the diagonal submatrices, the adaptive Block FSAI [Janna and Ferronato, 2011] of $P_1 A P_1^T$ is computed for the second factor:

$$P_2 = \begin{bmatrix} I & 0 \\ F & I \end{bmatrix} \quad (4.52)$$

The zero-level preconditioned matrix $M_0^{-1} A M_0^{-T}$ reads:

$$M_0^{-1} A M_0^{-T} = \begin{bmatrix} I & 0 \\ F & I \end{bmatrix} \begin{bmatrix} G K G^T & G B \\ B^T G^T & C \end{bmatrix} \begin{bmatrix} I & F^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} G K G^T & -R_F^T \\ -R_F & \tilde{S} \end{bmatrix} \quad (4.53)$$

where $R_F = -F (G K G^T) - B^T G^T$ is the residual on F , i.e., R_F approaches the null matrix as the accuracy in the computation on F increases. Finally, the $(2, 2)$ block of $M_0^{-1} A M_0^{-T}$ is the approximation of the first-level Schur complement:

$$\tilde{S} = C + F G B + B^T G^T F^T + F G K G^T F^T \quad (4.54)$$

that becomes the new matrix for the next level. Since $\tilde{S}$ in (4.54) is SPD for any F and G (see Corollary (4.3.2)), no breakdown is possible. Operations required for building the robust MF preconditioner are provided in Algorithm 5.

Algorithm 5 MULTILEVEL FSAI SET-UP

```

1: procedure CPTMLFSAI( $A, n_l$ )
2:   Set  $A_0 = A$ 
3:   for  $l = 0, \dots, n_l - 2$  do
4:     Partition  $A_l$  as  $\begin{bmatrix} K & B \\ B^T & C \end{bmatrix}$ 
5:     Compute the adaptive FSAI approximation  $G$  of  $K$  such that  $G^T G \simeq K^{-1}$ 
6:     Set  $P_1 = \begin{bmatrix} G & 0 \\ 0 & I \end{bmatrix}$ 
7:     Compute  $P_2 = \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}$  the adaptive Block FSAI approximation of  $P_1 A_l P_1^T$ 
8:     Compute  $\tilde{S} = C + F G B + B^T G^T F^T + F G K G^T F^T$ 
9:     Form  $M_l^{-1} = P_2 P_1$ 
10:    Set  $A_{l+1} = \tilde{S}$ 
11:  end for
12:  Compute  $M_{n_l-1}^{-1}$  as the adaptive FSAI approximation of  $A_{n_l-1}$ 
13:  Set  $M^{-1} = \{M_0^{-1}, M_1^{-1}, \dots, M_{n_l-1}^{-1}\}$ 
14: end procedure

```

At every level of the MF preconditioner set-up, the partial factorization of the approximated Schur complement (4.54) is computed. This operation introduces additional approximations level after level, potentially yielding to a Schur complement quite different from the exact one. Consider the zero-level preconditioner M_0^{-1} of equation (4.49):

$$M_0^{-1} = \begin{bmatrix} G & 0 \\ FG & I \end{bmatrix} \quad (4.55)$$

The natural choice for the next level preconditioner is:

$$\widetilde{M}_1^{-1} = \begin{bmatrix} I & 0 \\ 0 & L_{\widetilde{S}}^{-1} \end{bmatrix} \quad (4.56)$$

where $L_{\widetilde{S}}L_{\widetilde{S}}^T = \widetilde{S}$. However, $\widetilde{S}$ is an approximated Schur complement. If S is available, one could use:

$$M_1^{-1} = \begin{bmatrix} I & 0 \\ 0 & L_S^{-1} \end{bmatrix} \quad (4.57)$$

with $L_S L_S^T = S$. The following results allow for assessing the quality of both $\widetilde{M}_1^{-1}$ and M_1^{-1} as the next level preconditioner of A .

Proposition 4.3.3 *The eigenvalues λ of the preconditioned matrix $\widetilde{M}^{-1}A\widetilde{M}^{-T}$, with $\widetilde{M}^{-1} = \widetilde{M}_1^{-1}M_0^{-1}$ as defined in equations (4.55) and (4.56), satisfy:*

$$|\lambda - 1| \leq \frac{\|E_K\| + \sqrt{\|E_K\|^2 + 4\|\widetilde{Q}^T\|\|\widetilde{Q}\|}}{2} \quad (4.58)$$

where $\widetilde{Q} = G(KG^T F^T + B)L_{\widetilde{S}}^{-T} = -R_F^T L_{\widetilde{S}}^{-T}$ and $E_K = GKG^T - I$, for any consistent matrix norm.

Proof. The preconditioned matrix $\widetilde{M}^{-1}A\widetilde{M}^{-T}$ reads:

$$\widetilde{M}^{-1}A\widetilde{M}^{-T} = \begin{bmatrix} GKG^T & G(KG^T F^T + B)L_{\widetilde{S}}^{-T} \\ L_{\widetilde{S}}^{-1}(FGK + B^T)G^T & I \end{bmatrix} = \begin{bmatrix} I + E_K & \widetilde{Q} \\ \widetilde{Q} & I \end{bmatrix} \quad (4.59)$$

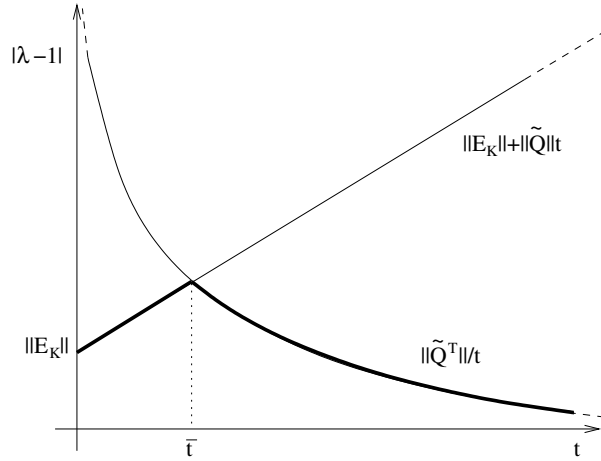


FIGURE 4.4: Schematic representation of the system of inequalities (4.63).

Its eigenpairs $(\lambda, \mathbf{w})$, $\mathbf{w} = [\mathbf{u}, \mathbf{v}]^T$, satisfy by definition the relationship:

$$\begin{bmatrix} I + E_K & \tilde{Q} \\ \tilde{Q} & I \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} = \lambda \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} \quad (4.60)$$

which is equivalent to:

$$\begin{bmatrix} E_K & \tilde{Q} \\ \tilde{Q} & 0 \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} = (\lambda - 1) \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} \quad (4.61)$$

Taking consistent norms at both sides of the first and second set of equations we have:

$$\begin{cases} \|E_K\| \|\mathbf{u}\| + \|\tilde{Q}\| \|\mathbf{v}\| \geq |\lambda - 1| \|\mathbf{u}\| \\ \|\tilde{Q}^T\| \|\mathbf{u}\| \geq |\lambda - 1| \|\mathbf{v}\| \end{cases} \quad (4.62)$$

which, by setting $t = \|\mathbf{v}\| / \|\mathbf{u}\|$, can be rearranged as:

$$\begin{cases} |\lambda - 1| \leq \|E_K\| + \|\tilde{Q}\| t \\ |\lambda - 1| \leq \|\tilde{Q}^T\| / t \end{cases} \quad (4.63)$$

When $\|\mathbf{u}\| = 0$, then trivially $\mathbf{v} \in \text{Ker}(\tilde{Q})$ and $\lambda = 1$, thus the inequality (4.58) is satisfied. The right-hand side of the first and second inequality in (4.63) increases and

decreases monotonically with t , respectively (see Fig. 4.4). The intersection point is:

$$\bar{t} = \frac{-\|E_K\| + \sqrt{\|E_K\|^2 + 4\|\tilde{Q}^T\|\|\tilde{Q}\|}}{2\|\tilde{Q}\|} \quad (4.64)$$

so, for any t , we have:

$$|\lambda - 1| \leq \frac{\|E_K\| + \sqrt{\|E_K\|^2 + 4\|\tilde{Q}^T\|\|\tilde{Q}\|}}{2} \quad (4.65)$$

□

Proposition 4.3.4 *The eigenvalues λ of the preconditioned matrix $M^{-1}AM^{-T}$, with $M^{-1} = M_1^{-1}M_0^{-1}$ as defined in equations (4.49) and (4.57), satisfy:*

$$|\lambda - 1| \leq \frac{\|E_K\| + \|Q^T\| \|(I + E_K)^{-1}\| \|Q\|}{2} + \frac{\sqrt{(\|E_K\| - \|Q^T\| \|(I + E_K)^{-1}\| \|Q\|)^2 + 4\|Q^T\| \|Q\|}}{2} \quad (4.66)$$

where $Q = G(KG^T F^T + B)L_S^{-T} = -R_F^T L_S^{-T}$ and $E_K = GKG^T - I$, for any consistent matrix norm.

Proof. Similarly to the proof of Proposition 4.3.3, it can be shown that the eigenpairs of $M^{-1}AM^{-T}$ satisfy:

$$\begin{bmatrix} I + E_K & Q \\ Q^T & I + Q^T(I + E_K)^{-1}Q \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} = \lambda \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} \quad (4.67)$$

which is equivalent to:

$$\begin{bmatrix} E_K & Q \\ Q^T & Q^T(I + E_K)^{-1}Q \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} = (\lambda - 1) \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} \quad (4.68)$$

Again, taking consistent norms at both sides of the first and second set of equations, and setting $t = \|\mathbf{v}\| / \|\mathbf{u}\|$, we get:

$$\begin{cases} |\lambda - 1| \leq \|E_K\| + \|Q\| t \\ |\lambda - 1| \leq \|Q^T\| / t + \|Q^T (I + E_K)^{-1} Q\| \end{cases} \quad (4.69)$$

If $\|\mathbf{u}\| = 0$, then trivially $\mathbf{v} \in \text{Ker}(Q)$ and $\lambda = 1$, thus satisfying the inequality (4.66).

Otherwise, denoting by $\bar{t}$ the intersection point between the right-hand sides of (4.69):

$$\begin{aligned} \bar{t} = & \frac{-\|E_K\| + \|Q^T (I + E_K)^{-1} Q\|}{2 \|Q\|} + \\ & + \frac{\sqrt{\left(\|E_K\| - \|Q^T (I + E_K)^{-1} Q\|\right)^2 + 4 \|Q^T\| \|Q\|}}{2 \|Q\|} \end{aligned} \quad (4.70)$$

for any t we have:

$$\begin{aligned} |\lambda - 1| \leq & \frac{\|E_K\| + \|Q^T\| \|(I + E_K)^{-1}\| \|Q\|}{2} + \\ & + \frac{\sqrt{\left(\|E_K\| - \|Q^T\| \|(I + E_K)^{-1}\| \|Q\|\right)^2 + 4 \|Q^T\| \|Q\|}}{2} \end{aligned} \quad (4.71)$$

□

The norms of E_K and Q or $\tilde{Q}$ are controlled by the accuracy in the computation of G and F , respectively. As $G^T G$ approaches K^{-1} , $\|E_K\| \rightarrow 0$, and in the limit the bounds (4.58) and (4.66) read:

$$|\lambda - 1| \leq \sqrt{\|\tilde{Q}\| \|\tilde{Q}^T\|} \quad (4.72)$$

and

$$|\lambda - 1| \leq \frac{1}{2} \left[\|Q\| \|Q^T\| + \sqrt{\|Q\| \|Q^T\| (4 + \|Q\| \|Q^T\|)} \right] \quad (4.73)$$

respectively. Similarly, as F approaches $-B^T K^{-1} G^{-1}$, $\|Q\|, \|Q^T\| \rightarrow 0$ and the bounds (4.58) and (4.66) trivially provide:

$$|\lambda - 1| \leq \|E_K\| \quad (4.74)$$

Using the matrix norm induced by the 2-norm of vectors, we have:

$$\|E_K\|_2 = \lambda_1(E_K) = \varepsilon_1, \quad \|Q\|_2 = \sqrt{\lambda_1(Q^T Q)} = \eta_1, \quad (4.75)$$

$$\|\tilde{Q}\|_2 = \sqrt{\lambda_1(\tilde{Q}^T \tilde{Q})} = \tilde{\eta}_1, \quad \|(I + E_K)^{-1}\|_2 = \lambda_n^{-1}(GKG^T) = \kappa_n^{-1}$$

and the bounds (4.58) and (4.66) respectively become:

$$|\lambda - 1| \leq \frac{\varepsilon_1 + \sqrt{\varepsilon_1^2 + 4\tilde{\eta}_1^2}}{2} \quad (4.76)$$

$$|\lambda - 1| \leq \frac{\varepsilon_1 + \eta_1^2 \kappa_n^{-1} + \sqrt{(\varepsilon_1 - \eta_1^2 \kappa_n^{-1})^2 + 4\eta_1^2}}{2} \quad (4.77)$$

Remark. When $\varepsilon_1 = 0$, it is easy to prove that the maximum and minimum eigenvalues of $\tilde{M}^{-1}A\tilde{M}^{-T}$ are $1 + \tilde{\eta}_1$ and $1 - \tilde{\eta}_1$, respectively, and that the maximum and minimum eigenvalues of $M^{-1}AM^{-T}$ are $1 + (\eta_1^2 + \sqrt{\eta_1^4 + 4\eta_1^2})/2$ and $1 + (\eta_1^2 - \sqrt{\eta_1^4 + 4\eta_1^2})/2$, respectively.

The following results suggest the use of $\tilde{M}^{-1}$, instead of M^{-1} , as the MF preconditioner of A .

Theorem 4.3.5 *For any choice of G and F in (4.55), the bound (4.76) is narrower than or equal to the bound (4.77).*

Proof. Using the arguments of Corollary 4.3.2, it follows that $\tilde{S} = S + \tilde{H}$, with $\tilde{H}$ a symmetric positive semidefinite matrix, hence $\|S\tilde{S}^{-1}\|_2 \leq 1$. In particular:

$$\begin{aligned} \tilde{S} &= C + FGKG^T F^T + FGB + B^T G^T F^T = \\ &= C - B^T K^{-1} B + (F + B^T K^{-1} G^{-1})(GKG^T)(F + B^T K^{-1} G^{-1})^T = \\ &= S + R_F (GKG^T)^{-1} R_F^T \end{aligned} \quad (4.78)$$

Moreover, the matrix $\tilde{Q}^T \tilde{Q}$ is similar to $L_S^T \tilde{S}^{-1} L_S Q^T Q$. In fact, recalling that $Q = -R_F^T L_{\tilde{S}}^{-T}$, we obtain:

$$R_F^T R_F = L_S Q^T Q L_S^T = L_S \tilde{Q}^T \tilde{Q} L_S^T \quad (4.79)$$

from which the similarity follows. Hence:

$$\left\| \tilde{Q}^T \tilde{Q} \right\|_2 = \left\| L_S^T \tilde{S}^{-1} L_S Q^T Q \right\|_2 \leq \left\| L_S^T \tilde{S}^{-1} L_S \right\|_2 \left\| Q^T Q \right\|_2 \leq \left\| Q^T Q \right\|_2 \quad (4.80)$$

because $L_S^T \tilde{S}^{-1} L_S$ is similar to $\tilde{S} \tilde{S}^{-1}$. As a consequence, $\tilde{\eta}_1 = \alpha \eta_1$ for some $\alpha \leq 1$. The thesis of the theorem reads:

$$\frac{\varepsilon_1 + \sqrt{\varepsilon_1^2 + 4\tilde{\eta}_1^2}}{2} \leq \frac{\varepsilon_1 + \eta_1^2 \kappa_n^{-1} + \sqrt{(\varepsilon_1 - \eta_1^2 \kappa_n^{-1})^2 + 4\eta_1^2}}{2} \quad (4.81)$$

Introducing $\tilde{\eta}_1 = \alpha \eta_1$ in (4.81), after some algebra, we obtain:

$$\alpha^2 \leq 1 + \frac{\sqrt{(\varepsilon_1 - \eta_1^2 \kappa_n^{-1})^2 + 4\eta_1^2} - (\varepsilon_1 - \eta_1^2 \kappa_n^{-1})}{2} \quad (4.82)$$

which holds true for any F and G . $\square$

Theorem 4.3.5 suggests that the use of $\tilde{S}$ in the MF preconditioner is likely to be more appropriate than the exact Schur complement S . In particular, the following result can be proven in the theoretical case of $G^T G = K^{-1}$.

Theorem 4.3.6 *If $E_K = 0$, the ratio between the conditioning number of the preconditioned matrices (4.67) and (4.59) is:*

$$r(\eta_1) = \frac{\left(1 + \frac{\eta_1^2 + \sqrt{\eta_1^4 + 4\eta_1^2}}{2} \right)^2}{1 + 2\eta_1^2 + 2\sqrt{\eta_1^4 + \eta_1^2}} \quad (4.83)$$

Proof. If $E_K = 0$, it can be easily verified from equation (4.78) that $\tilde{S} = S + R_F R_F^T$ and:

$$L_S^{-1} \tilde{S} L_S^{-T} = I + L_S^{-1} R_F R_F^T L_S^{-T} = I + Q^T Q \quad (4.84)$$

Recalling from the proof of Theorem 4.3.5 that $\tilde{Q}^T \tilde{Q}$ is similar to $L_S^T \tilde{S}^{-1} L_S Q^T Q$, we have:

$$\left\| \tilde{Q}^T \tilde{Q} \right\|_2 = \left\| (I + Q^T Q)^{-1} Q^T Q \right\|_2 \quad (4.85)$$

Trivially, $(I + Q^T Q)^{-1} Q^T Q$ is symmetric positive definite and has the same eigenvectors as $Q^T Q$. Denoting by λ an eigenvalue of $Q^T Q$, the norm $\|(I + Q^T Q)^{-1} Q^T Q\|_2$ is the maximum of the function:

$$f(\lambda) = \frac{\lambda}{1 + \lambda} \quad (4.86)$$

As $f(\lambda)$ monotonically increases with λ , its maximum value is attained for the largest eigenvalue of $Q^T Q$, i.e., $\|Q^T Q\|_2 = \eta_1^2$. Hence:

$$\|\tilde{Q}^T \tilde{Q}\|_2 = \frac{\|Q^T Q\|_2}{1 + \|Q^T Q\|_2} \Rightarrow \tilde{\eta}_1 = \sqrt{\frac{\eta_1^2}{1 + \eta_1^2}} \quad (4.87)$$

The proof is completed by introducing equation (4.87) in the results of Remark (4.3.3).

