School of Civil and Mechanical Engineering

Computational modelling of compaction bands in geomaterials

Roberto Jesús Cier Honores

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University of Technology

October 2022

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature:

Roberto Jesús Cier Honores
October 19, 2022

Acknowledgements

This project was developed under the sponsorship of a Curtin International Postgraduate Research Scholarship (CIPRS) and the CSIRO Deep Earth Imaging Future Science Platform (DEI-FSP) Postgraduate Top-Up Scholarship, Position No. 50059239. Their support during the time of this work is gratefully acknowledged.

I express my most sincere gratitude to my main supervisor, Prof Victor Calo, for his support, advice and patience at each stage of this project. Working under his guidance and learning from him has been a great experience. I would also like to acknowledge the advice, help and encouragement from my CSIRO's external advisor, Dr Thomas Poulet, during the development of this project. Additionally, I would like to thank all my research colleagues who have been an essential source of collaboration, support and friendship during all these years: Quanling, Sergio, Nicolas, Judith, Santiago, Juan Felipe, Pouria and Marcos. A big thank you to each of them.

Finally, I would like to thank my family back in Peru for being my source of motivation, and Astrid, my partner, for showing me the true meaning of love and care through her company along this journey. This work is also by and for them.

Abstract

Historically, most structural geology studies focused on two localisation types: shear and extensional strain; these explain various crucial geological processes such as faults and joints. However, a third deformation type, compaction banding, has recently attracted attention due to the impacts on the location and extension of mineral deposits and oil-and-gas reservoirs. Compaction bands form narrow planar zones where properties (i.e., density, permeability) are markedly different from the rest of a body and often create periodic patterns. Recently, the cnoidal wave theory seeks to explain their triggering, propagation, and distribution by analogy with shallow waves in fluids usually described using Jacobi cn elliptic functions. While this qualitative theory can capture the compaction banding periodicity in solids, there is no quantitative link to the observed phenomena. The equation only has closed-form solutions for integer exponents, while simulating it is a challenge for standard techniques. Thus, we develop a robust numerical framework to analyse compaction band formation for a broader range of parameters and conditions. We model the governing cnoidal equation to understand the bands' emergence and distribution. Given the system's highly nonlinear behaviour, we propose a new automatically adaptive stabilised finite element method for this class of problems. We start the study with linear diffusion problems. Subsequently, we extend the framework to capture nonlinear features in the formulation, such as constraint enforcement. Then, we study numerically the cnoidal equation, where the adaptive technology shows significant advantages against classical finite elements. Finally, we introduce a consistent theoretical framework that generalises the $p - q$ space combining the modified Cam Clay model and Perzyna's viscoplasticity. Thus, we describe the compaction band formation analytically from first principles rather than weak analogies. Numerical experiments seek to reproduce detailed laboratory tests to verify our assumptions and validate our model predictions.

Contents

List of Figures

Chapter 1

Introduction

1.1 State of the art

Localisation phenomena are predominant in Earth Sciences as many exciting geological features fall within this category, including faults, folds, boudinage, landslides, and mineralisation, to name a few. Among all localised geological features, spatially repeating patterns are increasingly gaining popularity due to their impact in applications such as mining [Iophis et al., 2007], particularly when larger-scale and deeper mines are the norm. Localisation features are particularly relevant as they affect permeability, which plays a critical role in various fields, including radioactive waste deposition [Zeng et al., 2020] and mineral prospection [Hayward et al., 2018; Kelka et al., 2017], as well as unconventional resources exploration and exploitation [Regenauer-Lieb et al., 2016].

A mechanism responsible for the occurrence of these patterns affecting porosity and permeability is the formation of compaction bands. These bands are usually defined as narrow flat zones of deformation perpendicular to the maximum compressive principal stress [Mollema & Antonellini, 1996; Das et al., 2013], as Figure 1.1 shows. A succession of compacted zones of lower permeability in a higher permeability background is an intuitive way to imagine compaction bands [Holcomb & Olsson, 2003], with critical implications on the fluid flow as they create impermeable barriers and compartmentalise reservoirs and aquifers. These compaction bands, therefore, have a significant impact on fluid production or geological storage (CO_2, nuclear waste). Several models exist to describe the mechanism of its in-situ periodic occurrence (e.g. Cecinato & Gajo [2014]); however, other physical processes lead to regular bands under compression with

increased permeability, known as decompaction bands in the melt segregation field [Rabinowicz & Vigneresse, 2004].

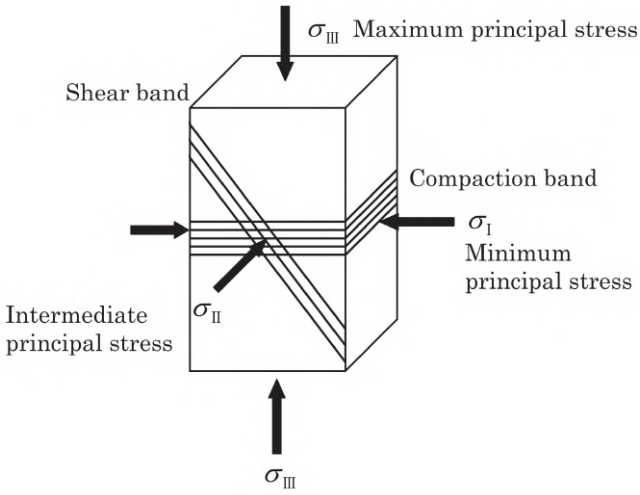


Figure 1.1: Compaction band in a specimen [Oka et al., 2011, reproduced with permission from *International Journal of Numerical and Analytical Methods in Geomechanics*].

In solid mechanics, attempts to theoretically describe deformation bands (principally of shear type) start early in the theory of plasticity [Hill, 1950], where material instabilities were associated with the stationary limit of the accelerating wave velocity in solids. Based on this early work, bifurcation criteria developed for geomaterials, the most widely used the proposed one by Rudnicki & Rice [1975]. This first framework established conditions for the shear banding inception as a bifurcation problem in brittle rocks under principal compressive stresses. Later, Olsson [1999] extended it for compaction banding scenarios, and finally Issen & Rudnicki [2001] formalised the theory for the onset of compaction bands in porous rocks incorporating a comprehensive bifurcation approach with a cap-type yield function. These approaches were formulated under strain-controlled conditions and in terms of the acoustic tensor, which depends simultaneously on the tangent constitutive tensor. On the other hand, a simple approach for stress-controlled scenarios was proposed by Vermeer [1982], limited only to two-dimensional conditions but of easier implementation compared to the first work of Rudnicki and Rice [Gutierrez, 2017]. Other viscoplastic approaches for the bifurcation problem derived from the controllability criterion proposed by Nova [1994], which establishes bifurcation conditions for mixed-mode loading cases. Consequently, Pisano & Prisco [2016] introduced a stability criterion for elasto-

viscoplastic constitutive laws from the spectral analysis of the resulting matrix of an ordinary partial differential equation derived from expressing a second-order form of the Perzyna [1966] constitutive equations.

The need for reconstruction of a tangent constitutive tensor represents a limitation for evaluating localisation processes in rate-dependent materials, given that classical viscoplastic theories were not expressed under consistency conditions, precluding the recovery of a viscoplastic constitutive tensor. In plasticity theory, it is well known that the plastic multiplier is a consequence of predefined constitutive assumptions [Simo & Taylor, 1985] rather than an explicit computation provided by ad-hoc definitions. Additionally, studies showed that when the (visco)plastic multiplier is computed straightforwardly using classical definitions such as Perzyna [1966] or Duvaut & Lions [1982] instead of computing it consistently, loading-unloading stress paths present energy dissipation [Heeres et al., 2002]. Consistent viscoplastic formulations re-signify the use of fundamental definitions to rewrite the yield function that becomes time-dependent. Nonetheless, an important part of the community still neglects consistency under viscoplastic scenarios, and they compute the viscoplastic component in a non-consistent way. A few researchers formulated a series of consistent viscoplastic approaches that allowed them to develop strain localisation analyses [Carosio et al., 2000; Wang et al., 1997], although uniquely focused on shear banding and using J_2 elastoplasticty.

All the above studies focused on the mechanical problem without considering multiphysical interactions to capture other contributions to the deformation process. In Veveakis & Regenauer-Lieb [2015], the authors overcame this issue and developed a wave mechanics approach that produces regularly spaced localisation bands of hydromechanical nature similar to those that appear in rocks under compaction. In this approach, the compaction banding phenomenon represents a material instability related to a propagating pressure wave (P-wave), analysed in the stationary limit, see Figure 1.2. The proposed wave description is radically different from classical approaches. The localised deformation generalises Terzaghi's linear consolidation theory to materials with nonlinear viscoplastic rheology and arbitrary internal mass transfer mechanisms. In this generalisation, the authors derived the following proxy equation for nonlinear consolidation, admitting material instabilities in the effective normalised stress σ' (for elastoviscoplastic

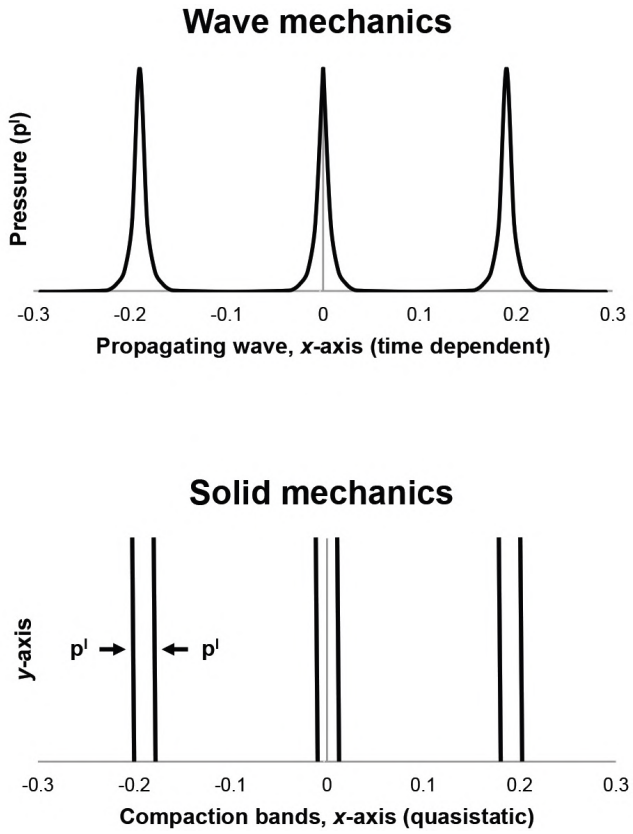


Figure 1.2: Compaction instability duality [Regenauer-Lieb et al., 2016, reproduced with permission from *The Leading Edge*, Springer Nature].

materials under hydromechanical loads):

$$\frac{\partial^2 \sigma'}{\partial z^2} - \lambda \left(\sigma'\right)^m = 0, \tag{1.1}$$

where λ is the ratio between different diffusive processes acting in the system (i.e., matrix mechanical deformation to internal mass exchange ratio) and m is a material-dependent pressure exponent. Under certain conditions, the solution (1.1) triggers instabilities when the loading rate is faster than the mass diffusion rate; thus, mass variations in the specimen cannot equilibrate, producing stress concentration zones that represent compaction bands (see Figure 1.3). These instabilities form periodic volumetric failure patterns, the cnoidal waves [Regenauer-Lieb et al., 2016; Veveakis & Regenauer-Lieb, 2015; Veveakis et al., 2015], due to their analogous counterparts in fluid dynamics. Cnoidal waves are solutions of the Korteweg-de Vries (KdV) equation [Korteweg & De Vries, 1895] expressed as the square Jacobi *cn* elliptic function. The KdV equation de-

scribes a travelling wave in shallow water surfaces, first observed by Russell [1844], an extensively studied process relevant to many physical phenomena related to wave mechanics.

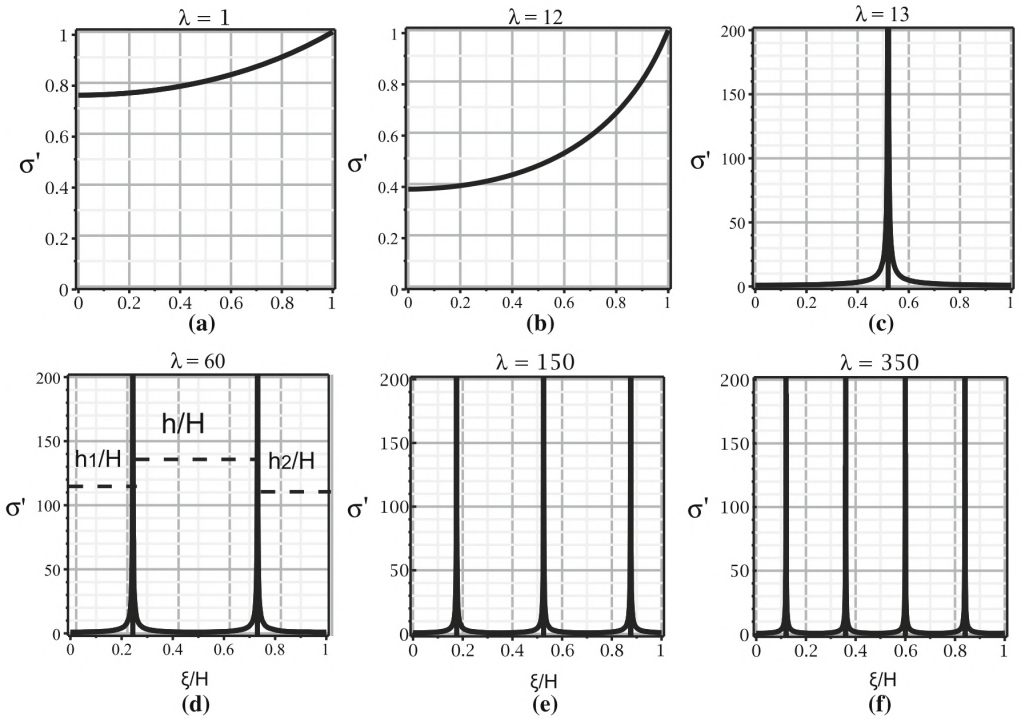


Figure 1.3: Effective stress (σ') configurations for different λ values in analytical solution of (1.1) for $m = 3$ [Alevizos et al., 2017, reproduced with permission from *Rock Mechanics and Rock Engineering*].

For solids, the cnoidal waves allow us to conclude that periodic instabilities are due to the volumetric response during failure, controlled by the deformation rate, not by the critical stress or hardening, as classical theories predict [Veveakis & Regenauer-Lieb, 2015]. Although anisotropy, heterogeneity, and boundary effects could produce periodic localisation in geomaterials [Das & Buscarnera, 2014; Shahin et al., 2019; Borja et al., 2020], the cnoidal theory forecasts that the compaction banding inception may occur for homogeneous materials. The cnoidal description shows that obtaining periodic patterns in a homogeneous material is feasible instead of relying on inhomogeneities to initiate the instabilities. Considering all the above, once the instabilities appear, they follow the local weakest inhomogeneities within the domain.

While the original theory [Veveakis & Regenauer-Lieb, 2015] captures the essence of hydromechanical instabilities, its simplifications lead to unbounded

stresses, which are unrealistic. Thus, this model requires a proper regularisation, as is common for shear bands (e.g., Vardoulakis [2000]), where higher-order theories in the deviatoric plastic flow law are common (Cosserat, viscous, gradient dependency). In the compaction band phenomenon, however, regularisations modify the flow's volumetric components, deviating the mass balance equation from the incompressible limit. Besides, crucial parametrisations such as the relationship between the thickness of compaction bands and the physical phenomena described are unclear. Follow-up publications that focused on the physics of the problem analysed these features. In particular, the complementary study by Alevizos et al. [2017] considered the effects of chemical reactions as a regularisation term $N_r(\sigma')$ in the equation, allowing the normalised effective stress to remain capped. Additionally, they show that the cnoidal theory can explain the compaction band thickness through an explicit dependence on the exact mass exchange mechanism assumed, with chemical dissolution/precipitation offering compaction bands as thick as cm-scale. As such, they generalise (1.1) to:

$$\frac{\partial^2 \sigma'}{\partial z^2} - \lambda (\sigma')^m + N_r(\sigma') = 0, \tag{1.2}$$

The Appendix A derives this equation in detail.

From a numerical perspective, compaction band simulations are few compared with the numerous studies on the formation of shear banding [Oka et al., 2011]. The most relevant works include Borja's [Borja & Aydin, 2004; Borja, 2004] computational modelling, where the onset conditions for deformation bands used a single hardening constitutive model [Kim & Lade, 1988]. Additionally, Oka et al. [2011] simulated with finite elements a diatomaceous mudstone with an elasto-viscoplastic model and compared them with triaxial test results. Although their models predicted compaction bands for higher confining pressures, they did not tackle the challenging problem of identifying the onset conditions of the phenomenon and its periodicity for a broader range of confinement stresses. Experimental studies also focused on sandstone specimens [Fortin et al., 2006; Wu et al., 2018] and limestone samples [Baxevanis et al., 2006; Wu et al., 2020], where grain breakage and packing along the fractures emerged during the unloading phase of triaxial compression tests suggest the formation of compaction bands. More recent numerical approaches in this topic include hypoelastic laws [Garavand et al., 2020] and gradient-dependent plasticity [Abdallah et al., 2020].

To date, one of the limitations in the theory of hydromechanical compaction instabilities proposed by Veveakis & Regenauer-Lieb [2015] is the lack of a robust numerical framework to solve the underlying equation. Although the physical and mathematical models are simplified following dimensionless assumptions that result in a one-dimensional cnoidal equation directly showing the term responsible for the material instability, the reduced equation is challenging due to the solution's localised behaviour. The complexity of this kind of phenomenon is well-understood in applied mathematics as the partial differential equation presented here is of the extinction type with elliptic poles, known for the presence of finite-time blow-up instabilities in its solution [Galaktionov & Vazquez, 1995; Galaktionov & Vázquez, 2002].

We capture the localised nature of the compaction band phenomenon using a new class of finite element method (FEM), one of the most well-known numerical methods for solving partial differential equations (PDE). FEM approximates a PDE solution over a specific domain solving a discrete system of equations. This approximation uses simple functions on each (finite) element that describes the physical domain. The domain's varying material properties may make the problem unstable; this produces inaccurate numerical approximations that produce unphysical solutions. Appendix B discusses the stability of finite elements.

Traditionally, we add stabilisation terms to the standard finite element approximation; these terms are extra-weighted residuals that improve the stability properties of the discrete solution. A precursor of these ideas is the Streamline Upwind Petrov-Galerkin (SUPG) method [Brooks & Hughes, 1982]. Later, the Galerkin/Least squares (GaLS) [Hughes et al., 1989], and variational multiscale (VMS) [Hughes & Sangalli, 2007]. In the advection-diffusion-reaction context, classical stabilised finite element methods can be generalised as a Galerkin formulation of the following way [Hughes et al., 2018]:

$$b_h(u_h, v_h) + (\tau(\mathcal{L}u_h - f), \mathbf{L}v_h) = \ell_h(v_h) \tag{1.3}$$

where $\mathbf{L}$ represents a *differential operator* such as

$$\begin{aligned}
\mathbf{L}v_h &= +\ \mathcal{L}_{adv}v_h = \beta \cdot \nabla v_h && \text{(SUPG)} \\
\mathbf{L}v_h &= +\ \mathcal{L}v_h = -\kappa\Delta v_h + \beta \cdot \nabla v_h + \sigma v_h && \text{(GaLS)} \\
\mathbf{L}v_h &= -\ \mathcal{L}^* v_h = \kappa\Delta v_h + \nabla \cdot (\beta v_h) - \sigma v_h && \text{(VMS)}
\end{aligned} \tag{1.4}$$

There exist many definitions for the stabilisation parameter τ, in general, they are derived from a scaling argument and can have the following form:

$$\tau = \left(C_1 \frac{\|\kappa\|_\infty}{h^2} + C_2 \frac{\|\beta\|_\infty}{h} + C_3 \|\sigma\|_\infty \right)^{-1}, \tag{1.5}$$

where C_1, C_2 and C_3 are problem-dependent coefficients.

Despite these efforts, the stabilisation process is still a significant challenge for the scientific community due to its weight on the numerical results. For instance, the quality of the discrete solution strongly depends on the appropriate selection of the penalisation parameter τ. Similarly, the solver performance is also susceptible to this parameter selection.

Remark 1. *Definitions in* (1.5) *pertain to linear problems defined in an Eulerian frame, such as fluid mechanics problems, where constitutive models are functions of constant parameters. However, constitutive models for accounting chemo-mechanical processes (such as curing) are also functions of the evolving geometric parameters in a finite-deformation context for solid mechanics. A series of works discuss this feature that could be extrapolated to strain localisation contexts and be of interest to the reader [Gajendran et al., 2018; Anguiano et al., 2020, 2022].*

Unlike the residual-based stabilised FEM, non-conforming schemes achieve numerical stability using a different approach. These techniques, such as the discontinuous Galerkin (dG) methods, reach a stable approximation by enforcing an element-by-element discretisation and introducing a suitable choice of numerical traces [Reed & Hill, 1973; Lesaint & Raviart, 1974; Johnson & Pitkäranta, 1986; Brezzi et al., 2004; Ern & Guermond, 2006; Cockburn et al., 2012]. Other conforming stabilised formulations are the minimal residual methods, such as the Least Squares Finite Element Method (LS-FEM) [Bochev & Gunzburger, 2009] or the Discontinuous Petrov-Galerkin (dPG) method [Demkowicz & Gopalakrishnan, 2010, 2011, 2014; Calo et al., 2014a; Demkowicz et al., 2012; Niemi et al., 2011a, 2013, 2011b], which minimise the discrete residual with respect to an artificial energy norm and, thus, attain the sought stability.

In Calo et al. [2020], the authors introduced a new class of stabilised finite element formulation in abstract form for any linear partial differential equation system. They analysed the method mathematically and presented numerical evidence supporting the analysis using advection-reaction problems as test cases.

The procedure builds two discrete spaces, solution space and enrichment, where they project the residuals. In Calo et al. [2020], the discrete solution lives in a conforming trial space while minimising the residual in a dual norm of a discontinuous test space that is inf-sup stable. The continuous trial space is a subspace of the discontinuous test space. This formulation endows the discrete, continuous solution with stability properties of the discontinuous Galerkin (dG) formulation used to define the dual norm. The residual minimisation leads to a saddle-point problem. Its solution includes a conforming approximation to the differential system and an on-the-fly error estimate to guide mesh adaptivity. The formulation builds on the non-conformity of the underlying dG method. This non-conformity, in turn, allows us to consider strong norms for the test space when the trial space has high regularity. Measuring the error in stronger norms is a distinguishing feature of this technology, particularly with other recent conforming stabilised formulations such as LS-FEM or the dPG method. This technique was successfully applied to a series of linear problems but lacked a proper extension to nonlinear scenarios.

Remark 2. *Recently, in Hasbani et al. [2021], the authors proposed an alternative enrichment strategy where the solution space is a conforming finite element space and the residual is projected onto a richer polynomial space. The method measures the error in the norm induced by the continuous interior penalty method [Burman & Hansbo, 2004; Burman & Ern, 2007]. This method has been applied to model dynamic fracture propagation and avoids mesh dependency [Labanda et al., 2022].*

Remark 3. *VMS methods open the possibility of overcoming issues related to the consistency associated with the derivation of stabilisation parameters for both linear [Masud et al., 2020] and nonlinear [Masud & Xia, 2006] problems. From a physical point of view, this aspect represents a highlight as typical scaling arguments do not work for geometrically nonlinear solids [Xia & Masud, 2009; Masud & Truster, 2013]. In that sense, in the context of the novel stabilised method proposed by Calo et al. [2018], an extension of this technique in a VMS context has been developed by Giraldo & Calo [2022] for singularly perturbed problems and opens the door for a promising synergy.*

1.2 Thesis overview

Predicting the emergence and features associated with this type of localised deformation is essential in recognising compaction bands in experimental tests and field observations. Thus, this research project aims to study compaction instabilities in geomaterials from a computational perspective.

1.2.1 Motivation and objectives

Our primary goal is to develop a consistent numerical framework that explains compaction banding in geomaterials through robust numerical techniques, considering the following specific aims for its completion:

1. Extend the automatically-adaptive stabilised finite element method to deal with highly nonlinear problems.

2. Simulate the governing equation of cnoidal waves in solids to reproduce the location, spacing, and thickness of stress concentrations.

3. Develop a theoretical framework that describes the strain localisation using a viscoplastic model to reproduce compaction banding for typical p-q stress paths in geomechanics.

4. Analyse compaction banding phenomenon in numerical experiments under triaxial compression, considering a simple viscoplastic model.

1.2.2 Significance

This research project leads to a better understanding of compaction bands in geomaterials, which is essential in various aspects. On the one hand, from a geomechanical perspective, this project enhances the current state of the art on the numerical modelling of the onset conditions and periodicity of compaction bands in geomaterials, which is a challenging topic due to the complexity of solving the equations and the current lack of suitable numerical tools. Besides, extending these results and obtaining numerical validation from experimental tests provide new insights into this problem. Furthermore, capturing this behaviour could be extended for developing a comprehensive re-analysis of classical failure processes related to the undrained strength of clays or the static liquefaction of sands. Additionally, a numerical significance is implied in solving this type of equation.

Providing an enhanced numerical formulation that can deal with this kind of problem may unlock other questions expressed by similar systems of equations within the field of wave mechanics.

On the other hand, from an industrial point of view, this type of localised deformation is crucial in the study of unconventional resources, which includes those subjected to volatile conditions, deeper and hotter than ever reached before and in a challenging environment for their extraction [Regenauer-Lieb et al., 2016]. Compaction bands play an essential role in the energy sector, for example, in shale gas and oil reservoir simulation [Alevizos et al., 2017] and the mineral exploration sector, as those bands can become pathways for mineralising fluids [Kelka et al., 2017; Poulet et al., 2017].

1.2.3 Structure of the thesis

We divide this thesis into two parts. The first part focuses on the numerical analysis of a recent equation for compaction band formation using a new finite-element-based method. The second part concerns a study of the compaction banding features in rate-dependent geomaterials. **Chapter 1** examines the computational modelling state-of-the-art of the compaction band phenomenon and details all the considerations made in terms of numerical approximations and constitutive modelling assumptions. **Chapter 2** introduces the new adaptive stabilised finite element framework in a direct application of the method in linear problems, specifically in heterogeneous and anisotropic diffusion cases, emphasising its more notable features and comparing it to classical techniques of stabilisation. **Chapter 3** develops an extension of this new numerical method for nonlinear problems, showing its efficiency in two specific kinds of problems: nonlinear weak constraint enforcement in advection-diffusion cases and highly nonlinear reaction-diffusion problems. **Chapter 4** concerns the numerical analysis of a novel one-dimensional governing equation for the compaction band phenomenon, first analysing the mathematical nature of the equation and then developing numerical simulations based on the new method previously described. **Chapter 5** contains an analytical and numerical study of the physical problem, considering a consistent constitutive model to reproduce the compaction band phenomenon and its essential features. The final chapter, **Chapter 6**, contains a general discussion and details our conclusions.

Chapter 2

Adaptive stabilised finite element method for problems with diffusion

Advection-diffusion-reaction problems arise in a wide range of phenomena relevant to many areas of applied physics and engineering, such as flow in porous media (e.g., reservoir engineering [Ewing & Wang, 2001; Calo et al., 2014b], groundwater flow [Calo et al., 2014b; Ern et al., 2009]), and drug delivery [Hossain et al., 2012; Calo et al., 2008; Bazilevs et al., 2007]. These processes generally involve heterogeneous and highly anisotropic diffusion tensors, representing varying material properties (e.g., permeability, porosity) in the domain [Calo et al., 2016; Galvis et al., 2018; Calo et al., 2011]. Thus, the accuracy and stability of the numerical approximation have been the focus of intense research for several decades. Moreover, this problem behaves as a hyperbolic partial differential equation (PDE) in advection-dominated regimes, implying that an inaccurate numerical approximation could produce non-physical oscillatory discrete solutions on coarse meshes.

In this chapter[1], we explore the performance of the new adaptive stabilised finite element method (V_h^*–FEM) proposed by Calo et al. [2020] for this type of problem. We consider different scenarios for this class of elliptic problems, such as advection-dominated diffusion and heterogeneous, with locally vanishing or

[1]The content of this chapter is published in Cier, R. J., Rojas, S., & Calo, V. M. (2021). Automatically adaptive, stabilised finite element method via residual minimisation for heterogeneous, anisotropic advection–diffusion–reaction problems. *Computer Methods in Applied Mechanics and Engineering*, 385, 114027.

highly-anisotropic diffusivities in two and three dimensions.

2.1 The diffusion-advection-reaction problem

Let $\Omega \subset \mathbb{R}^d$, with dimension $d = 2, 3$, be an open and bounded Lipschitz domain with boundary $\Gamma := \partial\Omega$, and outward unit normal vector $\mathbf{n}$. Using the standard notation of Hilbert and Banach spaces, let $K \in [L^\infty(\Omega)]^{d,d}$ be a diffusion tensor, to be symmetric and positive definite in Ω. Let $\mathbf{b} \in [L^\infty(\Omega)]^d$ denote an advection coefficient, and $\sigma \in L^\infty(\Omega)$ be a reactive coefficient. We write the advection-diffusion-reaction problem as follows:

$$\begin{cases} \text{Find } u \text{ such that:} \\ A(u) = -\text{div}\,(K\nabla u) + \mathbf{b}\cdot\nabla u + \sigma\, u = f, \quad \text{in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad\quad u = 0, \quad \text{on } \Gamma, \end{cases} \tag{2.1}$$

where $A(\cdot)$ represents the continuous operator for the problem and $f \in L^2(\Omega)$ denotes a spatial source.

In the present work, we focus on two different types of advection-diffusion-reaction problems:

- Advection-dominated problems, that is, problems where $0 < \|K\|_\infty, \|\sigma\|_\infty << \|\mathbf{b}\|_\infty$. These problems lead to unstable solutions using the standard FEM on coarse meshes.
- Highly heterogeneous and anisotropic diffusion problems, that is, problems where the diffusion locally takes small values, leading to advection-dominated regimes.

These two scenarios lead to sharp inner and boundary layers, which are difficult to capture with standard FEM formulation as they induce spurious oscillations (see Codina [1998]; Hughes et al. [2018]).

2.1.1 Continuous weak variational formulation

The weak formulation of (2.1) reads:

$$\begin{cases} \text{Find } u \in H_0^1(\Omega), \text{ such that:} \\ b(u, v) = \ell(v), \quad \forall v \in H_0^1(\Omega), \end{cases} \tag{2.2}$$

with bilinear form $b(u,v) = (K\nabla u, \nabla v)_{0,\Omega} + (\mathbf{b}\cdot\nabla u, v)_{0,\Omega} + (\sigma u, v)_{0,\Omega}$, and linear form $\ell(v) = (f,v)_{0,\Omega}$, where $(\cdot,\cdot)_{0,\Omega}$ denotes the L^2-scalar product in Ω. In what follows, we assume that there is a real number $\sigma_0 > 0$ such that

$$\sigma - \frac{1}{2}\nabla\cdot\mathbf{b} \geq \sigma_0 \qquad \text{a.e. in } \Omega. \tag{2.3}$$

Furthermore, we assume that the smallest eigenvalue of K is bounded from below by a positive constant K_0. Then, owing to the Lax-Milgram Lemma, problem (2.2) is well-posed (see e.g., Ern et al. [2009]).