□

Remark. The ratio $r(\eta_1)$ monotonically increases with η_1 and takes value 1 for $\eta_1 = 0$, i.e., $F = -B^T K^{-1} G^{-1}$ and $\tilde{S} = S$. For $\eta_1 \rightarrow \infty$, r monotonically diverges to infinity as the second power of η_1 . Hence, the sparser or more inaccurate F the more important is using $\tilde{S}$ instead of S in the MF preconditioner.

4.3.4 Improving the MF performance with Low-Rank corrections

To our knowledge, the first attempt to directly use low-rank representations in preconditioning is discussed in Li and Saad [2013] in the context of a divide and conquer strategy. The idea of using low-rank corrections in a multilevel framework has been introduced in [Xi et al., 2016], giving rise to the robust Multilevel Schur complement-based Low-Rank (MSLR) preconditioner. The basic concept is the one reported below. Define the matrix:

$$Y = L_C^{-1} B^T K^{-1} B L_C^{-T} = L_C^{-1} (C - S) L_C^{-T} \quad (4.88)$$

where L_C is the exact lower factor of C , i.e., $C = L_C L_C^T$. It is easy to recognize that the eigenvalues σ_i of Y are such that:

$$0 \leq \sigma_{n_2} \leq \dots \leq \sigma_1 \leq 1 \quad (4.89)$$

The separation of the eigenvalues θ_i of $X = L_C^T (S^{-1} - C^{-1}) L_C$ is larger than that of Y , because:

$$\theta_i = \frac{\sigma_i}{1 - \sigma_i} \quad i = 1, \dots, n_2 \quad (4.90)$$

$$\theta_i - \theta_{i+1} = \frac{\sigma_i - \sigma_{i+1}}{(1 - \sigma_i)(1 - \sigma_{i+1})} \quad i = 1, \dots, n_2 - 1 \quad (4.91)$$

The separation of the eigenvalues of $L_C^{-T} X L_C^{-1} = S^{-1} - C^{-1}$ has a stronger impact on the performance of the MSLR preconditioner [Xi et al., 2016], however studying X is easier and the main results for X are at some extent still valid for $S^{-1} - C^{-1}$. Equation (4.91) suggests that approximating with a low-rank matrix $(S^{-1} - C^{-1})$ is easier than approximating $(S - C)$ because of the faster eigenvalue decay. A better approximation of S^{-1} can be computed as:

$$S^{-1} \simeq C^{-1} + W_k \Theta_k W_k^T \quad (4.92)$$

with $W_k \Theta_k W_k^T$ a rank- k approximation of $L_C^{-T} X L_C^{-1}$ which can be obtained from the eigendecomposition of Y . In fact, by retaining the k largest eigenvalues and corresponding eigenvectors of Y , we can write:

$$Y \simeq U_k \Sigma_k U_k^T \quad (4.93)$$

Noting that:

$$S^{-1} - C^{-1} = L_C^{-T} \left[(I - Y)^{-1} - I \right] L_C^{-1} = L_C^{-T} \left[Y (I - Y)^{-1} \right] L_C^{-1} \quad (4.94)$$

the rank- k correction to $S^{-1} - C^{-1}$ is found by setting:

$$\Theta_k = \Sigma_k (I - \Sigma_k)^{-1} \quad (4.95)$$

and $W_k = L_C^{-T} U_k$.

By distinction with the MSLR preconditioner, we use low-rank corrections to improve the action of $\tilde{S}^{-1}$. A consequence of Corollary 4.3.2 is that the eigenvalues σ_i of

the matrix:

$$\bar{Y} = L_{\tilde{S}}^{-1} \left(\tilde{S} - S \right) L_{\tilde{S}}^{-T} \quad (4.96)$$

satisfy the condition (4.89). Thus, following the procedure outlined above for C , we can compute the k largest eigenpairs of Y :

$$\bar{Y} \simeq \bar{U}_k \bar{\Sigma}_k \bar{U}_k^T \quad (4.97)$$

and get the expression of the corrected Schur complement inverse:

$$S^{-1} \simeq \tilde{S}^{-1} + \bar{W}_k \bar{\Theta}_k \bar{W}_k^T \quad (4.98)$$

where $\bar{\Theta}_k = \bar{\Sigma}_k (I - \bar{\Sigma}_k)^{-1}$ and $\bar{W} = L_{\tilde{S}}^{-T} \bar{U}$. Note that the computation of (4.97) may be quite expensive, since every multiplication by S requires a solution of a linear system with K . Moreover, according to Theorems 4.3.5 and 4.3.6, correcting $\tilde{S}^{-1}$ so as to resemble S^{-1} generally is not optimal. Even in the special case $E_K = 0$, only a moderate benefit can be achieved for very low values of $\|Q\|$, i.e., with a very accurate F .

Since we are working in a multilevel framework, $\tilde{S}^{-1}$ will not be used exactly. Rather, a new approximation, $\hat{S}^{-1} \simeq \tilde{S}^{-1}$, will be computed. Thus, $\hat{S}$ will be the new target of the low-rank correction. Moreover, as shown by equation (4.58), reducing $\|E_K\|$ is also useful for improving the convergence. Also this task can be performed by a low-rank correction. Therefore, we use two low-rank correction techniques for the preconditioner set-up:

- *Descending low-rank corrections*: computed at each level, from the first to the last, to reduce $\|E_K\|$;
- *Ascending low-rank corrections*: computed at each level, from the last to the first, to reduce the gap between $\hat{S}$ and $\tilde{S}$.

The aim of the Descending Low-Rank corrections is to enhance the approximation of the inverse of K . Starting from equation (4.96), we can define the matrix:

$$\bar{Y} = G \left[(G^T G)^{-1} - K \right] G^T = I - G K G^T \quad (4.99)$$

where $(G^T G)^{-1}$ and K replace $\tilde{S}$ and S , respectively. Moreover, we can compute its rank- k approximation:

$$\bar{Y} \simeq \bar{U}_k \bar{\Sigma}_k \bar{U}_k^T \quad (4.100)$$

Note that the computation of $\bar{U}_k$ and $\bar{\Sigma}_k$ is less expensive than in (4.97), because both G and K are explicitly known. The enhanced preconditioner for K reads:

$$K^{-1} \simeq G^T G + \bar{W}_k \bar{\Theta}_k \bar{W}_k^T \quad (4.101)$$

The eigenvalues of Y are bounded from above by 1 as GKG^T is positive definite, but there is not a lower bound. Actually this is not a problem, as we are mainly interested in the computation of the eigenvalue $\bar{\sigma}_i$ of $\bar{Y}$ closest to 1. From the implementational point of view, it is better to dispose of a symmetrically splitted operator. Hence, we define:

$$\tilde{G} = \left(I + \bar{U}_k \bar{\Psi}_k \bar{U}_k^T \right) G \quad (4.102)$$

in order to have:

$$\tilde{G}^T \tilde{G} = G^T \left(I + \bar{U}_k \bar{\Psi}_k \bar{U}_k^T \right) \left(I + \bar{U}_k \bar{\Psi}_k \bar{U}_k^T \right) G = G^T G + \bar{W}_k \bar{\Theta}_k \bar{W}_k^T \quad (4.103)$$

Recalling that $\bar{W}_k = G^T \bar{U}_k$, the diagonal $k \times k$ matrix $\bar{\Psi}_k$ is simply found by solving:

$$I + 2\bar{U}_k \bar{\Psi}_k \bar{U}_k^T + \bar{U}_k \bar{\Psi}_k^2 \bar{U}_k^T = I + \bar{U}_k \bar{\Theta}_k \bar{U}_k^T \quad (4.104)$$

Using equation (4.104), the entries of $\bar{\Theta}_k$ read:

$$\bar{\psi}_i = -1 + \sqrt{\frac{1}{1 - \bar{\sigma}_i}}, \quad i = 1, \dots, k \quad (4.105)$$

Since GKG^T is positive definite, $\sigma_i < 1$ and ψ_i is real for any i .

This correction can significantly reduce $\|E_K\|$. However, the update in $\tilde{G}$ may propagate in the other blocks of the preconditioned matrix potentially shattering the overall

procedure efficiency:

$$\begin{aligned}
 P_1 A P_1^T &= \begin{bmatrix} \tilde{G} K \tilde{G}^T & \tilde{G} B \\ B^T \tilde{G}^T & C \end{bmatrix} = \begin{bmatrix} G K G^T & G B \\ B^T G^T & C \end{bmatrix} = \\
 &= \begin{bmatrix} \bar{U}_k \bar{\Psi}_k \bar{U}_k^T G K G^T + G K G^T \bar{U}_k \bar{\Psi}_k \bar{U}_k^T + \bar{U}_k \bar{\Psi}_k \bar{U}_k^T G K G^T \bar{U}_k \bar{\Psi}_k \bar{U}_k^T & \bar{U}_k \bar{\Psi}_k \bar{U}_k^T G B \\ B^T G^T \bar{U}_k \bar{\Psi}_k \bar{U}_k^T & C \end{bmatrix}
 \end{aligned} \tag{4.106}$$

Actually, this is not the case. In fact, the Block FSAI $\tilde{F}$ computed as the (2,1) block of P_2 when using $\tilde{G}$ is the approximate solution of the multiple right-hand-side system:

$$\tilde{F}^T \simeq - \left(\tilde{G} K \tilde{G}^T \right)^{-1} \tilde{G} B = -\tilde{G}^T K^{-1} B \tag{4.107}$$

By expanding equation (4.107) with the $\tilde{G}$ definition, we note that $\tilde{F}$ can be easily found as:

$$\tilde{F} = F \left(I + \bar{U}_k \bar{\Psi}_k \bar{U}_k^T \right)^{-1} \tag{4.108}$$

where F is the standard Block FSAI computed using G . Moreover, from (4.108) and (4.102) we notice that $FG = \tilde{F}\tilde{G}$. This implies that the approximate Schur complement (4.54) is not affected by the use of $\tilde{F}$ and $\tilde{G}$:

$$\tilde{S} = C + \tilde{F} \tilde{G} B + B^T \tilde{G}^T \tilde{F}^T + \tilde{F} \tilde{G} K \tilde{G}^T \tilde{F}^T = C + F G B + B^T G^T F^T + F G K G^T F^T \tag{4.109}$$

As a consequence, descending low-rank corrections on G have only a local impact with no changes for the following levels.

For the Ascending Low-Rank corrections, we define the matrix $\hat{Y}$ and compute its rank- k approximation:

$$\hat{Y} = I - \hat{G} \hat{S} \hat{G}^T \simeq \hat{U}_k \hat{\Sigma}_k \hat{U}_k^T \tag{4.110}$$

where $\hat{G}$ is the lower inverse factor of $\hat{S}$, i.e. $\left(\hat{G}^T \hat{G} \right)^{-1} = \hat{S}$, which is explicitly available by the approximation of lower levels. The computation of (4.110) is relatively cheap, because the explicit expression of every matrix is known. The new approximation to $\tilde{S}$ is given by:

$$\tilde{S}^{-1} = \hat{G}^T \hat{G}^T + \hat{W}_k \hat{\Theta}_k \hat{W}_k^T \tag{4.111}$$

where $\widehat{W}_k = \widehat{G}^T \widehat{U}_k$ and $\widehat{\Theta}_k = \widehat{\Sigma}_k (I - \widehat{\Sigma}_k)^{-1}$. Notice that, during the set-up, we use a splitted update, as done for the descending low-rank corrections, because it is operatively necessary to compute $\widehat{G} \widetilde{S} \widehat{G}^T$. However, during the preconditioner application, the use of equation (4.111) is more efficient as it requires only one update.

The Multilevel FSAI preconditioner with Low-Rank corrections (MFLR) can be implemented and applied through a recursive function (see Algorithms 6 and 7, respectively). The preconditioner is stored using the lists Q_{alr} and Q_M and the parallelization is performed equally by the partition of each level among the available cores using the OpenMP directives.

Algorithm 6 MFLR SET-UP

```

1: recursive procedure MFLR_SETUP( $l, A, n_l, alr, dlr$ )
2: if  $l < (n_l - 1)$  then
3:   Partition  $A_l$  as  $\begin{bmatrix} K & B \\ B^T & C \end{bmatrix}$ 
4:   Compute the adaptive FSAI approximation  $G$  of  $K$  such that  $G^T G \simeq K^{-1}$ 
5:   Set  $P_1 = \begin{bmatrix} G & 0 \\ 0 & I \end{bmatrix}$ 
6:   Compute  $P_2 = \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}$  the adaptive Block FSAI approximation of  $P_1 A_l P_1^T$ 
7:   Compute  $\widetilde{S} = C + FGB + B^T G^T F^T + FGKG^T F^T$ 
8:   Compute a rank  $k$  approximation  $\overline{U}_k \overline{\Theta}_k \overline{U}_k^T$  of  $\overline{Y} = I - GKG^T$ 
9:   Set  $\widetilde{G} = (I + \overline{U}_k \overline{\Psi}_k \overline{U}_k^T) G$  with  $\overline{\Psi}_k = (I - \overline{\Sigma}_k)^{-1/2} - I$ 
10:  Set  $P = \begin{bmatrix} \widetilde{G} & 0 \\ FG & I \end{bmatrix}$ 
11:   $[Q_{alr}, Q_M] = \text{MFLR\_SETUP}(l + 1, \widetilde{S}, n_l, alr, dlr)$ 
12:  Use  $Q_{alr}$  and  $Q_M$  to compute the rank  $k$  correction  $\widehat{W}_k$  and  $\widehat{\Theta}_k$  as in (4.111)
13:  Set  $\widetilde{P} = \begin{bmatrix} I & 0 \\ 0 & I + \widehat{W}_k \widehat{\Theta}_k \widehat{W}_k^T \end{bmatrix}$ 
14:  Push  $\widetilde{P}$  in the head of  $Q_{alr}$ 
15:  Push  $P$  in the head of  $Q_M$ 
16:  return  $Q_{alr}, Q_M$ 
17: else
18:   Compute the adaptive FSAI approximation  $G$  of  $A$  such that  $G^T G \simeq A^{-1}$ 
19:   Set  $Q_{alr} = \emptyset$ 
20:   Set  $Q_M = \{G\}$ 
21:   return  $Q_{alr}, Q_M$ 
22: end if

```

Algorithm 7 MFLR APPLY

```

1: recursive procedure MFLR_APPLY( $l, Q_{alr}, Q_M, \mathbf{x}$ )
2: if  $l < (n_l - 1)$  then
3:   Pop  $P$  from the head of  $Q_M$ 
4:   Pop  $\tilde{P}$  from the head of  $Q_{alr}$ 
5:   Compute  $\mathbf{y} = P\mathbf{x}$ 
6:   Partition  $\mathbf{y} = \{\mathbf{y}_1^T, \mathbf{y}_2^T\}^T$ 
7:   Compute  $\mathbf{z}_2 = \text{MFLR\_APPLY}(l + 1, Q_{alr}, Q_M, \mathbf{y}_2)$ 
8:   Compose  $\mathbf{z} = \{\mathbf{y}_1^T, \mathbf{z}_2^T\}^T$ 
9:   Compute  $\mathbf{y} = \tilde{P}\mathbf{z}$ 
10:  Compute  $\mathbf{y} = P^T\mathbf{y}$ 
11:  Push  $P$  in the head of  $Q_M$ 
12:  Push  $\tilde{P}$  in the head of  $Q_{alr}$ 
13:  return  $\mathbf{y}$ 
14: else
15:   Pop  $G$  from the head of  $Q_M$ 
16:   Compute  $\mathbf{y} = G^T G\mathbf{x}$ 
17:   Push  $G$  in the head of  $Q_M$ 
18:   return  $\mathbf{y}$ 
19: end if

```

4.4 Numerical performances for the SPD matrix

A numerical investigation of the preconditioners developed for the approximation of the SPD leading block A is presented in this section. The three approaches are compared with respect to the classic adaptive FSAI.

4.4.1 Block Tridiagonal FSAI and Domain Decomposition FSAI

The computational performance is examined through the total number of iterations needed for convergence n_{it} and the wall clock time needed for building the preconditioner T_p and solving the linear system T_s . These data are analysed for the simulation carried out in serial using a local cluster equipped with two Intel Xeon E5-2643 processors and 256 Gbytes of RAM memory. On the other hand, the scalability tests were performed on a single node of the IBM Blue Gene/Q FERMI supercomputer at the CINECA (Centre for High Performance Computing), which is equipped with a 16-core IBM Power A2 processor at 1.6 GHz and 16 Gbytes of RAM memory with 42.6 Gbytes/s bandwidth.

First, as a model problem we consider the solution of a 3D Poisson test case defined on a cube. This problem is solved by means of finite differences on a uniform grid containing 150 nodes in each direction, thus creating an SPD matrix consisting of 3,375,000 rows and 23,490,000 nonzeros.

As first step, we find the optimal configuration parameters for the native FSAI preconditioner, i.e. the one providing the lowest total wall-clock time, and set this configuration as the reference one. The search for the best setup of the preconditioner is accomplished by using the guidelines suggested in [Janna and Ferronato \[2011\]](#) and [Janna et al. \[2015b\]](#). The same input parameters are used for building also the BTFSAI and DDFSAI preconditioners, with the only additional information being the number of blocks or partitions, respectively. Figure 4.5 explains how this parameter affects the solution process, showing the number of iterations needed for convergence and the times for building these preconditioners and solving the linear systems with a number of blocks ranging between 50 and 400. For the BTFSAI preconditioner, it is possible to note that the number of iterations needed for convergence decreases with respect to the reference case up to about 200 blocks, then it starts to increase again. This behaviour is probably caused by the quality of the approximated Schur complements, which decreases as the number of blocks gets higher. Considering the DDFSAI preconditioner, we see that the number of iterations grows moderately with respect to both FSAI and BTFSAI preconditioners, but remains approximately constant changing the number of blocks, thus showing a very stable behaviour with n . In this problem, we see an increase of the times needed for computing and applying the preconditioner with respect to the reference case, however this 3D Poisson problem is so simple that a wall-clock time comparison is not significant.

The performance and scalability of the preconditioners BTFSAI and DDFSAI are analyzed considering a set of SPD matrices representing different engineering applications, such as structural mechanics, geomechanics, electromagnetism and multiphase flow. Table 4.1 lists the details about the chosen matrices.

Similarly to Figure 4.5, Figure 4.6 shows how the performance of both preconditioners vary according to the number of blocks or partitions used in their construction.