2.1.2 Discrete setting

Let $\{\mathscr{P}_h\}$ be a conforming and shape-regular family of simplicial meshes of Ω and, for simplicity, we assume that any mesh exactly represents Ω in $\mathscr{P}_h$, that is, Ω is a polygon or a polyhedron (cf. Burman & Ern [2007]). Let T be a generic element in $\mathscr{P}_h$, and denote by ∂T its boundary, by $h_T := \max_{x,y\in T}|x-y|$ its diameter, and by $\mathbf{n}_T$ its outward unit normal. We set $h = \max_{T\in\mathscr{P}_h} h_T$. For a given polynomial degree $p \geq 1$, we define the classical dG and FEM approximation spaces respectively given by:

$$\begin{aligned} V_h &:= \{v_h \in L^2(\Omega) \mid \forall T \in \mathscr{P}_h, v_h|_T \in \mathbb{P}^p(T)\},\\ U_h &:= V_h \cap C_0(\overline{\Omega}), \end{aligned} \tag{2.4}$$

where $\mathbb{P}^p(T)$ denotes the space of polynomial functions, defined over T, with a degree smaller or equal to p. Additionally, we define the extended space as:

$$V_{h,\#} = H^2(\mathscr{P}_h) + V_h \tag{2.5}$$

for convenience. We say that F is an "interior face" if there exist two elements $\{T^-(F), T^+(F)\} \in \mathscr{P}_h$, such that $T^-(F)\cap T^+(F) = F$ and F has nonzero measure, whereas F is a "boundary face" if there is $T(F) \in \mathscr{P}_h$ such that $F = T(F)\cap\Gamma$. We collect all the interior faces F of $\mathscr{P}_h$ into the set $\mathscr{S}_h^0 = \bigcup_{T\in\mathscr{P}_h} F$. Similarly, we denote by $\mathscr{S}_h^\partial$ the boundary skeleton, such that $\Gamma = \mathscr{S}_h^0 \cup \mathscr{S}_h^\partial$. Over each $F \in \mathscr{S}_h$, we set $\mathbf{n}_F$ as a predefined unit normal, oriented from $T^-(F)$ to $T^+(F)$, being coincident with $\mathbf{n}$ when $F \in \mathscr{S}_h^\partial$, and h_F as the diameter of the face F. On interior faces, any function $v_h \in V_h$ is two-valued, with values v_h^+ and v_h^-, defined with respect to the predefined normal $\mathbf{n}_F$. Thus, the jump $[\![v_h]\!]_F$ and

the weighted average $\{\!\{v_h\}\!\}_\omega$ functions are defined as:

$$[\![v_h]\!]_F := v_h^+ - v_h^-, \qquad \{\!\{v_h\}\!\}_\omega := \omega^- v_h^- + \omega^+ v_h^+,$$

where the weights satisfy $\omega^- + \omega^+ = 1$, with $\omega^-, \omega^+ \geq 0$. In particular, when considering heterogeneous tensorial diffusivities, we choose the weights accounting for the diffusivity structure:

$$\omega^- = \frac{\delta_{Kn}^+}{\delta_{Kn}^+ + \delta_{Kn}^-}, \qquad \omega^+ = \frac{\delta_{Kn}^-}{\delta_{Kn}^+ + \delta_{Kn}^-}, \tag{2.6}$$

with $\delta_{Kn}^\mp = \mathbf{n}_F \cdot K^\mp \mathbf{n}_F$ if $F \in \mathscr{S}_h^0$, and $\delta_{Kn} = \mathbf{n}_F \cdot K \mathbf{n}_F$ if $F \in \mathscr{S}_h^\partial$. When K is a continuous tensor (homogeneous diffusion), the weights reduce to $\omega^- = \omega^+ = 1/2$. Finally, on a boundary face $F \in \mathscr{S}_h^\partial$, we set $[\![v_h]\!]_F = \{\!\{v_h\}\!\}_F = v_h|_F$. We omit the subscript F in the jump and weighted average functions unless there is ambiguity.

2.1.3 Discontinuous Galerkin (dG) formulation

We briefly discuss a stable discontinuous Galerkin (dG) formulation for problem (2.1). It combines the Symmetry Weighted Interior Penalty (SWIP) scheme [Di Pietro et al., 2008; Ern et al., 2009] that handles general diffusivities, combined with the Upwinding (UPW) method [Brezzi et al., 2004; Di Pietro & Ern, 2012] that handles the advection-reaction contribution.

Considering the discrete setting described in § 2.1.2, the dG formulation of problem (2.2) reads:

$$\begin{cases} \text{Find } u_h \in V_h, \text{ such that:} \\ b_h(u_h^{\mathrm{dG}}, v_h) := b_h^{\mathrm{diff}}(u_h^{\mathrm{dG}}, v_h) + b_h^{\mathrm{adv}}(u_h^{\mathrm{dG}}, v_h) = \ell_h(v_h), \quad \forall v_h \in V_h, \end{cases} \tag{2.7}$$

where the bilinear and linear forms are

$$\begin{aligned} b_h^{\mathrm{diff}}(u_h, v_h) \;:=\; & \sum_{T \in \mathscr{P}_h} (K \nabla u_h \,,\, \nabla v_h)_{0,T} \\ & + \sum_{F \in \mathscr{S}_h} \Big[- ([\![u_h]\!] \,,\, \mathbf{n}_F \cdot \{\!\{K \nabla v_h\}\!\}_\omega)_{0,F} - (\mathbf{n}_F \cdot \{\!\{K \nabla u_h\}\!\}_\omega \,,\, [\![v_h]\!])_{0,F} \Big] \\ & + \sum_{F \in \mathscr{S}_h} \gamma_F \, ([\![u_h]\!] \,,\, [\![v_h]\!])_{0,F} \,, \end{aligned}$$

$$b_h^{\text{adv}}(u_h, v_h) := \sum_{T\in\mathscr{P}_h} (\,\mathbf{b}\cdot\nabla u_h + \sigma\, u_h\,,\, v_h)_{0,T} + \sum_{F\in\mathscr{S}_h^\partial} \left((\,\mathbf{b}\cdot\mathbf{n}_F)^\ominus u_h, v_h\right)_{0,F} + \sum_{F\in\mathscr{S}_h^0} \left[\frac{1}{2}\,(\,|\,\mathbf{b}\cdot\mathbf{n}_F|\,[\![u_h]\!]\,,\,[\![v_h]\!])_{0,F} - (\,\mathbf{b}\cdot\mathbf{n}_F\,[\![u_h]\!]\,,\,\{\!\{v_h\}\!\})_{0,F}\right],$$

and

$$\ell_h(v_h) := \sum_{T\in\mathscr{P}_h} (f, v_h)_{0,T}\,.$$

In the above, $(\cdot)^\ominus$ denotes the negative part of x (i.e., $x^\ominus := \frac{1}{2}\,(|x| - x)$ for any real number x). For problems with diffusion, there exist many suitable choices for the penalty parameter γ_F (e.g., Shahbazi [2005]; Epshteyn & Rivière [2007] analysed the penalty parameters and their dependence on the polynomial order of approximation; Ern et al. [2009] defined and analysed the impact of γ_K as the harmonic average of the "normal" permeabilities; Hartmann & Houston [2008] introduced a mesh-dependent penalty parameter). Aside from dG formulations, Masud et al. [2020] introduced a flexible-scale basis in a VMS context for locally adjusting the penalty parameter. In this work, following Ern et al. [2009]; Shahbazi [2005]; Bastian et al. [2012], we set the penalty parameter γ_F as $\gamma_F = \eta\,\gamma_K$. Here, $\eta > 0$ represents an element-wise parameter defined as:

$$\eta = \frac{1}{2}\frac{(p+1)(p+d)}{d}\left(\frac{\mathscr{A}(\partial T^+)}{\mathscr{V}(T^+)} + \frac{\mathscr{A}(\partial T^-)}{\mathscr{V}(T^-)}\right), \qquad \forall F\in\mathscr{S}_h^0, \tag{2.8}$$

$$\eta = \frac{(p+1)(p+d)}{d}\frac{\mathscr{A}(\partial T)}{\mathscr{V}(T)}, \qquad \forall F\in\mathscr{S}_h^\partial \tag{2.9}$$

where p is the polynomial order of the test space V_h, d is the dimension, and $\mathscr{A}$ and $\mathscr{V}$ denote area and volume, respectively, for $d = 3$, and length and area, respectively, for $d = 2$. We define γ_K as follows:

$$\gamma_K = (\omega^-)^2\delta_{Kn}^- + (\omega^+)^2\delta_{Kn}^+, \qquad \forall F\in\mathscr{S}_h^0, \tag{2.10}$$

$$\gamma_K = \delta_{Kn}, \qquad \forall F\in\mathscr{S}_h^\partial, \tag{2.11}$$

thus, recalling weights definition (2.6), we derive that

$$\gamma_K = \frac{\delta_{Kn}^+\delta_{Kn}^-}{\delta_{Kn}^+ + \delta_{Kn}^-} \qquad \forall F\in\mathscr{S}_h^0. \tag{2.12}$$

When we consider scalar diffusivities, that is, $K = \varkappa I$ for some scalar function $\varkappa : \Omega \to \mathbb{R}$, we recover the symmetric interior penalty (IP or SIP)

method [Wheeler, 1978; Arnold, 1982; Arnold et al., 2002] given that the penalty parameters reduces to $\gamma_F = \eta\varkappa$.

Remark 4. *The dG formulation allows us to weakly impose non-homogeneous Dirichlet boundary conditions through modifying the right-hand side of (2.7) as follows: if we look for a solution of problem (2.1) satisfying $u = g_D$ on Γ, being $g_D \in H^{1/2}(\Gamma)$ a boundary source, then we rewrite the linear form $\ell_h(v_h)$ as:*

$$\ell_h(v_h) := \sum_{T\in\mathscr{P}_h} (f, v_h)_{0,T} + \sum_{F\in\mathscr{S}_h^\partial} \Big[-(g_D, \mathbf{n}_F \cdot K\nabla v_h)_{0,F} + \gamma_F (g_D, v_h)_{0,F} + \big((\mathbf{b}\cdot\mathbf{n}_F)^\ominus g_D, v_h\big)_{0,F} \Big].$$

Additionally, we provide the discrete space V_h with the following norm:

$$\|w\|_{V_h}^2 := \|w\|_{\text{adv}}^2 + \|w\|_{\text{diff}}^2 \tag{2.13}$$

with

$$\begin{aligned} \|w\|_{\text{adv}}^2 &:= \|w\|_{0,\Omega}^2 + \frac{1}{2}\| \,|\,\mathbf{b}\cdot\mathbf{n}\,|^{\frac{1}{2}} w\|_{0,\Gamma}^2 + \frac{1}{2}\sum_{F\in\mathscr{S}_h^0} (|\,\mathbf{b}\cdot\mathbf{n}_F|\,[\![w]\!], [\![w]\!])_{0,F} \\ &\quad + \sum_{T\in\mathscr{P}_h} h_T \| \,\mathbf{b}\cdot\nabla w\,\|_{0,T}^2, \\ \|w\|_{\text{diff}}^2 &:= \| \,\kappa\nabla w\,\|_{0,\Omega}^2 + \sum_{F\in\mathscr{S}_h} (\gamma_F [\![w]\!], [\![w]\!])_{0,F}\,, \end{aligned} \tag{2.14}$$

where κ denotes the (unique) symmetric positive definite tensor-valued field such that $\kappa^2 = K$ a.e. in Ω. Following Ern et al. [2009], we define:

$$|w|_{V_{h,\beta}}^2 := \sum_{T\in\mathscr{P}_h} h_T \| \,\mathbf{b}\cdot\nabla w\,\|_{0,T}^2, \tag{2.15}$$

which represents the last component of the norm $\|w\|_{\text{adv}}^2$ that controls the advective derivative error for small diffusivities. Finally, we define the following extended norm $\|w\|_{V_{h,\#}}$:

$$\|w\|_{V_{h,\#}} := \|w\|_{V_h} + \left(\sum_{T\in\mathscr{P}_h} \|w\|_{0,\partial T}^2\right)^{\frac{1}{2}} + \left(\sum_{T\in\mathscr{P}_h} h_T \|\kappa\nabla w\|_{0,\partial T}^2\right)^{\frac{1}{2}}. \tag{2.16}$$

In the remainder, the symbol $\lesssim$ indicates an inequality involving a positive constant C independent of h, K, $\mathbf{b}$ and σ. The norms we define above imply that the following theorem holds (see [Ern et al., 2009, §3 & §4]):

Theorem 1 (Well-posedness and a priori error estimate of the dG formulation)**.** *The following propositions hold true:*

(a) Inf-sup stability: There exists a constant $C_{\text{sta}} = C\Delta_K^{-1}$*, with* $C > 0$*, uniform with respect to the mesh size, such that:*

$$\sup_{v_h \in V_h \setminus \{0\}} \frac{b_h(w_h, v_h)}{\|v_h\|_{V_h}} \geq C_{\text{sta}} \|w_h\|_{V_h}, \quad \forall w_h \in V_h,$$

where $\forall\, T \in \mathscr{P}_h$*,* $\Delta_K = \max_{T \in \mathscr{P}_h} \Delta_{K,T}$*, and*

$$\Delta_{K,T} = \begin{cases} 1 & \text{if } \|\mathbf{b}\|_{[L^\infty(T)]^d} \gtrsim \dfrac{\lambda_{M,T}}{h_T}, \\ \dfrac{\lambda_{M,T}}{\lambda_{m,T}} & \text{otherwise,} \end{cases}$$

with $\lambda_{M,T}/\lambda_{m,T}$ *as the maximum/minimum eigenvalues of* $K|_T$*, respectively.*

(b) Boundedness: There exists a mesh-independent constant $C_{\text{bnd}} < \infty$*, s.t.:*

$$b_h(z, v_h) \leq C_{\text{bnd}} \|z\|_{V_{h,\#}} \|v_h\|_{V_h}, \quad \forall (z, v_h) \in V_{h,\#} \times V_h.$$

(c) Consistency: Let u *be the solution of* (2.2)*. If* $u \in H_0^1(\Omega) \cap H^2(\mathscr{P}_h)$*, then*

$$b_h(u, v_h) = \ell_h(v_h), \quad \forall v_h \in V_h.$$

Henceforth, u^{dG}*, solution of problem* (2.7)*, is unique (cf. Ern et al. [2009]; Di Pietro & Ern [2012]). Moreover, the following a priori error estimate is satisfied*

$$\inf_{y_h \in V_h} \|u - y_h\|_{V_h} \leq \|u - u_h^{\text{dG}}\|_{V_h} \leq \left(1 + \frac{C_{\text{bnd}}}{C_{\text{sta}}}\right) \inf_{v_h \in V_h} \|u - v_h\|_{V_{h,\#}}. \tag{2.17}$$

We relate the convergence rates of both left- and right-hand sides of the error estimate (2.17) through the following definition (cf. [Di Pietro & Ern, 2012, §1.4.4]):

Definition 1 (Optimality, quasi-optimality and suboptimality of the error estimate)**.** *The error estimate* (2.17) *is*

1. optimal *if* $\|\cdot\|_{V_h} \simeq \|\cdot\|_{V_{h,\#}}$,
2. quasi-optimal *if the two norms are different, but the lower and upper bounds in* (2.17) *converge, for smooth* u*, at the same convergence rate as* $h \to 0$,
3. suboptimal *if the upper bound has lower convergence rate than the other.*

2.2 Adaptive stabilised method via residual minimisation on dual discontinuous Galerkin norms (V_h^*–FEM)

We now introduce and explain the automatic adaptive stabilised finite element method via residual minimisation on dual discontinuous Galerkin (dG) norms, hereafter, V_h^*–FEM, devised by Calo et al. [2020], in an abstract setting.

2.2.1 Method procedure

In (2.7), u_h^{dG} discretely approximates the solution of (2.2) belonging to a *discontinuous* discrete space. In Calo et al. [2020], the formulation approximates u in a discrete space that possesses additional properties, for instance, continuity and, possibly, higher smoothness. Rather than solving problem (2.7), the V_h^*–FEM method implies:

(a) For a given polynomial degree $p \geq 1$ and a given mesh of size h, considering the discrete space V_h as the test space, and $U_h \subset V_h$ as the trial space (see Eq. (2.4)).

(b) Obtaining the discrete solution $u_h \in U_h$ by solving the following residual minimisation problem:

$$\begin{cases} \text{Find } u_h \in U_h \subset V_h, \text{ such that:} \\ u_h = \underset{z_h \in U_h}{\arg\min} \dfrac{1}{2}\|\ell_h - B_h\, z_h\|_{V_h^*}^2 = \underset{z_h \in U_h}{\arg\min} \dfrac{1}{2}\|R_{V_h}^{-1}(\ell_h - B_h z_h)\|_{V_h}^2, \end{cases} \tag{2.18}$$

where $B_h : V_h, \# \to V_h^*$ is defined as:

$$\langle B_h w_h, v_h \rangle_{V_h^* \times V_h} := b_h(w_h, v_h), \tag{2.19}$$

and $R_{V_h}^{-1}$ denotes the inverse of the Riesz map:

$$R_{V_h} : V_h \to V_h^* \tag{2.20}$$

$$v_h \to \langle R_{V_h} y_h, v_h \rangle_{V_h^* \times V_h} := (y_h, v_h)_{V_h}, \tag{2.21}$$

with $(\cdot, \cdot)_{V_h}$ denotes the inner product inducing the discrete norm $\|\cdot\|_{V_h}$ (i.e., $\|\cdot\|_{V_h} = (\cdot, \cdot)_{V_h}^{1/2}$).

The second equality in (2.18) is due to the fact that Riesz map (2.20) is an isometric isomorphism, therefore $\|\cdot\|_{V_h^*}$ is equivalent to $\|R_{V_h}^{-1}(\cdot)\|_{V_h}$ (cf., Oden & Demkowicz [2017], Theorem 6.4.1). Thus, problem (2.18) is equivalent to the following saddle-point problem (see Cohen et al. [2012]):

$$\begin{cases} \text{Find } (\varepsilon_h, u_h) \in V_h \times U_h, \text{ such that:} \\ (\varepsilon_h, v_h)_{V_h} + b_h(u_h, v_h) = \ell_h(v_h), \quad \forall v_h \in V_h, \\ \qquad\qquad b_h(z_h, \varepsilon_h) = 0, \qquad\quad \forall z_h \in U_h, \end{cases} \tag{2.22}$$

where ε_h is a residual representative in V_h. Indeed, the first identity in (2.22) implies that:

$$\varepsilon_h = R_{V_h}^{-1}(\ell_h - B_h u_h) = R_{V_h}^{-1} B_h (u_h^{\mathrm{dG}} - u_h), \tag{2.23}$$

where the second identity in (2.23) comes from (2.7).

The saddle-point formulation has several desirable properties for numerical approximations. Firstly, the matrix associated with the inner product $(\varepsilon_h, v_h)_{V_h}$ in (2.22) is always symmetric and positive-definite, independently of the nature of the chosen dG formulation; thus, several well-known iterative solvers are effective on the resulting saddle-point problem. Moreover, the discrete approximation $u_h \in U_h$ inherits the discrete stability of the dG formulation. Finally, $\varepsilon_h \in V_h$ is a residual representative that is an efficient error estimator which, under an adequate saturation assumption, also becomes reliable and thus a robust error estimate. Below, we summarise the last properties (see Calo et al. [2020] for

details):

Theorem 2 (Well posedness and a priori error bound estimates for the saddle–point problem)**.** *The solution* $(\varepsilon_h, u_h) \in V_h \times U_h$ *of the saddle-point problem* (2.22) *is unique and the following a priori bound applies:*

$$\|\varepsilon_h\| \leq \|\ell_h\|_{V_h^*} \qquad \textit{and} \qquad \|u_h\|_{V_h} \leq \frac{1}{C_{\text{sta}}}\|\ell_h\|_{V_h^*}, \tag{2.24}$$

and the following a priori error estimate holds:

$$\|u - u_h\|_{V_h} \leq \left(1 + \frac{C_{\text{bnd}}}{C_{\text{sta}}}\right) \inf_{z_h \in U_h} \|u - z_h\|_{V_{h,\#}}, \tag{2.25}$$

where $u \in X_\#$ *is the exact solution to the continuous problem* (2.1)*.*

Proposition 1 (Efficiency of the residual representative)**.** *Under the same hypotheses of Theorem* (2)*, the following holds:*

$$\|\varepsilon_h\|_{V_h} \leq C_{\text{bnd}} \, \|u - u_h\|_{V_h,\#}. \tag{2.26}$$

Assumption 1 (Saturation)**.** *Let* $u_h \in U_h$ *be the second component of the pair* $(\varepsilon_h, u_h) \in V_h \times U_h$ *solving the saddle-point problem* (2.22)*. Let* $u_h^{\text{dG}} \in V_h$ *be the unique solution to* (2.7)*. There exists a real number* $\delta \in [0, 1)$*, uniform with respect to the mesh size, such that* $\|u - u_h^{\text{dG}}\|_{V_h} \leq \delta \|u - u_h\|_{V_h}$*.*

Proposition 2 (Reliability of the residual representative)**.** *Let* $u_h \in U_h$ *be the second component of* $(\varepsilon_h, u_h) \in V_h \times U_h$ *solving the saddle-point problem* (2.22)*. Let* $u_h^{dG} \in V_h$ *be the unique solution to* (2.7)*. Then the following holds true:*

$$\|u_h - u_h^{\text{dG}}\|_{V_h} \leq \frac{1}{C_{\text{sta}}}\|\varepsilon_h\|_{V_h}. \tag{2.27}$$

Moreover, if the saturation Assumption 1 is satisfied, then the following a posteriori error estimate holds true:

$$\|u - u_h\|_{V_h} \leq \frac{1}{(1-\delta)C_{\text{sta}}}\|\varepsilon_h\|_{V_h}. \tag{2.28}$$

We now state the requirements for an efficient and reliable error estimate. From (1) and (2.28), we have:

$$\|u - u_h\|_{V_h} \lesssim \|\varepsilon_h\|_{V_h} \lesssim \|u - u_h\|_{V_{h,\#}}. \tag{2.29}$$

Thus, to ensure the usefulness of the residual representative, we need at least quasi-optimality, which means that the left-hand side should decay at the same rate as the right-hand side of (2.29) (see Definition 1).

2.2.2 Convergence rates

In the above framework, we recover some insightful results related to the convergence rates. Similarly as done in Calo et al. [2020, Appendix B] (see also Karakashian & Pascal [2003]; Burman & Ern [2007]; Ern & Guermond [2017]), under mesh regularity assumptions of § 2.1.2, we can prove:

$$\inf_{v_h \in U_h} \|u - v_h\|_{V_{h,\#}} \lesssim \inf_{v_h \in V_h} \|u - v_h\|_{V_{h,\#}}, \forall u \in V_{h,\#}. \tag{2.30}$$

In particular, the residual minimisation method delivers a discrete solution with the same quality as the dG formulation as a consequence of (2.25) and (2.29). Thus, if the solution is regular enough, it follows:

$$\inf_{v_h \in V_h} \|u - v_h\|_{V_{h,\#}} \lesssim h^p. \tag{2.31}$$

Additionally, from (2.31) we can deduce a bound for the advective component $|\cdot|_{V_h,\beta}$, defined in (2.15). Indeed, we can easily infer that:

$$\| \mathbf{b} \cdot \nabla(u - u_h) \|_{0,T}^2 \lesssim \| \kappa \nabla(u - u_h) \|_{0,\Omega}^2 \lesssim \|u - u_h\|_{V_{h,\#}}^2, \tag{2.32}$$

where u and u_h are defined as in Theorem 2. Combining (2.30), (2.32) and (2.31), multiplying both sides by the mesh size, and taking the square root, we can recover the optimal convergence rate for $|u - u_h|_{V_{h,\beta}}$, which reads:

$$|u - u_h|_{V_{h,\beta}} = \left(\sum_{T \in \mathscr{P}_h} h_T \| \mathbf{b} \cdot \nabla(u - u_h) \|_{0,T} \right)^{1/2} \lesssim h^{1/2} \|u - u_h\|_{V_{h,\#}} \lesssim h^{p+1/2}. \tag{2.33}$$

Similarly, we can deduce a bound for the L^2-norm error in the following way:

$$\|u - u_h\|_{0,\Omega} \lesssim \|u - u_h\|_{V_{h,\#}} \lesssim \inf_{v_h \in V_h} \|u - v_h\|_{V_{h,\#}} \lesssim h^p, \tag{2.34}$$

which is suboptimal. Indeed, we show through numerical examples that the method could deliver suboptimal solutions in L^2.

2.3 Numerical examples

In this section, we discuss some implementation aspects and describe the test cases that show the method's performance under a wide range of scenarios.

2.3.1 Implementation aspects

We use FEniCS [Alnæs et al., 2015] to perform all the numerical simulations. We show convergence plots of the error measured in the L^2 and V_h norms versus the number of degrees of freedom (DOFs) (i.e., $\dim(U_h) + \dim(V_h)$ for (2.22)). As stated in § 2.1.3, we use the Symmetric Weighted Interior Penalty (SWIP) method for the diffusive part of the bilinear form, which extends the classical Symmetric Interior Penalty (SIP) formulation. Extensive numerical testing shows that other dG formulations have similar computational cost and convergence rates; thus, we detail the method's performance using the SWIP formulation for brevity.

2.3.1.1 Adaptive mesh refinement

We use $\varepsilon_h \in V_h$ to estimate the error and drive the adaptive mesh refinement process [Calo et al., 2020]. We follow a standard adaptive procedure, which considers an iterative loop where each level of refinement, we perform the following four steps:

$$\text{SOLVE} \rightarrow \text{ESTIMATE} \rightarrow \text{MARK} \rightarrow \text{REFINE}.$$

That is, given a mesh partition, we first solve the saddle-point problem (2.22). Later, we use a localised version of the inner product (2.13) (evaluated in each mesh cell T) as error estimator E_T:

$$E_T^2 := \|\varepsilon_h\|_{loc,T}^2 + \frac{1}{2}|\varepsilon_h|_{loc,F}^2\,, \tag{2.35}$$

with

$$\begin{aligned}\|\varepsilon_h\|_{loc,T}^2 := \; & \|\varepsilon_h\|_{0,T}^2 + \|\kappa\nabla\varepsilon_h\|_T^2 + h_T\|\,\mathbf{b}\cdot\nabla\varepsilon_h\|_{0,T}^2 \\ & + \sum_{F\in\mathscr{S}_h^\partial}\left(\frac{1}{2}\,|\mathbf{b}\cdot\mathbf{n}| + \gamma_F\right)(\varepsilon_h,\varepsilon_h)_{0,F}\,,\end{aligned} \tag{2.36}$$

and

$$|\varepsilon_h|^2_{loc,F} := \sum_{F \in \mathscr{S}_h^0} \left(\frac{1}{2} |\mathbf{b} \cdot \mathbf{n}| + \gamma_F \right) ([\![\varepsilon_h]\!], [\![\varepsilon_h]\!])_{0,F} . \tag{2.37}$$

We then mark elements for refinement using the Dörfler bulk-chasing criterion [Dörfler, 1996] with a dynamic marking strategy proposed by Arnold [2012, § 9.5]. We select all elements for which the cumulative sum of the local values E_T in a decreasing order remains below a user-defined fraction of the total estimated error ($\|\varepsilon_h\|_{V_h}$) and above a given percentage of the maximum local contribution. We keep looping until we reach the desired fraction of the total estimated error, decreasing the percentage of the maximum local contribution in each iteration. We set the total error estimation fraction (`part`) to be 0.25 for $d = 2$ and 0.125 for $d = 3$, whereas the threshold of the maximum local contribution (`frac`) starts on 95%, and decreases (`delfrac`) 5% per iteration. Finally, we bisect each marked element [Rivara, 1984; Bank et al., 1983] to obtain the refined mesh to use in the next step. We show pseudo-code sketching of this procedure in Algorithm 1.

Algorithm 1 Dynamic marking strategy [Arnold, 2012]

Initialise `frac`, `delfrac` and `part`;
Compute E_t from definition in §4.1.1;
Compute the maximum ($E_{t_{max}}$) and the sum of E_t (ΣE_t) ;
Initialise `marked` = `False`;
Initialise $\Sigma E_t^{marked} = 0$;
while $\Sigma E_t^{marked} <$ part $* \Sigma E_t$ **do**
    Compute `new_marked` = ($\neg$ `marked`) `and` ($E_t >$ `frac` $* E_{t_{max}}$)
    Update ΣE_t += ΣE_t[`new_marked`]
    Update `marked` += `new_marked`
    Update `frac` −= `delfrac`
    Attach `marked` to a cell function
end while

2.3.1.2 Iterative solver

The matrix system that problem (2.22) induces has the following form:

$$\begin{bmatrix} G & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\varepsilon} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{L} \\ \mathbf{0} \end{bmatrix} . \tag{2.38}$$

where the superindex T denotes transpose. Following Calo et al. [2020, §5], we apply the iterative algorithm proposed by Bank et al. [1989]. Denoting by $\widehat{G}$ a preconditioner for the Gram matrix G, and by $\widehat{S}$ a preconditioner for the reduced Schur complement $B^T\widehat{G}^{-1}B$, the iterative scheme becomes

$$\begin{bmatrix} \boldsymbol{\varepsilon}_{i+1} \\ \mathbf{u}_{i+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_{i} \\ \mathbf{u}_{i} \end{bmatrix} + \begin{bmatrix} \widehat{G} & B \\ B^T & \widehat{C} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \mathbf{L} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} G & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\varepsilon}_{i} \\ \mathbf{u}_{i} \end{bmatrix} \right\}, \tag{2.39}$$

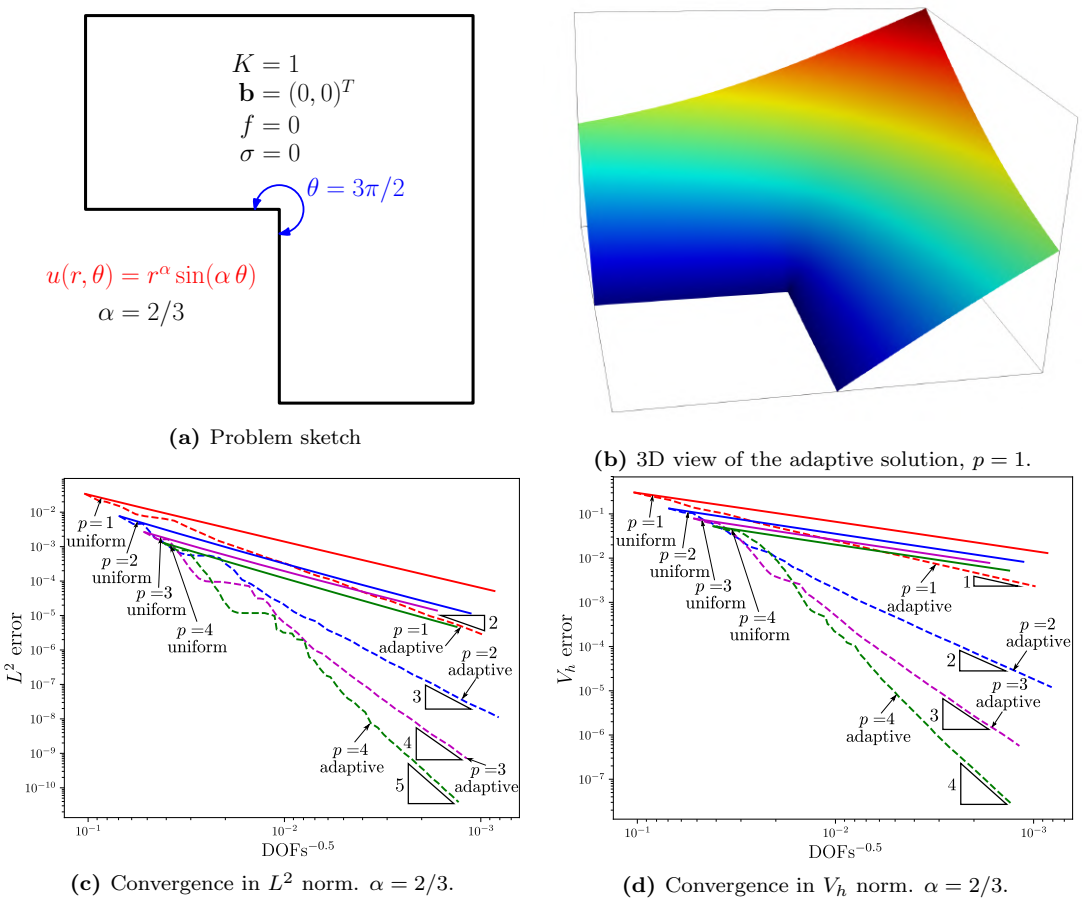
with $\widehat{C} = B^T\widehat{G}^{-1}B - \widehat{S}$. Let $\mathbf{r}_i = \mathbf{L} - G\,\boldsymbol{\varepsilon}_i - B\,\mathbf{u}_i$ and $\mathbf{s}_i = -B^T\boldsymbol{\varepsilon}_i$ be the residuals at the outer iteration i for $\boldsymbol{\varepsilon}$ and $\mathbf{u}$ respectively. Then, the scheme requires the resolution of two interior problems for these increments:

$$\boldsymbol{\eta}_{i+1} := \mathbf{u}_{i+1} - \mathbf{u}_i = \widehat{S}^{-1}\left(B^T\left(\widehat{G}^{-1}\mathbf{r}_i\right) - \mathbf{s}_i\right), \tag{2.40}$$

and

$$\boldsymbol{\delta}_{i+1} := \boldsymbol{\varepsilon}_{i+1} - \boldsymbol{\varepsilon}_i = \widehat{G}^{-1}\left(\mathbf{r}_i - B\,\boldsymbol{\eta}_{i+1}\right). \tag{2.41}$$

We construct an accurate approximation of G^{-1}, which delivers the best computational performance, as low-quality approximations lead to poor conditioning of the reduced Schur complement in (2.40). A relaxed G's approximation, e.g., through conjugate gradients, could be adopted (see, for instance, Bank et al. [1989]); nevertheless, we use an outer iteration since our problem is stiff. That is, we approximate G^{-1} with a sparse Cholesky factorization (e.g., using the module "sksparse.cholmod", see Chen et al. [2008]). Moreover, we choose preconditioning (2.40) with an approximate Schur complement built as $\widehat{S} = B^T(\mathrm{diag}(G))^{-1}B$, where $\mathrm{diag}(G)$ is the main diagonal of G, and the inverse is approximated through the same procedure used for G. We use the LGMRES algorithm (e.g., from Scipy sparse linear algebra package) to solve. On the coarsest mesh, our initial guess is zero, whereas, on the next refinements, our guess is the solution of the previous level of refinement.



(a) Problem sketch

(b) 3D view of the adaptive solution, $p = 1$.

(c) Convergence in L^2 norm. $\alpha = 2/3$.

(d) Convergence in V_h norm. $\alpha = 2/3$.

Figure 2.1: Convergence rates for the Poisson problem in an L-shape domain. $\alpha = 2/3$.

2.3.2 Pure diffusion on L-shape domain

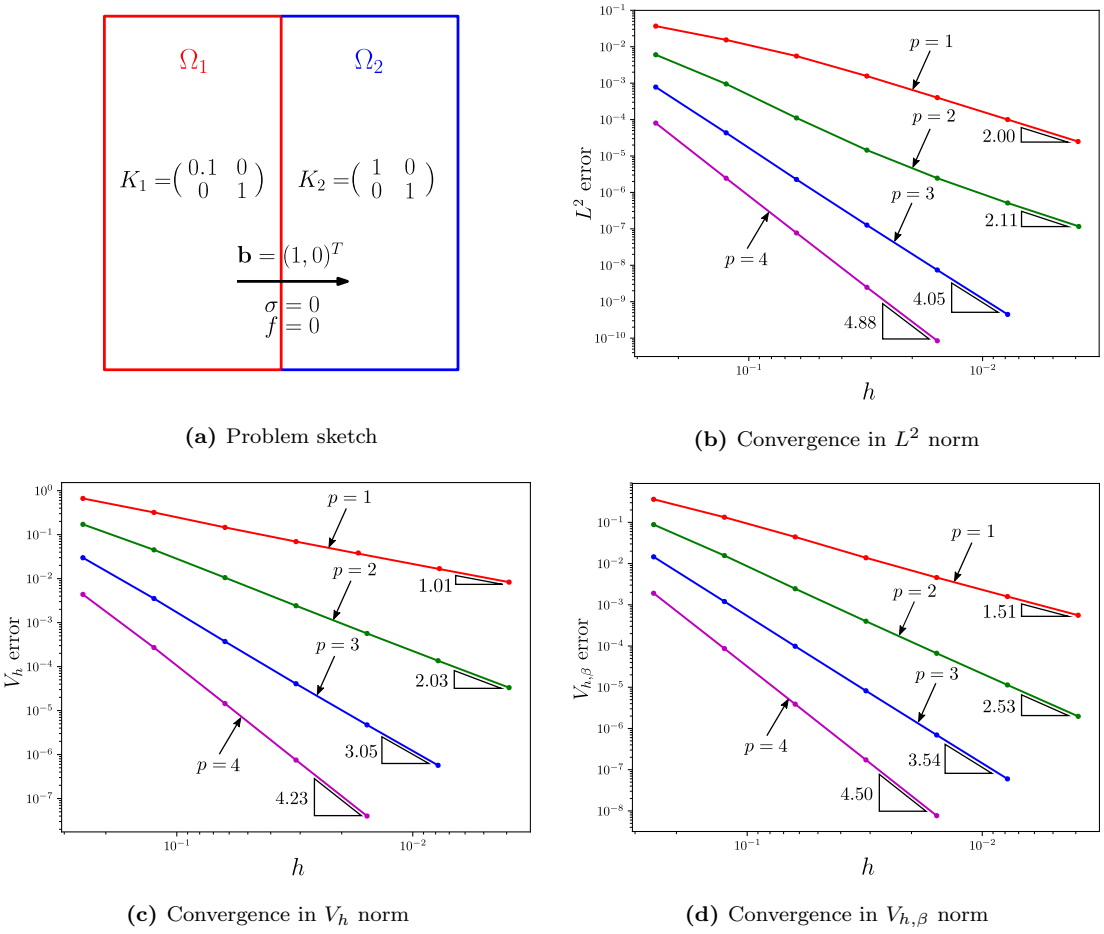
As a first example, we consider the L-shape domain $\Omega := (-1, 1)^2 \setminus (-1, 0]^2$, and the following Poisson problem:

$$\begin{aligned} \Delta u &= 0, \quad \text{in } \Omega, \\ u &= g_D, \quad \text{on } \partial\Omega, \end{aligned} \tag{2.42}$$

where g_D corresponds to the Dirichlet trace of the analytical function in polar coordinates $u(r, \theta) = r^\alpha \sin(\alpha\,\theta)$, with $\theta = 3\pi/2$ for our case. This particular problem is known as reentrant corner problem [Mitchell, 2013], where the solution has a singularity at the corner, and its solution belongs to $H^{1+\alpha-\epsilon}$, $\forall \epsilon > 0$ [Oden & Patra, 1995]. The dG variational formulation we use in problem (2.42) is the formulation (2.7) with $K = 1$, $\mathbf{b} = (0, 0)^T$, $\sigma = 0$ and $f = 0$. Figure 2.1

shows the convergence plots for $\alpha = 2/3$. A uniform refinement strategy cannot deal with the corner singularity achieving similar convergence rates for increasing polynomial orders. However, the adaptive stabilised methodology overcomes this limitation in this case, recovering optimal convergence rates on both L^2 and V_h error norms.

2.3.3 2D problem with heterogeneous diffusion



(a) Problem sketch

(b) Convergence in L^2 norm

(c) Convergence in V_h norm

(d) Convergence in $V_{h,\beta}$ norm

Figure 2.2: Heterogeneous diffusion problem

We solve an advection-diffusion problem with discontinuous diffusion coefficients based on a test from Burman & Zunino [2006]; this problem shows the method performance with heterogeneous diffusion. We split the domain in two subdomains, $\Omega_1 = [0, \frac{1}{2}] \times [0, 1]$ and $\Omega_2 = [\frac{1}{2}, 1] \times [0, 1]$ and use a constant diffu-

sivity tensor in each subdomain

$$K_i(x,y) = \begin{pmatrix} \epsilon_i(x) & 0 \\ 0 & 1.0 \end{pmatrix}$$

where $\epsilon_i(x)$ represent discontinuous values across the interface $x = \frac{1}{2}$ as Figure 2.2a displays.

We set $\epsilon_1(x) = 1 \times 10^{-1}$ and $\epsilon_2(x) = 1.0$ with $\mathbf{b} = (1,0)^T$, $\sigma = 0$ and $f = 0$. For this parameter choice, the exact solution is exponential with respect to the x-coordinate (i.e., independent of the y-coordinate). At the interface, the solution satisfies the following conditions:

$$\lim_{x\to\frac{1}{2}^-} u(x,y) = \lim_{x\to\frac{1}{2}^+} u(x,y),$$
$$\lim_{x\to\frac{1}{2}^-} -\epsilon(x)\partial_x u(x,y) = \lim_{x\to\frac{1}{2}^+} -\epsilon(x)\partial_x u(x,y).$$

We set $u(0,y) = 0$, $u(1,y) = 1$, and by consequence of the matching conditions, we obtain

$$u\left(\frac{1}{2},y\right) = \frac{\dfrac{u(0,y)\exp\left(\frac{1}{2\epsilon_1}\right)}{1-\exp\left(\frac{1}{2\epsilon_1}\right)} + \dfrac{u(1,y)}{1-\exp\left(\frac{1}{2\epsilon_2}\right)}}{\dfrac{\exp\left(\frac{1}{2\epsilon_1}\right)}{1-\exp\left(\frac{1}{2\epsilon_1}\right)} + \dfrac{1}{1-\exp\left(\frac{1}{2\epsilon_2}\right)}}.$$

Thus, the exact solution in each subdomain becomes:

$$u_1(x,y) = \frac{u\left(\frac{1}{2},y\right) - \exp\left(\frac{1}{2\epsilon_1}\right)u(0,y) + \left[u(0,y) - u\left(\frac{1}{2},y\right)\right]\exp\left(\frac{x}{\epsilon_1}\right)}{1-\exp\left(\frac{1}{2\epsilon_1}\right)},$$
$$u_2(x,y) = \frac{u\left(1,y\right) - \exp\left(\frac{1}{2\epsilon_2}\right)u(\frac{1}{2},y) + \left[u(\frac{1}{2},y) - u\left(1,y\right)\right]\exp\left(\frac{x-\frac{1}{2}}{\epsilon_2}\right)}{1-\exp\left(\frac{1}{2\epsilon_2}\right)}.$$

The convergence plots in Figure 2.2 result from using both trial and test space functions of the same polynomial degree, for $p = 1,2,3,4$, and considering a uniform refinement scheme. Thus, the proposed method recovers the same convergence rates as the original dG scheme: h^p in the V_h error norm and $h^{p+1/2}$ in

the $V_{h,\beta}$ error norm, as § 2.2.2 states. Similarly, in the L^2 error norm, we recover optimal convergence rates for odd polynomial degrees and lose half an order in even polynomial degrees. These results indicate that the scheme adequately approximates problems with discontinuous coefficients and efficiently captures the inner layer in the interface region like the original dG formulation.

2.3.4 2D problem with high-contrast anisotropic diffusion

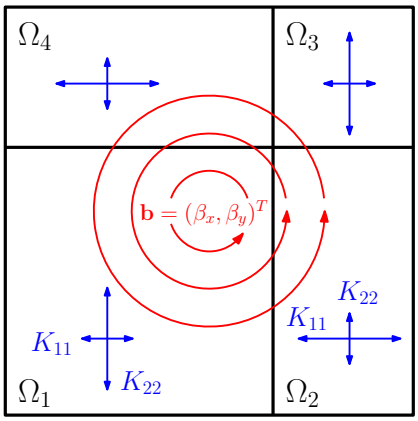


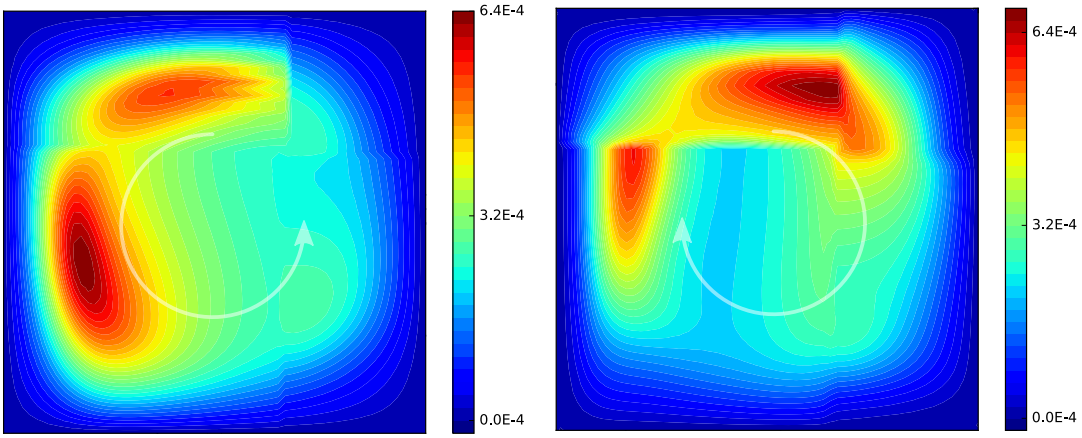
Figure 2.3: Anisotropic diffusion problem sketch. Counterclockwise advection field.

In this example, we consider an anisotropic problem with a vanishing viscosity, following Ern et al. [2009]. We consider the unit square $\Omega = [0,1]\times[0,1]$ split into four subdomains: $\Omega_1 = \left[0,\frac{2}{3}\right] \times \left[0,\frac{2}{3}\right]$, $\Omega_2 = \left[\frac{2}{3},1\right] \times \left[0,\frac{2}{3}\right]$, $\Omega_3 = \left[\frac{2}{3},1\right] \times \left[\frac{2}{3},1\right]$ and $\Omega_4 = \left[0,\frac{2}{3}\right] \times \left[\frac{2}{3},1\right]$. The diffusivity tensor takes different values in each subdomain,

$$K_i(x,y) = \begin{pmatrix} 10^{-6} & 0 \\ 0 & 1.0 \end{pmatrix}, \quad \text{for } i = 1,3 \;\; \forall (x,y) \in \Omega_1, \Omega_3,$$

$$K_i(x,y) = \begin{pmatrix} 1.0 & 0 \\ 0 & 10^{-6} \end{pmatrix}, \quad \text{for } i = 2,4 \;\; \forall (x,y) \in \Omega_2, \Omega_4,$$

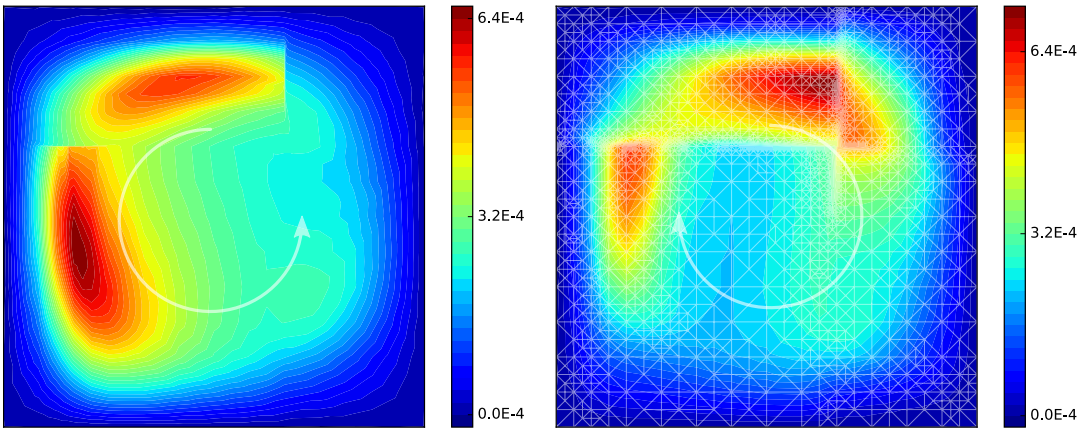
Figure 2.3 sketches the problem setup. The advection field is solenoidal $\mathbf{b} = (\beta_x, \beta_y)^T$, with $\beta_x = 40x(2y-1)(x-1)$ and $\beta_y = -40y(2x-1)(y-1)$ for the counterclockwise case, and $\mathbf{b} = -(\beta_x, \beta_y)^T$ for the clockwise case. Unlike the previous example, the advective field is neither constant nor orthogonal to the discontinuities in the diffusivity K. However, its orientation is still along the direction of increasing diffusivity; for that reason, internal layers develop in



(a) Counterclockwise advection field

(b) Clockwise advection field

Figure 2.4: Anisotropic diffusion problem. Uniform mesh (25.3 K DOFs).



(a) Solution on refined mesh, counterclockwise advection field
(Refinement level 12: 23.7 K DOFs)

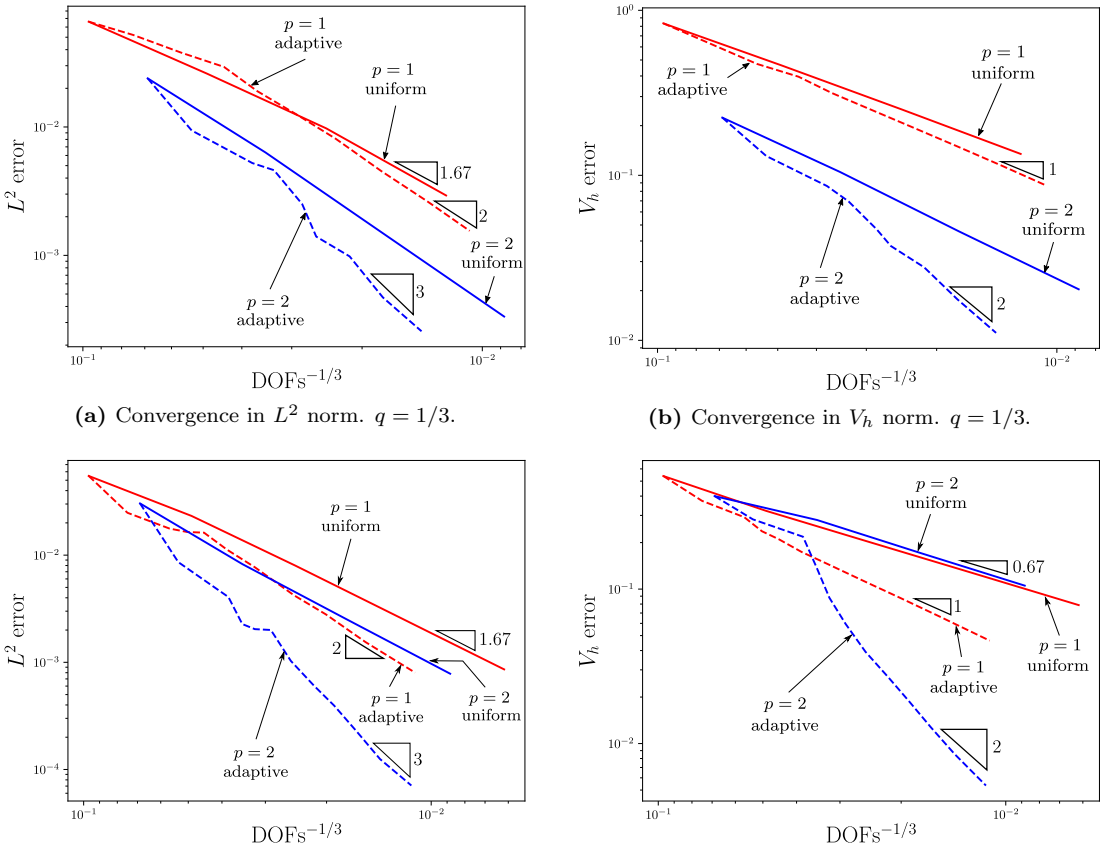
(b) Computed solution and refined mesh, clockwise advection field
(Refinement level 10: 22.3 K DOFs)

Figure 2.5: Anisotropic diffusion problem. Adaptive refinement.

the solution. The forcing term is $f(x, y) = 10^{-2} \exp(-(r - 0.35)^2/0.005)$ with $r^2 = (x - 0.5)^2 + (y - 0.5)^2$, corresponding to a Gaussian hill with center at $r = 0.35$. Finally, we set $\sigma = 1$ for the reaction term, and $g_D = 0$ on $\partial\Omega$ for the boundary condition. We consider two subcases, the first on a quasi-uniform mesh with $h = 0.024$, conforming to the discontinuities of K, and the second through an adaptive scheme starting from a uniform triangular mesh ($h = 0.177$). We solve both cases using the same polynomial degree $p = 1$ for trial and test spaces. Figure 2.4 shows results obtained for a uniform mesh, while Figure 2.5 shows

results for adaptive mesh refinement with fewer degrees of freedom than in the uniform mesh case. The SWIP formulation considers the principal directions of the diffusivity to compute the jump penalty; thus, SWIP avoids overshoots and undershoots near the material interfaces, which is not the case for standard interior penalty schemes [Ern et al., 2009]. Finally, the adaptive strategy concentrates refinement at the inner layer without losing the approximation quality.

2.3.5 3D Fichera corner problem with vertex singularity



(a) Convergence in L^2 norm. $q = 1/3$.

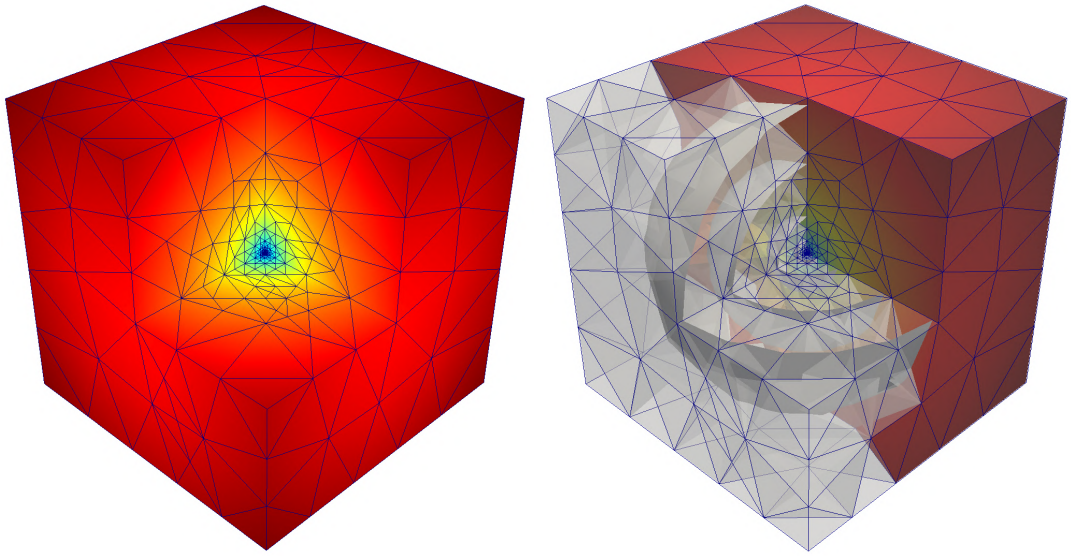
(b) Convergence in V_h norm. $q = 1/3$.

(c) Convergence in L^2 norm. $q = 1/10$.

(d) Convergence in V_h norm. $q = 1/10$.

Figure 2.6: 3D Fichera corner problem. Convergence rates for $q = 1/3$ and $q = 1/10$.

As a first example in three dimensions, we consider a similar problem than in § 2.3.2, considering a polyhedral convex domain $\Omega = (-1\,,\,1)^3 \setminus [\,0,1)^3 \subset \mathbb{R}^3$, with a vertex singularity in the origin, known as Fichera corner problem. Similar to § 2.3.2, we consider (2.1) with $K = 1$, $\mathbf{b} = (0,0,0)^T$ and $\sigma = 0$. We analyse the



(a) Discrete solution. **(b)** Refined mesh and inner contours.

Figure 2.7: 3D Fichera corner problem. Adaptive mesh and discrete solution for $q = 1/10$. Level 7 for $p = 2$: 97,174 DOFs.

following exact solution:

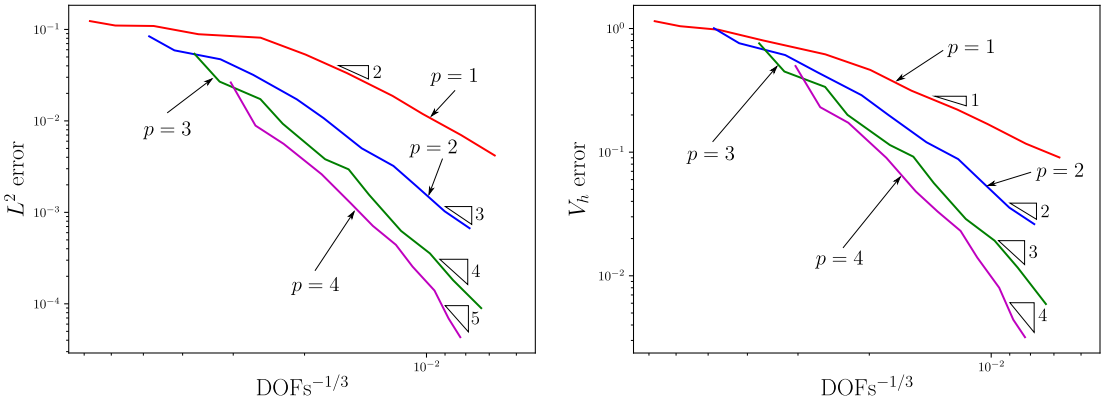
$$u(x, y, z) = \left(\sqrt{x^2 + y^2 + z^2}\right)^q, \tag{2.43}$$

where q is a constant value. For $q > -1/2$, the right-hand side that corresponds to (2.43) reads (cf. Beilina et al. [2005]):

$$f(x, y, z) = -q(q + 1)\left(x^2 + y^2 + z^2\right)^{q/2\,-\,1}.$$

On $\partial\Omega$, we impose the exact solution (2.43) as Dirichlet boundary condition. We solve the problem for two case: $q = 1/3$ and $q = 1/10$. The results reflect the method's robustness, capturing the problem's main features and the solution's spherical nature in 3D. Figure 2.6 shows that the uniform refinement does not improve with p-refinement; our adaptive procedure recovers optimal convergence rates in both L^2 and V_h error norms, regardless of the values of q and the regularity of the solution. Besides, Figure 2.7 shows the resultant adaptive mesh for the convex polyhedron, with accumulated refinement at the origin, where the vertex singularity appears.

2.3.6 3D Eriksson-Johnson problem



(a) Convergence in L^2 norm

(b) Convergence in V_h norm

Figure 2.8: 3D Eriksson-Johnson problem. Triangles show optimal convergence rates.

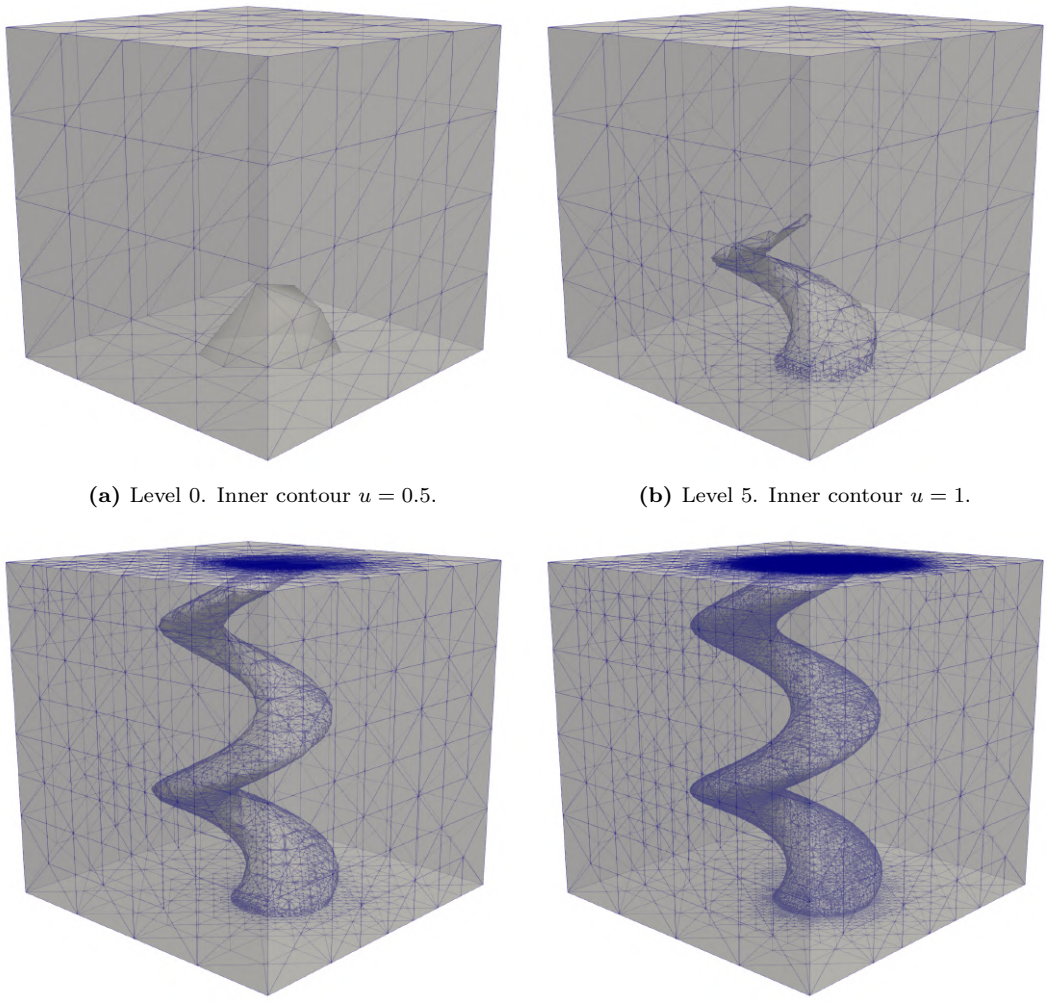
We extend the modified 2D Eriksson-Johnson advection-diffusion problem from Chan & Evans [2013] to analyse the performance of our scheme for advection-diffusion problems in 3D. We extrude the problem in the z direction to obtain a 3D setup; thus, the solution and boundary conditions extend from 2D to 3D. The advection-diffusion problem considers (2.1) in a unit cube as the domain $\Omega = (0\,,\,1)^3 \subset \mathbb{R}^3$, with an advective field of $\mathbf{b} = (1, 0, 0)^T$, a constant scalar diffusion of $K = 10^{-2}$ and a source term of $f = 0$. The problem presents an analytical solution of the following form:

$$u(x, y, z) = \frac{\exp(r_1(x-1)) - \exp(r_2(x-1))}{\exp(-r_1) - \exp(-r_2)} \sin(\pi y),$$

with $r_{1,2} = 1 \pm \sqrt{1 + 4K^2\pi^2}/2K$. We impose at the boundary the value of the analytical solution on $\partial\Omega$. Figure 2.8 shows the convergence for L^2 and V_h for adaptive meshes and the optimal slopes for degrees $p = 1, 2, 3, 4$. As in the heterogeneous 2D problem, our scheme recovers optimal convergence in V_h and L^2, although the theoretical analysis does not guarantee optimality in L^2.