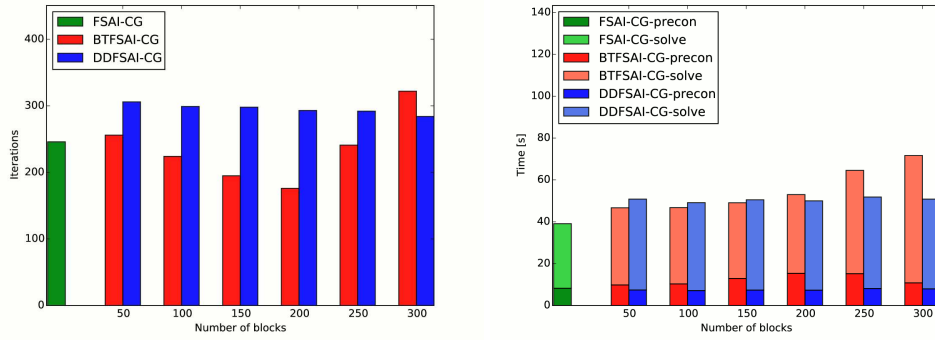


FIGURE 4.5: Number of iterations needed for convergence as a function of the number of blocks used, left. Wall clock time needed for computing the preconditioner and solving the linear system again as a function of the number of blocks used, right. These figures show the results obtained with DDFSAI and BTFSAI.

TABLE 4.1: BTFSAI - DDFSAI: Test matrices.

Name	Rows	Nonzeros	Description
Ecology2	999,999	4,995,991	2D ecology problem
Tmt-sym	726,713	5,080,961	Electromagnetic problem
Cylinder	372,960	16,473,150	3D geomechanical problem
Hook1498	1,498,023	59,374,451	3D structural problem

Except for the last test case, the BTFSAI preconditioner always shows a gain in performance with respect to the FSAI preconditioner, though with a performance that is sensitive to the number of selected blocks. It is possible to note that there is a problem dependent optimal number of blocks for building this preconditioner. More blocks involve the approximation of too many Schur complements with a performance degradation. On the other hand, a few blocks do not exhibit thoroughly the block tridiagonal structure of the matrix, but still can perform better than the pure FSAI. The DDFSAI preconditioner can also decrease the solution time and the total number of iterations, although this effect is not so pronounced as in the BTFSAI case. Its behaviour, however, is always very stable with n . When the time needed for building a good preconditioner is relatively large, like in the `Hook1498` case, the use of DDFSAI is more convenient since the inverse of the Schur complement can be calculated with a low accuracy showing no significant loss of effectiveness.

The strong scalability profiles of DDFSAI and BTFSAI for the first three test cases are given in Figure 4.7. For each preconditioner, we consider two values for the number

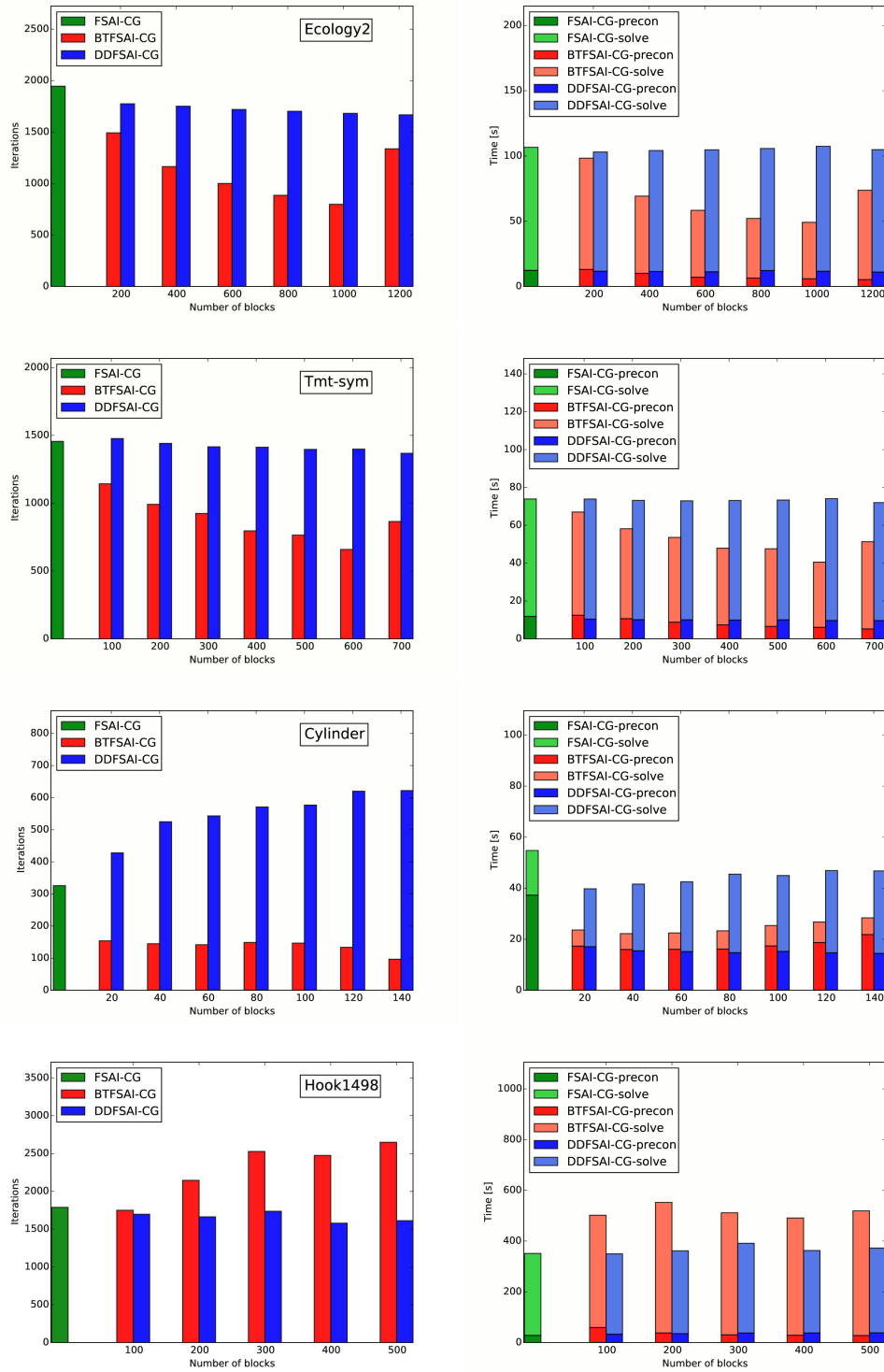


FIGURE 4.6: Convergence results for the matrices Ecology2, Tmt-sym, Cylinder and Hook1498, respectively.

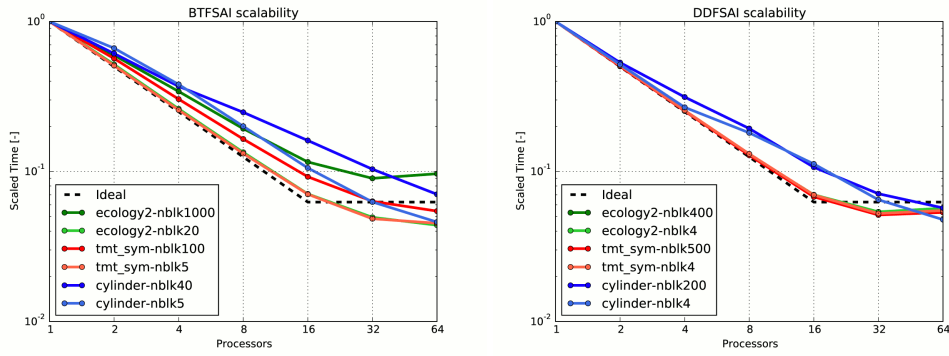


FIGURE 4.7: Scalability profiles of DDFSAl and BTFSAl, respectively.

of blocks, i.e., a small value and the one giving the best total time for the solution. We show the scalability profiles up to 64 threads since the IBM Blue Gene/Q is a computer designed for parallel computations that supports up to 4 threads per core with negligible overhead. According to the figure on the left, DDFSAl shows a good scalability close to the ideal one for the test cases `ecology2` and `tmt-sym` in both configurations, i.e., with many partitions and just a few. The scalability for the `Cylinder` is a little bit worse, probably because of the small size of this linear system. Looking at the leftmost plot of Figure 4.7, we note that the scalability of BTFSAl is not as good. This is because the algorithm involves more synchronization points than DDFSAl, thus decreasing the level of concurrency especially when increasing the number of blocks. However, the BTFSAl can still provide a good scalability, such as for the `tmt-sym` matrix.

4.4.2 Multilevel FSAI with Low Rank corrections

For the sake of clarity, we recall the user-specified parameters that control the MFLR quality and performance:

1. n_l : number of levels. Therefore, $(n_l - 1)$ is the number of computed Schur complements;
2. $k_{G,max}$: maximum number of steps for the adaptive computation of G ;
3. ϵ_G : tolerance for the adaptive computation of G ;
4. $k_{F,max}$: maximum number of steps for the adaptive computation of F ;
5. ϵ_F : tolerance for the adaptive computation of F ;

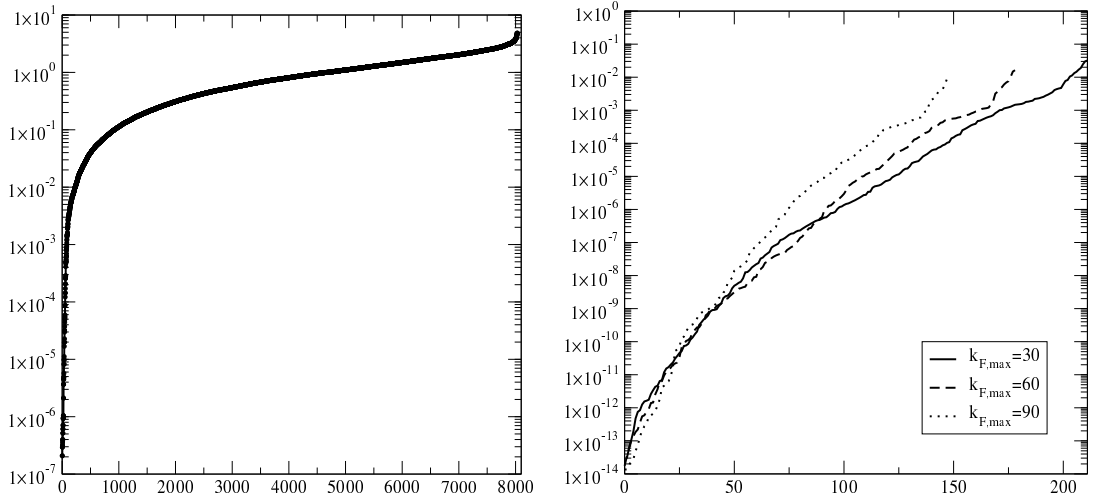


FIGURE 4.8: `bcsstk38` test case: eigenspectrum of A (left) and $(\tilde{S} - S)$ (right) for different Block FSAI F .

6. *dlr*: Descending Low-Rank correction size, i.e., number of eigenpairs used to enrich G . This is a local improvement;
7. *alr*: Ascending Low-Rank correction size, i.e., number of eigenpairs used to enrich $\hat{S}^{-1}$. This is a global improvement, ensuring that the Schur complement size grows up to the size of the zero-level partition.

First of all we analyze the `bcsstk38` matrix from the University of Florida Sparse Matrix Collection [Davis and Hu, 2011]. This SPD matrix has 8,032 rows and 355,460 nonzeros, and has been scaled in order to have a unitary diagonal. Its eigenspectrum is provided in the leftmost frame of Figure 4.8. The matrix is uniformly partitioned into two levels ($n_1 = n_2 = 4,016$). The approximate Schur complement $\tilde{S}$ is computed using equation (4.54), where F is the Block FSAI of A obtained using the dynamic strategy introduced in Janna and Ferronato [2011] and a variable number $k_{F,max}$ of entries retained per row. Other parameters used in the computation of F are $\epsilon_F = 0$, $s = 1$ and $\tau = 0$. The inverse of the leading block is computed exactly through a direct solver. The nonzero eigenspectrum of $(\tilde{S} - S)$ is shown in the rightmost frame of Figure 4.8 and is strictly positive for any F , as expected from Corollary 4.3.2.

The ratio between the conditioning numbers of $M^{-1}AM^{-T}$ and $\tilde{M}^{-1}A\tilde{M}^{-T}$, as obtained in Theorem 4.3.6, is reported in Table 4.2 for different Block FSAI F . Decreasing

TABLE 4.2: bcsstk38 test case: comparison between the conditioning numbers of (4.59) and (4.67) varying F with $\|E_K\| = 0$. λ and $\tilde{\lambda}$ denote the eigenvalues of $M^{-1}AM^{-T}$ and $\tilde{M}^{-1}\tilde{A}\tilde{M}^{-T}$, respectively.

$k_{F,max}$	η_1	$\tilde{\eta}_1$	$[\lambda_n, \lambda_1]$	$[\tilde{\lambda}_n, \tilde{\lambda}_1]$	$r(\eta_1)$
30	4.395	0.975	[0.0470, 21.273]	[0.0249, 1.975]	5.709
60	2.867	0.944	[0.0988, 10.121]	[0.0558, 1.944]	2.939
90	2.248	0.914	[0.1448, 6.9069]	[0.0863, 1.914]	2.153

TABLE 4.3: bcsstk38 test case: eigenvalue distribution of (4.59) and (4.67) varying F and G . The same notation as in Table 4.2 is used.

$k_{G,max}$	$k_{F,max}$	ε_1	η_1	$\tilde{\eta}_1$	$[\lambda_n, \lambda_1]$	$[\tilde{\lambda}_n, \tilde{\lambda}_1]$	bound (4.77)	bound (4.76)
30	30	1.076	4.505	0.812	[1.8e-04, 47.3]	[1.7e-04, 2.1]	19473.3	2.512
30	60	1.076	3.181	0.790	[1.8e-04, 33.2]	[1.8e-04, 2.1]	9707.6	2.494
30	90	1.076	2.617	0.648	[1.8e-04, 29.1]	[1.8e-04, 2.1]	6570.8	2.380
60	30	1.240	4.393	0.718	[3.9e-04, 40.0]	[3.9e-04, 2.2]	8291.3	2.569
60	60	1.240	3.037	0.641	[4.0e-04, 29.5]	[4.0e-04, 2.2]	3963.8	2.512
60	90	1.240	2.466	0.589	[4.0e-04, 25.8]	[4.0e-04, 2.2]	2613.6	2.475
90	30	1.092	4.300	0.738	[6.6e-04, 36.0]	[6.5e-04, 2.1]	5353.9	2.464
90	60	1.092	2.935	0.651	[6.8e-04, 26.8]	[6.7e-04, 2.1]	2495.6	2.396
90	90	1.092	2.434	0.620	[6.9e-04, 23.3]	[6.8e-04, 2.1]	1716.8	2.372

the quality of F , $\eta_1 = \|Q\|_2$ and $\tilde{\eta}_1 = \|\tilde{Q}\|_2$ increase. As a consequence, the effectiveness of $\tilde{M}^{-1}$ improves with respect to M^{-1} . In this test problem, the ratio $r(\eta_1)$ between the conditioning numbers of (4.67) and (4.59) increases up to 5.709.

Next, we introduce also the approximation $G^T G$ for K^{-1} . The matrix G is computed as the FSAI of K using the adaptive strategy implemented in the FSAIPACK software package [Janna et al., 2015b]. The quality of G is controlled by the maximum number of entries $k_{G,max}$ retained per row. First of all, notice that $\tilde{S}$ computed as in equation (4.40) might be indefinite if G is not accurate enough. For instance, even with $k_{G,max} = 250$, the approximate Schur complement $(C - B^T G^T G B)$ is still characterized by 1 negative eigenvalue. The representativeness of the theoretical bounds (4.76) and (4.77) is investigated in Table 4.3 for different choices of G and F . It can be observed that the role played by ε_1 and $\tilde{\eta}_1$, i.e., a measure of the quality of G and F , respectively, has a similar impact on the actual eigenvalue distribution of $\tilde{M}^{-1}\tilde{A}\tilde{M}^{-T}$, as expected from the right-hand side of inequality (4.76). Hence, one may argue that improving both G and F is essential for a better MF performance. On the other hand, the bound (4.77) is less significant because of the presence of κ_n^{-1} , which can be quite

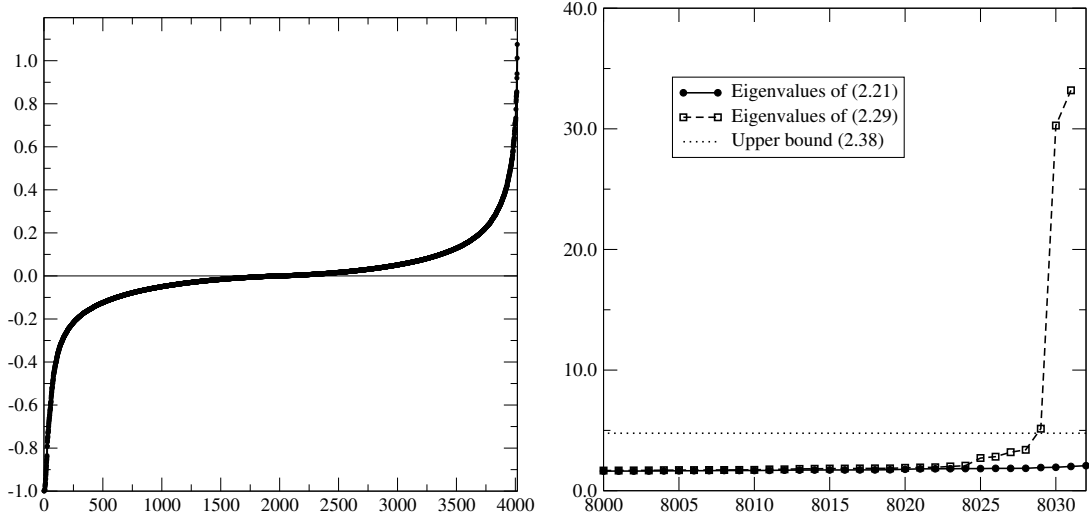


FIGURE 4.9: bcsstk38 test case: eigenspectrum of E_K (left) and right-most eigenvalues of the preconditioned matrices (4.67) and (4.59) (right) for $k_{G,max} = 30$ and $k_{F,max} = 60$. The upper bound (4.76) is also shown.

large. Nonetheless, the actual eigenvalue distribution of $M^{-1}AM^{-T}$ is always worse than that of $\widetilde{M}^{-1}\widetilde{A}\widetilde{M}^{-T}$.

As an example, Figure 4.9 shows the eigenvalue distribution of E_K for $k_{G,max} = 30$ and the preconditioned matrices (4.67) and (4.59) for the same $k_{G,max}$ and $k_{F,max} = 60$. The matrix E_K is indefinite, with the eigenvalues almost equally distributed between negative and positive values approximately in the interval $[-1, 1]$. The matrix $\widetilde{M}^{-1}\widetilde{A}\widetilde{M}^{-T}$ has a more favorable eigenvalues distribution than $M^{-1}AM^{-T}$.

The effect of Descending and Ascending Low-Rank corrections on the quality of the MFLR preconditioner is finally shown in Table 4.4, where the $\widetilde{M}^{-1}\widetilde{A}\widetilde{M}^{-T}$ conditioning number for $k_{G,max} = 30$ and $k_{F,max} = 60$ is reported for a different number of eigenpairs computed to correct either G (*dln*) or $\widehat{S}^{-1}$ (*aln*). The use of both low-rank strategies allows a great reduction of the conditioning number. For example, in this case, the retention of 20 eigenpairs for both corrections yields a reduction of the conditioning number of about two orders of magnitude, i.e., from about 24,000 to 500. By distinction, correcting $\widetilde{S}^{-1}$ towards S^{-1} is not as effective as expected from the MFLR theoretical properties. It is also interesting to observe that the largest eigenvalue is the most sensitive to the selection of the target matrix for the Ascending Low-Rank corrections, while it is insensitive to the Descending Low-Rank corrections, which mainly affect the smallest eigenvalue.

TABLE 4.4: bcsstk38 test case: effect of Descending and Ascending Low-Rank Corrections, dlr and alr , respectively, with $k_{G,max} = 30$ and $k_{F,max} = 60$. The target of Ascending Low-Rank correction is $\tilde{S}^{-1}$ and S^{-1} in the upper and lower table, respectively.

alr target	alr	dlr	$[\lambda_n, \lambda_1]$	λ_1/λ_n
$\tilde{S}^{-1}$	0	0	[1.682e-04, 2.337e+00]	2.389e+04
	0	10	[4.391e-04, 2.337e+00]	5.321e+03
	0	20	[4.438e-04, 2.337e+00]	5.265e+03
	10	0	[1.682e-04, 2.076e+00]	1.234e+04
	10	10	[4.391e-04, 2.077e+00]	4.731e+03
	10	20	[4.438e-04, 2.137e+00]	4.816e+03
	20	0	[1.744e-04, 2.076e+00]	1.190e+04
	20	10	[3.380e-03, 2.077e+00]	6.145e+02
	20	20	[4.413e-03, 2.137e+00]	4.843e+02
S^{-1}	10	0	[1.753e-04, 1.467e+01]	8.371e+04
	10	10	[4.342e-03, 1.482e+01]	3.414e+03
	10	20	[6.308e-03, 1.493e+01]	2.367e+03
	20	0	[1.772e-04, 3.220e+01]	1.817e+05
	20	10	[5.475e-03, 3.270e+01]	5.973e+03
	20	20	[9.065e-03, 3.290e+01]	3.630e+03

To test the influence of the user-specified parameters that control the MFLR behavior, we use the medium-size matrix *Cube* arising from a homogeneous and uniformly discretized structural problem by $P1$ finite elements, with 190,581 unknowns and 7,531,389 non-zero entries. The linear system (4.1) is solved by using an MFLR-preconditioned Conjugate Gradient (PCG) method, with an exit tolerance on the relative residual equal to 10^{-8} . The right-hand side $\mathbf{b}$ is such that the solution $\mathbf{x}$ is the vector with components $x_j = j + 1, j = 1, \dots, n$. All the tests here reported are obtained using a machine equipped with Intel®Xeon®E5-2680 v2 processors at 2.80 GHz and 256 Gbyte of RAM. Each CPU has 10 cores. For these tests, just one thread is used. Indeed, the main goal of this work is to prove the robustness and effectiveness of the multilevel framework in explicit preconditioning. The MFLR preconditioner is implemented in Fortran90, with the code compiled by the Intel Fortran compiler using the -O3 optimization level. We also used BLAS and LAPACK routines [Anderson et al., 1990] from the Intel Math Kernel Library [Wang et al., 2014]. For the computation of the eigenpairs needed by the Low-Rank corrections, we used the *Laneig* software, that is part of the *Filtlan* package [Fang and Saad, 2012].

TABLE 4.5: Cube test case: MFLR performance varying n_l and ϵ_G ($\epsilon_F = 10^{-2}$, $dlr = alr = 0$).

ϵ_G	n_l	n_{it}	ρ	T_p	T_s	ϵ_G	n_l	n_{it}	ρ	T_p	T_s
10^{-1}	10	930	0.43	14.5	26.8	10^{-2}	10	640	0.77	34.5	21.9
	20	843	0.73	132.0	34.3		20	591	1.17	265.9	29.3
	50	664	1.47	349.9	49.1		50	503	2.04	441.7	41.6
	100	612	2.78	515.1	79.2		100	464	3.67	612.9	67.0
10^{-3}	10	446	1.97	268.4	24.0	10^{-4}	10	362	3.83	2232.9	37.3
	20	416	2.53	881.2	29.5		20	338	4.67	3741.6	35.7
	50	386	3.45	829.9	41.8		50	329	5.74	2404.2	47.4
	100	367	5.71	956.9	66.3		100	318	8.88	2053.8	73.1

The Cube test matrix is reordered with the Reverse Cuthill-McKee algorithm and uniformly partitioned into equal-size levels. The results are evaluated in terms of number of iterations, n_{it} , time needed to compute the preconditioner, T_p , the time spent in the PCG iterations, T_s , and the preconditioner density, ρ , defined as:

$$\rho = \frac{1}{\text{nnz}(A)} \sum_{i=0}^{n_l-1} (\text{nnz}(Q_{alr}^i) + \text{nnz}(Q_M^i)) \quad (4.112)$$

where $\text{nnz}(\cdot)$ gives the number of non-zero entries stored for the operator $(\cdot)$. All the remaining symbols are defined as in Algorithm (6). For these preliminary tests, the matrix-matrix products are fully computed, i.e., any use of thresholds and/or levels of fill is avoided.

Table 4.5 shows the results obtained by varying the number of levels n_l and the tolerance ϵ_G . The other tolerance ϵ_F is set to 10^{-2} and no low-rank corrections are applied, i.e., $dlr = alr = 0$. As expected, the iteration count decreases progressively as ϵ_G decreases, i.e., G is more accurate, and n_l grows. The preconditioner density also increases, so that the set-up burden and the cost per iteration grows. Notice that, in particular, the set-up time can become very heavy. Although such a cost could be reduced by introducing thresholds and level-of-fills in the matrix-matrix computations with no substantial loss in the PCG acceleration, it appears that the proposed multilevel approach is of interest whenever the preconditioner can be re-used several times, e.g. in eigensolvers or in some transient simulations, so that its set-up cost can be properly amortized and set apart in the present analysis. Hence, we focus on the solution time T_s only. With the most efficient ϵ_G value, i.e., 10^{-2} , we vary ϵ_F and n_l . The results,

TABLE 4.6: Cube test case: MFLR performance varying n_l and ϵ_F ($\epsilon_G = 10^{-2}$, $dlr = alr = 0$).

ϵ_F	n_l	n_{it}	ρ	T_p	T_s	ϵ_F	n_l	n_{it}	ρ	T_p	T_s
10^{-1}	10	660	0.65	29.1	22.8	10^{-2}	10	640	0.77	34.5	21.9
	20	632	0.94	216.8	29.5		20	591	1.17	265.9	29.3
	50	601	1.61	369.8	47.8		50	503	2.04	441.7	41.6
	100	579	2.81	498.0	78.7		100	464	3.67	612.9	67.0
10^{-3}	10	621	1.13	85.7	27.0	10^{-4}	10	585	2.304	1163.3	35.1
	20	544	1.79	639.2	32.9		20	460	3.445	4513.5	39.3
	50	403	3.21	1129.2	43.0		50	309	5.550	4020.9	43.6
	100	348	6.98	2278.7	71.2		100	255	11.192	6029.8	68.6

TABLE 4.7: Cube test case: MFLR performance varying either dlr or alr ($\epsilon_G = \epsilon_F = 10^{-2}$, $n_l = 10$).

dlr	n_{it}	ρ	T_p	T_s	alr	n_{it}	ρ	T_p	T_s
5	534	0.90	71.3	29.3	5	289	1.35	95.1	17.0
10	485	1.01	66.1	29.4	10	213	1.92	78.0	15.1
20	450	1.24	65.9	25.5	20	196	3.06	88.1	14.1
50	405	1.92	75.0	30.1	50	186	6.48	214.5	36.7

provided in Table 4.6, show that the MFLR preconditioner appears to be less sensitive to the quality of F than of G . In fact, the iteration count is more stable than in Table 4.5.

With $\epsilon_G = 10^{-2}$, $\epsilon_F = 10^{-2}$ and $n_l = 10$, we test the effects of Descending and Ascending Low-Rank corrections. Table 4.7 shows the impact of using either Descending Low-Rank or Ascending Low-Rank corrections. Increasing the rank size, the number of iterations decreases for both approaches, but Ascending Low-Rank corrections have a stronger impact on the solution time T_s . This is somewhat expected because Ascending Low-Rank corrections have a global effect on the overall preconditioner, while Descending Low-Rank corrections improve locally the current level approximation of K^{-1} . Finally, the combined effect of both corrections is investigated in table 4.8. This allows for obtaining the best result in terms of both iteration count and solution time T_s decrease. In particular, the latter is more than halved with respect to the case with no corrections with at least $dlr = alr = 10$. The solver acceleration is paid with a bigger set-up cost, that makes the MFLR approach interesting when the preconditioner can be recycled.

The computational performance of the MFLR preconditioner is evaluated in a set of large size SPD matrices taken from the University of Florida Sparse Matrix Collection [Davis and Hu, 2011]. The main properties of the selected test cases are provided in

TABLE 4.8: Cube test case: MFLR performance varying both dlr and alr ($\epsilon_G = \epsilon_F = 10^{-2}$, $n_l = 10$).

dlr	alr	n_{it}	ρ	T_p	T_s
1	1	525	0.92	76.3	34.4
1	10	204	1.95	87.7	14.8
1	20	188	3.09	112.5	18.5
10	1	381	1.13	71.7	25.0
10	10	148	2.15	83.6	10.9
10	20	134	3.29	100.0	11.3
20	1	341	1.35	86.6	23.9
20	10	132	2.38	100.5	10.4
20	20	123	3.52	100.1	10.8

TABLE 4.9: MFLR: Test matrices.

Name	Rows	Nonzeros	Description
af_shell3	504,855	17,588,875	structural problem
af_shell8	504,855	17,579,155	structural problem
Emilia_923	923,136	40,373,538	structural problem
Geo_1438	1,437,960	60,236,322	structural problem
StocF_1465	1,465,137	21,005,389	computational fluid dynamics problem

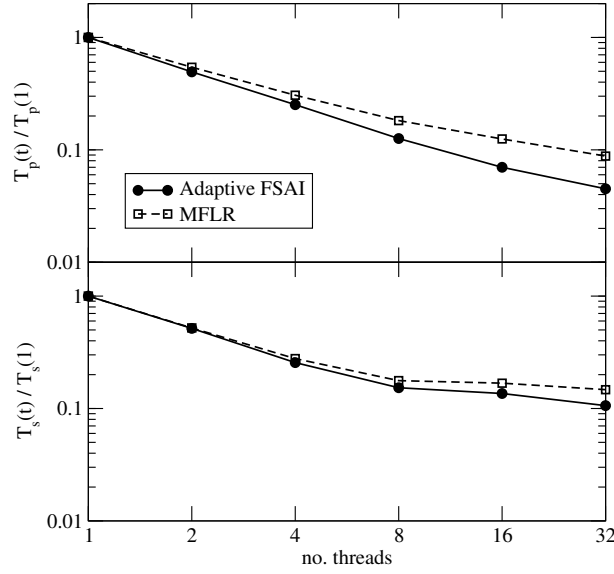
Table 4.9, while the best results obtained by the MFLR preconditioner in terms of iteration and CPU time T_s are given in Table 4.10. As a benchmark, Table 4.10 also shows the performance provided by the native adaptive FSAI algorithm.

The linear system (4.1) is solved by using an MFLR-preconditioned Conjugate Gradient (PCG) method, with the same right-hand side and exit tolerance as before. The local cluster previously described is used.

As already observed, the MFLR preconditioner set-up can be quite expensive, especially because of the computation of the eigenpairs needed by the Low-Rank correction procedures. However, its effectiveness in the iteration count and CPU time can be quite significant. For instance, in the af_shell3 and af_shell8 test cases, n_{it} is

TABLE 4.10: Computational performance of the adaptive FSAI and MFLR preconditioners for the test matrices of Table 4.9.

	Adaptive FSAI				MFLR			
	n_{it}	ρ	T_p [s]	T_s [s]	n_{it}	ρ	T_p [s]	T_s [s]
af_shell3	963	0.89	88.8	63.9	126	4.50	237.8	31.5
af_shell8	1033	0.86	81.2	68.0	131	3.22	174.8	28.6
Emilia_923	1513	0.11	5.0	128.5	575	0.25	82.8	80.4
Geo_1423	491	0.31	59.4	89.6	456	0.49	159.8	110.4
StocF_1465	937	0.93	51.3	133.4	936	1.66	260.4	222.2

FIGURE 4.10: Strong scalability test for a regular 300^3 Laplacian.

approximately reduced by a factor 10 and T_s is more than halved. Again, the use of the MFLR preconditioner can be of great interest whenever the set-up time can be amortized along several linear solves (e.g. eigenvalues computation [Sleijpen and Van der Vorst, 2000]). On the other hand, when the reduction of the number of iterations is marginal, such as in the `Geo_1438` and `StocF_1465` test cases, the Adaptive FSAI is more efficient than the MFLR preconditioner. Finally, to show the potential parallelism of the MFLR preconditioner, a scalability test has been carried out on a discrete Laplacian computed over a regular $300 \times 300 \times 300$ grid. The numerical experiment is performed on the Marconi cluster at the CINECA Center of High Performance Computing, Bologna, Italy. Currently, the cluster consists of 1,512 nodes, each one with 128 Gbyte of RAM memory. Every node is equipped with 2 Intel®Xeon®E5-2697v4 Broadwell processors at 2.30 GHz with 36 cores. This preliminary implementation of MFLR makes use of shared-memory parallelism through the OpenMP directives, hence the strong scalability test provided in Figure 4.10 is performed using a single node, with up to 32 threads. The strong scalability of the MFLR preconditioner is compared to that of the Adaptive FSAI algorithm as far as both the set-up and the iteration time, T_p and T_s , is concerned. In particular, Figure 4.10 provides the speed-ups:

$$S_p = \frac{T_p(t)}{T_p(1)}, \quad S_s = \frac{T_s(t)}{T_s(1)} \quad (4.113)$$

with $T_p(t)$ and $T_s(t)$ the set-up and iteration wall-clock times measured with t threads. It has been already verified that the native Adaptive FSAI is practically characterized by an ideal speed-up [Janna et al., 2015b], according to the hardware properties of the specific computational architecture used for the numerical experiments. In the case of the multi-core processors of the Marconi cluster, it is well-known that it is virtually impossible to obtain ideal speed-ups with iterative solvers because these algorithms are bandwidth limited being characterized by a low bit per flop ratio. The MFLR preconditioner is theoretically less parallel than FSAI because of the intrinsic sequentiality introduced by the multilevel framework. Nevertheless, Figure 4.10 shows that the strong scalability is only marginally affected by such a sequentiality so that the MFLR preconditioner still preserves a good degree of parallelism.

4.5 The Inexact Constraint Preconditioning framework

Finally, we consider the overall performance of the inexact constraint preconditioning (ICP) for solving the system (4.2).

As J in (4.2) is non-symmetric and indefinite, we elect to use the Bi-Conjugate Gradient Stabilized (Bi-CGStab, [Van der Vorst, 1992]) as a Krylov solver. If properly preconditioned, Bi-CGStab has proved to be a robust and efficient solver for large and sparse non-symmetric systems of equations, with an intrinsic higher parallel degree than other non-symmetric solvers, such as GMRES [Saad and Schultz, 1986]. The scheme of the Preconditioned Bi-CGStab is provided in Algorithm 8. The iterative procedure can be stopped whenever the 2-norm of the current residual vector $\mathbf{r}_{k+1}$ falls below a prescribed tolerance. It is well-known that the inner linear steps of a Newton scheme do not need to be solved with a high accuracy far from convergence. Typically, the exit tolerance TOL of the linear solver can be decreased as the outer Newton residual approaches zero.

The key for the success of the Algorithm 8 is the choice of the preconditioner. Recalling equation (4.2), the computation of $\mathbf{v} = [\mathbf{v}_u, \mathbf{v}_\lambda]^T$ resulting from the application

Algorithm 8 PRECONDITIONED BI-CGSTAB

```

1: procedure BICGSTAB( $A, \mathbf{b}, M^{-1}, \mathbf{x}_0, TOL$ )
2:   Compute  $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$  for some initial guess  $\mathbf{x}_0$ 
3:   Choose  $\mathbf{r}^*$  (for example,  $\mathbf{r}^* = \mathbf{r}_0$ )
4:    $\mathbf{p}_0 = \mathbf{r}_0$ 
5:    $\rho_0 = \mathbf{r}^{*T} \mathbf{r}_0$ 
6:   for  $k = 0, 1, 2, \dots$  do
7:      $\mathbf{p}'_k = M_2^{-1} \mathbf{p}_k$ 
8:      $\mathbf{v} = A\mathbf{p}'_k$ 
9:      $\mathbf{v} \leftarrow M_1^{-1} \mathbf{v}$ 
10:     $\alpha_k = \rho_k / \mathbf{r}^{*T} \mathbf{v}$ 
11:     $\mathbf{s}_k = \mathbf{r}_k - \alpha_k \mathbf{v}$ 
12:     $\mathbf{s}'_k = M_2^{-1} \mathbf{s}_k$ 
13:     $\mathbf{t} = A\mathbf{s}'_k$ 
14:     $\mathbf{t} \leftarrow M_1^{-1} \mathbf{t}$ 
15:     $\omega_k = \mathbf{t}^T \mathbf{s}_k / \mathbf{t}^T \mathbf{t}$ 
16:     $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}'_k + \omega_k \mathbf{s}'_k$ 
17:     $\mathbf{r}_{k+1} = \mathbf{s}_k - \omega_k \mathbf{t}$ 
18:    if  $\|\mathbf{r}_{k+1}\|_2 < TOL$  then
19:      return  $x_{k+1}$ 
20:    else
21:       $\rho_{k+1} = \mathbf{r}^{*T} \mathbf{r}_{k+1}$ 
22:       $\beta_k = (\rho_{k+1} \alpha_k) / (\rho_k \omega_k)$ 
23:       $\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_k \mathbf{p}'_k - \omega_k \beta_k \mathbf{v}$ 
24:    end if
25:  end for
26:  return  $x_{k+1}$ 
27: end procedure

```

of J^{-1} to a vector $\mathbf{r} = [\mathbf{r}_u, \mathbf{r}_\lambda]^T$ can be done by solving the following set of linear equations:

$$\begin{cases} A\mathbf{v}_u + B_1\mathbf{v}_\lambda = \mathbf{r}_u \\ B_2^T\mathbf{v}_u = \mathbf{r}_\lambda \end{cases} \quad (4.114)$$

Solving for $\mathbf{v}_u$ in the upper set of equations and introducing the result in the lower set provides:

$$\begin{cases} \mathbf{v}_\lambda = S^{-1}(\mathbf{r}_\lambda - B_2^T A^{-1} \mathbf{r}_u) \\ \mathbf{v}_u = A^{-1}(\mathbf{r}_u - B_1 \mathbf{v}_\lambda) \end{cases} \quad (4.115)$$

where the matrix $S \in \mathbb{R}^{n_\lambda \times n_\lambda}$:

$$S = -B_2^T A^{-1} B_1 \quad (4.116)$$

is the Schur complement of J . Note that S is generally non-symmetric because of the non-symmetry of J . The constraint preconditioning operator M_C^{-1} can be defined as an incomplete computation of equation (4.114) where A^{-1} and S^{-1} are replaced by some approximations, i.e., by preconditioners of A and S . The approximate computation of the Schur complement S can be naturally obtained by replacing A^{-1} with $G_1^T G_1$:

$$S \simeq \tilde{S} = -B_2^T G_1^T G_1 B_1 \quad (4.117)$$

However, the approximate S of equation (4.117) is non-symmetric. Preconditioners for non-symmetric matrices based on aFSAI have been recently developed, e.g., [Ferronato et al. \[2014, 2016\]](#), but they are generally less robust than those for SPD problems.