2.3.7 3D advection-dominated diffusion

After showing the method's performance in terms of adaptivity and convergence in 3D, we show its capabilities in more challenging problems where no analytic solution is available. Thus, we consider a 3D



(a) Level 0. Inner contour $u = 0.5$.

(b) Level 5. Inner contour $u = 1$.

(c) Level 9. Inner contour $u = 1$.

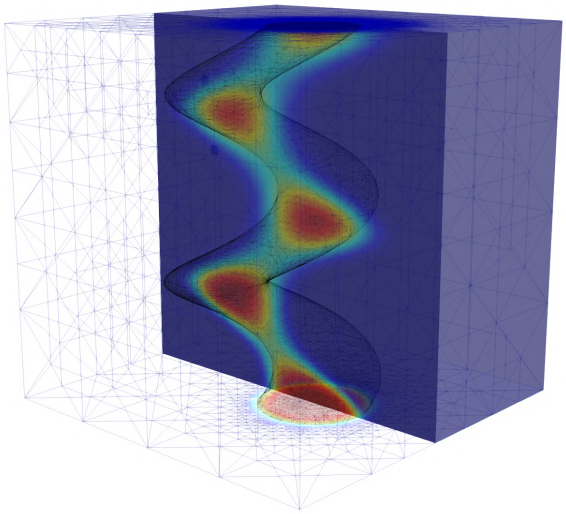
(d) Level 13. Inner contour $u = 1$.

Figure 2.9: 3D advection-dominated diffusion problem. Adaptive mesh evolution.

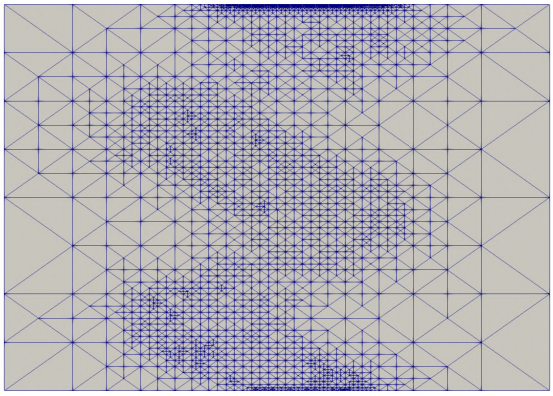
advection-diffusion problem in the unit cube $\Omega = (0\,,\,1)^3 \subset \mathbb{R}^3$. We set the source term $f = 0$, the diffusion $K = 10^{-3}$, the spiral-type advective field $\mathbf{b} = (\beta_x, \beta_y, \beta_z)^T = (-0.15\,\sin(4\pi z)\,,\,0.15\,\cos(4\pi z)\,,\,1)^T$, and the inflow boundary datum g as

$$g = \begin{cases} 1 + \tanh\left[M\left(0.15^2 - (x-0.6)^2 - (y-0.5)^2\right)\right] & \text{on } z = 0, \\ 0 & \text{elsewhere on } \Gamma, \end{cases}$$

These parameters produce a solution that presents an interior layer that starts at the bottom of the unit cube ($z = 0$) for $M \gg 1$ at the inflow boundary datum



(a) Slice at $y = 0.5$.



(b) Mesh diagonal cross section.

Figure 2.10: 3D advection-dominated diffusion problem. Level 13: 4'858,125 DOFs.

that propagates into the domain following the spiral flow. At the top of the unit cube ($z = 1$) (out-flow boundary), the solution exhibits a boundary layer due to the advection-dominant regime and the homogeneous boundary condition we impose at that surface. Figure 2.9 shows the evolution of the 3D mesh as the refinement strategy progresses. Here, Figure 2.9a displays the initial spiral's progress forming the inner layer inside the unit cube, whereas Figures 2.9b, 2.9c, and 2.9d follow the refinement process to capture the sharp internal layer induced by the advective field, as well as the boundary layer that appears at the outlet due

to the small diffusion. Finally, Figure 2.10 shows the discrete solution for a refined mesh displaying the interior mesh and a cut of the solution. The method not only captures the boundary layers via adaptivity due to the robustness of the error estimator but also shows non-oscillatory continuous solutions in each refinement level due to its inherited stability. These two features represent the highlights of this method, as they merge automatically in the framework, compared to classical stabilised schemes (SUPG, VMS, GLS) or even typical dG formulations, where additional error-estimate constructions are necessary if adaptivity is required. Similarly, compared to more recent schemes such as dPG, the proposed approach needs neither the insertion of extra variables in the traces/fluxes (known as "ultra-weak" formulations) nor test space enrichment (i.e., increasing the polynomial degree in the test functions).

This chapter describes an adaptive stabilised conforming finite element method that minimises the residual on dual discontinuous Galerkin (dG) norms for advection-diffusion-reaction problems. This method recovers the optimal convergence rates for h-adaptive schemes in the dG norm in the context of advection-diffusion-reaction problems. Besides, the method captures sharp boundary and internal layers and overcomes the classical overshooting and undershooting problems. With these ideas, we now develop an extension of the method to deal with non-linearities and then employ this technique to resolve the cnoidal governing equation.

Chapter 3

Extension of V_h^*–FEM for nonlinear problems

Motivated by the nonlinear nature of the cnoidal equation, we extend the V_h^*–FEM framework developed by Calo et al. [2020] to solve nonlinear problems. In this chapter[1], we show two first extensions developed for this method. The first weakly enforces constraints in advection-diffusion-reaction problems, whereas the second deals with highly nonlinear problems like Bratu's equation.

3.1 Abstract setting

We consider a well-posed variational formulation for a general nonlinear problem in an abstract setting. For an open set $\Omega \neq 0$, and Hilbert spaces U (trial) and V (test), let N be a differentiable nonlinear map with Fréchet derivative $DN(u)$ at $u \in \Omega$. We associate the nonlinear map $n : U \times V \to \mathbb{R}$, $n(u; v) := \langle N(u), v\rangle$, where $\langle \cdot, \cdot \rangle$ represents the duality pairing in V. Let $n'(u; z, v)$ denote $DN(u)$,

[1] Parts of the content of this chapter are published in:

- Cier, R. J., Rojas, S., & Calo, V. M. (2021). A nonlinear weak constraint enforcement method for advection-dominated diffusion problems. *Mechanics Research Communications*, 112, 103602.

- Cier, R. J., Rojas, S., & Calo, V. M. (2021). Automatically adaptive, stabilised finite element method via residual minimisation for heterogeneous, anisotropic advection–diffusion–reaction problems. *Computer Methods in Applied Mechanics and Engineering*, 385, 114027.

around a known value u, and in the direction of an increment z:

$$n'(u;z,v) := \langle DN(u;z), v\rangle = \frac{d}{d\epsilon} n(u+\epsilon z; v)\Big|_{\epsilon=0} \quad \text{for } u \in \Omega, z \in U, v \in V. \tag{3.1}$$

We finally set $\ell(\cdot) : V \to \mathbb{R}$ as a continuous linear form. Hence, the weak formulation for a nonlinear problem reads:

$$\begin{cases} \text{Find } u \in U, \text{ such that:} \\ \quad n(u;v) = \ell(v), \quad \forall v \in V, \end{cases} \tag{3.2}$$

3.2 Nonlinear V_h^*–FEM

As shown in § 2.2.1, V_h^*–FEM delivers a mixed problem, with a saddle-point structure, in the case of linear problems. In this section, we develop the discrete formulation for the continuous problem (3.2), which reads:

$$N_h(u_h) = \ell_h(\cdot), \tag{3.3}$$

where $N_h : U_h \to V_h^*$ represents the discrete nonlinear map with $\langle N_h(z_h), v_h\rangle_{V_h^* \times V_h} := n_h(z_h; v_h)$, being $n_h(\cdot\,;\cdot)$ the discretisation of the nonlinear form $n(\cdot\,;\cdot)$. Similar to the linear problem, the solution u_h is then computed through minimizing the residual $\ell_h(\cdot) - N_h(z_h)$ associated to (3.3) in the norm of V_h^*:

$$u_h = \underset{z_h \in U_h}{\operatorname{argmin}} \frac{1}{2}\|\ell_h(\cdot) - N_h(z_h)\|^2_{V_h^*} = \underset{z_h \in U_h}{\operatorname{argmin}} \frac{1}{2}\|R^{-1}_{V_h}(\ell_h(\cdot) - N_h(z_h))\|^2_{V_h}, \tag{3.4}$$

with $\|\cdot\|_{V_h}$ the norm of the discrete space, and $R^{-1}_{V_h}$ the inverse of the Riesz map. Following Calo et al. [2020], the nonlinear problem seeks a critical point of the minimising functional, which becomes a linear problem: Find $u_h \in U_h$ such that:

$$(R^{-1}_{V_h}(\ell_h - N_h(u_h)), R^{-1}_{V_h} DN_h(u_h; z_h))_{V_h} = 0, \quad \forall z_h \in U_h, \tag{3.5}$$

being $DN_h(u_h; z_h)$ the discrete form of the derivative (cf., (3.1)). As noticed by Cohen et al. [2012], problem (3.5) can be equivalently written as the following

saddle-point problem:

$$\begin{cases} \text{Find } (\varepsilon_h, u_h) \in V_h \times U_h, \text{ such that:} \\ (\varepsilon_h, v_h)_{V_h} + n_h(u_h; v_h) = \ell_h(v_h), \qquad \forall v_h \in V_h, \\ n_h'(u_h; z_h, \varepsilon_h) = 0, \qquad \forall z_h \in U_h. \end{cases} \tag{3.6}$$

The system (3.6) simultaneously delivers a stable and continuous approximation $u_h \in U_h$ of the dG formulation, and a residual representation $\varepsilon_h \in V_h$ that guides the adaptive mesh refinement. The first line represents the nonlinear problem associated with the residual, whereas the second line constrains the system to remain in the tangent space built from the linearised form.

3.3 Nonlinear solver

We use Newton's method for solving the nonlinear problem. Given the discrete solution pair (ε_h^i, u_h^i) of an iterative step i, we solve for the increment $(\delta\varepsilon_h, \delta u_h)$ of the next iteration, and set $u_h^{i+1} = u_h^i + t^i \delta u_h$, and $\varepsilon_h^{i+1} = \varepsilon_h^i + t^i \delta\varepsilon_h$, where t^i represents a relaxation parameter that controls the increment size. The method seeks for the solution pair $(\varepsilon_h^{i+1}, u_h^{i+1})$ that satisfies (3.6). For the i+1-th iteration, the linearization of (3.6) reads:

$$\begin{cases} \text{Given the pair } (\varepsilon_h^i, u_h^i), \text{ find } (\delta\varepsilon_h, \delta u_h) \in V_h \times U_h, \text{ such that:} \\ (\delta\varepsilon_h, v_h)_{V_h} + n_h'(u_h^i; \delta u_h, v_h) = \ell_h(v_h) - (\varepsilon_h^i, v_h)_{V_h} - n_h(u_h^i; v_h), \quad \forall v_h \in V_h, \\ n_h'(u_h^i; z_h, \delta\varepsilon_h) = -n_h'(u_h^i; z_h, \varepsilon_h^i), \qquad \forall z_h \in U_h. \end{cases} \tag{3.7}$$

In matrix form, formulation (3.7) reads:

$$\begin{bmatrix} G & B_u \\ B_u^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta\varepsilon_h} \\ \boldsymbol{\delta u_h} \end{bmatrix} = \begin{bmatrix} \mathbf{L} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} G\boldsymbol{\varepsilon_h^i} + N(\boldsymbol{u_h^i}) \\ B_u^T \boldsymbol{\varepsilon_h^i} \end{bmatrix} \tag{3.8}$$

where G is the Grammian matrix of the inner product that induces the norm in the discrete space V_h, $N(\boldsymbol{u^i})$ is the vector associated to the nonlinear form $n_h(u_h; v_h)$ and B_u is the matrix associated with its linearisation $n_h'(u_h^i; \delta u_h, v_h)$. The residual representative ε_h is an implicit function of u_h. We define the pair $\boldsymbol{x_h} = (\boldsymbol{\varepsilon_h}, \boldsymbol{u_h})$ that comprises both the solution and the residual representative,

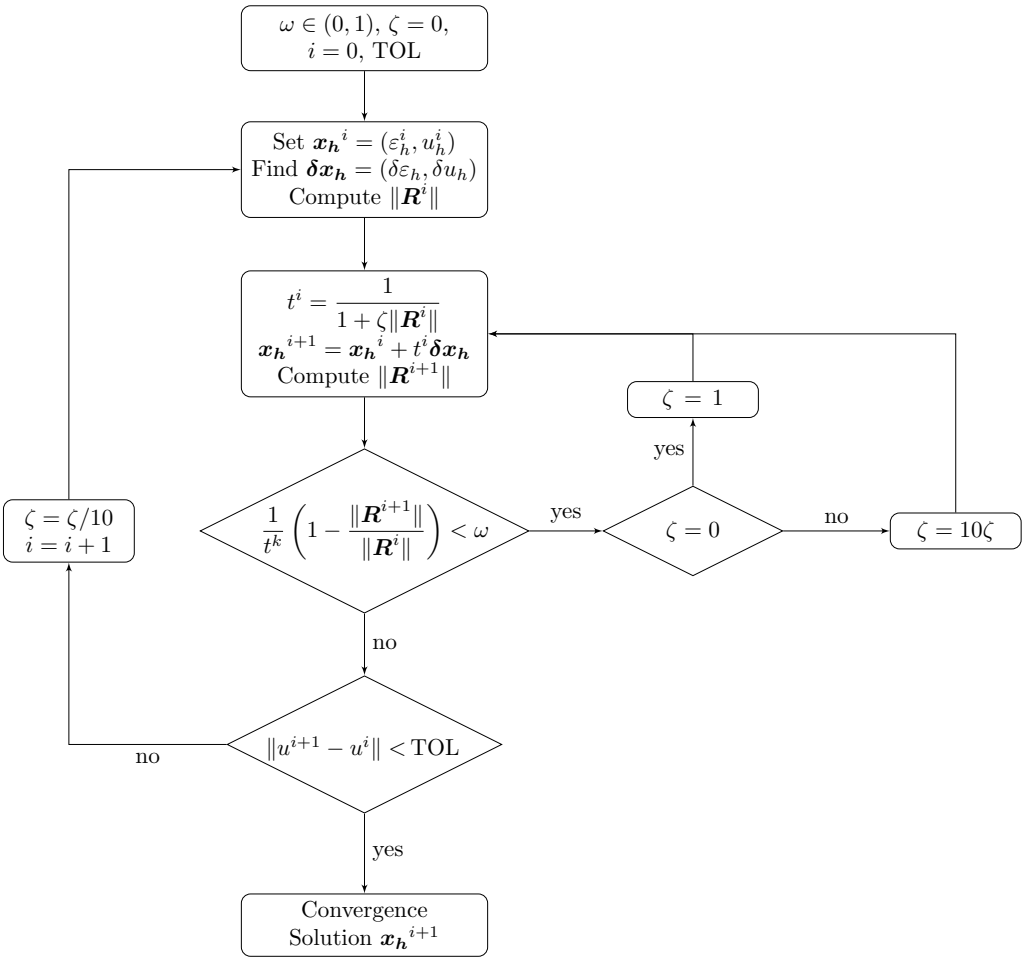


Figure 3.1: Damped Newton's algorithm flow chart.

being valid also for the increments, which allows us to rewrite (3.8) as:

$$J^i\,\boldsymbol{\delta x_h} = \boldsymbol{R}^i,$$

where

$$J^i = \begin{bmatrix} G & B_u \\ B_u^T & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{R}^i = \begin{bmatrix} \mathbf{L} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} G\boldsymbol{\varepsilon_h^i} + N(\boldsymbol{u_h^i}) \\ B_u^T\boldsymbol{\varepsilon_h^i} \end{bmatrix}$$

Each iteration step size controls the method's convergence through the relaxation parameter t^i. Figure 3.1 sketches the damped Newton's method [Bank & Rose, 1981] we use for that purpose.

3.4 Weak constraint enforcement

We start with a linear problem subject to constraints. Thus, we develop a non-linear V_h^*-FEM for weak constraint enforcement in advection-diffusion problems. Although stabilised formulations improve the robustness and accuracy of the numerical solutions, spurious undershoots and overshoots can still be present, especially in low-resolution meshes. These oscillations are a drawback as many engineering applications (i.e., transport of density, concentration, or temperature) require them to remain within their physical range. Violating these bounds delivers unphysical simulation outputs. Thus, we seek to eliminate these overshoots or undershoots through proper constraint enforcement procedures. Therefore, many techniques to surmount this effect exist in the literature, mainly constructed from a stabilised formulation. One of these schemes incorporates shock-capturing terms to satisfy a discrete maximum principle [Burman & Ern, 2005; Mizukami & Hughes, 1985]. Also, flux-corrected methods [Kuzmin & Turek, 2002; Löhner et al., 1987] seek to impose the constraints by altering the system matrix. These methods are generally only first-order accurate. Higher-order schemes require terms to control and often reduce the method's dissipative response. More recent work showed that discontinuity capturing methods have roots in the stabilised methods, and can be derived using VMS notions with a more rigorous basis [Masud & Al-Naseem, 2018].

Our approach finds its origin in the work of Burman & Ern [2017], where the authors propose an alternative constraint imposition approach –more precisely, positivity preserving. The authors weakly satisfy the discrete maximum principle by adding a consistent penalty term to the variational formulation of a Galerkin least-squares (Ga-LS) finite element discretisation. This flexible method incorporates a priori lower and upper bounds on the discrete solution by adding the corresponding consistent penalty term to the discrete formulation. We combine this consistent penalisation with a new adaptive stabilised finite element framework that minimises the residual in dual norms of discontinuous Galerkin (dG) methods [Calo et al., 2020]. This formulation inherits the stability and accuracy of the underlying dG approximation. The formulation seeks a solution in a continuous trial function space, a proper subspace of the dG function space. The resulting saddle-point problem delivers stable formulations with continuous solutions with a robust *a posteriori* error estimate, which can be computed on the fly to drive optimal adaptive mesh refinements.

In this section, we develop a constraint enforcement technique that combines the ideas of the nonlinear penalty method by Burman & Ern [2017] and the residual minimisation technique by Calo et al. [2020], applied to advection-dominated diffusion problems. We construct it as follows. First, we modify the corresponding bilinear dG form by adding a nonlinear penalty term to enforce constraints weakly. Next, we solve a residual minimisation problem in a dG dual norm. The resulting technique minimises the violation of solution bounds and delivers a robust residual estimator to guide adaptive mesh refinement. The main advantage of this procedure is that it results in a nonlinear saddle-point problem with a symmetric Jacobian. Therefore, an extensive list of iterative solvers is available for each step of the Newton iteration (see, e.g., Benzi et al. [2005]). The idea of combining residual minimisation with nonlinear techniques was also considered by Muga et al. [2019] as an extension to advection-reaction problems in Banach spaces and by Houston et al. [2020] as a technique to remove the Gibbs' phenomenon in diffusion-advection-reaction problems. However, the main difference with this work is that the nonlinearity appears in the dual norm.

For the sake of simplicity, we assume that the aim is to enforce a positivity-preserving condition, that is, $u \geq 0$. However, we can extend the technique to impose an upper bound or another minimal value (see Remark 5).

3.4.1 Nonlinear consistent penalty method

Consider the following penalization term (see Remark 6):

$$\gamma = \gamma_0 \left(\frac{\| \beta \|_\ell}{h} + \frac{\|K\|_\infty}{h^2} + \|\sigma\|_\infty \right)^{-1}, \tag{3.9}$$

where $0 < \gamma_0 < 1$ is a user-defined constant real number. We define $\xi_\gamma : V_h \to \mathbb{R}$, as the function:

$$\xi_\gamma(v_h) := [v_h - \gamma(A(v_h) - f)]_-, \qquad \forall v_h \in V_h, \tag{3.10}$$

where $A(v_h)$ represents the discrete version of the operator defined in the original advection-diffusion-reaction problem (2.1), and $x_- = \frac{1}{2}(x - |x|)$ denotes the negative part of the real number x, satisfying $x_- = x$ if $x < 0$, and $x_- = 0$ if $x \geq 0$.

We define $b_h^\gamma(u_h; v_h)$, composed by the original bilinear form $b_h(u_h, v_h)$ and a

nonlinear penalty term, as follows:

$$b_h^\gamma(u_h; v_h) := b_h(u_h, v_h) + \langle \gamma^{-1} \xi_\gamma(u_h), v_h \rangle_h, \tag{3.11}$$

where

$$\langle x_h, y_h \rangle_h := \sum_{T \in \mathscr{T}_h} (x_h, y_h)_T.$$

By construction, the analytical solution satisfies that $\xi_\gamma(u) = 0$, since $A(u) = f$ and $u_- = 0$. We consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in V_h, \text{ such that:} \\ b_h^\gamma(u_h; v_h) = \ell_h(v_h), \quad \forall v_h \in V_h. \end{cases} \tag{3.12}$$

Since $\xi_\gamma(u)$ vanishes identically in Ω, exact consistency still holds for (3.12). Consistency still holds if we substitute the penalty parameter γ by a function taking uniformly positive values in Ω.

Remark 5. *The nonlinear form $b_h^\gamma(u_h; v_h)$ can also impose a constraint on the upper limit of the solution. For instance, if it is known that $u \in [u_{\min}, u_{\max}]$, $b_h^\gamma(u_h; v_h)$ can be written as:*

$$\begin{aligned} b_h^\gamma(u_h; v_h) := b_h(u_h, v_h) &+ \langle \gamma^{-1} \xi_\gamma^{\min}(u_h), v_h \rangle_h \\ &+ \langle \gamma^{-1} \xi_\gamma^{\max}(u_h), v_h \rangle_h, \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} \xi_\gamma^{\min}(u_h) &:= [(u_h - u_{\min}) - \gamma(A(u_h) - f)]_- \\ \xi_\gamma^{\max}(u_h) &:= [(u_{\max} - u_h) - \gamma(A(u_h) - f)]_- \end{aligned}$$

are penalty terms controlling the solution's lower and upper bounds, respectively.

Remark 6. *The election of the stabilization term (3.9) is motivated by the classical stabilization theory (SUPG, Ga-LS, VMS) for diffusive problems (see Codina [2000]), and for advective problems Burman & Ern [2017]. Naive elections of the stabilisation term, such as γ constant, affect the convergence of the discrete solution.*

3.4.1.1 Discrete norms

We definie a norm built from the inf-sup stable version of the $\|w\|_{\text{adr}}^2$ norm:

$$|||w|||^2 := \|w\|_{\text{adr}\sharp}^2 + C_i^{-2} h_T^2 \|A(w_h)\|_h^2, \qquad \forall T \in \mathscr{T}_h, \tag{3.14}$$

where C_i is independent of the mesh size.

Proposition 3 (Norm equivalence)**.** *There exists a constant $C_{\text{adr}} > 0$, such that:*

$$\left(1 - \frac{1}{1 + C_{\text{adr}}}\right) |||w_h|||^2 \leq \|w_h\|_{\text{adr}\sharp}^2 \leq |||w_h|||^2 \tag{3.15}$$

Proof. The upper bound is trivial by definition. For the lower bound, we recall the triangle inequality:

$$C_i^{-2} h_T^2 \|A(w_h)\|_h^2 \leq C_i^{-2} h_T^2 (\|K\Delta w_h\|_h^2 + \|\beta \cdot \nabla w_h\|_h^2 + \|\sigma w_h\|_h^2). \tag{3.16}$$

We bound the diffusion part in (3.16) using the following inverse inequality [Ern & Guermond, 2013],

$$\|\nabla w_h\|_T \leq C_i h_T^{-1} \|w_h\|_T, \forall \quad T \in \mathscr{T}_h, \tag{3.17}$$

we may then write

$$\|\Delta w_h\|_h^2 \leq C_i^2 h_T^{-2} \|\nabla w_h\|_h^2, \tag{3.18}$$

then it holds:

$$K^{-1} C_i^{-2} h_T^2 \|K\Delta w_h\|_h^2 \leq \|K^{\frac{1}{2}} \nabla w_h\|_h^2 \leq \|w_h\|_{\text{sip}}^2. \tag{3.19}$$

We bound the advection-reaction part of (3.16) recalling the Cauchy-Schwarz inequality, thus:

$$\begin{aligned} C_i^{-2} h_T^2 (C_\beta^{-1} \|\beta \cdot \nabla w_h\|_h^2 + C_\sigma^{-1} \|\sigma w_h\|_h^2) &\leq h_T \|\beta \cdot \nabla w_h\|_h^2 + \|\sigma\|_{L^\infty(\Omega)} \|w_h\|_{0,\Omega}^2 \\ &\leq \|w_h\|_{\text{up}\sharp}^2, \end{aligned} \tag{3.20}$$

where C_β and C_σ are constants independent of h.

Combining (3.19) and (3.20), we can establish that there exists $C_{\text{adr}} =$

$\min\{1, K^{-1}, C_\beta^{-1}, C_\sigma^{-1}\}$, such that

$$C_{\text{adr}} C_i^{-2} h_T^2 \|A(w_h)\|_h^2 \leq \|w_h\|_{\text{adr}\sharp}^2. \tag{3.21}$$

From (3.21), we can also establish a bound for the term associated with the operator $A(\cdot)$ in the triple norm as follows:

$$C_i^{-2} h_T^2 \|A(w_h)\|_h^2 \leq \frac{1}{1 + C_{\text{adr}}} |||w_h|||^2. \tag{3.22}$$

Adding $C_{\text{adr}} \|w_h\|_{\text{adr}\sharp}^2$ at each side of the inequality (3.21) and rearranging the terms, we finally obtain

$$\left(1 - \frac{1}{1 + C_{\text{adr}}}\right) |||w_h|||^2 \leq \|w_h\|_{\text{adr}\sharp}^2, \tag{3.23}$$

completing the proof. □

In order to establish the monotonicity properties of $b_h^\gamma(\cdot;\cdot)$, let us first define the following norm $\|w\|_{\text{adr}\natural}$ as

$$\|w\|_{\text{adr}\natural} := \|w\|_{0,\Omega}^2 + \frac{1}{2} \||\beta \cdot n|^{\frac{1}{2}} w\|_{0,\Gamma}^2 + \frac{1}{2} \sum_{F \in \mathscr{F}_h^i} \langle |\beta \cdot n_F| [\![w]\!], [\![w]\!] \rangle_{0,F} + \|w\|_{\text{sip}}^2. \tag{3.24}$$

We build the norm from $\|w_h\|_{\text{adr}\flat}$ preserving the L^2 part for cases where $\sigma = 0$.

Lemma 1 (Discrete coercivity)**.** *The bilinear form $b_h(v_h, v_h)$ is coercive respect to the norm $\|w\|_{adr\natural}$:*

$$b_h(v_h, v_h) \geq \|v_h\|_{adr\natural}^2, \quad \forall v_h \in V_h. \tag{3.25}$$

Proof. See Di Pietro & Ern [2012], Lemma 4.59. □

Proposition 4 (Upper bound of the norm of the operator)**.** *There exists a constant $C_\natural = \min\{1, K^{-1}, C_\eta^{-1}, C_\sigma^{-1}\}$, such that:*

$$C_\natural C_i^{-2} h_T^2 \|A(w_h)\|_h^2 \leq \|w_h\|_{\text{adr}\natural}^2, \tag{3.26}$$

Proof. The part of the norm associated to the diffusion, $\|w_h\|_{\text{sip}}$, is bounded

through (3.19). For the advective part, we recall the inverse inequality

$$C_i^{-2} h_T^2 \|\beta \cdot \nabla w_h\|_h^2 \leq C_\eta \|w_h\|_h^2 \leq C_\eta \|w_h\|_{\mathrm{up}\natural}^2, \tag{3.27}$$

and for the reactive part

$$C_i^{-2} h_T^2 \|\sigma w_h\|_h^2 \leq C_\sigma \|w_h\|_h^2 \leq C_\sigma \|w_h\|_{\mathrm{up}\natural}^2, \tag{3.28}$$

where C_η and C_σ are constants h independent. Combining these two, (3.26) is proofed. □

3.4.1.2 Monotonicity and well-posedness

Now, we establish that $b_h^\gamma(u_h; v_h)$ has reasonable monotonicity properties.

Lemma 2 (Monotonicity)**.** *Assume that*

$$0 \leq \gamma \leq C_i^{-2} h_T^2. \tag{3.29}$$

Then, the following holds for all $u_1, u_2 \in V$:

$$\frac{1}{2}\left(\|u_1 - u_2\|_{\mathrm{adr}\natural}^2 + \|\gamma^{-\frac{1}{2}}\left(\xi_\gamma(u_1) - \xi_\gamma(u_2)\right)\|_h^2\right) \leq b_h^\gamma(u_1; u_1 - u_2) - b_h^\gamma(u_2; u_1 - u_2) \tag{3.30}$$

$$\frac{1}{4}\left(\|u_1\|_{\mathrm{adr}\natural}^2 + \|\gamma^{-\frac{1}{2}}\xi_\gamma(u_1)\|_h^2\right) \leq b_h^\gamma(u_1; u_1 - u_2) + \|\gamma^{\frac{1}{2}} f\|_\Omega^2 \tag{3.31}$$

Proof. We first prove (3.30). We observe that

$$\begin{aligned} b_h^\gamma(u_1; u_1 - u_2) - b_h^\gamma(u_2; u_1 - u_2) = \\ = b_h(u_1 - u_2, u_1 - u_2) + \langle \gamma^{-1}(\xi_\gamma(u_1) - \xi_\gamma(u_2)), u_1 - u_2\rangle_h \\ \geq \|u_1 - u_2\|_{\mathrm{adr}\natural}^2 + \langle \gamma^{-1}(\xi_\gamma(u_1) - \xi_\gamma(u_2)), u_1 - u_2\rangle_h, \end{aligned}$$

where we have used (3.25). Moreover, using the fact that

$$|x_- - y_-|^2 \leq (x_- - y_-)(x - y), \quad \forall x, y \in \mathbb{R}, \tag{3.32}$$

we have

$$\begin{aligned}\langle\gamma^{-1}(\xi_\gamma(u_1)-\xi_\gamma(u_2)),u_1-u_2\rangle_h &= \\ &= \langle\gamma^{-1}(\xi_\gamma(u_1)-\xi_\gamma(u_2)),u_1-\gamma(A(u_1)-f\rangle_h - \langle u_2-\gamma(A(u_2)-f))_h \\ &\qquad + (\xi_\gamma(u_1)-\xi_\gamma(u_2),A(u_1-u_2)\rangle_h \\ &\geq \|\gamma^{-\frac{1}{2}}(\xi_\gamma(u_1)-\xi_\gamma(u_2))\|_h^2 + \langle\xi_\gamma(u_1)-\xi_\gamma(u_2),A(u_1-u_2)\rangle_h.\end{aligned}$$

Using Young's inequality, we infer that:

$$\begin{aligned}\langle\xi_\gamma(u_1)-\xi_\gamma(u_2),A(u_1-u_2)\rangle_h &\geq \\ &\geq -\frac{1}{2}\|\gamma^{-\frac{1}{2}}(\xi_\gamma(u_1)-\xi_\gamma(u_2))\|_h^2 - \frac{1}{2}\|\gamma^{\frac{1}{2}}A(u_1-u_2)\|_h^2. \quad (3.33)\end{aligned}$$

We can bound the second term of the right hand side in (3.33), associated to the operator $A(\cdot)$, using the assumption $\gamma \leq C_i^{-2}h^2$ and recalling (3.26). Thus, (3.30) holds true. Finally, the proof of (3.31) follows from (3.30) by taking $u_2 = 0$, and using the fact that $\|\gamma^{-\frac{1}{2}}(\xi_\gamma(u_1)-\xi_\gamma(0))\|_h^2 \leq \|\gamma^{-\frac{1}{2}}\xi_\gamma(u_1)\|_h^2 + \|\gamma^{-\frac{1}{2}}\xi_\gamma(0)\|_h^2 \leq \|\gamma^{-\frac{1}{2}}\xi_\gamma(u_1)\|_h^2 + \|\gamma^{\frac{1}{2}}f\|_h^2$. □

We can now prove that the discrete problem (3.12) is well-posed.