The preconditioning operator resulting by the approximation of A^{-1} and S^{-1} in equation (4.115) is generally referred to in the literature as *Inexact Constraint Preconditioner* (ICP, see [Bergamaschi et al. \[2008\]](#); [Janna et al. \[2012b\]](#)). The word “constraint” recalls that this framework was originally developed for the saddle-point systems arising in constrained optimization problems [[Bergamaschi et al., 2004](#); [Benzi et al., 2005](#)], while the qualification of “inexact” identifies the fact that an approximate Schur complement is actually computed.

4.5.1 ICP description and algorithms

As widely discussed, we elect to replace the inverse of the SPD matrix A with FSAI or one of the FSAI-based multilevel approaches. For the sake of simplicity, the following developments assume to use the simplest factorized approximate inverse, i.e. the raw Adaptive FSAI. All other FSAI-based multilevel approaches can be used in the same way, due to the fact that they can be written in a factorized form. Thus, setting $G_1^T G_1 \simeq A^{-1}$, we have:

$$J = \begin{bmatrix} A & B_1 \\ B_2^T & 0 \end{bmatrix} \simeq \begin{bmatrix} G_1^{-1} & 0 \\ \widetilde{H}_2^T & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \widetilde{S} \end{bmatrix} \begin{bmatrix} G_1^{-T} & \widetilde{H}_1 \\ 0 & I \end{bmatrix} \quad (4.118)$$

where $\widetilde{H}_1 = G_1 B_1$, $\widetilde{H}_2 = G_1 B_2$ and $\widetilde{S} = -\widetilde{H}_2^T \widetilde{H}_1$. Equation (4.118) can be rewritten as:

$$\begin{bmatrix} G_1 & 0 \\ N_2^T & I \end{bmatrix} \begin{bmatrix} A & B_1 \\ B_2^T & 0 \end{bmatrix} \begin{bmatrix} G_1^T & N_1 \\ 0 & I \end{bmatrix} \simeq \begin{bmatrix} I & 0 \\ 0 & \widetilde{S} \end{bmatrix} \quad (4.119)$$

with $N_1 = -G_1^T G_1 B_1$ and $N_2 = -G_1^T G_1 B_2$. Developing the product on the left hand side of Equation (4.119), we obtain:

$$\widetilde{S} = N_2^T A N_1 + N_2^T B_1 + B_2^T N_1 = B_2^T G_1^T G_1 A G_1^T G B_1 - 2B_2^T G_1^T G_1 B_1 \quad (4.120)$$

Assuming $G_1 A G_1^T = I$, i.e. $A^{-1} = G_1^T G_1$, we have the exact Schur complement, as expected. Nevertheless, equation (4.120) states that approximating the Schur complement in this way is more robust than using its definition. This difference is the same detected between the BTFSAI and the MFLR approaches. Due to the fact that this approach is more expensive from a computational point of view, in this work, we use the first way, i.e. $\widetilde{S} = -\widetilde{H}_2^T \widetilde{H}_1$.

The ICP algorithm can be computed and applied in different ways.

The preconditioning operator M_C^{-1} can be applied to a vector $\mathbf{r}$ simply by replacing A^{-1} of (4.115) with $G_1^T G_1$. The sequence of operations needed to compute $\mathbf{v} = M_C^{-1} \mathbf{r}$ is provided in Algorithm 9.

The use of the Algorithm 9 allows for either a *right* or a *left* preconditioning. The

Algorithm 9 ICP APPLICATION

```

1: procedure APPL_PREC( $B_1, B_2, G_1, \tilde{S}, \mathbf{r}$ )
2:    $\mathbf{w}_u = G_1^T G_1 \mathbf{r}_u$ 
3:    $\mathbf{w}_\lambda = \mathbf{r}_\lambda - B_2^T \mathbf{w}_u$ 
4:    $\mathbf{v}_\lambda = \tilde{S}^{-1} \mathbf{w}_\lambda$ 
5:    $\mathbf{w}_u = \mathbf{r}_u - B_1 \mathbf{v}_\lambda$ 
6:    $\mathbf{v}_u = G_1^T G_1 \mathbf{w}_u$ 
7:    $\mathbf{v} = [\mathbf{v}_u^T, \mathbf{v}_\lambda^T]^T$ 
8:   return  $\mathbf{v}$ 
9: end procedure

```

right preconditioning is obtained by setting $M_1^{-1} = I$ and $M_2^{-1} = M_C^{-1}$ in Algorithm 8, while the left preconditioning is used with $M_1^{-1} = M_C^{-1}$ and $M_2^{-1} = I$. In norm-minimizing schemes, such as GMRES, the right preconditioning is to be usually preferred. With Bi-CGSTab the computational experience typically does not show significant differences between the two approaches.

Once a factorization (exact or approximate) of $\tilde{S} \simeq L_{\tilde{S}} U_{\tilde{S}}$ is available, the ICP application can be re-formulated so as to provide also a *split* preconditioning, i.e., both M_1^{-1} and M_2^{-1} are different from the identity. The application of the factors M_1^{-1} and M_2^{-1} to a vector $\mathbf{r}$ is summarized in Algorithm 10 and 11, respectively. Other splits are also possible, with additional approximations in the factors M_1^{-1} and M_2^{-1} , see for instance Ferronato et al. [2010a].

Algorithm 10 ICP SPLIT APPLICATION (LEFT)

```

1: procedure APPL_PREC_LEFT( $B_2, G_1, L_{\tilde{S}}, \mathbf{r}$ )
2:    $\mathbf{v}_{1,u} = G_1 \mathbf{r}_u$ 
3:    $\mathbf{w}_u = G_1^T \mathbf{v}_{1,u}$ 
4:    $\mathbf{w}_\lambda = \mathbf{r}_\lambda - B_2^T \mathbf{w}_u$ 
5:    $\mathbf{v}_{1,\lambda} = L_{\tilde{S}}^{-1} \mathbf{w}_\lambda$ 
6:    $\mathbf{v}_1 = [\mathbf{v}_{1,u}^T, \mathbf{v}_{1,\lambda}^T]^T$ 
7:   return  $\mathbf{v}_1$ 
8: end procedure

```

4.5.2 Schur complement approximations and inversion

The ICP algorithm may differ according to the way the Schur complement is computed. Here we consider three alternatives:

Algorithm 11 ICP SPLIT APPLICATION (RIGHT)

```

1: procedure APPL_PREC_RIGHT( $B_1, G_1, U_{\tilde{S}}, \mathbf{r}$ )
2:    $\mathbf{v}_{2,\lambda} = U_{\tilde{S}}^{-1} \mathbf{r}_\lambda$ 
3:    $\mathbf{w}_u = B_1 \mathbf{v}_{2,\lambda}$ 
4:    $\mathbf{w}_u \leftarrow \mathbf{r}_u - G_1 \mathbf{w}_u$ 
5:    $\mathbf{v}_{2,u} = G_1^T \mathbf{w}_u$ 
6:    $\mathbf{v}_2 = [\mathbf{v}_{2,u}^T, \mathbf{v}_{2,\lambda}^T]^T$ 
7:   return  $\mathbf{v}_2$ 
8: end procedure

```

1. Fully non-symmetric way, i.e. $\tilde{S} = -\tilde{H}_2^T \tilde{H}_1$. To approximate the inverse of $\tilde{S}$ we can use an incomplete LU factorization [Gupta, 2017] or a generalized approximate inverse for non-symmetric matrices [Ferronato et al., 2014].
2. Symmetric way, i.e. considering only the upper triangular block matrix (4.2). We compute $B = B_1 = C - F$ and we assume $B_2 = B_1$. Such an assumption is justified if $F \simeq 0$. This allows the Schur complement to be symmetric and negative definite, so a FSAI or a multilevel FSAI can be used to build an operator applying the inverse of $\tilde{S}$, i.e. $\tilde{S} \simeq -G_2^{-1} G_2^{-T}$. In Algorithms 10 and 11, $L_{\tilde{S}} = -G_2^{-1}$ and $L_{\tilde{U}} = G_2^{-T}$. Note that, with this choice, no triangular solvers are needed.
3. A third way is to exploit some physically-based assumptions related to the locality of deformation, i.e. a variation of the stress on a fault interface element produces a variation of displacements only on the close vicinity of that element. This is very similar to what was developed by Aagaard et al. [2013]; Jha and Juanes [2014], who used this assumption to compute the increment in the Newton-Raphson algorithm, while we use it for the preconditioner. A similar approach is proposed in the field of mixed finite elements applied to coupled poromechanics by Castelletto et al. [2016]; White et al. [2016]. This method allows us to obtain either a symmetric or a non-symmetric matrix, according to the fact that $B = B_1 = C - F$ and $B_2 = B_1$ or $B_1 = C - F$ and $B_2 = C$, respectively. In this case, to apply exactly the inverse of $\tilde{S}$ through a direct solver is the best choice, given the Schur complement block structure.

Finally, it is always possible to apply exactly the inverse of $\tilde{S}$, through a direct solver, e.g. MUMPS [Amestoy et al., 2000], PARDISO [Schenk and Gärtner, 2004; Kalinkin

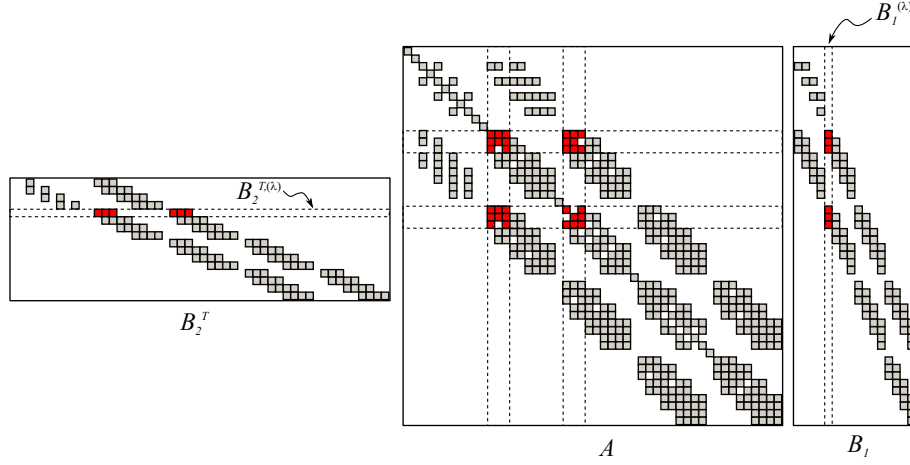


FIGURE 4.11: Non-zero entries (in red) of B_2^T , A , and B_1 required in the approximate computation of $B_2^{T,(λ)} A^{-1} B_1^{(λ)}$. Adapted from [Castelletto et al. \[2016\]](#).

[et al., 2014\]](#), SuperLU_dist [[Li and Demmel, 2003](#)] and WMPS [[Gupta et al., 2001](#)]. In this case, in Algorithms 10 and 11, $L_{\tilde{S}} = \tilde{S}^{-1}$ and $U_{\tilde{S}} = I$ or $L_{\tilde{S}} = I$ and $U_{\tilde{S}} = \tilde{S}^{-1}$.

The procedure to compute the physically-based Schur complement approximation consists in the clusterization of the three unknowns (i.e. λ_ζ , λ_ξ and λ_η) related to each Lagrangian multiplier λ into supernodes (for a definition of supernode see, e.g., [Janna et al. \[2015a\]](#)). For each supernode, the corresponding submatrices of B_2^T , A and B_1 are collected. With reference to Fig. 4.11, $B_2^{T,(λ)}$ is a three-row matrix, $A^{(λ)}$ is an SPD submatrix and $B_1^{(λ)}$ is a three-column matrix. Then, for each supernode we compute:

$$\tilde{S}^{(λ)} = -B_2^{T,(λ)} \left[A^{(λ)} \right]^{-1} B_1^{(λ)} \quad (4.121)$$

and store it in the corresponding position of $\tilde{S}$, that becomes a block diagonal matrix, with 3×3 blocks, that can be exactly inverted.

4.6 Numerical results

First of all, we have to notice that the matrix $F(\mathbf{u}, \boldsymbol{\lambda})$, despite ensuring a quadratic convergence of the non-linear algorithm, has a negligible magnitude. For instance, for Case C of Chapter 3, we computed the Frobenius norm of the Jacobian with the largest

degree of non-symmetry, i.e. when the largest fault activation is reached, obtaining:

$$\|J\|_F = 1.36 \cdot 10^{14} \quad \text{and} \quad \|J - J^T\|_F = 8.44 \quad (4.122)$$

From (4.122), we have that the non-symmetric part the Jacobian has a norm that is about 13 order of magnitude less than the one of the matrix itself. The same behavior is observed with the Jacobians derived from other applications.

Due to (4.122), a symmetric preconditioner appears to be justified. Summarizing the alternatives listed above, we have three options:

1. FSAI + FSAI: the leading block is preconditioned with FSAI, the approximate Schur complement $\tilde{S} = -\tilde{H}^T \tilde{H}$ is symmetric and negative definite and FSAI is used also for the second block;
2. FSAI + Direct: only the leading block is preconditioned with FSAI. Once the Schur complement is computed as before, its inverse is applied exactly.
3. FSAI + block diagonal Schur (BDSchur): as before, FSAI is used for the leading block, while the approximate Schur complement is computed through the block diagonal approach.

All tests reported in this section are obtained using a local cluster equipped with Intel®Xeon®E5-2680 v2 processors at 2.80 GHz and 256 Gbyte of RAM. Each CPU has 10 cores and used up to 32 threads, with hyper-threading enabled.

The procedure to tune the aFSAI parameters is described in Janna et al. [2015b]. The most important aFSAI parameter is ϵ . The total number of steps k_{max} is fixed to 100. The other two parameters, τ and s , are set to the default value, 0.0 and 1, respectively (Table 4.11). We vary only ϵ , denoting with ϵ_1 the parameter for the FSAI of the leading block and with ϵ_2 the parameter for the FSAI of the Schur complement, used only with approach 1. The matrix-matrix products are computed retaining all entries.

The main properties of the selected test cases are provided in Table 4.12.

The linear system (4.1) is solved by using the Bi-Conjugate Gradient Stabilized (Bi-CGstab) method, with an exit tolerance on the relative residual equal to 10^{-8} . The right-hand side $\mathbf{b}$ is such that the solution $\mathbf{x}$ is the vector with components $x_j = 1, j = 1, \dots, n$.

TABLE 4.11: FSAI set-up parameters.

Parameter	Value / range
k_{max}	100
s	1
ϵ	$10^{-1} \div 10^{-3}$
τ	0.0

TABLE 4.12: Size and number of non-zeroes of the test matrices.

	Total size	$n_1 + n_2$	Number of non-zeroes
Case B	183,177	171,150 + 12,027	8,315,517
Case C	72,666	69,909 + 2,757	2,982,756
Case D	1,180,764	1,142,655 + 38,109	50,354,166

In table 4.13, T_p , T_s and T_t are preconditioning time, solve time and total time, respectively. The definition of the preconditioner density ρ changes definition according to the three options outlined above for the preconditioner structure:

$$\begin{aligned}
 \rho_1 &= \frac{\text{nnz}(F) + \text{nnz}(H) + \text{nnz}(G)}{\text{nnz}(A)} \\
 \rho_2 &= \frac{\text{nnz}(F) + \text{nnz}(H) + \text{nnz}(L_S)}{\text{nnz}(A)} \\
 \rho_3 &= \frac{\text{nnz}(F) + \text{nnz}(H) + \text{nnz}(S)}{\text{nnz}(A)}
 \end{aligned} \tag{4.123}$$

Tables 4.13, 4.14 and 4.15 collect the best results obtained by tuning the parameter ϵ_1 (and ϵ_2 in the first case).

Slight changes in the parameters of the optimal configuration do not modify the results. Thus, we can state that the different approaches of the presented preconditioner are robust.

Comparing tables 4.13, 4.14 and 4.15, we can note that approaches 1 and 2 are comparable, while approach 3 is significantly weaker. In particular, with Case D it does

TABLE 4.13: Best numerical results with approach 1, FSAI + FSAI.

Case	ϵ_1	ϵ_2	n_{iter}	ρ	T_p [s]	T_s [s]	T_t [s]
B	$5 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	48	0.379	1.222	0.508	1.730
C	$5 \cdot 10^{-1}$	$1 \cdot 10^{-3}$	64	0.096	0.082	0.081	0.163
D	$1 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	277	0.252	2.557	13.073	15.630

TABLE 4.14: Best numerical results with approach 2, FSAI + Direct.

Case	ϵ_1	n_{iter}	ρ	T_p [s]	T_s [s]	T_t [s]
B	$5 \cdot 10^{-2}$	70	0.730	0.625	2.227	2.852
C	$5 \cdot 10^{-1}$	65	0.101	0.244	0.163	0.408
D	$5 \cdot 10^{-2}$	263	0.169	1.046	10.489	11.535

TABLE 4.15: Best numerical results with approach 3, FSAI + BDSchur.

Case	ϵ_1	n_{iter}	ρ	T_p [s]	T_s [s]	T_t [s]
B	$1 \cdot 10^{-2}$	159	0.312	1.467	1.709	3.176
C	$5 \cdot 10^{-3}$	211	0.219	0.364	0.491	0.855
D	—	> 3000	—	—	—	—

not achieve the convergence, even if varying the parameter ϵ_1 . Generally, considering both the total time to converge and the memory footprint, i.e. ρ , we can say that the approach 1 seems to be the best for small problems, indeed, for Cases B and C, it needs less time and memory to reach the convergence. For big problems, as Case D, the approach 2 becomes the best, employing less time and memory to converge. Usually direct solvers have a larger memory footprint than preconditioners. In this specific case, however, the overall density is the combination of a two level procedure. In the Approach 1, the optimal configuration requires a denser FSAI with respect to the Approach 2, making ρ_1 higher than ρ_2 , despite of the FSAI approximation of the Schur complement has significantly less non-zeroes, 304,626, than its exact Cholesky factor, 1,188,480.

Finally, to show the parallelism of the proposed constraint preconditioner, a scalability test has been carried out on the Case D matrix. The numerical experiment is performed on the already described local cluster. This implementation of the preconditioner makes use of shared-memory parallelism through the OpenMP directives, hence in the strong scalability test provided in Figure 4.12, up to 32 threads are used. The set-up time, the iteration time and the total wall-clock time, T_p , T_s and T_t respectively, are reported both for the strong scalability of the preconditioner and the the ideal profile. In particular, Figure 4.12 provides the speed-ups:

$$S_p = \frac{T_p(t)}{T_p(1)}, \quad S_s = \frac{T_s(t)}{T_s(1)} \quad \text{and} \quad S_t = \frac{T_t(t)}{T_t(1)} \quad (4.124)$$

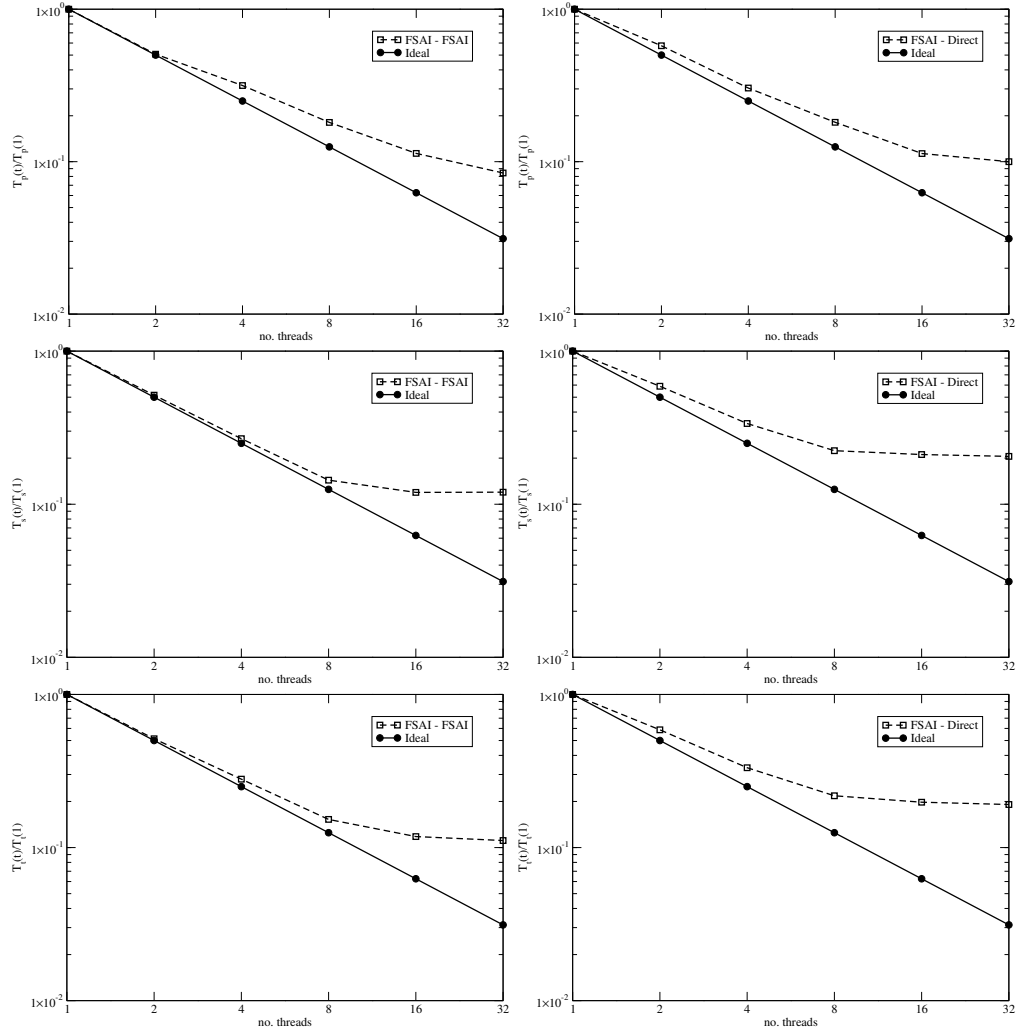


FIGURE 4.12: Strong scalability test for Case D matrix. Approaches 1 and 2 are used.

with $T_p(t)$, $T_s(t)$ and $T_t(t)$ the wall-clock times measured with t threads. Comparing approaches 1 and 2, we can note that the first has a better strong scalability, mainly due to the scalability of the FSAI of the Schur complement.

Chapter 5

Conclusions

A novel formulation based on the use of Lagrange multipliers is developed for the stable and robust numerical modeling of fault mechanics. A fault or fracture is simulated as a pair of inner surfaces included in a 3D geological formation where Lagrange multipliers are used to prescribe the contact constraint. The activation of a fault is controlled by the classical Mohr-Coulomb failure criterion. The standard variational formulation of the contact problem with Lagrange multipliers is modified to take into account the energy dissipated by the frictional work along the activated fault portion. This term is computed by making use of the principle of maximum plastic dissipation, whose application defines the direction of the limiting shear stress vector.

The numerical model is developed in the framework of the FE method and gives rise to a non-linear system of algebraic equations. The non-linear system is solved by a Newton-Raphson method where the Jacobian matrix is non-symmetric and has a generalized saddle-point structure. The proposed formulation is validated and investigated in some test cases, using piecewise linear and element-wise constant approximations for displacements and contact stresses, respectively. The following results are worth summarizing:

- the novel formulation has been successfully validated against analytical solutions;
- the non-linear algorithm exhibits a quadratic convergence even in numerically challenging configurations, providing stable and smooth solutions;

- the numerical contribution arising from the application of the principle of maximum plastic dissipation to compute the frictional work on the activated fault portion is decisive to ensure the convergence of the non-linear algorithm;
- by distinction with the classical penalty approach, the numerical solution appears to be much more regular, avoiding local stress oscillations.

The proposed approach has been tested on real-world problems. To verify the theoretical behaviour of the model, the mechanics of two blocks is analyzed. Then, two different geometries for an over-exploited aquifer subject to ground fissures are studied. Finally, we worked on a reservoir simulation, with the prediction of the possible induced seismicity. In particular, with our model we wanted to predict the fault motion and use these values to estimate the possible earthquake magnitude triggered by the human activities.

The solution of the large and sparse linear system of equations arising from the fault discretization can be challenging. The solver used in this work is a preconditioned Krylov subspace method. The selected preconditioner is based on the constraint approach, that is specifically designed for saddle-point matrices. The linear system has a block structure, with a symmetric positive definite matrix as leading block. First of all, we discussed existing preconditioners for SPD matrices based on factorized approximate inverses and advanced three new approaches in a multilevel framework. Then, different Schur complement approximations are presented and compared.

The development of a multilevel framework is often useful in several applications. However, the use of the FSAI preconditioner as basic kernel in a standard multilevel approach may give rise to some difficulties related with the approximations introduced in the computation of the Schur complement at each level. In SPD problems such a Schur complement might be indefinite, thus causing a breakdown of the multilevel algorithm.

The present work develops a robust multilevel framework for SPD matrices based on the use of Adaptive FSAI as main kernel. An alternative way of computing the Schur complement is introduced so as to guarantee its positive definiteness independently of

the preconditioner sparsity. A theoretical analysis is formulated with the aim of providing appropriate bounds for the eigenspectrum of the preconditioned matrix. The multilevel FSAI preconditioner is further enhanced by introducing Low-Rank corrections at both a local and a global level, namely Descending and Ascending Low-Rank corrections, respectively, thus producing the MFLR preconditioning framework.

The MFLR preconditioner has been investigated in a set of test problems to analyze: (i) the relative influence of the user-specified parameters controlling the algorithm setup, and (ii) the computational performance and potential scalability in a parallel environment. The numerical results show that the proposed approach is generally able to significantly accelerate the solver convergence rate still preserving a good degree of parallelism. At the present time, the solver acceleration is paid off by a large set-up cost, which is mainly due to the computation of the eigenpairs needed by the Low-Rank corrections. The increase of the cost for building the preconditioner with respect to the native Adaptive FSAI makes this approach attractive especially for those applications where the preconditioner can be effectively recycled along a number of linear solves. Further investigations will be devoted in the development of a faster set-up stage, by using for instance randomized approaches while computing low rank corrections [Halko et al., 2011], and enforcing sparsity in the resulting factors.

Finally, some different approximations for the Schur complement of the overall saddle-point matrix are tested. A scalability test is performed to show the parallelism of the proposed constraint preconditioner. The results shows that the degree of non-symmetry of the resulting Jacobian is usually very small and a symmetric preconditioner may suffice and prove more effective than a full non-symmetric approach. When the size of interfaces in the 3D domain is small, also the exact application of the Schur complement inverse with the aid of a direct method appears to be a viable approach.

Bibliography

- Aagaard, B. T., Knepley, M. G., and Williams, C. A. A domain decomposition approach to implementing fault slip in finite-element models of quasi-static and dynamic crustal deformation. *Journal of Geophysical Research: Solid Earth*, 118(6):3059–3079, 2013.
- Ahmad, M. U. and Smith, J. A. Earthquakes, injection wells, and the Perry nuclear power plant, Cleveland, Ohio. *Geology*, 16(8):739–742, 1988.
- Ake, J., Mahrer, K., O’Connell, D., and Block, L. Deep-injection and closely monitored induced seismicity at Paradox Valley, Colorado. *Bulletin of the Seismological Society of America*, 95(2):664–683, 2005.
- Amestoy, P., Ashcraft, C., Boiteau, O., Buttari, A., L’Excellent, J.-Y., and Weisbecker, C. Improving multifrontal methods by means of block low-rank representations. *SIAM Journal on Scientific Computing*, 37(3):A1451–A1474, 2015.
- Amestoy, P. R., Duff, I. S., L’Excellent, J.-Y., and Koster, J. MUMPS: a general purpose distributed memory sparse solver. In *International Workshop on Applied Parallel Computing*, pages 121–130. Springer, 2000.
- Anderson, E. and Saad, Y. Solving sparse triangular linear systems on parallel computers. *International Journal of High Speed Computing*, 1(01):73–95, 1989.
- Anderson, E., Bai, Z., Dongarra, J., Greenbaum, A., McKenney, A., Du Croz, J., Hammerling, S., Demmel, J., Bischof, C., and Sorensen, D. LAPACK: A portable linear algebra library for high-performance computers. In *Proceedings of the 1990 ACM/IEEE conference on Supercomputing*, pages 2–11. IEEE Computer Society Press, 1990.

- Annalarapu, C., Hautefeuille, M., and Dolbow, J. E. A Nitsche stabilized finite element method for frictional sliding on embedded interfaces. Part II: intersecting interfaces. *Computer Methods in Applied Mechanics and Engineering*, 267:318–341, 2013.
- Annalarapu, C., Hautefeuille, M., and Dolbow, J. E. A Nitsche stabilized finite element method for frictional sliding on embedded interfaces. Part I: single interface. *Computer Methods in Applied Mechanics and Engineering*, 268:417–436, 2014.
- Axelsson, O. and Neytcheva, M. Preconditioning methods for linear systems arising in constrained optimization problems. *Numerical linear algebra with applications*, 10(1-2): 3–31, 2003.
- Bai, M. and Abousleiman, Y. Thermoporoelastic coupling with application to consolidation. *International Journal for Numerical and analytical methods in Geomechanics*, 21(2):121–132, 1997.
- Bai, Z.-Z., Ng, M. K., and Wang, Z.-Q. Constraint preconditioners for symmetric indefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(2):410–433, 2009.
- Bailey, I. W. and Ben-Zion, Y. Statistics of earthquake stress drops on a heterogeneous fault in an elastic half-space. *Bulletin of the Seismological Society of America*, 99(3):1786–1800, 2009.
- Bank, R. E., Welfert, B. D., and Yserentant, H. A class of iterative methods for solving saddle point problems. *Numerische Mathematik*, 56(7):645–666, 1989.
- Bathe, K.-J. *Finite element procedures*. Klaus-Jurgen Bathe, 2006.
- Baù, D., Ferronato, M., Gambolati, G., and Teatini, P. Basin-scale compressibility of the Northern Adriatic by the radioactive marker technique. *Geotechnique*, 52(8):605–616, 2002.
- Bear, J. *Dynamics of fluids in porous media*. Courier Corporation, 2013.
- Beer, G. An isoparametric joint/interface element for finite element analysis. *International journal for numerical methods in engineering*, 21(4):585–600, 1985.

- Benkhira, E.-H., Essoufi, E.-H., and Fakhar, R. On convergence of the penalty method for a static unilateral contact problem with nonlocal friction in electro-elasticity. *European Journal of Applied Mathematics*, 27(01):1–22, 2016.
- Benzi, M. Preconditioning techniques for large linear systems: A survey. *Journal of computational Physics*, 182(2):418–477, 2002.
- Benzi, M., Meyer, C. D., and Tuma, M. A sparse approximate inverse preconditioner for the conjugate gradient method. *SIAM Journal on Scientific Computing*, 17(5):1135–1149, 1996.
- Benzi, M., Kouhia, R., and Tuma, M. Stabilized and block approximate inverse preconditioners for problems in solid and structural mechanics. *Computer methods in applied mechanics and engineering*, 190(49):6533–6554, 2001.
- Benzi, M., Golub, G. H., and Liesen, J. Numerical solution of saddle point problems. *Acta numerica*, 14:1–137, 2005.
- Benzi, M., Bini, D., Kressner, D., Munthe-Kaas, H., and Van Loan, C. *Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications: Cetraro, Italy 2015*, volume 2173. Springer, 2017.
- Bergamaschi, L. On eigenvalue distribution of constraint-preconditioned symmetric saddle point matrices. *Numerical Linear Algebra with Applications*, 19(4):754–772, 2012.
- Bergamaschi, L., Gondzio, J., and Zilli, G. Preconditioning indefinite systems in interior point methods for optimization. *Computational Optimization and Applications*, 28(2):149–171, 2004.
- Bergamaschi, L., Ferronato, M., and Gambolati, G. Novel preconditioners for the iterative solution to FE-discretized coupled consolidation equations. *Computer Methods in Applied Mechanics and Engineering*, 196(25):2647–2656, 2007.
- Bergamaschi, L., Ferronato, M., and Gambolati, G. Mixed constraint preconditioners for the iterative solution of FE coupled consolidation equations. *Journal of Computational Physics*, 227(23):9885–9897, 2008.

- Bertsekas, D. P. *Constrained optimization and Lagrange multiplier methods*. Academic press, 1982.
- Biot, M. A. General theory of three-dimensional consolidation. *Journal of applied physics*, 12(2):155–164, 1941.
- Birkholzer, J. T. and Zhou, Q. Basin-scale hydrogeologic impacts of CO₂ storage: capacity and regulatory implications. *International Journal of Greenhouse Gas Control*, 3(6):745–756, 2009.
- Bishop, A. W. and Bjerrum, L. *The principle of effective stress*. Norges Geotekniske Inst., 1960.
- Bizzarri, A. How to promote earthquake ruptures: Different nucleation strategies in a dynamic model with slip-weakening friction. *Bulletin of the Seismological Society of America*, 100(3):923–940, 2010.
- Bollhöfer, M. and Saad, Y. Multilevel preconditioners constructed from inverse-based ILUs. *SIAM Journal on Scientific Computing*, 27(5):1627–1650, 2006.
- Bourne, S. J., Oates, S. J., van Elk, J., and Doornhof, D. A seismological model for earthquakes induced by fluid extraction from a subsurface reservoir. *Journal of Geophysical Research: Solid Earth*, 119(12):8991–9015, 2014.
- Brodsky, E. E. and Lajoie, L. J. Anthropogenic seismicity rates and operational parameters at the Salton Sea Geothermal Field. *Science*, 341(6145):543–546, 2013.
- Bu, Y., Carpentieri, B., Shen, Z., and Huang, T.-Z. A hybrid recursive multilevel incomplete factorization preconditioner for solving general linear systems. *Applied Numerical Mathematics*, 104:141–157, 2016.
- Burman, E. and Ern, A. A nonlinear consistent penalty method for positivity preservation in the finite element approximation of the transport equation. preprint, 2016.
- Cappa, F. and Rutqvist, J. Impact of CO₂ geological sequestration on the nucleation of earthquakes. *Geophysical Research Letters*, 38(17), 2011a.