Proposition 5 (Well-posedness)**.** *Assume that γ satisfies (3.29). Then, the discrete problem (3.12) admits one and only one solution.*

Proof. Uniqueness is an immediate consequence of (3.30). If u_1 and u_2 are both solution to (3.12), then

$$b_h^\gamma(u_1;u_1-u_2) - b_h^\gamma(u_2;u_1-u_2) = 0 \quad (3.34)$$

and, from the left-hand side of (3.30), we conclude that $\|u_1-u_2\|_{\text{adr}\natural} = 0$ and hence $u_1 \equiv u_2$.

To prove existence, we use Brouwer's fixed point theorem (see, for instance, Temam [2001], Chapter 2, Lemma 1.4). Let $N := \dim V_h$, and let $G : \mathbb{R}^N \to \mathbb{R}^N$ be the map defined by $(G(U),V)_{\mathbb{R}^N} := b_h^\gamma(u_h;v_h) - \ell_h(v_h)$, where $U, V \in \mathbb{R}^N$ are the component vectors associated with the functions u_h, v_h in the Lagrange basis of V_h. Since Cauchy-Schwarz inequality implies that

$|\ell_h(v_h)| \leq \Lambda \|v_h\|_{\text{adr}\natural}$, with $\Lambda = (\|f\|_\Omega + \|K^{\frac{1}{2}}\nabla g\|_{\partial\Omega} + \||\beta\cdot n|^{\frac{1}{2}} g\|_{\Gamma_-})$, using (3.31) we conclude:

$$(G(U),U)_{\mathbb{R}^N} = b_h^\gamma(u_h;v_h) - \ell_h(v_h) \geq$$
$$\geq \frac{1}{4}(\|u_h\|^2_{\text{adr}\natural} + \|\gamma^{-\frac{1}{2}}\xi_\gamma(u_h)\|_h^2) - \|\gamma^{\frac{1}{2}} f\|_h^2 - \Lambda\|u_h\|_{\text{adr}\natural}.$$

Last proves that there is a real number, say Λ', such that $(G(U),U)_{\mathbb{R}^N} > 0$, for all $U \in \mathbb{R}^N$ with $\|U\|_{\mathbb{R}^N} \geq \Lambda'$. Indeed, using norm equivalence on discrete spaces, we infer that there exists $C_N > 0$, such that $C_N\|U\|_{\mathbb{R}^N} \leq \|u_h\|_{\text{adr}\natural}$, for all $U \in \mathbb{R}^N$ with associated discrete function $u_h \in V_h$. This statement leads to:

$$(G(U),U)_{\mathbb{R}^N} \geq \frac{1}{8}\|u_h\|^2_{\text{adr}\natural} - \|\gamma^{\frac{1}{2}} f\|_h^2 - 2\Lambda^2 \geq \frac{1}{8}C_N^2\|U\|^2_{\mathbb{R}^N} - \|\gamma^{\frac{1}{2}} f\|_h^2 - 2\Lambda^2.$$

Therefore, the expected inequality holds with

$$\Lambda' = \frac{\sqrt{8}}{C_N}\sqrt{\|\gamma^{\frac{1}{2}} f\|_h^2 + 2\Lambda^2 + 1}.$$

Existence is a direct consequence of well-known arguments (see, e.g., Temam [2001]). □

3.4.2 Penalty method using nonlinear V_h^*–FEM

We extend the discrete formulation to solve a nonlinear problem of the form: $N_h(u_h) = \ell_h$, where $N_h : U_h \to V_h^*$ represents the operator that includes the nonlinear penalty term, defined as $\langle N_h(z_h), v_h\rangle_{V_h^*\times V_h} := b_h^\gamma(z_h;v_h)$. Given that $b_h^\gamma(z_h;v_h)$ is built from the original bilinear form, the discrete problem (3.12) presents unique solution.

At the discrete level, we seek a minimizer $u_h \in U_h \subset V_h$ for the residual $\ell_h - N_h(z_h)$ associated to (3.12):

$$\begin{cases} \text{Find } u_h \in U_h \subset V_h, \text{ such that:} \\ u_h = \arg\min\limits_{z_h\in U_h} \dfrac{1}{2}\|\ell_h - N_h(z_h)\|^2_{V_h^*} \\ \quad = \arg\min\limits_{z_h\in U_h} \dfrac{1}{2}\|R_{V_h}^{-1}(\ell_h - N_h(z_h))\|^2_{V_h}, \end{cases} \tag{3.35}$$

Extending ideas of § 2.2.1, in particular (2.18), we solve the nonlinear problem

by finding critical points of the functional we minimise; this results in the following linearised problem:

$$\begin{cases} \text{Find } u_h \in U_h \subset V_h, \text{ such that:} \\ (R_{V_h}^{-1}(\ell_h - N_h(u_h)), R_{V_h}^{-1} DN_h(u_h; z_h)) = 0, \forall z_h \in U_h. \end{cases} \tag{3.36}$$

$DN_h : U_h \to V_h^*$ is defined as:

$$\langle DN_h(u_h; z_h), v_h \rangle_{V_h^* \times V_h} := db_h^\gamma(u_h; z_h, v_h), \tag{3.37}$$

where $db_h^\gamma(u_h; z_h, v_h)$ represents the derivative of the nonlinear form $b_h^\gamma(u_h; v_h)$ in the direction of an increment z_h:

$$db_h^\gamma(u_h; z_h, v_h) := \frac{d}{d\epsilon} b_h^\gamma(u_h + \epsilon z_h; v_h)\big|_{\epsilon=0}, \tag{3.38}$$

for instance, if we can impose a positivity-preserving condition through the penalty term, the derivative reads:

$$db_h^\gamma(u_h; z_h, v_h) := b_h(z_h, v_h) + \langle \frac{1}{\gamma} d\xi_\gamma(u_h; z_h), v_h \rangle_h \tag{3.39}$$

where $d\xi_\gamma(u_h; z_h) = \frac{1}{2}[1 - \text{sgn}(u_h - \gamma(Au_h - f))][z_h - \gamma A z_h]$. Hence, the modified discrete formulation reads:

$$\begin{cases} \text{Find } (\varepsilon_h, u_h) \in V_h \times U_h, \text{ such that:} \\ (\varepsilon_h, v_h)_{V_h} + b_h^\gamma(u_h; v_h) = \ell_h(v_h), \quad \forall v_h \in V_h, \\ \qquad db_h^\gamma(u_h; z_h, \varepsilon_h) = 0, \qquad \forall z_h \in U_h, \end{cases} \tag{3.40}$$

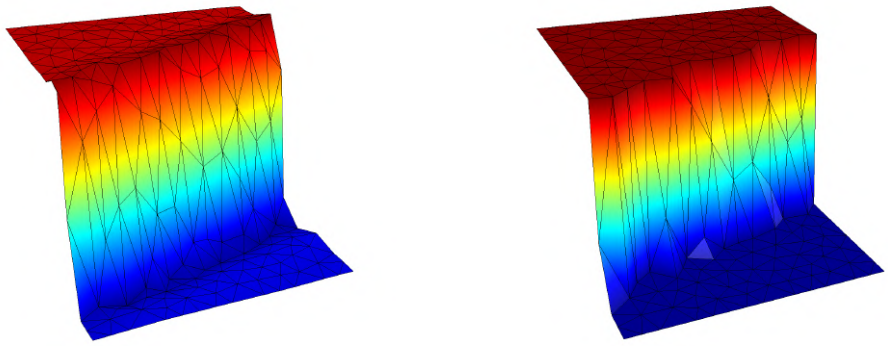
The first line of the system (3.40) represents the nonlinear problem to solve, whereas the second line defines the constraint subspace where we minimise the residual.

Remark 7. *In practice, solving* (3.40) *implies that a price in the energy norm may be paid to enforce the constraints since the residual minimisation method without penalty achieves the lowest possible variational residual for the linear problem (see Calo et al. [2020], Theorem 2).*

3.4.3 Numerical experiments

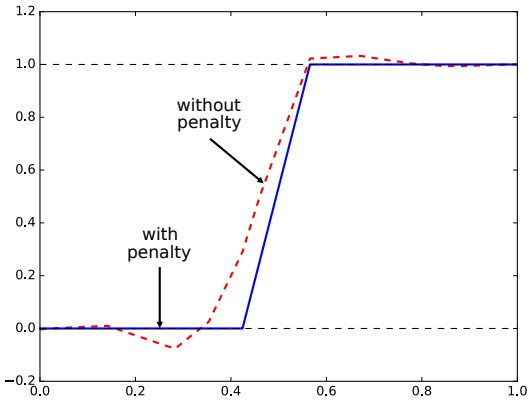
In this section, we implement the nonlinear constraint enforcement method to solve several numerical tests using FEniCS [Alnæs et al., 2015].

3.4.3.1 Advection problem over a quasi-uniform mesh



(a) Solution without penalty method.

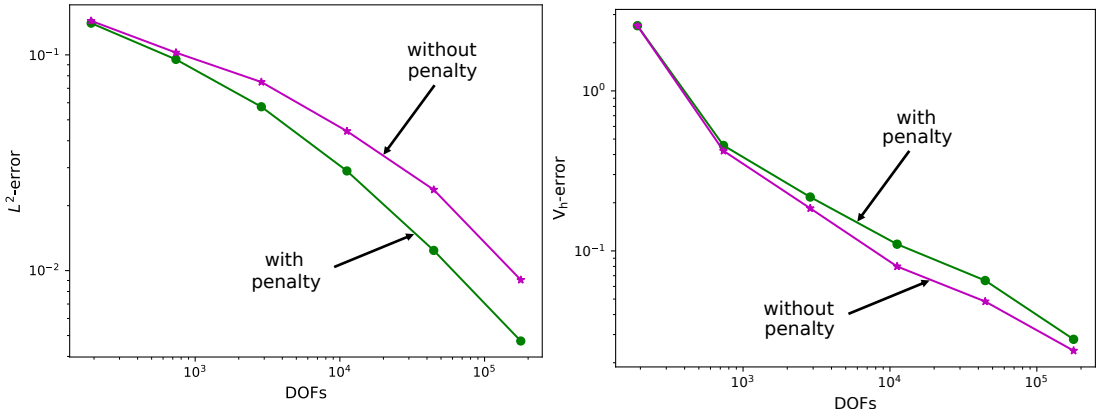
(b) Solution with penalty method.



(c) Cross section normal to β.

Figure 3.2: Advection problem over a quasi-uniform mesh.

We simulate a pure advection problem over a quasi-uniform mesh of size $h = 0.126$. We set $\Omega := (0,1) \times (0,1)$ and $\beta = (3/\sqrt{10}, 1/\sqrt{10})^T$, $K = 0$, $f = 0$. The unit advection field defines that Γ_- corresponds to the part where $xy = 0$. The exact solution is $u = \frac{1}{2}(\tanh((y - \frac{x}{3} - \frac{1}{4})/\epsilon) + 1.0)$, defining an inner layer in the solution of width ϵ. We compute solutions for a sharp layer ($\epsilon = 0.01$) using the stabilised method based on residual minimisation with the addition of the nonlinear penalty term. We consider linear ($p = 1$) finite elements. Given the source $f = 0$ and the boundary condition $0 \leq g \leq 1$, the solution is $0 \leq u \leq 1$.



(a) L_2 error norm vs DOFs.

(b) V_h error norm vs DOFs.

Figure 3.3: Convergence plots. Uniform refinement.

Thus, the penalty imposes both the lower and upper bounds. Using (3.9), we set $\gamma_0 = 10^{-5}$. We converge after 18 iterations using $TOL = 10^{-5}$.

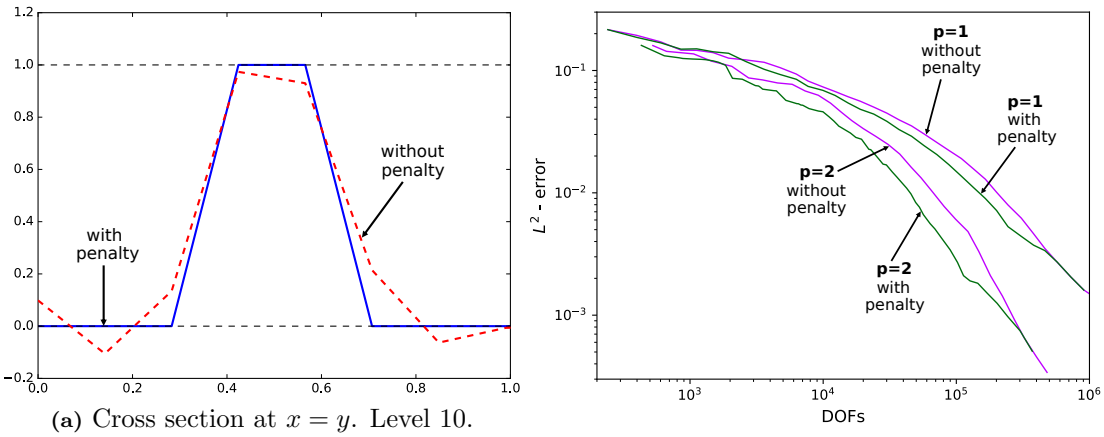
As Figures 3.2a & 3.2b show, the penalties consistently reduce the violation of the solution bounds up to the order of $10^{-3}\%$. Figure 3.2c displays a cross-section, normal to the advective field. The formulation with penalty significantly improves the bound preservation of the solution, removing the over- and undershoots that appear in the stabilised formulation. Finally, in Figures 3.3a and 3.3b, we show the L^2 and V_h-error norm convergence, respectively, considering a sequence of uniform meshes. We note that the constraint enforcement asymptotically produces worse convergence in the V_h-norm, which is in line with Remark 7, while surprisingly producing an improvement in the L^2-norm.

3.4.3.2 Rotating flow: adaptive mesh

We now solve a pure-advection test problem proposed by Kuzmin & Möller [2010]. Let $\Omega := (0,1) \times (-1,1)$ with $\boldsymbol{b} = (-y, x)^T$, $K = 0$, $f = 0$. The convection field rotates counterclockwise, and defines an inflow boundary equals to $\Gamma_- = (0,1) \times \{0\} \cup (0,1) \times \{1\} \cup \{1\} \times (0,1) \cup \{0\} \times (-1,0)$.

We set the boundary condition g as:

$$g = \begin{cases} 0.5\{1 + \tanh[\epsilon\,(y - 0.35)]\} & \text{on } (0, 0.5) \times \{0\}, \\ 0.5\{1 + \tanh[\epsilon\,(0.65 - y)]\} & \text{on } (0.5, 1) \times \{0\}, \\ 0 & \text{elsewhere on } \Gamma_-, \end{cases}$$



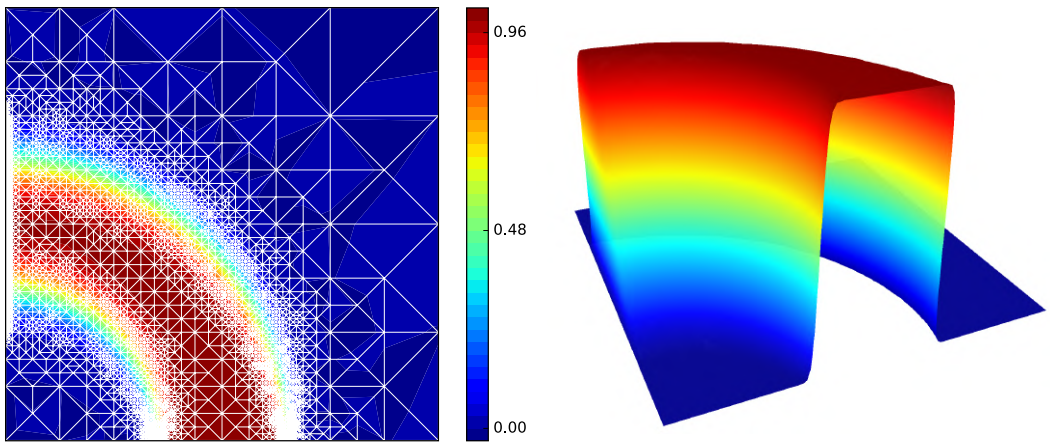
(a) Cross section at $x = y$. Level 10.

(b) Convergence plot. L^2 error norm vs DOFs.

Figure 3.4: Rotating flow over an adaptive mesh.

which produces an inner layer in the solution of width ϵ between 0.35 and 0.65. As in the previous case, we set $\epsilon = 0.01$. Figure 3.4a shows a cross-section with and without the inclusion of the penalty term. The bound penalty improves the constraint satisfaction and the inner layer slope. Figure 3.4b shows the convergence in L^2 and reflects a similar behaviour to the uniform mesh case, with the error norm for the penalty formulation solution higher than the one without penalty.

3.4.3.3 Advection-dominated diffusion problem: adaptive mesh



(a) Computational mesh and 2D solution.

(b) Solution in 3D. Level 25 (82k DOFs).

Figure 3.5: Advection-dominated diffusion problem (adaptive mesh).

We use the nonlinear penalty method to solve a version of the previous test with diffusion. All parameters as in § 3.4.3.2 except $K = 10^{-3}$. This modification

induces a boundary layer at $x = 0$ in the solution due to the contribution of the diffusion part. Our initial mesh is structured and has 4×4 triangular elements. We set $\gamma_0 = 10^{-4}$. Both trial and test functions are of degree $p = 1$—the penalty constrains the lower and upper bounds. Figure 3.5 shows that the adaptive scheme with the nonlinear penalty method captures the boundary layer through a proper error estimate, minimising the bound violation on each refinement level and, thus, delivering physically meaningful solutions at each level.

3.5 Nonlinear reaction problems: formulation

3.5.1 Bratu's equation

We extend our methodology to tackle the well-known Bratu's equation, a highly nonlinear problem with many applications such as radiative heat transfer, hydrodynamics, thermal-reaction processes, etc. In a two-dimensional context, for a $\Omega = (0, 1)^2$, the problem reads:

$$\begin{cases} \text{Find } u \text{ such that:} \\ \Delta u + \lambda \exp(u) = 0, \quad \text{in } \Omega, \\ \qquad\qquad u = 0, \quad \text{on } \partial\Omega, \end{cases} \tag{3.41}$$

where λ is a constant scalar. Although the original equation is highly nonlinear, the linearised structure of the problem has a reaction-diffusion structure. Thus, the dG formulation for this problem reads:

$$\begin{cases} \text{Find } u_h \in V_h, \text{ such that:} \\ n_h(u_h; v_h) = \ell_h(v_h), \quad \forall v_h \in V_h, \end{cases} \tag{3.42}$$

with the non-linear form $n_h(u_h; v_h)$ and the linear form $\ell_h(v_h)$ defined as:

$$\begin{aligned} n_h(u_h; v_h) := & \sum_{T \in \mathscr{P}_h} (\nabla u_h \,,\, \nabla v_h)_{0,T} - \sum_{T \in \mathscr{P}_h} (\lambda \exp(u_h) \,,\, v_h)_{0,T} \\ & + \sum_{F \in \mathscr{S}_h} \Big[(\llbracket u_h \rrbracket \,,\, \{\nabla v_h\} \cdot \mathbf{n}_F)_{0,F} - (\{\nabla u_h\} \cdot \mathbf{n}_F \,,\, \llbracket v_h \rrbracket)_{0,F} \Big] \\ & + \sum_{F \in \mathscr{S}_h} \gamma_F \, (\llbracket u_h \rrbracket, \llbracket v_h \rrbracket)_{0,F} \,, \end{aligned}$$

and

$$\ell_h(v_h) := \sum_{T\in\mathscr{P}_h} (f, v_h)_{0,T}.$$

Above, the constant γ_F uses the same definition given in (2.8). Besides, the discrete Frechét derivative is:

$$\begin{aligned} n_h'(u_h; z_h, v_h) &:= \frac{d}{d\epsilon} n_h(u_h + \epsilon z_h; v_h)\big|_{\epsilon=0} \\ &= \sum_{T\in\mathscr{P}_h} (\nabla z_h\,,\, \nabla v_h)_{0,T} + \sum_{T\in\mathscr{P}_h} (\lambda \exp(u_h)\, z_h\,,\, v_h)_{0,T} \\ &+ \sum_{F\in\mathscr{S}_h} \Big[(\llbracket z_h \rrbracket\,,\, \{\nabla v_h\}\cdot \mathbf{n}_F)_{0,F} - (\{\nabla z_h\}\cdot \mathbf{n}_F\,,\, \llbracket v_h \rrbracket)_{0,F} \Big] \\ &+ \sum_{F\in\mathscr{S}_h} \gamma_F \left(\llbracket z_h \rrbracket, \llbracket v_h \rrbracket\right)_{0,F}. \end{aligned} \tag{3.43}$$

We solve the linearised form (3.43) by interpreting the system as a reaction-diffusion problem in each increment z_h, for that reason, we provide the discrete space V_h with a diffusion-type norm:

$$\|w\|_{V_h}^2 := \|w\|_{0,\Omega}^2 + \|\nabla w\|_{0,\Omega}^2 + \sum_{F\in\mathscr{S}_h} (\gamma_F \llbracket w \rrbracket, \llbracket w \rrbracket)_{0,F}. \tag{3.44}$$

We use Newton's method to solve the nonlinear problem. Given the discrete solution pair (ε_h^i, u_h^i) of an iterative step i, we look for the increment $(\delta\varepsilon_h, \delta u_h)$ of the next iteration, and we set $u_h^{i+1} = u_h^i + t^i \delta u_h$, and $\varepsilon_h^{i+1} = \varepsilon_h^i + t^i \delta\varepsilon_h$, where t^i represents a relaxation parameter to control the increment size. For the $i+1$-th iteration, the linearisation reads:

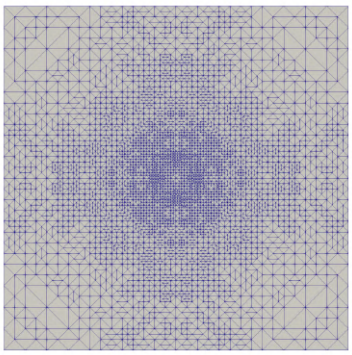
$$\begin{cases} \text{Given the pair } (\varepsilon_h^i, u_h^i),\ \text{find } (\delta\varepsilon_h, \delta u_h) \in V_h \times U_h,\ \text{such that:} \\ (\delta\varepsilon_h, v_h)_{V_h} + n_h'\left(u_h^i; \delta u_h, v_h\right) = \ell_h(v_h) - (\varepsilon_h^i, v_h)_{V_h} - n_h(u_h^i; v_h), & \forall v_h \in V_h, \\ n_h'\left(u_h^i; z_h, \delta\varepsilon_h\right) = -n_h'(u_h^i; z_h, \varepsilon_h^i), & \forall z_h \in U_h. \end{cases} \tag{3.45}$$

Similar to §3.3, in matrix form, formulation (3.45) reads [Cier et al., 2020]:

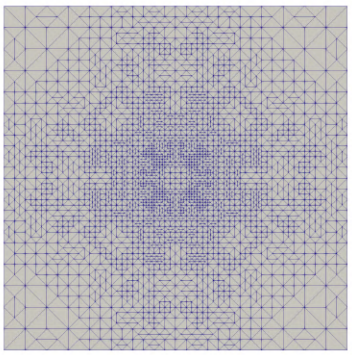
$$\begin{bmatrix} G & B_u \\ B_u^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta\varepsilon_h} \\ \boldsymbol{\delta u_h} \end{bmatrix} = \begin{bmatrix} \mathbf{L} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} G\boldsymbol{\varepsilon_h^i} + N(\boldsymbol{u_h^i}) \\ B_u^T \boldsymbol{\varepsilon_h^i} \end{bmatrix}. \tag{3.46}$$

Then, we use the nonlinear solver previously introduced in §3.3.

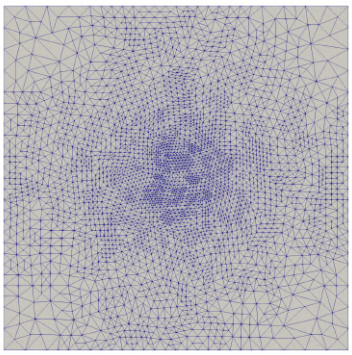
3.5.2 Numerical experiments



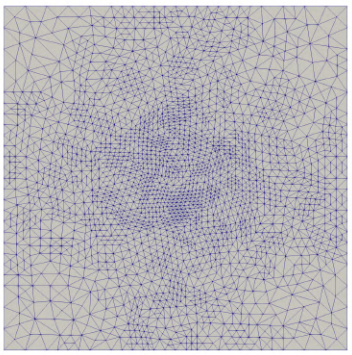
(a) Structured mesh: $\lambda = 1$ (DOFs = 65,889)



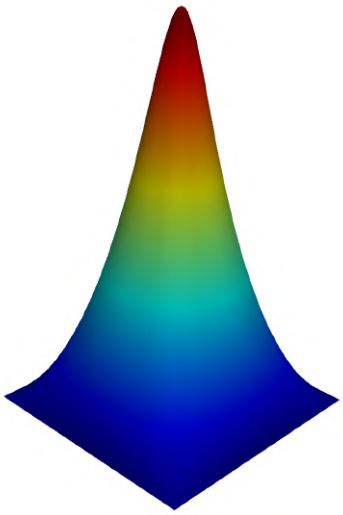
(b) Structured mesh: $\lambda = 2$ (DOFs = 43,609)



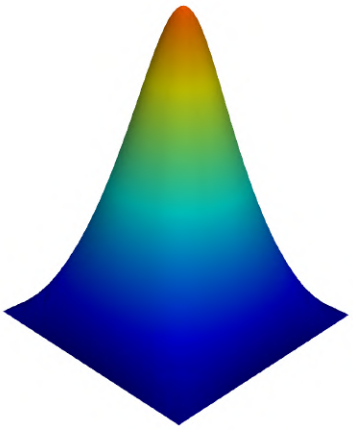
(c) Unstructured mesh: $\lambda = 1$ (DOFs = 61,918)



(d) Unstructured mesh: $\lambda = 2$ (DOFs = 42,462)



(e) Upper solution at $\lambda = 1$ ($u(0.5, 0.5) = 6.5489$)



(f) Upper solution at $\lambda = 2$ ($u(0.5, 0.5) = 5.0725$)

Figure 3.6: 2D Bratu's equation. Adaptive upper solutions for $\lambda = 1$ and $\lambda = 2$

The two-dimensional Bratu equation presents two solution branches for each

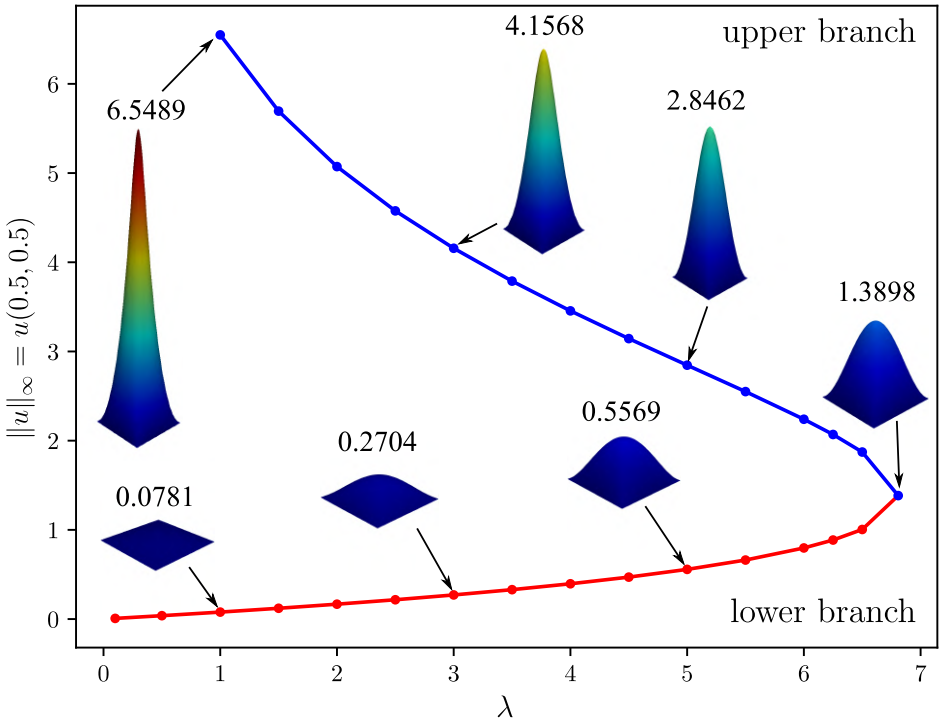


Figure 3.7: 2D Bratu's equation. Bifurcation map built using our method

value of $\lambda < \lambda_c = 6.808124423$ [Moore & Spence, 1980]. The efficient representation of this bifurcated nature still represents a numerical challenge, especially near the singular point at λ_c, where high resolution is needed; our method overcomes one of the issues: obtaining the upper and lower branches of the solution for a given λ. As example, we solve (3.41) using $u_L(x, y) = \lambda(x - x^2)(y - y^2)$ and $u_U(x, y) = \frac{50(2+\lambda)}{\lambda}(x - x^2)(y - y^2)$ for the lower and upper branches, respectively, following Hajipour et al. [2018]. We start from a 4×4 crossed-triangular mesh ($h = 0.25$) and polynomial degree $p = 2$. Figure 3.6 shows the solutions for upper cases with $\lambda = 1$ and 2, along with both the structured and unstructured adaptive meshes obtained when convergence is reached, measured as $TOL = \|\boldsymbol{R}^i\|_0 < 10^{-10}$. Additionally, Figure 3.7 shows the bifurcation map of discrete solutions, each of them obtained by separate with the method –no continuation method was used– for different values of λ in both branches, including the solution obtained for the critical value of $\lambda = \lambda_c = 6.808124423$. The robustness of our scheme allows automatic refinement depending on the solution's nature and the particular value of λ used, regardless of the initial mesh structure.

Chapter 4

Numerical study of one-dimensional compaction banding

The cnoidal wave approach in solids seeks to predict the formation of specific localised deformation bands more readily than alternative explanations provided by classical theories. While the cnoidal approach offers a new perspective to the localisation phenomenon, some points related to the solution of its governing equation need to be addressed before a detailed study is possible. Equation (1.2) has known analytical solutions only for the integer values of $m = 1, 2, 3$. However, solutions for higher or non-integer values of m need numerical treatment, and, to date, there has been no successful attempt to solve this equation satisfactorily numerically. The lack of a robust numerical tool for exploring the family of solutions for this equation is related to the complexity of the treatment of this class of nonlinear problems. In this chapter[1], we seek to overcome this issue from the numerical perspective by developing a consistent discrete solution using the proposed nonlinear extension of V_h^*–FEM.

[1]The content of this chapter is published in: Cier, R. J., Poulet, T., Rojas, S., Veveakis, M., & Calo, V. M. (2021). Automatically adaptive stabilised finite elements and continuation analysis for compaction banding in geomaterials. *International Journal for Numerical Methods in Engineering*, 122(21), 6234-6252.