- Cappa, F. and Rutqvist, J. Impact of CO₂ geological sequestration on the nucleation of seismic fault ruptures. In *45th US Rock Mechanics/Geomechanics Symposium*. American Rock Mechanics Association, 2011b.
- Cappa, F. and Rutqvist, J. Modeling of coupled deformation and permeability evolution during fault reactivation induced by deep underground injection of CO₂. *International Journal of Greenhouse Gas Control*, 5(2):336–346, 2011c.
- Carpenter, M. C. Continuous monitoring of an earth fissure in Chino, California, USA—a management tool. *Proceedings of the International Association of Hydrological Sciences*, 372:291, 2015.
- Carpentieri, B., Duff, I. S., Giraud, L., and Sylvand, G. Combining fast multipole techniques and an approximate inverse preconditioner for large electromagnetism calculations. *SIAM Journal on Scientific Computing*, 27(3):774–792, 2005.
- Carreón-Freyre, D., Cerca, M., and Hernández Marín, M. Propagation of fracturing related to land subsidence in the Valley of Queretaro, Mexico. In *Proc. 7th Int. Symp. on Land Subsidence*, volume 1, pages 155–164, 2005.
- Castelletto, N., Ferronato, M., Gambolati, G., Janna, C., and Teatini, P. Numerical modeling of rock/casing interaction in radioactive-marker boreholes of the Northern Adriatic basin, Italy. *SPE Reservoir Evaluation & Engineering*, 13(06):906–913, 2010.
- Castelletto, N., Ferronato, M., Gambolati, G., Janna, C., Marzorati, D., and Teatini, P. Can natural fluid pore pressure be safely exceeded in storing gas underground? *Engineering Geology*, 153:35–44, 2013.
- Castelletto, N., White, J. A., and Ferronato, M. Scalable algorithms for three-field mixed finite element coupled poromechanics. *Journal of Computational Physics*, 327:894–918, 2016.
- Cescotto, S. and Charlier, R. Frictional contact finite elements based on mixed variational principles. *International Journal for numerical methods in engineering*, 36(10):1681–1701, 1993.

- Chan, T. F. and Mathew, T. P. Domain decomposition algorithms. *Acta numerica*, 3: 61–143, 1994.
- Chow, E. A priori sparsity patterns for parallel sparse approximate inverse preconditioners. *SIAM Journal on Scientific Computing*, 21(5):1804–1822, 2000.
- Chow, E. Parallel implementation and practical use of sparse approximate inverse preconditioners with a priori sparsity patterns. *The international Journal of High Performance Computing Applications*, 15(1):56–74, 2001.
- Chow, E. and Saad, Y. Approximate inverse preconditioners via sparse-sparse iterations. *SIAM Journal on Scientific Computing*, 19(3):995–1023, 1998.
- Craig, E. J. The N-Step Iteration Procedures. *Studies in Applied Mathematics*, 34(1-4): 64–73, 1955.
- Cueto-Felgueroso, L. and Juanes, R. Forecasting long-term gas production from shale. *Proceedings of the National Academy of Sciences*, 110(49):19660–19661, 2013.
- Cuthill, E. and McKee, J. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pages 157–172. ACM, 1969.
- Da-Yu, G. and Li, Z. Ground fissure hazards in USA and China. *Acta Seismologica Sinica*, 13(4):466–476, 2000.
- Dahm, T., Krüger, F., Stammler, K., Klinge, K., Kind, R., Wylegalla, K., and Grasso, J.-R. The 2004 Mw 4.4 Rotenburg, Northern Germany, earthquake and its possible relationship with gas recovery. *Bulletin of the Seismological Society of America*, 97(3): 691–704, 2007.
- Dahm, T., Cesca, S., Hainzl, S., Braun, T., and Krüger, F. Discrimination between induced, triggered, and natural earthquakes close to hydrocarbon reservoirs: A probabilistic approach based on the modeling of depletion-induced stress changes and seismological source parameters. *Journal of Geophysical Research: Solid Earth*, 120(4): 2491–2509, 2015.

- Davies, R., Foulger, G., Bindley, A., and Styles, P. Induced seismicity and hydraulic fracturing for the recovery of hydrocarbons. *Marine and Petroleum Geology*, 45:171–185, 2013.
- Davis, S. D. and Pennington, W. D. Induced seismic deformation in the Cogdell oil field of west Texas. *Bulletin of the Seismological Society of America*, 79(5):1477–1495, 1989.
- Davis, T. A. and Hu, Y. The University of Florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS)*, 38(1):1, 2011.
- De Wiest, R. J. M. On the storage coefficient and the equations of groundwater flow. *Journal of Geophysical Research*, 71(4):1117–1122, 1966.
- Demaison, G. and Huizinga, B. J. Genetic Classification of Petroleum Systems. *AAPG bulletin*, 75(10):1626–1643, 1991.
- Dolean, V., Jolivet, P., and Nataf, F. *An introduction to domain decomposition methods: algorithms, theory, and parallel implementation*. SIAM, 2015.
- Dyn, N. and Ferguson, W. E. The numerical solution of equality constrained quadratic programming problems. *Mathematics of Computation*, 41(163):165–170, 1983.
- Ellsworth, W. L. Injection-induced earthquakes. *Science*, 341(6142):1225942, 2013.
- Elman, H. C., Silvester, D. J., and Wathen, A. J. Performance and analysis of saddle point preconditioners for the discrete steady-state Navier-Stokes equations. *Numerische Mathematik*, 90(4):665–688, 2002.
- Engelder, T. Capillary tension and imbibition sequester frack fluid in Marcellus gas shale. *Proceedings of the National Academy of Sciences*, 109(52):E3625–E3625, 2012.
- Eslamizadeh, A. and Samanirad, S. Land Subsidence and Fissuring Due to Ground Water Withdrawal in Yazd-Ardakan Basin, Central Iran. *International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering*, 4(11):535–538, 2010.

- Ewing, R. E., Lazarov, R. D., Lu, P., and Vassilevski, P. S. Preconditioning indefinite systems arising from mixed finite element discretization of second-order elliptic problems. In *Preconditioned conjugate gradient methods*, pages 28–43. Springer, 1990.
- Fang, H.-R. and Saad, Y. A filtered Lanczos procedure for extreme and interior eigenvalue problems. *SIAM Journal on Scientific Computing*, 34(4):A2220–A2246, 2012.
- Ferronato, M. Preconditioning for sparse linear systems at the dawn of the 21st century: History, current developments, and future perspectives. *ISRN Applied Mathematics*, 2012, 2012.
- Ferronato, M., Gambolati, G., and Teatini, P. Ill-conditioning of finite element poroelasticity equations. *International Journal of Solids and Structures*, 38(34):5995–6014, 2001.
- Ferronato, M., Gambolati, G., Janna, C., and Teatini, P. Numerical modelling of regional faults in land subsidence prediction above gas/oil reservoirs. *International journal for numerical and analytical methods in geomechanics*, 32(6):633–657, 2008a.
- Ferronato, M., Janna, C., and Gambolati, G. Mixed constraint preconditioning in computational contact mechanics. *Computer Methods in Applied Mechanics and Engineering*, 197(45):3922–3931, 2008b.
- Ferronato, M., Pini, G., and Gambolati, G. The role of preconditioning in the solution to FE coupled consolidation equations by Krylov subspace methods. *International journal for numerical and analytical methods in geomechanics*, 33(3):405–423, 2009.
- Ferronato, M., Bergamaschi, L., and Gambolati, G. Performance and robustness of block constraint preconditioners in finite element coupled consolidation problems. *International journal for numerical methods in engineering*, 81(3):381–402, 2010a.
- Ferronato, M., Gambolati, G., Janna, C., and Teatini, P. Geomechanical issues of anthropogenic CO₂ sequestration in exploited gas fields. *Energy Conversion and Management*, 51(10):1918–1928, 2010b.
- Ferronato, M., Janna, C., and Pini, G. Parallel solution to ill-conditioned FE geomechanical problems. *International Journal for Numerical and Analytical Methods in Geomechanics*, 36(4):422–437, 2012.

- Ferronato, M., Castelletto, N., Gambolati, G., Janna, C., and Teatini, P. II cycle compressibility from satellite measurements. *Geotechnique*, 63(6):479, 2013.
- Ferronato, M., Janna, C., and Pini, G. Parallel Jacobi–Davidson with block FSAI preconditioning and controlled inner iterations. *Numerical Linear Algebra with Applications*, 2016.
- Ferronato, M., Janna, C., and Pini, G. A generalized Block FSAI preconditioner for nonsymmetric linear systems. *Journal of Computational and Applied Mathematics*, 256: 230–241, 2014.
- Figueiredo, B., Tsang, C.-F., Rutqvist, J., Bensabat, J., and Niemi, A. Coupled hydro-mechanical processes and fault reactivation induced by CO₂ injection in a three-layer storage formation. *International Journal of Greenhouse Gas Control*, 39:432–448, 2015.
- Fjær, E., Holt, R. M., Horsrud, P., Raaen, A. M., and Risnes, R. Petroleum Related Rock Mechanics. *Developments in Petroleum Science*, 53, 2008.
- Fletcher, R. Conjugate gradient methods for indefinite systems. In *Numerical analysis*, pages 73–89. Springer, 1976.
- Franceschini, A., Teatini, P., Janna, C., Ferronato, M., Gambolati, G., Ye, S.-J., and Carreón-Freyre, D. Modelling ground rupture due to groundwater withdrawal: applications to test cases in China and Mexico. *Proceedings of the International Association of Hydrological Sciences*, 372:63, 2015.
- Frederickson, P. O. Fast approximate inversion of large sparse linear systems. *Math. Report*, 7, 1975.
- Fredrich, J. T., Arguello, J. G., Deitrick, G. L., and De Rouffignac, E. P. Geomechanical modeling of reservoir compaction, surface subsidence, and casing damage at the Belridge diatomite field. *SPE Reservoir Evaluation & Engineering*, 3(04):348–359, 2000.
- Freund, R. W. and Nachtigal, N. M. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numerische mathematik*, 60(1):315–339, 1991.
- Freund, R. W. and Nachtigal, N. M. Software for simplified Lanczos and QMR algorithms. *Applied Numerical Mathematics*, 19(3):319–341, 1995.

- Galloway, D. L. and Burbey, T. J. Review: regional land subsidence accompanying groundwater extraction. *Hydrogeology Journal*, 19(8):1459–1486, 2011.
- Gambolati, G. and Freeze, R. A. Mathematical simulation of the subsidence of Venice: 1. Theory. *Water Resources Research*, 9(3):721–733, 1973.
- Gambolati, G. and Teatini, P. Geomechanics of subsurface water withdrawal and injection. *Water Resources Research*, 51(6):3922–3955, 2015.
- Gambolati, G., Putti, M., and Teatini, P. Land subsidence. In *Hydrology of Disasters*, pages 231–268. Springer, 1996.
- Gambolati, G., Teatini, P., Baú, D., and Ferronato, M. Importance of poroelastic coupling in dynamically active aquifers of the Po river basin, Italy. *Water resources research*, 36(9):2443–2459, 2000.
- Gambolati, G., Ferronato, M., Teatini, P., Deidda, R., and Lecca, G. Finite element analysis of land subsidence above depleted reservoirs with pore pressure gradient and total stress formulations. *International journal for numerical and analytical methods in geomechanics*, 25(4):307–327, 2001.
- Gan, Q. and Elsworth, D. Analysis of fluid injection-induced fault reactivation and seismic slip in geothermal reservoirs. *Journal of Geophysical Research: Solid Earth*, 119(4):3340–3353, 2014.
- Geertsma, J. Problems of rock mechanics in petroleum production engineering. In *1st ISRM Congress*. International Society for Rock Mechanics, 1966.
- Golub, G. H. and Varga, R. S. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods. *Numerische Mathematik*, 3(1):147–156, 1961.
- González, P. J., Tiampo, K. F., Palano, M., Cannavó, F., and Fernández, J. The 2011 Lorca earthquake slip distribution controlled by groundwater crustal unloading. *Nature Geoscience*, 5(11):821–825, 2012.
- Goodman, R. E., Taylor, R. L., and Brekke, T. L. A model for the mechanics of jointed rocks. *Journal of Soil Mechanics & Foundations Div.*, 1968.

- Gould, N. I., Hribar, M. E., and Nocedal, J. On the solution of equality constrained quadratic programming problems arising in optimization. *SIAM Journal on Scientific Computing*, 23(4):1376–1395, 2001.
- Gould, N. I. M. and Simoncini, V. Spectral analysis of saddle point matrices with indefinite leading blocks. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1152–1171, 2009.
- Grote, M. J. and Huckle, T. Parallel preconditioning with sparse approximate inverses. *SIAM Journal on Scientific Computing*, 18(3):838–853, 1997.
- Gupta, A. Enhancing Performance and Robustness of ILU Preconditioners by Blocking and Selective Transposition. *SIAM Journal on Scientific Computing*, 39(1):A303–A332, 2017.
- Gupta, A., Joshi, M., and Kumar, V. WSMP: A high-performance shared- and distributed-memory parallel sparse linear equation solver. *IBM Research Division RC*, 22038, 2001.
- Hackbusch, W. A sparse matrix arithmetic based on H-Matrices. Part I: Introduction to H-Matrices. *Computing*, 62(2):89–108, 1999.
- Hadjefandiari, A. R. and Dargush, G. F. Couple stress theory for solids. *International Journal of Solids and Structures*, 48(18):2496–2510, 2011.
- Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- Hanks, T. C. and Kanamori, H. A moment magnitude scale. *Journal of Geophysical Research: Solid Earth*, 84(B5):2348–2350, 1979.
- Hansbo, P. Nitsche’s method for interface problems in computational mechanics. *GAMM-Mitteilungen*, 28(2):183–206, 2005.
- Healy, J. H., Rubey, W. W., Griggs, D. T., and Raleigh, C. B. The denver earthquakes. *Science*, 161(3848):1301–1310, 1968.

- Hestenes, M. R. and Stiefel, E. Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, 49(6), 1952.
- Holzer, T. L., Davis, S. N., and Lofgren, B. E. Faulting caused by groundwater extraction in southcentral Arizona. *Journal of Geophysical Research: Solid Earth*, 84(B2):603–612, 1979.
- Huckle, T. Approximate sparsity patterns for the inverse of a matrix and preconditioning. *Applied Numerical Mathematics*, 30(2-3):291–303, 1999.
- Huckle, T. Factorized sparse approximate inverses for preconditioning. *The Journal of Supercomputing*, 25(2):109–117, 2003.
- Hysom, D. and Pothen, A. Efficient parallel computation of ILU (k) preconditioners. In *Proceedings of the 1999 ACM/IEEE conference on Supercomputing*, page 29. ACM, 1999.
- Janna, C. and Ferronato, M. Adaptive pattern research for block FSAI preconditioning. *SIAM Journal on Scientific Computing*, 33(6):3357–3380, 2011.
- Janna, C., Ferronato, M., and Gambolati, G. Multi-level incomplete factorizations for the iterative solution of non-linear finite element problems. *International journal for numerical methods in engineering*, 80(5):651–670, 2009.
- Janna, C., Ferronato, M., and Gambolati, G. A block FSAI-ILU parallel preconditioner for symmetric positive definite linear systems. *SIAM Journal on Scientific Computing*, 32(5):2468–2484, 2010a.
- Janna, C., Ferronato, M., Gambolati, G., and Teatini, P. Simulation of ground failure due to groundwater pumping. *IAHS-AISH publication*, pages 133–135, 2010b.
- Janna, C., Castelletto, N., Ferronato, M., Gambolati, G., and Teatini, P. A geomechanical transversely isotropic model of the Po River basin using PSInSAR derived horizontal displacement. *International Journal of Rock Mechanics and Mining Sciences*, 51:105–118, 2012a.
- Janna, C., Ferronato, M., and Gambolati, G. Parallel inexact constraint preconditioning for ill-conditioned consolidation problems. *Computational Geosciences*, 16(3):661–675, 2012b.

- Janna, C., Ferronato, M., and Gambolati, G. Enhanced block FSAI preconditioning using domain decomposition techniques. *SIAM Journal on Scientific Computing*, 35(5):S229–S249, 2013.
- Janna, C., Ferronato, M., and Gambolati, G. The use of supernodes in factored sparse approximate inverse preconditioning. *SIAM Journal on Scientific Computing*, 37(1):C72–C94, 2015a.
- Janna, C., Ferronato, M., Sartoretto, F., and Gambolati, G. FSAIPACK: A Software Package for High-Performance Factored Sparse Approximate Inverse Preconditioning. *ACM Transactions on Mathematical Software (TOMS)*, 41(2):10, 2015b.
- Jha, B. and Juanes, R. Coupled multiphase flow and poromechanics: A computational model of pore pressure effects on fault slip and earthquake triggering. *Water Resources Research*, 50:3776–3808, 2014.
- Jovanović, B. S. and Süli, E. *Analysis of finite difference schemes: for linear partial differential equations with generalized solutions*, volume 46. Springer Science & Business Media, 2013.
- Juanes, R., Samper, J., and Molinero, J. A general and efficient formulation of fractures and boundary conditions in the finite element method. *International Journal for Numerical Methods in Engineering*, 54(12):1751–1774, 2002.
- Kalinkin, A., Anders, A., and Anders, R. Intel® Math Kernel Library Parallel Direct Sparse Solver for Clusters. In *EAGE Workshop on High Performance Computing for Upstream*, 2014.
- Kanamori, H. The energy release in great earthquakes. *Journal of geophysical research*, 82(20):2981–2987, 1977.
- Kanamori, H. and Anderson, D. L. Theoretical basis of some empirical relations in seismology. *Bulletin of the Seismological Society of America*, 65(5):1073–1095, 1975.
- Kaporin, I. E. A preconditioned conjugate-gradient method for solving discrete analogs of differential problems. *Differential Equations*, 26(7):897–906, 1990.

- Kaporin, I. E. New convergence results and preconditioning strategies for the conjugate gradient method. *Numerical linear algebra with applications*, 1(2):179–210, 1994.
- Karypis, G. and Kumar, V. Multilevel k-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed computing*, 48(1):96–129, 1998.
- Kato, N. Earthquake Cycles in a Model of Interacting Fault Patches: Complex Behavior at Transition from Seismic to Aseismic Slip. *Bulletin of the Seismological Society of America*, 106(4):1772–1787, 2016.
- Keller, C., Gould, N. I., and Wathen, A. J. Constraint preconditioning for indefinite linear systems. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1300–1317, 2000.
- Keranen, K. M., Savage, H. M., Abers, G. A., and Cochran, E. S. Potentially induced earthquakes in Oklahoma, USA: Links between wastewater injection and the 2011 Mw 5.7 earthquake sequence. *Geology*, 41(6):699–702, 2013.
- Khoei, A. R. and Lewis, R. W. Adaptive finite element remeshing in a large deformation analysis of metal powder forming. *International Journal for Numerical Methods in Engineering*, 45(7):801–820, 1999.
- Kim, W.-Y. Induced seismicity associated with fluid injection into a deep well in Youngstown, Ohio. *Journal of Geophysical Research: Solid Earth*, 118(7):3506–3518, 2013.
- King, G. C. P., Stein, R. S., and Lin, J. Static stress changes and the triggering of earthquakes. *Bulletin of the Seismological Society of America*, 84(3):935–953, 1994.
- Kolotilina, L. Y. and Yeremin, A. Y. Factorized sparse approximate inverse preconditionings I. Theory. *SIAM Journal on Matrix Analysis and Applications*, 14(1):45–58, 1993.
- Kostrov, V. V. Seismic moment and energy of earthquakes, and seismic flow of rocks. *Izv. Acad. Sci. USSR Phys. Solid Earth*, 1:23–44, 1974. Eng. Transl.
- Labuz, J. F. and Zang, A. Mohr-Coulomb failure criterion. *Rock Mechanics and Rock Engineering*, 45(6):975–979, 2012.