4.1 Mathematical nature of the equation

The governing equation of the localisation phenomenon, also known as *cnoidal equation* is a nonlinear reaction-diffusion equation. In its one-dimensional reduced form, the equation is similar to a quasilinear heat equation [Galaktionov & Vazquez, 1995]. From a practical point of view, we seek to determine the specific set of conditions, both in the initial conditions and the parameters, to produce instabilities in the solution. In mathematics, these instabilities are known as singularities, and the finite time they occur is called blow-up time. The analysis of this type of behaviour is a specific topic in numerical analysis, and many references can be found [Bandle & Brunner, 1998; Hu, 2011]. Blow-up phenomena are treated case by case, which means that their onset conditions need to be understood before being applied to a new one. Within the field of quasilinear heat equations, we can find insights from equations with nonlinear absorption terms:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 \phi(u)}{\partial z^2} - f(u),$$

where $f(u)$ represents the nonlinear absorption term that competes with the diffusive part expressed by a function $\phi(u)$. This type of equation, however, is being studied in detail only for conditions leading to extinction processes (where $u \equiv 0$) [Galaktionov & Vazquez, 1995; Galaktionov & Vázquez, 2002]. Thus, for the blow-up phenomenon, the problem remains open to a comprehensive analysis that allows us to determine the necessary and sufficient conditions for the onset of singularities in the solution.

Since this work focuses on the numerical simulation of the mechanical process implied by this equation, we develop a numerical approximation that allows us to analyse the onset conditions of localisation phenomena, noting that future work may address a rigorous mathematical analysis.

4.2 Numerical simulation

This problem is still open; nevertheless, we can look for numerical solutions to better understand the equation's nonlinearity behaviour. Therefore, we use the new adaptive stabilised finite element method based on residual minimisation, developed by Calo et al. [2020], and extend its application to this specific nonlinear problem. We use this formulation instead of alternative FEM approximations

because of its stability properties and built-in adaptive mesh refinements. These features represent a crucial aspect for analysing the localisation of instabilities.

4.2.1 Weak variational formulation

We state the variational formulation for the cnoidal equation as stated by Alevizos et al. [2017]. We set $\Omega = [0, 1]$, with boundary $\partial\Omega = \{0, 1\}$. Following Veveakis & Regenauer-Lieb [2015], we define the boundary conditions as $\sigma' = 1$ on $\partial\Omega$. To derive the continuous formulation of (1.2), we split $\sigma' = u + 1$, and the steady state equation then reads:

$$\begin{cases} \text{Find } u = \sigma' - 1 \text{ such that:} \\ \Delta u - \mathcal{F}(u) = 0 \quad \text{in } \Omega, \\ \qquad\qquad u = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{4.1}$$

where Δ represents the Laplacian operator and $\mathcal{F}(u) = \lambda\,(1+u)^m - \mu \exp(\beta u)$. In (4.1), $\mu \exp(\beta u)$ is equivalent to the regularization term $N(\sigma')$ from (1.2), which avoids the unbounded stress growth of the exponent $m > 1$ capping the value of u to finite values and making $\mathcal{F}(u)$ to remain positive.

Multiplying (4.1) by a test function v and integrating by parts, we obtain the following weak variational formulation:

$$\begin{cases} \text{Find } u \in H_0^1(\Omega), \text{ such that:} \\ n(u; v) = \ell(v), \quad \forall v \in H_0^1(\Omega), \end{cases} \tag{4.2}$$

with $n(u; v) = (\nabla u, \nabla v)_{0,\Omega} + (\mathcal{F}(u), v)_{0,\Omega}$ and $\ell(v) = (f, v)_{0,\Omega}$, where $(\cdot, \cdot)_{0,\Omega}$ represents the L^2 scalar product in Ω.

4.2.1.1 Nonlinear discontinuous Galerkin formulation

Considering the well-known dG discrete setting, we build the dG formulation for the continuous weak variational formulation of (4.2) as

$$\begin{cases} \text{Find } u_h \in V_h, \text{ such that:} \\ n_h(u_h; v_h) = \ell_h(v_h), \quad \forall v_h \in V_h, \end{cases} \tag{4.3}$$

with

$$
\begin{aligned}
n_h(u_h; v_h) := & \sum_{T\in\mathscr{P}_h} (\nabla u_h \,,\, \nabla v_h)_{0,T} + \sum_{0,T\in\mathscr{P}_h} (\mathcal{F}(u_h) \,,\, v_h)_{0,T} \\
& + \sum_{F\in\mathscr{S}_h} \Big[(\llbracket u_h \rrbracket \,,\, \{\nabla v_h\}\cdot \mathbf{n}_F)_{0,F} - (\{\nabla u_h\}\cdot \mathbf{n}_{0,F} \,,\, \llbracket v_h \rrbracket)_{0,F} \Big] \\
& + \sum_{F\in\mathscr{S}_h} \left[\frac{\gamma}{h_F} (\llbracket u_h \rrbracket, \llbracket v_h \rrbracket)_{0,F} \right],
\end{aligned}
$$

and

$$
\ell_h(v_h) := \sum_{T\in\mathscr{P}_h} (f, v_h)_{0,T}.
$$

In the above, $\gamma > 0$ is a user-defined constant that we set as $\gamma = 3(k+1)(k+2)$, being k the polynomial degree of the test space. Besides, we recall (3.1) and set the discrete derivative as:

$$
\begin{aligned}
n_h'(u_h; z_h, v_h) := & \frac{d}{d\epsilon} n_h(u_h + \epsilon z_h; v_h)\big|_{\epsilon=0} \\
= & \sum_{T\in\mathscr{P}_h} (\nabla z_h \,,\, \nabla v_h)_{0,T} + \sum_{T\in\mathscr{P}_h} (\mathcal{F}'(u_h)\, z_h \,,\, v_h)_{0,T} \\
& + \sum_{F\in\mathscr{S}_h} \Big[(\llbracket z_h \rrbracket \,,\, \{\nabla v_h\}\cdot \mathbf{n}_F)_{0,F} - (\{\nabla z_h\}\cdot \mathbf{n}_F \,,\, \llbracket v_h \rrbracket)_{0,F} \Big] \\
& + \sum_{F\in\mathscr{S}_h} \frac{\gamma}{h_F} (\llbracket z_h \rrbracket, \llbracket v_h \rrbracket)_{0,F} \,,
\end{aligned}
\tag{4.4}
$$

where

$$
\mathcal{F}'(u_h) = \lambda\, m\, (1+u_h)^{m-1} - \mu\, \beta \exp(\beta u_h). \tag{4.5}
$$

We build our resolution scheme using (4.4). This linearised form can be seen as a reaction-diffusion form in each increment z_h; for that reason, we provide the discrete space V_h with a diffusion-type norm:

$$
\|w\|_{V_h}^2 := \theta\ \|w\|_{0,\Omega}^2 + \|\nabla w\|_{0,\Omega}^2 + \sum_{F\in\mathscr{S}_h} \left(\frac{\gamma}{h_F} \llbracket w \rrbracket, \llbracket w \rrbracket \right)_{0,F}. \tag{4.6}
$$

with $\theta = \lambda\, m\, A^{m-1}$, where $A > 0$ is a given constant associated with the maximum value of the normalised stress in the cnoidal solution.

Equivalently, we write the problem as the following saddle-point problem:

$$\begin{cases} \text{Find } (\varepsilon_h, u_h) \in V_h \times U_h, \text{ such that:} \\ (\varepsilon_h, v_h)_{V_h} + n_h(u_h; v_h) = \ell_h(v_h), \quad \forall v_h \in V_h, \\ \qquad\qquad n_h'(u_h; z_h, \varepsilon_h) = 0, \qquad \forall z_h \in U_h. \end{cases} \tag{4.7}$$

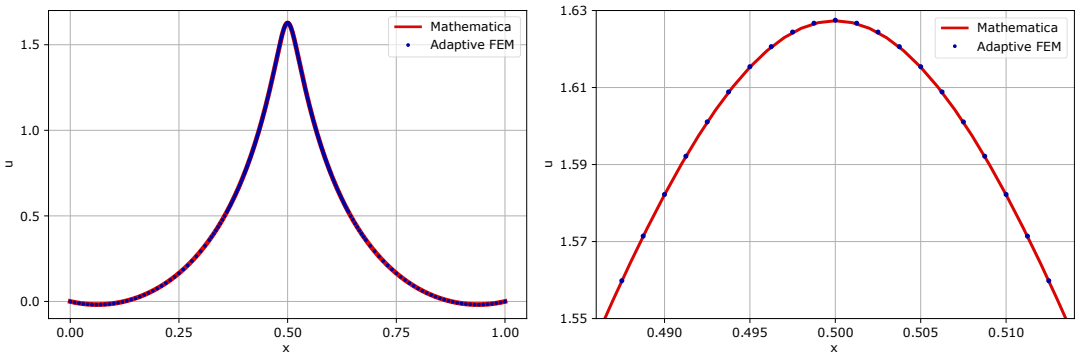
The system (4.7) delivers simultaneously a stable and continuous approximation $u_h \in U_h$ of the dG formulation and a residual representation $\varepsilon_h \in V_h$ that guides the adaptive mesh refinement. The first line represents the residual projection of the nonlinear problem, whereas the second line imposes the constraint on the tangent space built from the linearised form.

4.3 Numerical tests

In this section, we describe several 1D numerical examples that illustrate the performance of the adaptive stabilised finite element method in the context of the cnoidal equation. We use FEniCS [Alnæs et al., 2015] to produce the simulation results. The main drawback of standard FEM implementations for this kind of problem lies in resolving the localised peak and finding their location, which delivers low-quality solutions. These limitations severely restrict their usage, forcing the initial guess to be close enough to the solution for the algorithm to converge, which is impractical. To overcome these limitations in a nonlinear framework, we seek an algorithm that automatically finds the peaks' locations. In practice, we can start from an arbitrary initial trial solution with peaks far from the final configuration. The numerical examples demonstrate that V_h^*–FEM can easily overcome these issues. This section develops numerical examples using a range of possible rock parameters, although it does not replicate any former experimental work. A validation of the cnoidal approach, using realistic rock parameters, was developed in the primal theoretical work [Veveakis & Regenauer-Lieb, 2015, §6].

4.3.1 Single peak solution

The framework's enhanced stability allows us to solve (4.1) and retrieves the expected peak solution for appropriate parameters. Figure 4.1 compares the semi-analytical solution computed with Mathematica [Wolfram et al., 1999] and



(a) Whole solution. **(b)** Zoom around $x = 0.5$.

Figure 4.1: Semi-analytic procedure (Mathematica) vs adaptive framework. Comparison shows good match between results for $\lambda = 10$, starting from the initial guess $u_{IG} = 2\exp(-100(x-0.5)^2)$.

the results obtained with our approach for $\lambda = 10$, $m = 3$, $\mu = 10^{-4}$ and $\beta = 10$, starting from an initial guess $u_{IG} = 2\exp(-100(x-0.5)^2)$ on a regular mesh of 100 nodes, getting to 273 nodes after four levels of adaptivity. We observe an excellent match, including the peak location, shape, and intensity, as shown in Figure 4.1b.

4.3.2 Multiple peak solution

We can recover known semi-analytic solutions with the method, so we explore a more challenging problem considering more peaks. In this context, standard FEM is not appropriate due to its lack of stability. For the following numerical examples, we define

$$u_{IG}(x) := A_0 \left[\frac{\exp(-1250(x-x_0)^2)}{\sin(x_0\pi)} + \frac{\exp(-1250(x-(1-x_0))^2)}{\sin((1-x_0)\pi)} \right] \sin(x\pi) \tag{4.8}$$

as the initial guess function, being x_0 the arbitrary location of the first peak ($0 \leq x_0 \leq 0.5$). This initial guess choice implies that the second peak location is at $1 - x_0$. Besides, $A_0 > 0$ is an arbitrary number that coincides with the values of u at the peak locations in the initial guess. Table 4.1 shows the set parameters for the examples in this subsection.

Table 4.1: Parameters for 1D numerical examples with two-peak solution

Example	λ	m	μ	β	A_0	x_0
3.2.1	40	3	10^{-4}	10	1.80	0.200
3.2.2	40	3	10^{-4}	10	1.80	0.175
3.2.3	40	3	10^{-4}	10	1.80	0.425
3.2.4	40	π	10^{-4}	10	2.70	0.350

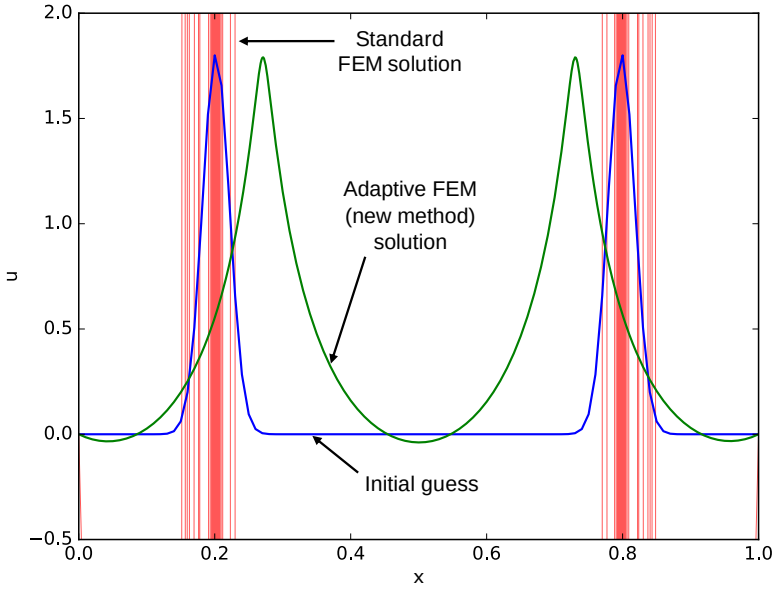


Figure 4.2: Standard finite element solution vs adaptive stabilised method for initial guess with two misplaced peaks. The stabilised method converges, whereas the standard FEM approach leads to spurious oscillations.

4.3.2.1 Comparison against standard FEM

As a first example, Figure 4.2 shows the results comparing the discrete solution obtained with the standard FEM formulation and the new adaptive stabilised method for the same initial guess. For this two-peak example, we set the arbitrary location as $x_0 = 0.2$ (see Table 4.1) that defines

$$u_{IG} := 1.8 \left[\frac{\exp(-1250(x-0.2)^2)}{\sin(0.2\pi)} + \frac{\exp(-1250(x-0.8)^2)}{\sin(0.8\pi)} \right] \sin(x\pi)$$

as the initial guess. We use cubic trial functions ($\mathbb{P}_3$) for both methods. Still, we take advantage of the possibility of enriching the test space in the adaptive stabilised method, using test functions one degree higher ($\mathbb{P}_4$). Finally, we use a fixed mesh for the standard finite element solution of mesh size of $h = 10^{-6}$,

to develop a fair comparison with the final refined mesh obtained through the adaptive method, which starts from a mesh size of 100 elements ($h = 0.01$) and gets no finer than $h = 10^{-6}$ locally. We can observe that this new technique properly captures the peaks' final locations at $x \approx \{0.27, 0.73\}$, whereas the standard method delivers spurious oscillations.

4.3.2.2 Initial guess close to the boundaries

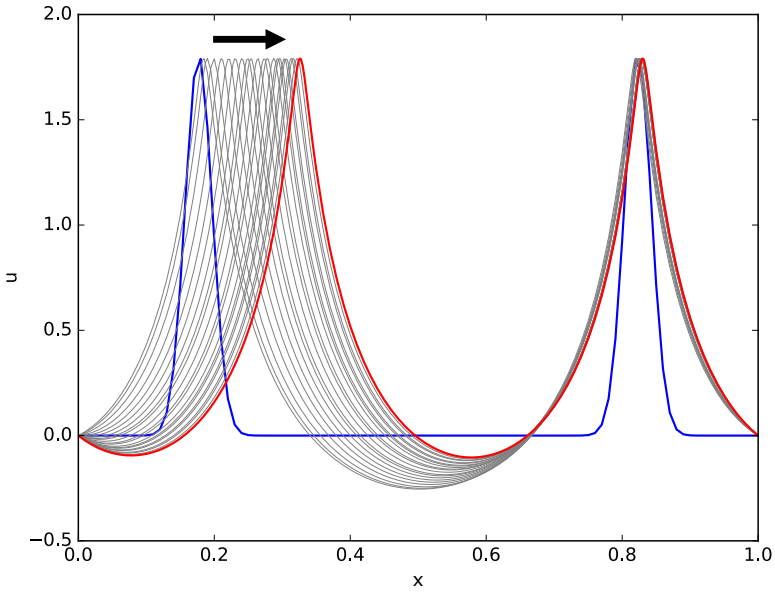


Figure 4.3: Profile evolution at intermediate refinement steps: Initial guess at $x_0 = 0.175$

We now investigate examples with different conditions to show that the new method converges robustly with respect to the initial condition. The distance between the initial and final peak locations using the adaptive method is significantly larger than standard FEM on a fine mesh. In this example, we locate the peaks close to the boundaries (see Table 4.1). Figure 4.3 shows the iterative solutions obtained at each refinement step for an initial guess using $x_0 = 0.175$ (in blue) showing convergence to an appropriate (asymmetrical) solution (in red). The adaptive method converges when standard FEM fails, even with an order of magnitude finer mesh ($h = 10^{-7}$). The adaptive approach corrects the peak locations at each refinement level, starting from a mesh size of $h = 0.01$. After 47 refinement levels (approximately, $24,000$ iterations), we obtain a solution with a

final residual norm $\|\boldsymbol{R}^{i+1}\|_0 < 10^{-9}$. Incidentally, this example also shows that the solution can be asymmetric.

4.3.2.3 Initial guess close to the centre

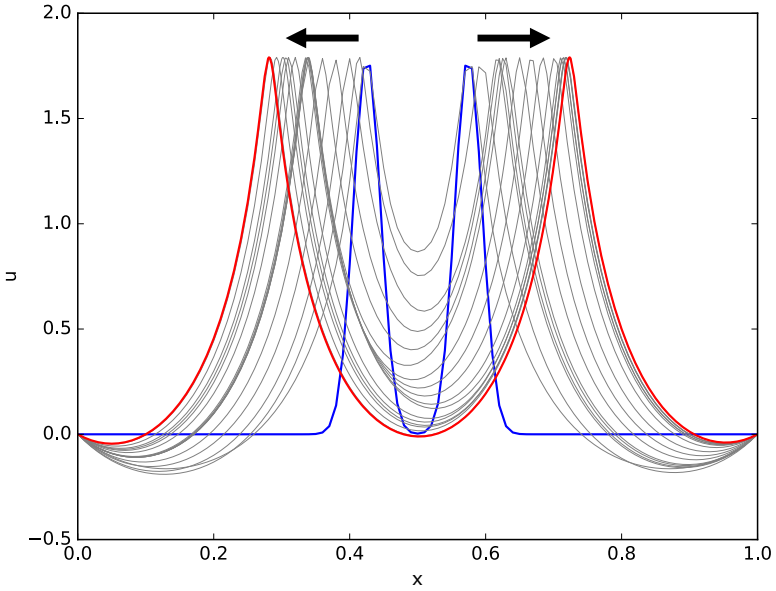


Figure 4.4: Profile evolution at intermediate refinement steps: Initial guess at $x_0 = 0.425$

In this example, we locate the initial peaks close to the centre (see Table 4.1). In that instance, the converges to the solution of Figure 4.4 using the adaptive stabilised method after 37 refinement levels (approximately $7,400$ iterations) for the same tolerance in the previous case. The final peak locations are $x \approx \{0.27, 0.73\}$ for an initial guess using $x_0 = 0.425$ (in blue), showing convergence to an appropriate (symmetrical) solution (in red). FEM is not able to converge using this initial guess either. From our experience, which we do not report for brevity, we find that FEM simulations require initial guesses sufficiently close to the final solution for the method to converge. In practice, FEM requires the distance between the initial and the final solution peak to be at least an order of magnitude smaller than the adaptive stabilised method admits.

4.3.2.4 Non-integer exponent

Finally, we simulate a scenario with a non-integer exponent $m = \pi$ to show the robustness of the adaptive stabilised method for an irrational exponent. As we

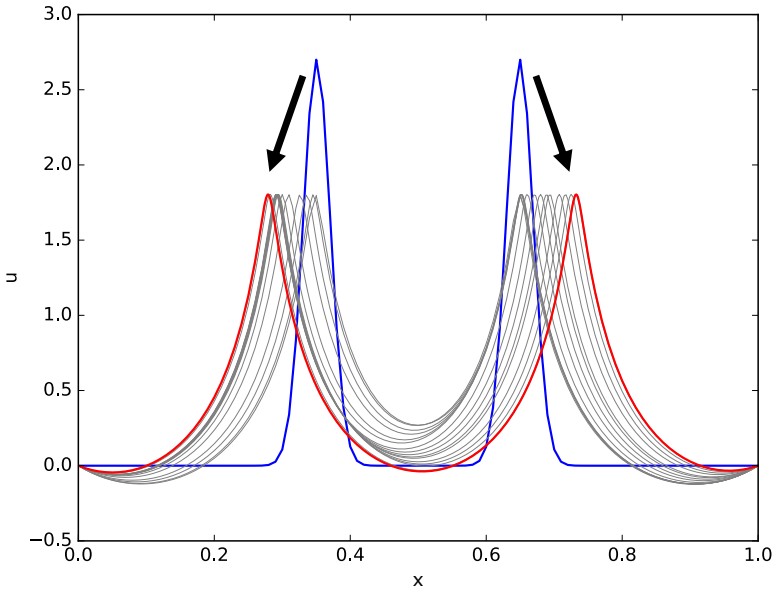


Figure 4.5: Profile evolution at intermediate refinement steps: for $m = \pi$ & initial guess with peaks $u_0(0.35) = u_0(0.65) = 2.7$

mentioned earlier, the analytical approach of the cnoidal equation does not provide solutions for this class of exponents. We consider the initial guess (4.8) with $x_0 = 0.35$ and we set higher peaks values than in past examples (see Table 4.1). Figure 4.5 displays the evolution of the solution profile, showing that the method can robustly simulate irrational exponents larger than 3, a limitation of the analytical resolution approach [Veveakis & Regenauer-Lieb, 2015]. Also, the adaptive stabilised scheme corrects the height of the peak values. Numerical simulations not presented here also showed a good performance for even higher exponent values up to 7.

4.4 Periodic conditions in the cnoidal equation

Although the original work of Veveakis & Regenauer-Lieb [2015] considers Dirichlet boundary conditions in the one-dimensional governing equation, we consider that, given the wave-mechanics nature of the expected solutions, periodic boundary conditions allow us to study the influence of the material parameters in the onset and distribution of the compaction banding phenomenon. We modify the original V_h^*–FEM formulation in this section to consider periodic boundary con-

ditions. Then we study the influence of the diffusivity ratio, λ, in the governing equation's solution behaviour.

4.4.1 Modification to the original V_h^*–FEM formulation

Following Vemaganti [2007], we implement periodic boundary conditions that constrain our test space to link our domain's left and right-hand edges. For this, we construct a `PeriodicBoundary` class that restricts the approximation space V_h. Thus, we could state the space with periodic boundary conditions V_h^{per} as:

$$V_h^{per} := \{v_h \in L^2(\Omega) \mid \forall\, T \in \mathscr{P}_h, v_h|_T \in \mathbb{P}^p(T) \ \wedge \ v_h(0) = v_h(1)\},$$

As an example of the versatility of using periodic boundary conditions for the cnoidal equation, we display the asymmetrical outcome from an initial guess with the difference in amplitude in the peaks in Figure 4.6. In this picture, the intermediate steps show the sequence of splitting of the higher-amplitude peak, whereas the lower one vanishes, which generates the asymmetry. Under typical Dirichlet boundary conditions, this approximation is equally reproducible; however, it becomes computationally expensive because it represents a more restrictive scenario. Moreover, the behaviour implies that the cnoidal equation with periodic boundary conditions presents a threshold for the localised concentrations. Hence, we use this feature to analyse the influence of the diffusivity ratio λ in the response using this new set of boundary conditions.

4.4.2 Influence of the diffusivity ratio λ

Using periodic boundary conditions for the cnoidal problem, we study the influence of the diffusivity ratio λ in our solution's evolution of the number of peaks. We start from the same initial guess used in §4.3.2.1, and we vary the value of λ. Figure 4.7 shows the increase of the number of peaks as the diffusivity ratio increases, which is qualitatively consistent with the theoretical framework. As stated by Veveakis & Regenauer-Lieb [2015], for scenarios of low diffusivity ratio (in our case, shown by $\lambda = 5$), the loading rate is slower than the mass diffusion rate, and the specimen has time enough to diffuse away any pressure variations induced by the loading conditions, resulting in homogeneous deformation. Contrary to that, in high-diffusivity scenarios (in our example, $\lambda = 20, 100$), the

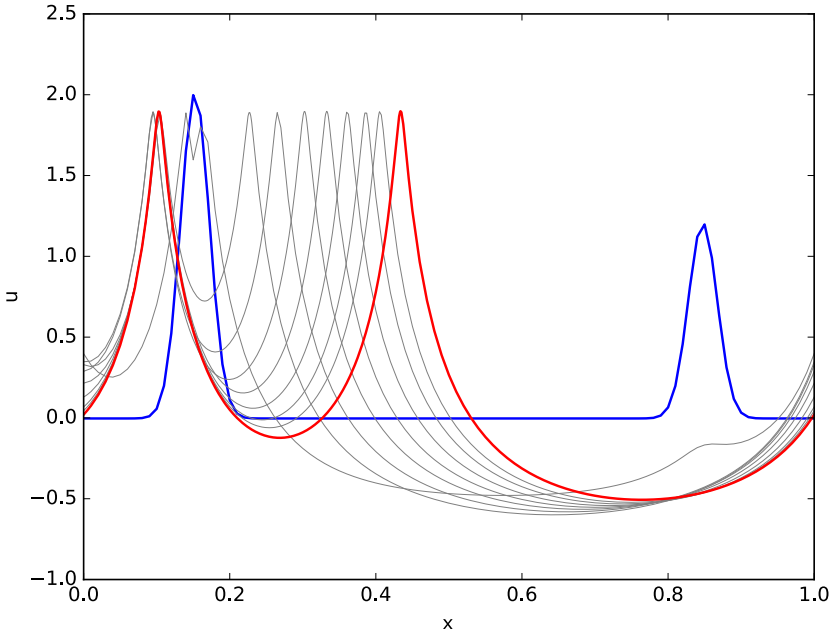


Figure 4.6: Profile evolution at intermediate refinement steps for periodic boundary conditions: Initial guess at $x = \{0.15, 0.85\}$

loading rate is faster than the mass diffusion rate, which localises.

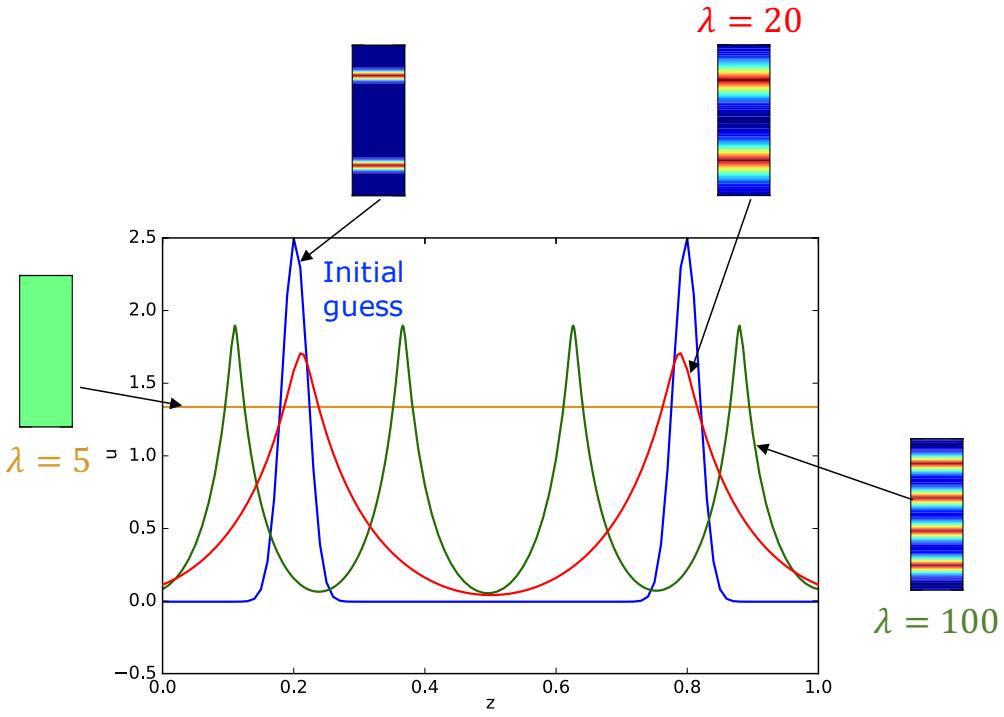


Figure 4.7: Peak number evolution as λ increases for periodic boundary conditions

4.5 Advantages of V_h^*–FEM in cnoidal problem

In summary, V_h^*–FEM can robustly and efficiently predict the solution patterns for many relevant configurations for the cnoidal wave problem. This seemingly insurmountable problem is now solvable since our technique automatically enriches the test space, improving the approximation and giving better solution behaviour at each level. Also, the adaptive mesh refinement scheme eliminates oscillations by reducing the local error. This robust adaptivity represents an essential feature of the method, primarily because of the solution's localised nature. Using this adaptive stabilised method, we can build an indicator from the residual representative to correct the peak location at each refinement level.

Regarding the chemical effects in the physical formulation, following the work from Alevizos et al. [2017] provides a bounded solution of the original problem of cnoidal waves in solids [Veveakis & Regenauer-Lieb, 2015; Regenauer-Lieb et al., 2013]. We solve the resulting nonlinear equation using an adaptive stabilised finite element framework [Calo et al., 2020] for a wide spectrum of scenarios. Beyond the validation of the cnoidal approach, using realistic rock parameters developed in the original theoretical work [Veveakis & Regenauer-Lieb, 2015, §6], our results open the door to testable hypotheses for laboratory experiments, based on the definition of λ (see Appendix A).

Remark 8. *As previously explained, the cnoidal theory represents a completely novel approach to explain, from a mathematical point of view, the occurrence of the compaction banding phenomenon usually attributed to a different set of features. According to the original theory by Veveakis & Regenauer-Lieb [2015], localisation in the form of compaction bands is linked to the interplay between loading rates and internal mass transfer processes occurring in porous media. Their occurrence is not induced by specific nucleation points or weak zones within the domain but by the mathematical response of the governing equation under a given set of parameters representing the occurring diffusion process. Although some aspects of the physical explanation of the derivation require further special attention, the original approach represented a complex enough problem to motivate a series of developments in the line of the numerics. Hence, further work should be focused on reconciling typical mechanical approaches with the wave-mechanics cnoidal approach and developing a careful review of the introduced numerical techniques and their meaning in the general localisation context.*

In the future, one of the aspects that will require further attention is to clarify that the adaptive mesh refinement algorithm is not self-biasing when the physical problem is inherently prone to instability. Indications that this is not occurring can be inferred from the results for multiple peak scenarios, where the peaks correct through the refinement process to reduce the residual error of the discretised version of the nonlinear problem.

Based on the outcomes of this first part, we extend to higher-dimensional hydro-mechanical simulations. This extension is a complex task in higher dimensions because it indirectly involves resolving the numerical instabilities. The blow-up term is hidden in the plastic behaviour instead of directly appearing in the system of equations. Nonetheless, the rich information that the appropriate numerical scheme delivers to a simple generalisation of the consolidation theory in this stage suggests that using this scheme in more elaborate elasto-viscoplastic formulations could enhance the mechanical solution with additional modes of localisation stemming from the volumetric part of the plastic increment.

Chapter 5

Analytical and numerical study of compaction banding phenomenon

As stated in Chapter 1, the current state of the art for the study of compaction banding presents a series of limitations from the mechanical point of view. In this chapter[1], we propose a new analytical-numerical analysis of this phenomenon based on a consistent axiomatic formulation. We build our theoretical framework with the minimum amount of ingredients needed for a viscoplastic model. We base our model on six principles that allow deriving new versions from studying compaction band localisation triggered by viscous effects. Then, we study different stress states analytically to prove the conditions for finding compaction bands.

Furthermore, motivated by different laboratory investigations that show evidence of the appearance of this phenomenon in porous rocks, we develop a series of numerical experiments to reproduce this phenomenon under triaxial compression conditions. For this, we use a version of the constitutive model that accounts for creep based on Perzyna's viscoplasticity. The obtained results confirm the transitional effect of the confinement pressure reported in the literature and open the discussion for analysing the periodicity and spacing of the bands and their dependence on the material parameters.

[1]The content of this chapter is published in Cier, R. J., Labanda, N.A., & Calo, V. M. (2022). Compaction band localisation in geomaterials: a mechanically consistent failure criterion. Submitted to *International Journal for Numerical and Analytical Methods in Geomechanics*

5.1 Theoretical framework

In this section, we use Cambridge's notation for stress invariants [Roscoe et al., 1958]. That is, for principal effective stresses, the mean and the deviatoric stresses are:

$$\begin{aligned} p' &= \frac{\sigma'_{11}+\sigma'_{22}+\sigma'_{33}}{3} \quad , \\ q &= \frac{1}{\sqrt{2}}\sqrt{(\sigma'_{11}-\sigma'_{22})^2+(\sigma'_{22}-\sigma'_{33})^2+(\sigma'_{22}-\sigma'_{33})^2+\sigma'^2_{12}+\sigma'^2_{23}+\sigma'^2_{31}}. \end{aligned} \tag{5.1}$$

Moreover, we assume the samples undergo straight stress paths; we consider the pair (p, q) follows a known incremental stress ratio η, such that $q = \eta(p - p_r)$, where p_r is the reference mean stress. For instance, the shearing stage of an isotropically consolidated drained compression (CIDC) triaxial test in geomaterials considers $p_r = \sigma'_3$ and $\eta = 3$, whereas in an isotropic compression test $p_r = 0$ and $\eta = 0$ (see Figure 5.1 for a sketch of these ideas). This assumption allows us to state the problem exclusively regarding the mean stresses and the incremental stress ratio.

5.1.1 Model axiomatic statement

In this work, we propose a framework that can consistently formulate any visco-elastoplastic model by specifying the following *six* features:

(i) An elastic constitutive behaviour:

$$\sigma'_{ij} = \frac{\partial \psi^e\left(\varepsilon^e\right)}{\partial \varepsilon^e_{kl}} = \mathbb{C}^e_{ijkl}\varepsilon^e_{kl} \ ,$$

where $\psi^e\left(\varepsilon^e\right)$ is the scalar elastic potential, and $\mathbb{C}^e_{ijkl} = \frac{\partial \psi^e(\varepsilon^e)}{\partial \varepsilon^e_{ij} \otimes \varepsilon^e_{kl}}$ is the elastic constitutive tensor.

(ii) A kinematic compatibility condition between reversible and irreversible strains:

$$\dot{\varepsilon}_{ij} = \dot{\varepsilon}^e_{ij} + \dot{\varepsilon}^{vp}_{ij} \ .$$

(iii) The existence of an elastic region E, bounded by a yield surface F.

(iv) A viscoplastic strain evolution usually expressed as:

$$\dot{\varepsilon}_{ij}^{vp} = \dot{\lambda}\frac{\partial G}{\partial \sigma'_{ij}} \ .$$

where G is a plastic potential function.

(v) An evolution law for internal variables (hardening-softening rules). In this theory, the only internal variable is the preconsolidation pressure p_c;

(vi) An overstress or superload surface *active only above the yield surface* F, that results in a time-dependent yield function $\hat{F}$, with the following general structure:

$$\hat{F} = F - \dot{\lambda}S \ ,$$

where S depends on F at the current state; this feature distinguishes elasto-viscoplastic models from elastoplastic ones, which require only the first five features. Usually, we define S using classical viscoplastic definitions; thus, we follow Perzyna's definition that uses a time-dependent yield function $\hat{F}$ to compute consistency conditions to simulate the viscous effect without spurious dissipation.

We now introduce a theoretical framework based on the axiomatic structure to obtain a simple formulation. We define the remaining functions F, G, S and the time evolution of the preconsolidation pressure $\dot{p}_c$.

5.1.2 Elastic behaviour

We define an elastic potential as:

$$\psi^e\left(\varepsilon^e\right) = \frac{1}{2}\left(K_{ur} - \frac{2}{3}G_{ur}\right)\mathrm{tr}^2\varepsilon_{ij}^e + G_{ur}\varepsilon_{ij}^e\varepsilon_{ij}^e \ , \tag{5.2}$$

and consequently, the elastic constitutive tensor reads:

$$\mathbb{C}_{ijkl}^e = \frac{\partial \psi^e\left(\varepsilon^e\right)}{\partial \varepsilon_{ij}^e \otimes \varepsilon_{kl}^e} = \left(K_{ur} - \frac{2}{3}G_{ur}\right)\mathbf{I}\otimes\mathbf{I} + 2G_{ur}\mathbb{I} \ , \tag{5.3}$$

where $\mathbf{I} = \delta_{ij}$ is the second-order tensor identity, and $\mathbb{I} = \frac{1}{2}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right)$ represents the fourth-order tensor identity. Two stiffness parameters define the elastic response: the unloading-reloading bulk modulus K_{ur} and the unloading-reloading

shear modulus G_{ur}. These definitions allow for a straightforward computation of the consistent viscoplastic constitutive tensor; then, we analyse its spectral properties given its definition.

5.1.3 Yield function F and plastic potential function G

Using the modified Cam-Clay (MCC) model ideas [Roscoe & Burland, 1968], we express the yield function $F(p\left(\sigma'_{ij}\right), q, p_c)$ as:

$$F(p\left(\sigma'_{ij}\right), p_c) = \frac{q^2}{M^2 p} + p - p_c = \left(\frac{\eta}{M}\right)^2 \frac{(p - p_r)^2}{p} + p - p_c \ , \tag{5.4}$$

where M represents the slope of the critical state line (CSL), p_r is the reference pressure, and p_c refers to the preconsolidation pressure. Our derivations assume that (p, q) is a post-yield state, which implies that there exists a non-zero viscoplastic deformation. In what follows, we consider an associative viscoplastic flux, that is, $G \equiv F$.

5.1.4 Viscous evolution law S

Following Perzyna's overstress ideas [Perzyna, 1966], we assume there exists a super-loading post-yield surface $\hat{F}(p\left(\sigma'_{ij}\right), q, p_c) \leq 0$ at some instant t, such that:

$$\hat{F}(p\left(\sigma'_{ij}\right), p_c, \dot{\lambda}) = F(p\left(\sigma'_{ij}\right), p_c) - \dot{\lambda} S \ , \tag{5.5}$$

where S (scaling factor) estimates the overstress with respect to the yield function:

$$S = \frac{\langle F\left(p\left(\sigma'_{ij}\right), p_c\right)\rangle^m}{\mu} \ , \tag{5.6}$$

where $\langle\cdot\rangle$ stands for the Macaulay bracket, m depends on the compression index λ^*, the swelling index κ^* and the adimensional viscosity parameter μ^*, typical indices from critical-state theories in geomaterials. Besides, $\mu = \mu^*/\tau$ represents the viscosity rate with units s^{-1}, measured as the strain produced in a reference time-frame τ. For the sake of simplicity and without loss of generality, we fix the exponent value to be $m = 1$.

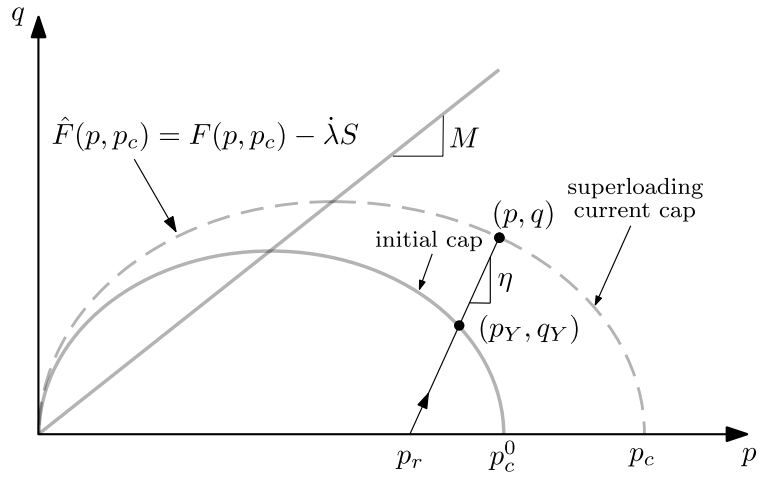


Figure 5.1: Problem statement for a modified Cam-Clay-type cap surface.

Our assumptions define the viscoplastic strain rate as follows:

$$\dot{\varepsilon}_{ij}^{vp} = \dot{\lambda} \mathbf{N}_{ij} \ , \tag{5.7}$$

where $\mathbf{N}_{ij}$ is the partial derivative of the plastic potential G with respect to the effective stress σ'_{ij}, the (visco)plastic flux. By associativity, we define the plastic potential as the yield function F, leading to the following definition of the plastic flux:

$$\mathbf{N}_{ij} := \frac{\partial F}{\partial \sigma'_{ij}} \ . \tag{5.8}$$

Additionally, we split the flux $\mathbf{N}$, into two orthogonal directions, the deviatoric and volumetric components (i.e., $\mathbf{N}_{\mathrm{d}}$ and N_{v}):

$$\dot{\boldsymbol{\varepsilon}}^{vp} = \dot{\lambda} \mathbf{N} = \dot{\lambda} (\mathbf{N}_{\mathrm{d}} + N_{\mathrm{v}} \mathbf{I}) \ ; \tag{5.9}$$

the deviatoric and volumetric strain rates then become:

$$\dot{\boldsymbol{\varepsilon}}_{\mathrm{d}}^{vp} = \dot{\lambda} \mathbf{N}_d \quad ; \quad \dot{\varepsilon}_{\mathrm{v}}^{vp} = \dot{\lambda} N_{\mathrm{v}} \ . \tag{5.10}$$

The above definition of the plastic flux follows classical critical-state assumptions in the Modified Cam-Clay model, and defines a compressive zone ($\dot{\boldsymbol{\varepsilon}}_{\mathrm{v}}^{vp} < 0$) for $p > p_c/2$, a dilatant zone ($\dot{\boldsymbol{\varepsilon}}_{\mathrm{v}}^{vp} > 0$) for $p < p_c/2$ (supercritical states) and an isochoric zone ($\dot{\boldsymbol{\varepsilon}}_{\mathrm{v}}^{vp} = 0$) at $p = p_c/2$ (critical state).

5.1.5 Volumetric hardening law: preconsolidation pressure evolution

A hardening law for the preconsolidation stress increment $\dot{p_c}$ in terms of viscoplastic strain-rate volumetric contribution $\dot{\varepsilon}_{\mathrm{v}}^{vp}$ is:

$$\dot{p_c} = H\dot{\varepsilon}_{\mathrm{v}}^{vp} = H\dot{\lambda}N_{\mathrm{v}} \ , \tag{5.11}$$

where H represents the hardening parameter that follows classical critical-state-based models:

$$H = \frac{p_c}{\lambda^* - \kappa^*} \ , \tag{5.12}$$

and

$$N_{\mathrm{v}} = 1 - \left(\frac{q}{Mp}\right)^2 = 1 - \left(\frac{\eta\left(p - p_r\right)}{Mp}\right)^2 \ , \tag{5.13}$$

is the volumetric contribution of the plastic flux as in (5.10).

5.1.6 Viscoplastic constraint and explicit preconsolidation evolution

These conditions are equivalent to the Prager's consistency condition in terms of $\hat{F}(p, p_c, \dot{\lambda})$, which reads:

$$\dot{\hat{F}}(p\left(\sigma'_{ij}\right), p_c, \dot{\lambda}) = \frac{\partial \hat{F}}{\partial \sigma'_{ij}}\dot{\sigma}'_{ij} + \frac{\partial \hat{F}}{\partial p_c}\dot{p_c} + \frac{\partial \hat{F}}{\partial \dot{\lambda}}\ddot{\lambda} = 0. \tag{5.14}$$

Using the volumetric hardening law (5.11), we rewrite (5.14) as:

$$\dot{\hat{F}}(p\left(\sigma'_{ij}\right), p_c, \dot{\lambda}) = \frac{\partial \hat{F}}{\partial \sigma'_{ij}}\dot{\sigma}'_{ij} + \frac{\partial \hat{F}}{\partial p_c}\dot{\lambda}HN_{\mathrm{v}} + \frac{\partial \hat{F}}{\partial \dot{\lambda}}\ddot{\lambda} = 0. \tag{5.15}$$

where partial derivatives are:

$$\begin{aligned} \frac{\partial \hat{F}}{\partial \sigma'_{ij}} &= \mathbf{N}_{ij}, \\ \frac{\partial \hat{F}}{\partial p_c} &= -1, \\ \frac{\partial \hat{F}}{\partial \dot{\lambda}} &= -S. \end{aligned} \tag{5.16}$$

The consistency condition allows us to compute the exact solution of the plastic multiplier and the consistent tangent constitutive tensor.

5.2 Compaction banding localisation analysis

5.2.1 Viscoplastic constitutive tensor recovery

We recover the viscoplastic constitutive tensor following similar previous approaches [Wang et al., 1997; Carosio et al., 2000]. For this task, we assume multi-axial stress compatibility. Then, starting from (5.15), the consistency condition reads:

$$\begin{aligned}
\dot{\hat{F}}(\sigma_{ij}, p_c, \dot{\lambda}) &= \frac{\partial \hat{F}}{\partial \sigma_{ij}} \dot{\sigma}_{ij} + \frac{\partial \hat{F}}{\partial p_c} \dot{\lambda} H N_{\mathrm{v}} + \frac{\partial \hat{F}}{\partial \dot{\lambda}} \ddot{\lambda} = 0 \ , \\
&= \frac{\partial \hat{F}}{\partial \sigma_{ij}} \mathbb{C}^{e}_{ijkl} \left(\dot{\varepsilon}_{ij} - \dot{\varepsilon}^{vp}_{ij} \right) + \frac{\partial \hat{F}}{\partial p_c} \dot{\lambda} H N_{\mathrm{v}} + \frac{\partial \hat{F}}{\partial \dot{\lambda}} \ddot{\lambda} = 0 \ , \\
&= \frac{\partial \hat{F}}{\partial \sigma_{ij}} \mathbb{C}^{e}_{ijkl} \left(\dot{\varepsilon}_{ij} - \dot{\lambda} \mathbf{N}_{ij} \right) + \frac{\partial \hat{F}}{\partial p_c} \dot{\lambda} H N_{\mathrm{v}} - S \ddot{\lambda} = 0 \ , \\
&= \mathbf{N} : \mathbb{C}^{e} : \dot{\varepsilon} - \left(\mathbf{N} : \mathbb{C}^{e} : \mathbf{N} - \frac{\partial \hat{F}}{\partial p_c} H N_{\mathrm{v}} \right) \dot{\lambda} - S \ddot{\lambda} = 0 \ , \\
&= a + b \dot{\lambda} + c \ddot{\lambda} = 0 \ .
\end{aligned} \tag{5.17}$$

which results in a first-order differential equation, with exact solution. We parametrise the overstress function S in terms of the trial pressure p_0 and the previous known preconsolidation pressure p_{c0} as follows:

$$S = \frac{\langle F\left(p_0, p_{c0}\right) \rangle}{\mu} \ , \tag{5.18}$$

Assuming frozen coefficients at the current increment after linearisation, we state the solution of (5.17) in exponential form as:

$$\dot{\lambda} = \left(\dot{\lambda}_0 + \frac{a}{b} \right) e^{-\frac{b}{c} t} - \frac{a}{b} \ , \tag{5.19}$$

where the coefficients correspond to the following expressions:

$$\begin{aligned} a &= \mathbf{N} : \mathbb{C}^e : \dot{\varepsilon} \,, \\ b &= -\mathbf{N} : \mathbb{C}^e : \mathbf{N} - H N_{\mathrm{v}} \,, \\ c &= -S \,. \end{aligned} \tag{5.20}$$

and $\dot{\lambda}_0 = \dot{\lambda}(t=0)$. Analysing (5.19), we can see that in the limit when $t \to \infty$, we recover the plastic multiplier along the lines of elastoplasticity. We then replace the viscoplastic multiplier in the incremental stress in the following way:

$$\begin{aligned} \dot{\sigma}_{ij} &= \mathbb{C}^e_{ijkl} \left(\dot{\varepsilon}_{kl} - \dot{\varepsilon}^{vp}_{kl} \right) = \mathbb{C}^e_{ijkl} \left(\dot{\varepsilon}_{kl} - \dot{\lambda} \mathbf{N}_{kl} \right) \,, \\ &= \mathbb{C}^e_{ijkl} \dot{\varepsilon}_{kl} - \mathbb{C}^e_{ijkl} \left[\left(\dot{\lambda}_0 + \frac{a}{b} \right) e^{-\frac{b}{c} t} - \frac{a}{b} \right] \mathbf{N}_{kl} \,, \\ &= \mathbb{C}^e_{ijkl} \dot{\varepsilon}_{kl} - \mathbb{C}^e_{ijkl} \mathbf{N}_{kl} \dot{\lambda}_0 e^{-\frac{b}{c} t} - \frac{\mathbb{C}^e_{ijmn} \mathbf{N}_{mn} \mathbf{N}_{pq} \mathbb{C}^e_{pqkl}}{\mathbf{N}_{ij} \mathbb{C}^e_{ijkl} \mathbf{N}_{kl} + H N_{\mathrm{v}}} \left(1 - e^{-\frac{\mathbf{N}_{ij} \mathbb{C}^e_{ijkl} \mathbf{N}_{kl} + H N_{\mathrm{v}}}{S} t} \right) \dot{\varepsilon}_{kl} \,. \end{aligned} \tag{5.21}$$

For the sake of simplicity, we define $H_p := \mathbf{N}_{ij} \mathbb{C}^e_{ijkl} \mathbf{N}_{kl} + H N_{\mathrm{v}}$. Finally, assuming a virgin initial state, i.e., $\dot{\lambda}_0 = 0$, the tangent viscoplastic constitutive tensor reads:

$$\dot{\sigma}_{ij} = \mathbb{C}^{vp}_{ijkl} \dot{\varepsilon}_{kl} \,, \tag{5.22}$$

with:

$$\mathbb{C}^{vp}_{ijkl} = \mathbb{C}^e_{ijkl} - \mathbb{C}^d_{ijkl}, \text{ with } \mathbb{C}^d_{ijkl} = \frac{\mathbb{C}^e_{ijmn} \mathbf{N}_{mn} \mathbf{N}_{pq} \mathbb{C}^e_{pqkl}}{H_p} \left(1 - e^{-\frac{H_p}{S} t} \right) \,, \tag{5.23}$$

In (5.23), whether $t \to 0$, the viscoplastic constitutive tensor tends to its elastic counterpart

$$\mathbb{C}^{vp}_{ijkl} \to \mathbb{C}^e_{ijkl} \,,$$

whereas when $t \to \infty$, it tends to the elastoplastic one $\mathbb{C}^{vp}_{ijkl} \to \mathbb{C}^{ep}_{ijkl}$. Finally, we compute the flux tensor $\mathbf{N}_{ij}$ in terms of the stress invariants for a generalised stress state, assuming the MCC yield surface, as follows:

$$\mathbf{N}_{ij} = \begin{bmatrix} -\frac{1}{3} + \frac{q^2 - 9p(\sigma'_{11} - p)}{3(Mp)^2} & -\frac{\sigma'_{12}}{M^2 p} & -\frac{\sigma'_{13}}{M^2 p} \\ -\frac{\sigma'_{12}}{M^2 p} & -\frac{1}{3} + \frac{q^2 - 9p(\sigma'_{22} - p)}{3(Mp)^2} & -\frac{\sigma'_{12}}{M^2 p} \\ -\frac{\sigma'_{13}}{M^2 p} & -\frac{\sigma'_{23}}{M^2 p} & \frac{1}{3} + -\frac{q^2 - 9p(\sigma'_{33} - p)}{3(Mp)^2} \end{bmatrix} \,. \tag{5.24}$$

5.2.2 Acoustic tensor as a bifurcation indicator

We use the classical bifurcation tensor, based on the spectral properties of the constitutive tensor [Rice, 1976; Rice & Rudnicki, 1980], which determines the admissibility condition for a discontinuity. Maxwell's restriction formulates that a jump in the strain increment must follow [Thomas, 1961; Rice & Rudnicki, 1980]:

$$[\![d\boldsymbol{\varepsilon}]\!] = d\boldsymbol{\gamma} \otimes^{s} \boldsymbol{n} \ , \tag{5.25}$$

where $\boldsymbol{n}$ represents the unit vector normal to the surface where we evaluate the stress state, and $d\boldsymbol{\gamma}$ is a vector that defines the discontinuity direction in the localisation. Besides, $[\![d\boldsymbol{\varepsilon}]\!]$ represents a jump in the strain increment between two points located on opposite sides of the discontinuity surface:

$$d\boldsymbol{\varepsilon}^{+} = d\boldsymbol{\varepsilon}^{-} + [\![d\boldsymbol{\varepsilon}]\!] \ . \tag{5.26}$$

The mechanical constitutive law in incremental form is:

$$d\boldsymbol{\sigma} = \mathbb{C} : d\boldsymbol{\varepsilon} \ , \tag{5.27}$$

where $\mathbb{C}$ is the tangent constitutive tensor, see (5.23). Equilibrium along the discontinuity surface Γ, considering continuity in the projected stresses, reads:

$$[\![d\boldsymbol{T}]\!] = [\![d\boldsymbol{\sigma} \cdot \boldsymbol{n}]\!] = d\boldsymbol{T}^{+} - d\boldsymbol{T}^{-} = (\mathbb{C} : [\![d\boldsymbol{\varepsilon}]\!]) \cdot \boldsymbol{n} = 0 \ . \tag{5.28}$$

From (5.25) and the symmetry properties of the constitutive tensor $\mathbb{C}$, we can rewrite (5.28) as:

$$(\mathbb{C} \cdot \boldsymbol{n}) \cdot d\boldsymbol{\gamma} \cdot \boldsymbol{n} = \mathbb{Q}(\boldsymbol{n}) \cdot d\boldsymbol{\gamma} = 0 \ , \tag{5.29}$$

where $\mathbb{Q}(\boldsymbol{n})$ is the acoustic tensor:

$$\mathbb{Q}(\boldsymbol{n}) = \boldsymbol{n} \cdot \mathbb{C} \cdot \boldsymbol{n} = 0 \ , \tag{5.30}$$

conventionally, we solve an eigenvalue problem to determine non-trivial solutions for $d\boldsymbol{\gamma} \neq 0$ in (5.29):

$$\det(\mathbb{Q}(\boldsymbol{n})) = 0 \ . \tag{5.31}$$

We expand the acoustic tensor in index notation in Appendix C.

5.2.3 Necessary condition for strain localisation

Heretofore, we interpret (5.31) as a statement of the condition for non-homogeneous localization [Olsson, 1999; Bésuelle, 2001], which is incomplete. Thus, we add necessary conditions on the strain rate from energetic considerations. For this, we recall Hill's instability condition Hill [1958] on the second-order work density d^2W under incremental perturbation:

$$d^2W = 0 \implies \dot{\boldsymbol{\sigma}} : \dot{\boldsymbol{\varepsilon}} = 0 \ . \tag{5.32}$$

We rewrite (5.32) in terms of the elastic strain using index notation; thus, the instability condition becomes:

$$\mathbb{C}^e_{ijkl} \left(\dot{\varepsilon}_{kl} - \dot{\varepsilon}^{vp}_{kl} \right) \dot{\varepsilon}_{ij} = 0 \ . \tag{5.33}$$

The above identity is valid for any $\dot{\varepsilon}_{ij}$, regardless of its direction. This feature, along with the positive definiteness of $\mathbb{C}^e_{ijkl}$, allows us to state that:

$$\dot{\varepsilon}_{kl} \to \dot{\varepsilon}^{vp}_{kl} \tag{5.34}$$

in the localisation onset.

Alternatively, we express the stress rate in (5.32) in terms of the total strain rate and the viscoplastic constitutive tensor as in (5.22), as follows:

$$\dot{\boldsymbol{\sigma}} : \dot{\boldsymbol{\varepsilon}} = \dot{\varepsilon}_{ij} \mathbb{C}^{vp}_{ijkl} \dot{\varepsilon}_{kl} = 0 \ , \tag{5.35}$$

and, considering (5.7) and (5.34), expression (5.35) reads:

$$\dot{\varepsilon}_{ij} \mathbb{C}^{vp}_{ijkl} \dot{\varepsilon}^{vp}_{kl} = \dot{\lambda} \ \dot{\varepsilon}_{ij} \mathbb{C}^{vp}_{ijkl} \mathbf{N}_{kl} = 0 \ , \tag{5.36}$$

We then construct a tensorial quantity that becomes an indicator of the localisation in a more robust way than the acoustic tensor, especially under isotropic conditions, which reads:

$$\mathbb{L}_{ij} := \mathbb{C}^{vp}_{ijkl} \mathbf{N}_{kl} \ . \tag{5.37}$$

Unlike typical approaches such as Olsson [1999], the localisation onset occurs when an eigenvalue in $\mathbb{L}$ becomes zero, and the localisation direction is parallel to the eigenvector associated with the zero eigenvalue. We illustrate the usefulness

of this definition in the following section.

5.2.4 Stress scenarios analysis

This section analyses the localisation conditions that trigger the onset of compaction banding for several well-known stress conditions in geomechanics. For this, we express the normal direction $\boldsymbol{n}$ of the localisation plane in terms of two angles θ and φ as follows:

$$\boldsymbol{n} = \begin{bmatrix} \cos\theta \sin\varphi \\ \cos\theta \cos\varphi \\ \sin\theta \end{bmatrix} , \tag{5.38}$$

as Figure 5.2 shows. Under triaxial states, we can define an infinite number of localisation planes for all values of the angle φ. We use a fixed value of $\varphi = \pi/2$, which allows us to define the normal n only by the angle θ, measured from the horizontal line.

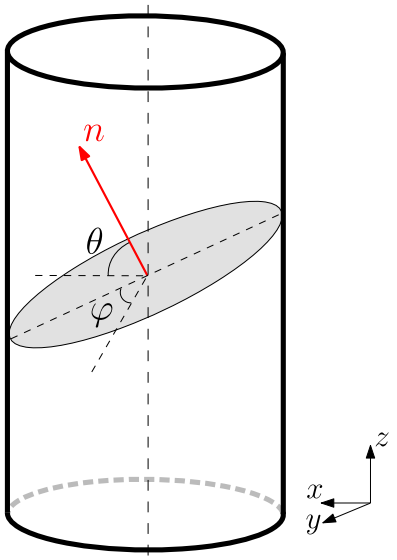
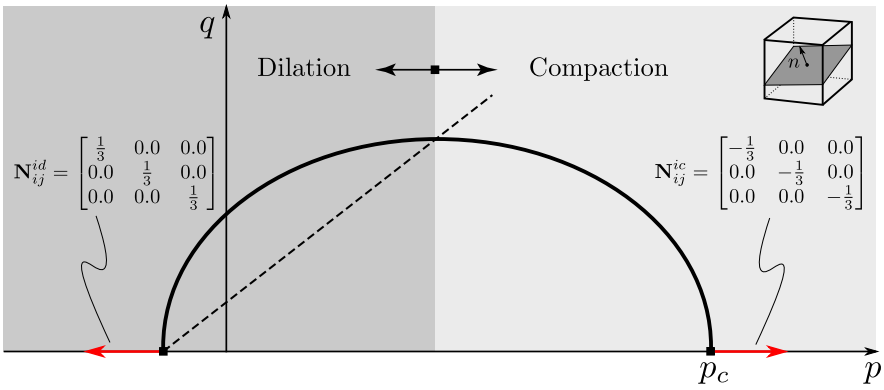


Figure 5.2: Localisation plane with normal direction $\boldsymbol{n}$ expressed by their angular components.

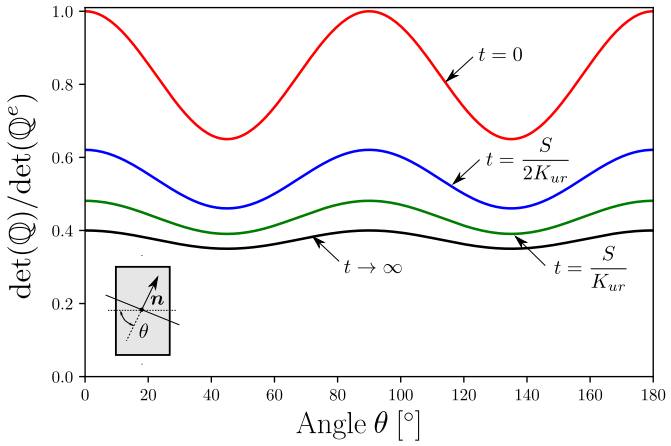
Using this notation, we can associate the localisation planes of the compaction bands that appear in contractive viscoplastic strain regimes with an angle θ close to 90°. We analyse the onset of this phenomenon in both compression and extension scenarios under stress-controlled states.

5.2.4.1 Isotropic compression/extension

Proposition 6. *Isotropic pressure states (compression or extension) in isotropic geomaterials localises at all directions* $\boldsymbol{n}$ *simultaneously.*



(a) Stress path and plastic flux evolution.



(b) Isotropic compression/extension.

Figure 5.3: Isotropic compression/extension.