- Lanczos, C. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49(1):33–53, 1952.
- Lei, X., Ma, S., Chen, W., Pang, C., Zeng, J., and Jiang, B. A detailed view of the injection-induced seismicity in a natural gas reservoir in Zigong, southwestern Sichuan Basin, China. *Journal of Geophysical Research: Solid Earth*, 118(8):4296–4311, 2013.
- Li, R. and Saad, Y. Divide and conquer low-rank preconditioners for symmetric matrices. *SIAM Journal on Scientific Computing*, 35(4):A2069–A2095, 2013.
- Li, R., Xi, Y., and Saad, Y. Schur complement-based domain decomposition preconditioners with low-rank corrections. *Numerical Linear Algebra with Applications*, 23(4):706–729, 2016.
- Li, X. S. and Demmel, J. W. SuperLU_DIST: A scalable distributed-memory sparse direct solver for unsymmetric linear systems. *ACM Transactions on Mathematical Software (TOMS)*, 29(2):110–140, 2003.
- Li, Z., Saad, Y., and Sosonkina, M. pARMS: a parallel version of the algebraic recursive multilevel solver. *Numerical linear algebra with applications*, 10(5-6):485–509, 2003.
- Liesen, J. A note on the eigenvalues of saddle point matrices. Technical report, Tech. Rep. 10-2006, TU Berlin, 2006.
- Luksan, L. and Vlcek, J. Indefinitely preconditioned inexact Newton method for large sparse equality constrained nonlinear programming problems. *Numerical linear algebra with applications*, 5(3):219–247, 1998.
- Majer, E. L. and Peterson, J. E. The impact of injection on seismicity at The Geysers, California Geothermal Field. *International Journal of Rock Mechanics and Mining Sciences*, 44(8):1079–1090, 2007.
- Mazzoldi, A., Rinaldi, A. P., Borgia, A., and Rutqvist, J. Induced seismicity within geological carbon sequestration projects: maximum earthquake magnitude and leakage potential from undetected faults. *International Journal of Greenhouse Gas Control*, 10:434–442, 2012.

- McBryan, O. A., Frederickson, P. O., Lindenand, J., Schüller, A., Solchenbach, K., Stüben, K., Thole, C.-A., and Trottenberg, U. Multigrid methods on parallel computers-a survey of recent developments. *IMPACT of Computing in Science and Engineering*, 3(1):1–75, 1991.
- McGarr, A. Dependence of magnitude statistics on strain rate. *Bulletin of the Seismological Society of America*, 66(1):33–44, 1976.
- McGarr, A. On a possible connection between three major earthquakes in California and oil production. *Bulletin of the Seismological Society of America*, 81(3):948–970, 1991.
- McInnes, L. C., Smith, B., Zhang, H., and Mills, R. T. Hierarchical Krylov and nested Krylov methods for extreme-scale computing. *Parallel Computing*, 40(1):17–31, 2014.
- Mihajlović, M. D. and Silvester, D. J. Efficient parallel solvers for the biharmonic equation. *Parallel Computing*, 30(1):35–55, 2004.
- Murad, M. A. and Loula, A. F. D. On stability and convergence of finite element approximations of Biot’s consolidation problem. *International Journal for Numerical Methods in Engineering*, 37(4):645–667, 1994.
- Nicholson, C. and Wesson, R. L. Earthquake hazard associated with deep well injection; a report to the US Environmental Protection Agency. Technical report, US Geological Survey, 1990.
- Nicholson, C. and Wesson, R. L. Triggered earthquakes and deep well activities. *Pure and applied Geophysics*, 139(3-4):561–578, 1992.
- Nielsen, S., Spagnuolo, E., Smith, S. A. F., Violay, M., Di Toro, G., and Bistacchi, A. Scaling in natural and laboratory earthquakes. *Geophysical Research Letters*, 2016.
- Onate, E. and Rojek, J. Combination of discrete element and finite element methods for dynamic analysis of geomechanics problems. *Computer methods in applied mechanics and engineering*, 193(27):3087–3128, 2004.
- Paige, C. C. and Saunders, M. A. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

- Paludetto Magri, V. A., Franceschini, A., Ferronato, M., and Janna, C. Multilevel Approaches for FSAI Preconditioning. *Numerical Linear Algebra with Applications*, 2017. Submitted.
- Pennington, W. D., Davis, S. D., Carlson, S. M., DuPree, J., and Ewing, T. E. The evolution of seismic barriers and asperities caused by the depressuring of fault planes in oil and gas fields of south Texas. *Bulletin of the Seismological Society of America*, 76(4): 939–948, 1986.
- Perugia, I. and Simoncini, V. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. *Numerical linear algebra with applications*, 7(7-8): 585–616, 2000.
- PetroWiki. Reservoir geology, 2015. URL http://petrowiki.org/Reservoir_geology. Accessed: 2017-05-18.
- Phan, A.-V., Napier, J. A. L., Gray, L. J., and Kaplan, T. Symmetric-Galerkin BEM simulation of fracture with frictional contact. *International journal for numerical methods in engineering*, 57(6):835–851, 2003.
- Pratt, W. E. and Johnson, D. W. Local subsidence of the Goose Creek oil field. *The Journal of Geology*, 34(7, Part 1):577–590, 1926.
- Raghavan, P. and Teranishi, K. Parallel hybrid preconditioning: incomplete factorization with selective sparse approximate inversion. *SIAM Journal on Scientific Computing*, 32(3):1323–1345, 2010.
- Raleigh, C. B., Healy, J. H., and Bredehoeft, J. D. An experiment in earthquake control at Rangely, Colorado. *Science*, 108(52):30, 1976.
- Rinaldi, A. P. and Rutqvist, J. Modeling of deep fracture zone opening and transient ground surface uplift at KB-502 CO₂ injection well, In Salah, Algeria. *International Journal of Greenhouse Gas Control*, 12:155–167, 2013.
- Roche, V. and Van der Baan, M. The role of lithological layering and pore pressure on fluid-induced microseismicity. *Journal of Geophysical Research: Solid Earth*, 120(2): 923–943, 2015.

- Rothenburg, L., Obah, A., and El Baruni, S. Horizontal ground movements due to water abstraction and formation of earth fissures. *International Association of Hydrological Sciences, Publication*, 234:239–249, 1995.
- Rozložník, M. and Simoncini, V. Krylov subspace methods for saddle point problems with indefinite preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 24(2):368–391, 2002.
- Rutqvist, J. and Tsang, C.-F. A study of caprock hydromechanical changes associated with CO₂-injection into a brine formation. *Environmental Geology*, 42(2-3):296–305, 2002.
- Rutqvist, J. and Tsang, C.-F. Coupled hydromechanical effects of CO₂ injection. *Developments in Water Science*, 52:649–679, 2005.
- Rutqvist, J., Birkholzer, J., Cappa, F., and Tsang, C.-F. Estimating maximum sustainable injection pressure during geological sequestration of CO₂ using coupled fluid flow and geomechanical fault-slip analysis. *Energy Conversion and Management*, 48(6):1798–1807, 2007.
- Rutqvist, J., Birkholzer, J. T., and Tsang, C.-F. Coupled reservoir–geomechanical analysis of the potential for tensile and shear failure associated with CO₂ injection in multilayered reservoir–caprock systems. *International Journal of Rock Mechanics and Mining Sciences*, 45(2):132–143, 2008.
- Rutqvist, J., Cappa, F., Mazzoldi, A., and Rinaldi, A. Geomechanical modeling of fault responses and the potential for notable seismic events during underground CO₂ injection. *Energy Procedia*, 37:4774–4784, 2013a.
- Rutqvist, J., Rinaldi, A. P., Cappa, F., and Moridis, G. J. Modeling of fault reactivation and induced seismicity during hydraulic fracturing of shale-gas reservoirs. *Journal of Petroleum Science and Engineering*, 107:31–44, 2013b.
- Rutqvist, J., Rinaldi, A. P., Cappa, F., Jeanne, P., Mazzoldi, A., Urpi, L., Guglielmi, Y., and Vilarrasa, V. Fault activation and induced seismicity in geological carbon

- storage—Lessons learned from recent modeling studies. *Journal of Rock Mechanics and Geotechnical Engineering*, 8(6):789–804, 2016.
- Saad, Y. ILUT: A dual threshold incomplete LU factorization. *Numerical linear algebra with applications*, 1(4):387–402, 1994.
- Saad, Y. *Iterative methods for sparse linear systems*. SIAM, 2003.
- Saad, Y. and Schultz, M. H. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, 7(3):856–869, 1986.
- Saad, Y. and Suchomel, B. ARMS: An algebraic recursive multilevel solver for general sparse linear systems. *Numerical linear algebra with applications*, 9(5):359–378, 2002.
- Sabetamal, H., Nazem, M., Sloan, S. W., and Carter, J. P. Frictionless contact formulation for dynamic analysis of nonlinear saturated porous media based on the mortar method. *International Journal for Numerical and Analytical Methods in Geomechanics*, 40(1):25–61, 2016.
- Schenk, O. and Gärtner, K. Solving unsymmetric sparse systems of linear equations with PARDISO. *Future Generation Computer Systems*, 20(3):475–487, 2004.
- Scott, J. and Tuma, M. On positive semidefinite modification schemes for incomplete Cholesky factorization. *SIAM Journal on Scientific computing*, 36(2):A609–A633, 2014.
- Segall, P. Stress and subsidence resulting from subsurface fluid withdrawal in the epicentral region of the 1983 Coalinga earthquake. *Journal of Geophysical Research: Solid Earth*, 90(B8):6801–6816, 1985.
- Segall, P. Earthquakes triggered by fluid extraction. *Geology*, 17(10):942–946, 1989.
- Segura, J. M., Fisher, Q. J., Crook, A. J. L., Dutko, M., Yu, J. G., Skachkov, S., Angus, D. A., Verdon, J. P., and Kendall, J. M. Reservoir stress path characterization and its implications for fluid-flow production simulations. *Petroleum Geoscience*, 17(4):335–344, 2011.

- Shi, X.-Q., Xue, Y.-Q., Ye, S.-J., Wu, J.-C., Zhang, Y., and Yu, J. Characterization of land subsidence induced by groundwater withdrawals in Su-Xi-Chang area, China. *Environmental geology*, 52(1):27, 2007.
- Shi, Y. Solving linear systems involved in constrained optimization. *Linear algebra and its applications*, 229:175–189, 1995.
- Silvester, D. J., Elman, H. C., Kay, D., and Wathen, A. J. Efficient preconditioning of the linearized Navier-Stokes equations for incompressible flow. *Journal of Computational and Applied Mathematics*, 128(1):261–279, 2001.
- Simo, J. C. and Hughes, T. J. R. *Computational Inelasticity*. Springer, 2006.
- Simo, J. C. and Laursen, T. A. An augmented Lagrangian treatment of contact problems involving friction. *Computers & Structures*, 42(1):97–116, 1992.
- Simpson, D. W., Leith, W. S., and Scholz, C. H. Two types of reservoir-induced seismicity. *Bulletin of the Seismological Society of America*, 78(6):2025–2040, 1988.
- Sleijpen, G. L. G. and Van der Vorst, H. A. A Jacobi-Davidson iteration method for linear eigenvalue problems. *Siam Review*, 42(2):267–293, 2000.
- Sonneveld, P. CGS, a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 10(1):36–52, 1989.
- Sonneveld, P. and Van Gijzen, M. B. IDR (s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations. *SIAM Journal on Scientific Computing*, 31(2):1035–1062, 2008.
- Stüben, K. and Trottenberg, U. Multigrid methods: Fundamental algorithms, model problem analysis and applications. In *Multigrid methods*, pages 1–176. Springer, 1982.
- Taylor, M. E. *Partial Differential Equations I*. Springer, 2011.
- Teatini, P., Castelletto, N., Ferronato, M., Gambolati, G., Janna, C., Cairo, E., Marzorati, D., Colombo, D., Ferretti, A., Bagliani, A., and Bottazzi, A. Geomechanical response to seasonal gas storage in depleted reservoirs: A case study in the Po River basin, Italy. *Journal of Geophysical Research: Earth Surface*, 116(F2), 2011.

- Teatini, P., Carreon-Freyre, D., Galloway, D. L., and Ye, S.-J. Mechanisms, Monitoring and Modeling Earth Fissure generation and Fault activation due to subsurface Fluid exploitation (M3EF3): A UNESCO-IGCP project in partnership with the UNESCO-IHP Working Group on Land Subsidence. In *AGU Fall Meeting Abstracts*, 2015.
- Terzaghi, K. Principles of soil mechanics, IV-Settlement and consolidation of clay. *Engineering News-Record*, 95(3):874–878, 1925.
- Tong, Z. and Sameh, A. On an iterative method for saddle point problems. *Numerische Mathematik*, 79(4):643–646, 1998.
- Trottenberg, U., Oosterlee, C. W., and Schuller, A. *Multigrid*. Academic press, 2000.
- Vafai, K. *Handbook of porous media*. Crc Press, 2015.
- Van der Elst, N. J., Savage, H. M., Keranen, K. M., and Abers, G. A. Enhanced remote earthquake triggering at fluid-injection sites in the midwestern United States. *Science*, 341(6142):164–167, 2013.
- Van der Knaap, W. Nonlinear behavior of elastic porous media. *Trans. AIME*, 216: 179–187, 1959.
- Van der Vorst, H. A. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on scientific and Statistical Computing*, 13(2):631–644, 1992.
- Van Eck, T., Goutbeek, F., Haak, H., and Dost, B. Seismic hazard due to small-magnitude, shallow-source, induced earthquakes in The Netherlands. *Engineering Geology*, 87(1):105–121, 2006.
- Van Thienen-Visser, K., Pruiksma, J. P., and Breunese, J. N. Compaction and subsidence of the Groningen gas field in the Netherlands. *Proceedings of the International Association of Hydrological Sciences*, 372:367, 2015.
- Verdon, J. P., Stork, A. L., Bissell, R. C., Bond, C. E., and Werner, M. J. Simulation of seismic events induced by CO₂ injection at In Salah, Algeria. *Earth and Planetary Science Letters*, 426:118–129, 2015.

- Verruijt, A. Elastic storage of aquifers. *Flow through porous media*, pages 331–376, 1969.
- Verruijt, A. *Theory and problems of poroelasticity*. Delft University of Technology, The Netherlands, 2013.
- Voigt, W. Theoretische Studien fiber die Elastizitatsverhiltnisse der Kristalle (Theoretical studies on the elasticity relationships of crystals). *Abh. Gesch. Wissenschaften*, 34: 424–428, 1887.
- Wan, J. *Stabilized finite element methods for coupled geomechanics and multiphase flow*. PhD thesis, Stanford university, 2002.
- Wang, E., Zhang, Q., Shen, B., Zhang, G., Lu, X., Wu, Q., and Wang, Y. Intel math kernel library. In *High-Performance Computing on the Intel® Xeon Phi™*, pages 167–188. Springer, 2014.
- Wang, G. Y., You, G., Shi, B., Yu, J., Li, H. Y., and Zong, K. H. Earth fissures triggered by groundwater withdrawal and coupled by geological structures in Jiangsu Province, China. *Environmental geology*, 57(5):1047–1054, 2009.
- Wang, S., Li, X. S., Xia, J., Situ, Y., and De Hoop, M. V. Efficient scalable algorithms for solving dense linear systems with hierarchically semiseparable structures. *SIAM Journal on Scientific Computing*, 35(6):C519–C544, 2013.
- Warner, N. R., Jackson, R. B., Darrah, T. H., Osborn, S. G., Down, A., Zhao, K., White, A., and Vengosh, A. Reply to Engelder: Potential for fluid migration from the Marcellus Formation remains possible. *Proceedings of the National Academy of Sciences*, 109(52): E3626–E3626, 2012.
- White, J. A., Castelletto, N., and Tchelepi, H. A. Block-partitioned solvers for coupled poromechanics: A unified framework. *Computer Methods in Applied Mechanics and Engineering*, 303:55–74, 2016.
- Wriggers, P. *Computational Contact Mechanics*. Springer, 2006.
- Wriggers, P. and Zavarise, G. Computational contact mechanics. In E. Stein, T. H., R. deBorst, editor, *Encyclopedia of Computational Mechanics*, chapter 6, pages 195–226. John Wiley & Sons, 2004.

- Xi, Y., Li, R., and Saad, Y. An algebraic multilevel preconditioner with low-rank corrections for sparse symmetric matrices. *SIAM Journal on Matrix Analysis and Applications*, 37(1):235–259, 2016.
- Xia, J. On the complexity of some hierarchical structured matrix algorithms. *SIAM Journal on Matrix Analysis and Applications*, 33(2):388–410, 2012.
- Xia, J. Efficient structured multifrontal factorization for general large sparse matrices. *SIAM Journal on Scientific Computing*, 35(2):A832–A860, 2013.
- Xia, J., Chandrasekaran, S., Gu, M., and Li, X. S. Fast algorithms for hierarchically semiseparable matrices. *Numerical Linear Algebra with Applications*, 17(6):953–976, 2010.
- Xu, J. and Zikatanov, T. L. Algebraic Multigrid Methods. *Acta Numerica*, pages 1–128, 2016.
- Ye, S.-J., Wang, Y., Wu, J., Teatini, P., Yu, J., Gong, X., and Wang, G. Characterization of earth fissures in South Jiangsu, China. *Proceedings of the International Association of Hydrological Sciences*, 372:249–253, 2015.
- Yerkes, R. F. and Castle, R. O. Seismicity and faulting attributable to fluid extraction. *Engineering Geology*, 10(2-4):151–167, 1976.
- Zbinden, D., Rinaldi, A. P., Urpi, L., and Wiemer, S. On the physics-based processes behind production-induced seismicity in natural gas fields. *Journal of Geophysical Research: Solid Earth*, 122, 2017.
- Zhang, Y., Wang, Z., Xue, Y., Wu, J., and Yu, J. Mechanisms for earth fissure formation due to groundwater extraction in the Su-Xi-Chang area, China. *Bulletin of Engineering Geology and the Environment*, 75(2):745–760, 2016.
- Zienkiewicz, O. C. and Taylor, R. L. *The finite element method: solid mechanics*, volume 2. Butterworth-heinemann, 2000.
- Zoback, M. D. and Zinke, J. C. Production-induced normal faulting in the Valhall and Ekofisk oil fields. *Pure and Applied Geophysics*, 159(1):403–420, 2002.