Proof. The principal stress tensor associated to an isotropic triaxial compression/extension test is:

$$\sigma'_{ij} = \begin{bmatrix} p' & 0 & 0 \\ 0 & p' & 0 \\ 0 & 0 & p' \end{bmatrix} \quad \text{with } \sigma'_{11} = \sigma'_{22} = \sigma'_{33} = p', \tag{5.39}$$

Particularising (5.24), the flux tensor becomes:

$$\mathbf{N}_{ij}^{ic} = \begin{bmatrix} -\frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & -\frac{1}{3} \end{bmatrix} \quad \text{and} \quad \mathbf{N}_{ij}^{id} = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}, \tag{5.40}$$

see Figure 5.3(a). Revisiting (5.23), as the plastic flow direction appears quadratically in all terms, the resulting instability condition is insensitive to the flow direction (i.e., compression vs extension). Then, assuming a material with bulk modulus K_{ur} and shear modulus G_{ur}, the isotropic elastic tensor reads:

$$\mathbb{C}_{ijkl}^{e} = \begin{bmatrix} \frac{4}{3}G_{ur} + K_{ur} & -\frac{2}{3}G_{ur} + K_{ur} & -\frac{2}{3}G_{ur} + K_{ur} & 0 & 0 & 0 \\ -\frac{2}{3}G_{ur} + K_{ur} & \frac{4}{3}G_{ur} + K_{ur} & -\frac{2}{3}G_{ur} + K_{ur} & 0 & 0 & 0 \\ -\frac{2}{3}G_{ur} + K_{ur} & -\frac{2}{3}G_{ur} + K_{ur} & \frac{4}{3}G_{ur} + K_{ur} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2G_{ur} & 0 & 0 \\ 0 & 0 & 0 & 0 & 2G_{ur} & 0 \\ 0 & 0 & 0 & 0 & 0 & 2G_{ur} \end{bmatrix}, \tag{5.41}$$

and computing (5.23), we obtain the following viscoplastic constitutive tensor:

$$\mathbb{C}_{ijkl}^{vp} = G_{ur} \begin{bmatrix} \frac{4}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & -\frac{2}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & -\frac{2}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & 0 & 0 & 0 \\ -\frac{2}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & \frac{4}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & -\frac{2}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & 0 & 0 & 0 \\ -\frac{2}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & -\frac{2}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & \frac{4}{3} + \frac{K_{ur}}{G_{ur}} e^{-\frac{K_{ur}t}{S}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}, \tag{5.42}$$

or in terms of Poisson modulus $\frac{K_{ur}}{G_{ur}} = \frac{2(1+\nu)}{3(1-2\nu)}$. If $t \to \infty$, matrix (5.42) becomes:

$$\mathbb{C}_{ijkl}^{vp}|_{t\to\infty} = G_{ur} \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} & -\frac{2}{3} & 0 & 0 & 0 \\ -\frac{2}{3} & \frac{4}{3} & -\frac{2}{3} & 0 & 0 & 0 \\ -\frac{2}{3} & -\frac{2}{3} & \frac{4}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}. \tag{5.43}$$

Computing the $\mathbb{L}_{ij}$ tensor for this case, we obtain

$$\mathbb{L}_{ij} = \begin{bmatrix} K_{ur}e^{-\frac{K_{ur}}{S}t} & 0 & 0 \\ 0 & K_{ur}e^{-\frac{K_{ur}}{S}t} & 0 \\ 0 & 0 & K_{ur}e^{-\frac{K_{ur}}{S}t} \end{bmatrix} , \tag{5.44}$$

where the eigenvalues ρ_i and eigenvectors $\boldsymbol{n}_i$ that represent the localisation are:

$$\rho_i = K_{ur}e^{-\frac{K_{ur}}{S}t}, \quad \text{and} \quad \boldsymbol{n}_i = K_{ur}e^{-\frac{K_{ur}}{S}t}\ e_i\ . \tag{5.45}$$

The eigenvalues show that the second-order energy density d^2W becomes zero for large time ($t \to \infty$). Besides, the localisation occurs in all directions $\boldsymbol{n}_i$. Thus, isotropic loadings do not have a preferential localisation direction, regardless of the sample drainage (drained or undrained). When a specimen is subject to isotropic pressure, the degradation is only induced on the volumetric stiffness until it completely vanishes, localising in all directions simultaneously. Then, no localisation occurs; the material collapses due to the complete loss of volumetric stiffness. This analysis also reveals that the acoustic tensor in (5.30) cannot detect this volumetric failure. Figure 5.3(b) displays this deficiency by showing that the determinant of the acoustic tensor degrades uniformly up to a value larger than zero as time grows ($t \to \infty$). □

5.2.4.2 Drained triaxial compression

Proposition 7. *In drained triaxial compression tests, the compaction band occurs when the stress path reaches a well-defined point on the yield surface where the plastic flow* $\boldsymbol{N}$ *is parallel to the maximum principal stress. Mathematically, the plastic flow components meet the condition* $\boldsymbol{N}_{22} \to 0$, $\boldsymbol{N}_{33} \to 0$ *and* $\boldsymbol{N}_{11} < 0$..

Proof. Assuming a reference pressure $p_r \leq p_c$, the stress tensor associated with deviatoric stress of a drained triaxial compression test reads:

$$\sigma'_{ij} = \begin{bmatrix} 3p' - 2p_r & 0 & 0 \\ 0 & p_r & 0 \\ 0 & 0 & p_r \end{bmatrix} , \tag{5.46}$$

with p' being the effective isotropic pressure of the sample. Particularising (5.24) to the stress state, and assuming that $q = 3(p' - p_r)$ in triaxial compression, the

flux tensor in this case is:

$$\mathbf{N}_{ij} = \begin{bmatrix} -\frac{1}{3} - \frac{3(p'^2 - p_r^2)}{(Mp')^2} & 0 & 0 \\ 0 & -\frac{1}{3} + \frac{3(p'-p_r)(2p'-p_r)}{(Mp')^2} & 0 \\ 0 & 0 & -\frac{1}{3} + \frac{3(p'-p_r)(2p'-p_r)}{(Mp')^2} \end{bmatrix} . \quad (5.47)$$

As in Section § 5.2.4.1, we particularise the elastic and viscoplastic constitutive tangent tensors to compute the acoustic tensor to find a localisation trigger.

First, we analyse a drained triaxial test with $p_r = p_c$, where the initial plastic flow $\mathbf{N}_{ij}^{tx}|_{t_0}$ is the one from Proposition 6, as Figure 5.4(a) shows. The plastic flow direction that triggers the compaction bands is approximately:

$$\mathbf{N}_{ij}^{tx}|_{t_l} = \begin{bmatrix} \rightarrow -0.76 & 0 & 0 \\ 0 & \rightarrow 0 & 0 \\ 0 & 0 & \rightarrow 0 \end{bmatrix} , \quad (5.48)$$

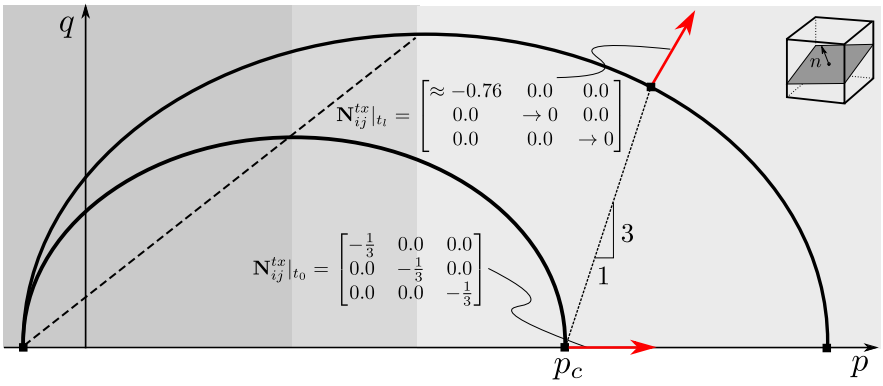
that occurs at time $t_l = 30\frac{S}{H_p}$. Figure 5.4(b) shows the acoustic tensor degradation as it reaches a localized state ($\det(\mathbb{Q}) = 0$) for $\theta = 90°$ at time t_l. Additionally, $\mathbf{N}_{11}^{tx} < 0$ and the plastic multiplier $\dot{\lambda} > 0$, implying that the viscoplastic strain is contractive and the localization corresponds to a compaction band.

Next, we analyse a sample where $p_c = 40\text{MPa}$ and $p_r = 22\text{MPa} < p_c$. Figure 5.5(a) shows the stress path. The stress path touches the yield surface, inducing the following initial plastic flow:

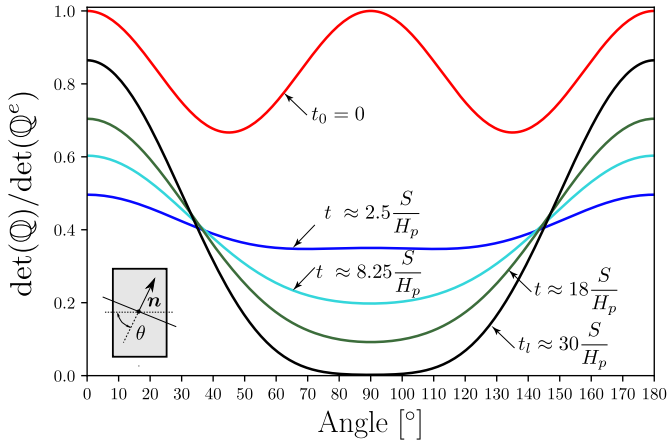
$$\mathbf{N}_{ij}^{tx}|_{t_0} = \begin{bmatrix} -0.73 & 0 & 0 \\ 0 & -0.03 & 0 \\ 0 & 0 & -0.03 \end{bmatrix} , \quad (5.49)$$

The stress progressively grows and localization happens at time $t_l \approx 0.77\frac{S}{H_p}$ in the same direction of (5.48). In this case, the compaction band appears earlier than in the normally consolidated case.

As a consequence, due to the construction of the acoustic tensor and the tangent viscoplastic constitutive tensor, the instability region where the compaction band occurs goes through a point where the plastic flow $\mathbf{N}_{ij}$ is parallel to the principal stress applied by the triaxial test. Figure 5.6 shows the plastic flow evolution in a drained triaxial test for the normally consolidated sample (a) and the over-consolidated one (b). In both cases, the radial components of the plastic



(a) Stress path and plastic flux evolution.



(b) Acoustic tensor degradation.

Figure 5.4: Drained triaxial compression test, starting from a reference pressure $p_r = p_c$.

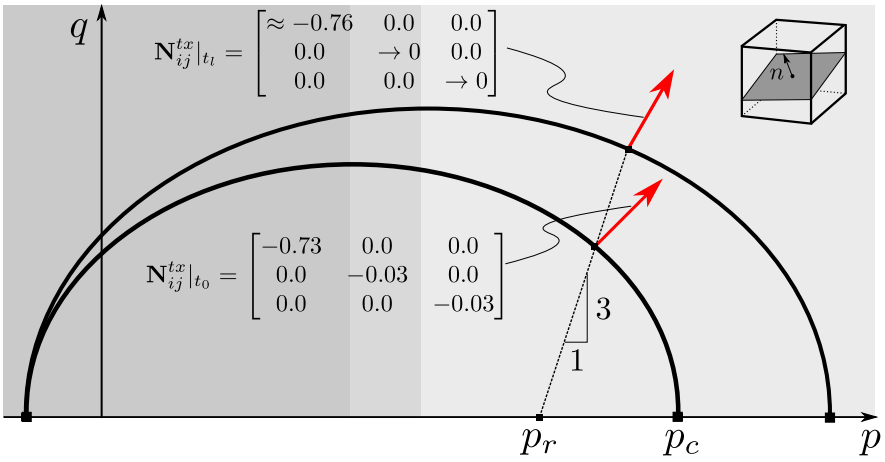
flow, $\mathbf{N}_{22}$ and $\mathbf{N}_{33}$, become zero when they reach the transition from compressive to extensive regimes, whereas the axial component $\mathbf{N}_{11}$ remains negative throughout the whole stress loading history.

□

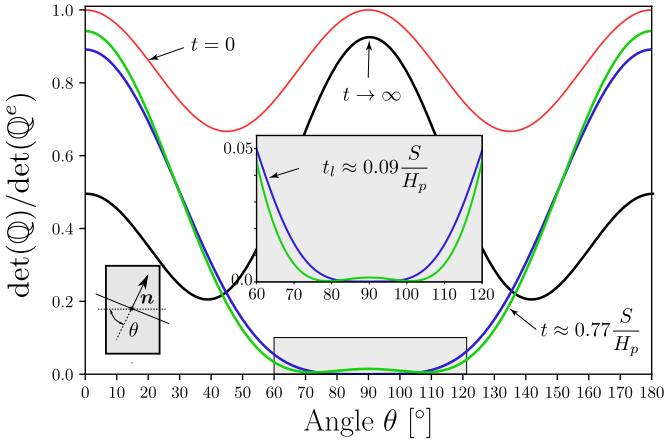
5.2.4.3 Drained triaxial extension test

Proposition 8. *In drained triaxial extension tests, the dilation band is triggered when the stress path touches a well-defined point on the yield surface where the plastic flux* $\boldsymbol{N}$ *is oriented to the maximum principal stress. Mathematically, plastic flux components meet the condition* $\boldsymbol{N}_{22} \to 0$, $\boldsymbol{N}_{33} \to 0$ *and* $\boldsymbol{N}_{11} > 0$.

Proof. Figure 5.7(a) shows the drained triaxial extension test stress path, for a sample with $p_r < p_c$, analogous to the previous case. Similarly to the compaction band, the dilation band appears when radial plastic flow reads $\mathbf{N}_{22} = \mathbf{N}_{33} = 0$



(a) Stress path and plastic flux evolution.



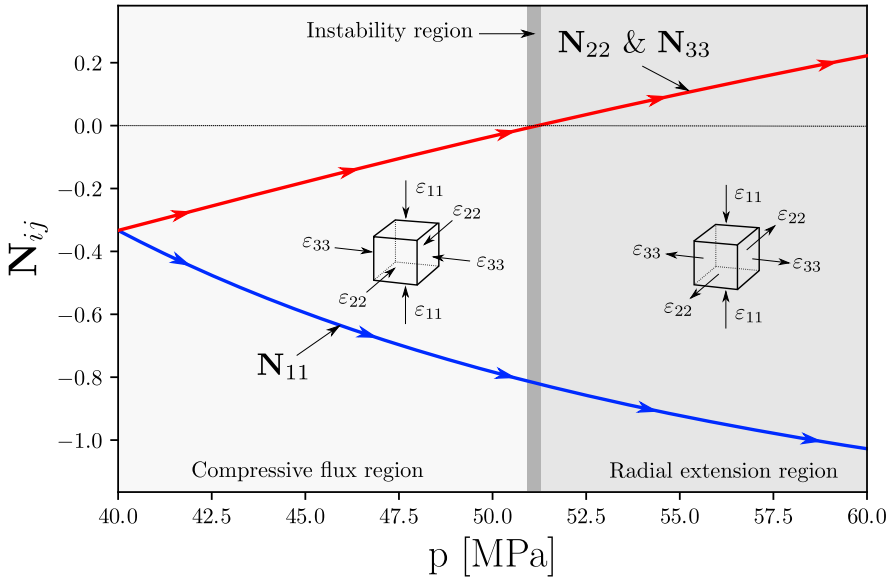
(b) Acoustic tensor degradation.

Figure 5.5: Drained triaxial compression test, starting from a reference pressure $p_r = 22$ MPa and preconsolidation pressure of $p_c = 40$ MPa.

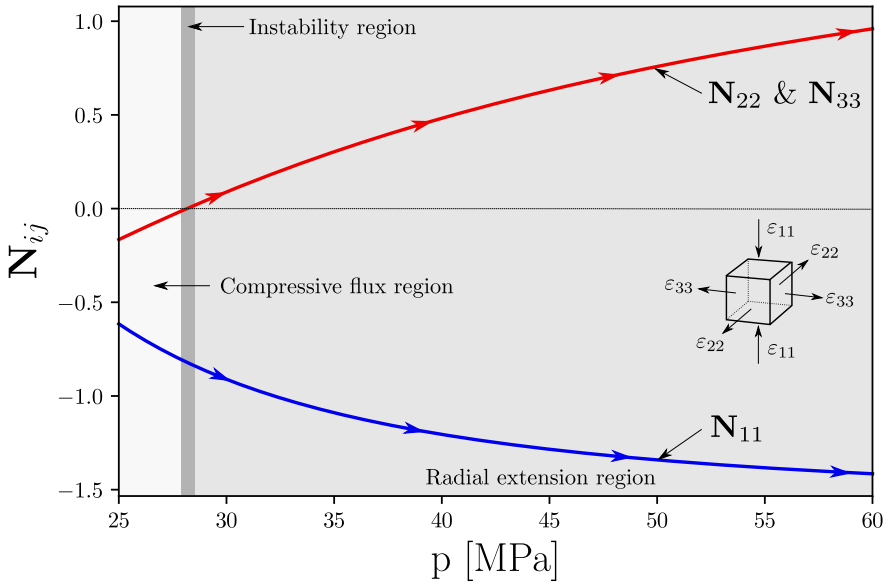
under the condition of having a positive plastic flow parallel to the principal stress. This stress path induces the following initial plastic flow, $\mathbf{N}_{ij}^{dtx}|_{t_0}$, and localization plastic flow, $\mathbf{N}_{ij}^{dtx}|_{t_l}$,

$$\mathbf{N}_{ij}^{dtx}|_{t_0} = \begin{bmatrix} 2.32 & 0 & 0 \\ 0 & -0.55 & 0 \\ 0 & 0 & -0.55 \end{bmatrix} \implies \mathbf{N}_{ij}^{dtx}|_{t_l} = \begin{bmatrix} 4.57 & 0 & 0 \\ 0 & \to 0 & 0 \\ 0 & 0 & \to 0 \end{bmatrix}. \tag{5.50}$$

Figure 5.7(b) shows that the final condition triggers a localization at $\theta = 90°$. However, as $\mathbf{N}_{11} > 0$, we observe dilative regime during the localization. Figure 5.7(c) shows the instability region for this scenario in terms of the plastic flow in the three directions.



(a) $\mathbf{N}_{ij}$ evolution for the normally consolidated case.



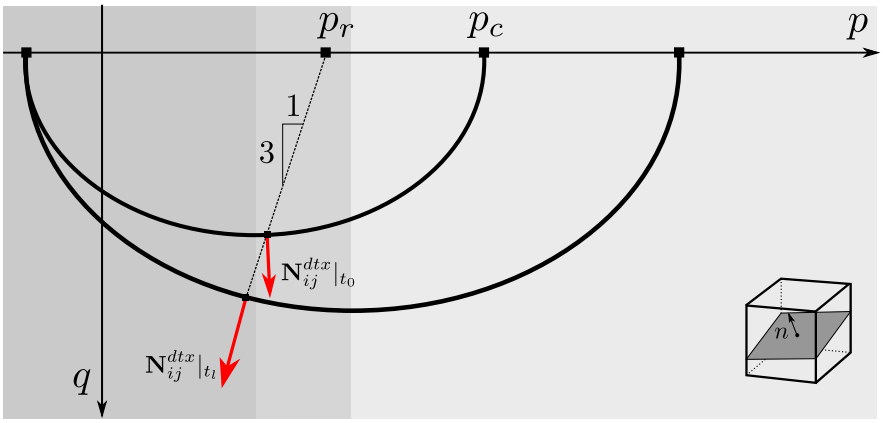
(b) $\mathbf{N}_{ij}$ evolution for the over consolidated case.

Figure 5.6: Plastic flux components evolution and region detection of compaction bands instability.

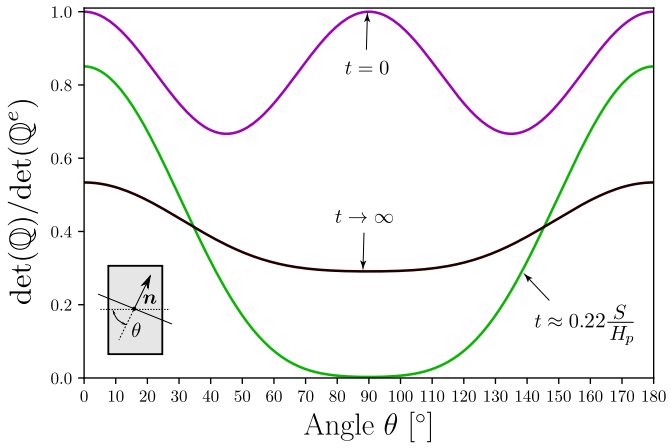
□

5.3 Numerical simulations

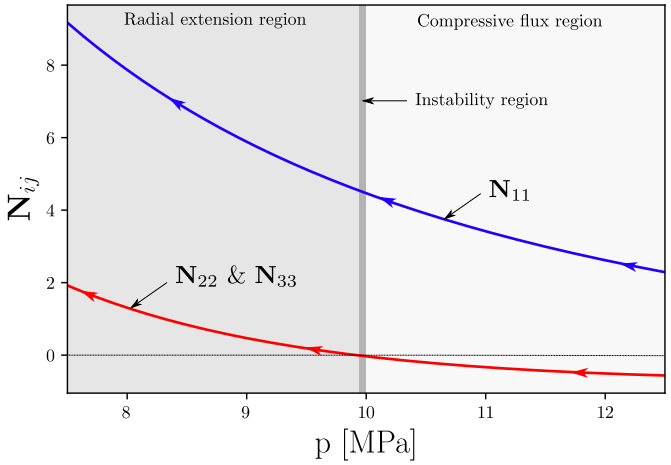
Our analysis framework seeks to predict the appearance of compaction bands in porous rocks processes in several laboratory tests [Arroyo et al., 2005; Fortin



(a) Stress path for triaxial unloading test.



(b) Localisation with acoustic tensor.



(c) $\mathbf{N}_{ij}$ evolution for the over consolidated case.

Figure 5.7: Dilation band setup.

et al., 2006; Oka et al., 2011; Abdallah et al., 2021; Leuthold et al., 2021]. Our numerical experiments induce localisation under different triaxial compression

conditions, using the Vermeer and Neher's [Vermeer & Neher, 1999] model, a modified overstress model based on Perzyna's viscoplasticity that can be understood as a particularisation of the constitutive framework of Section 5.1. Our results show that identical samples subject to different confinement pressures undergo different localisation processes. Effectively, the variation of the confinement pressure transitions the localisation from shear to compaction bands, as reported in the literature. Our experiments also analyse the bands' periodicity and spacing and their dependence on the material parameters.

5.3.1 Constitutive model

The [Vermeer & Neher, 1999] model incorporates rate-dependent effects into an elastoplastic constitutive model by generalising the logarithmic creep law for secondary compression [Bjerrum, 1967]. In a three-dimensional stress state, the constitutive model combines a perfectly-plastic Mohr-Coulomb yield surface to reproduce shear effects, along with an elliptic cap based on the Modified-Cam Clay (MCC) model introduced by Roscoe & Burland [1968] that allows simulating the compressive behaviour. Moreover, the model incorporates a hardening law that simulates the rate-dependent effect of the material, where all the inelastic strains are considered to be due to creep. Figure 5.8 shows the yield surface and the viscosity effect associated with the compressive cap.

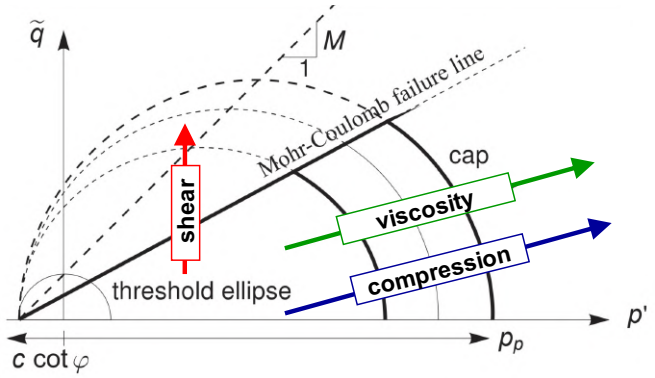


Figure 5.8: Vermeer & Neher [1999] yield surface and viscous hardening in compression.

The model introduces the following time-dependent yield function:

$$f = p^{eq} - p_p^{eq}[t] = p + \frac{q^2}{M^2(p + c\cot\varphi)} - p_p^{eq}[t], \tag{5.51}$$

where c is the cohesion, φ is the friction angle, and M is the critical state line's

slope (see Figure 5.8), $M = 6\sin\varphi_{cv}/(3-\sin\varphi_{cv})$, with φ_{cv} as the critical-state friction angle. Additionally, the superscript eq represents an equivalent three-dimensional generalisation from one-dimensional scenarios where the effective stress ratio K_0^{NC} is known. Thus, for instance, we can compute p^{eq} and p_p^{eq} from the one-dimensional effective stress σ' and the preconsolidation pressure σ_p respectively.

Assuming standard critical state considerations, this model considers the viscoplastic strain rate evolution entirely in the volumetric part $\dot{\varepsilon}_{\mathrm{v}}^{vp}$ that extends from the one-dimensional creep law, which reads:

$$\dot{\varepsilon}_{\mathrm{v}}^{vp} = -\frac{\mu^*}{\tau}\left(\frac{p^{eq}}{p_p^{eq}}\right)^{\frac{\lambda^*-\kappa^*}{\mu^*}} , \tag{5.52}$$

where τ is a reference time frame (generally 24 hours), and κ^*, λ^* and μ^* are indices related to the classical oedometric indices C_s, C_c and C_α by:

$$\kappa^* = \frac{2C_s}{2.3(1+e_0)} \quad , \quad \lambda^* = \frac{C_c}{2.3(1+e_0)} \quad , \quad \mu^* = \frac{C_\alpha}{2.3(1+e_0)} \quad , \tag{5.53}$$

with e_0 the initial void ratio. Finally, we can deduce the preconsolidation pressure p_p^{eq} in (5.52) from an MCC state equation modified that accounts for the viscoplastic strain in the following way:

$$p_p^{eq} = p_{p0}^{eq}\exp\left(-\frac{\varepsilon_{\mathrm{v}}^{vp}}{\lambda^*-\kappa^*}\right), \tag{5.54}$$

where p_{p0}^{eq} is an equivalent initial preconsolidation pressure at $t = 0$, considering that $\varepsilon_{\mathrm{v}}^{vp} = 0$.

Although the Vermeer & Neher [1999] model was not originally conceived for modelling rocks, we can find examples in the literature where this model is used for this type of geomaterials, especially in subsurface subsidence modelling [Volonté et al., 2017; Ghisi et al., 2021], given the model's simplicity and its small number of parameters.

5.3.2 Material parameters selection

Below, we use standard relationships between different indices to reproduce specific behaviours in this experiment. For our rock, we assume a porosity around

30%, implying an initial void ratio of $e_0 = 0.42$. In the literature, compression indices for porous rocks as sandstones typically have values in the range $C_c = 0.2 - 0.4$ [Hüpers & Kopf, 2012]. In range, we assume $\lambda^* = 0.1$. Then, we estimate the other indices based on well-known ratios [Vermeer & Neher, 1999]. Table 5.1 summarizes the parameters of the model.

Table 5.1: Model parameters for the modelled rock.

Parameter	**Symbol**	**Unit**	**Value**
Unit weight	γ	kN/m^3	22
Compression parameter	λ^*	-	0.1
Swelling parameter	κ^*	-	0.01
Creep parameter	μ^*	-	5e-4
Poisson's ratio	ν_{ur}	-	0.15
Cohesion	c	kPa	100
Friction angle	φ	-	38°
Effective stress ratio	K_0^{NC}	-	0.5239
Critical state line slope	M	-	1.563
Initial preconsolidation pressure	p'_{p0}	MPa	40

From this parameters selection, we compute a creep ratio (CR), an indirect measure of the viscous contribution in the Vermeer and Neher's [Vermeer & Neher, 1999] model, as

$$\text{CR} = \frac{\lambda^* - \kappa^*}{\mu^*}.$$

Thus, the creep ratio value is $CR = 180$, which is high enough to ignore possible rate-dependent effects. However, our numerical examples show that the viscous input induces a change in the strain-localisation behaviour in our sample.

5.3.3 Finite element analysis of triaxial compression tests

For the numerical experiments, we employ an axisymmetric strain model for a rectangular domain of $[0, 0.025] \times [0, 0.1]$ m^2, which is partitioned into a regular mesh composed of quadratic triangular elements of size $h = 0.0025$ m. The boundary conditions are such that displacements normal to $x = 0$ and $y = 0$ are null. Additionally, we impose a distributed load σ'_3 at $x = 0.1$ m and $y = 0.025$ m to simulate the isotropic compression load in the consolidation stage and the confinement pressure during the shearing stage and a time-dependent displacement u_y at $y = 0.1$ m in the shearing stage to reproduce the deviatoric deformation at the top of the sample. In this experiment, the confinement pressure takes

values of $\sigma_3' = 5, 10, 14, 22$ and 30 MPa, whereas the prescribed displacement is $u_y = 5 \times 10^{-3}$ m, such that it produces a vertical strain of $\varepsilon_{yy} = 5\%$. Figure 5.9 sketches the mesh and the boundary conditions considered for the numerical experiments.

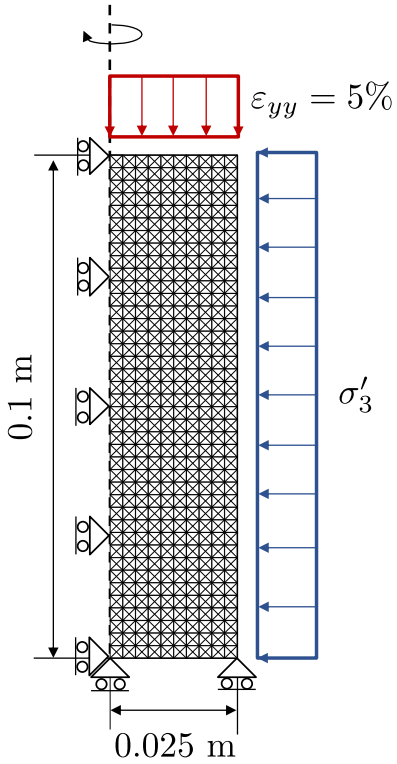


Figure 5.9: Mesh and boundary conditions for the finite element model of the shearing stage in the compression triaxial tests.

We simulate the triaxial compression test using a hydromechanical model that solves the equilibrium and continuity equations similarly to Biot's theory for coupled consolidation. We impose loading strain rate of $\dot{\varepsilon}_{yy} = 10^{-5}$ s^{-1}, with a time step of $\Delta t = 20$ s. We do not introduce a weak element that induces a preferential localisation in the sample for the experiments.

5.3.4 Results discussion

We also analyse the transition in the localisation behaviour through the stress paths from the tests, as Figure 5.11 shows. Here, the shear band occurrence (zone (1)) appears for the lowest confinement pressures because the stress path reaches the Mohr-Coulomb yield surface before the cap, implying that the localisation is strictly inviscid. For the intermediate confinement pressures, there exists an

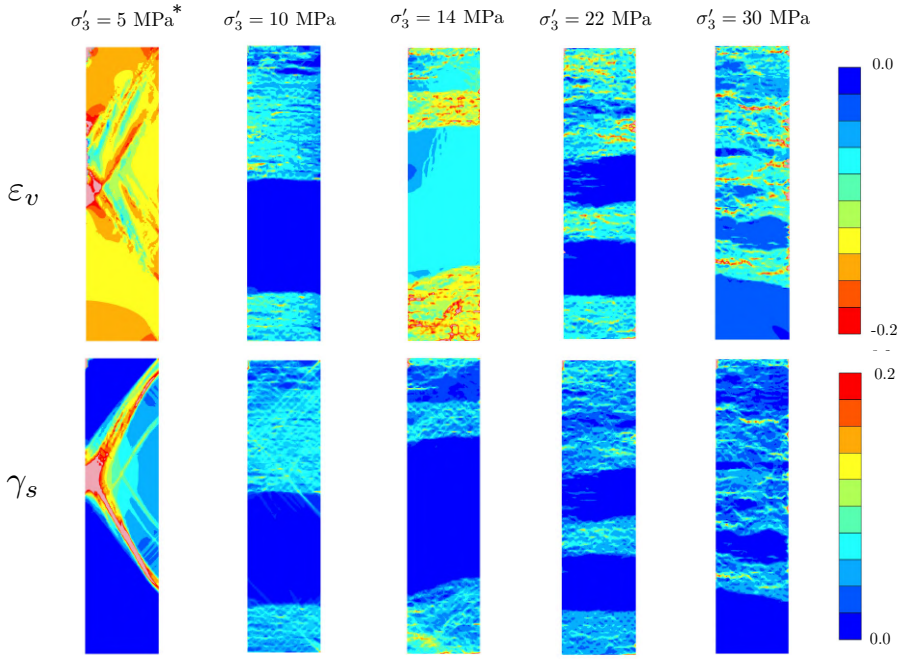


Figure 5.10: Transitional effect in the volumetric (ε_{v}) and deviatoric (ε_{d}) strains for confinement pressure increase. Note: at 5MPa, the values for ε_v are 20 times smaller than for the rest of the cases, and the colour bar must be read considering this.

interplay between the viscous effect produced by pushing the cap (zone (2)) and the failure associated with reaching the Mohr-Coulomb yield surface, producing a compounded (transitional) shear/compaction effect in the sample. Higher confinements generate stress paths that yield a more significant visco-plastic strain inducing the samples to localise purely by compaction (zone (3)). This transition occurs not only through the strain components (see Figure 5.10) but also through the effective mean stress (p'), where the phenomenon evolves from a shear failure, in low confinement pressures, to a well-defined and rich set of mean stress accumulation zones in high confinement scenarios. Finally, these results explain experimental observations obtained under similar loading conditions [Sari et al., 2022] and validate our analytical findings.

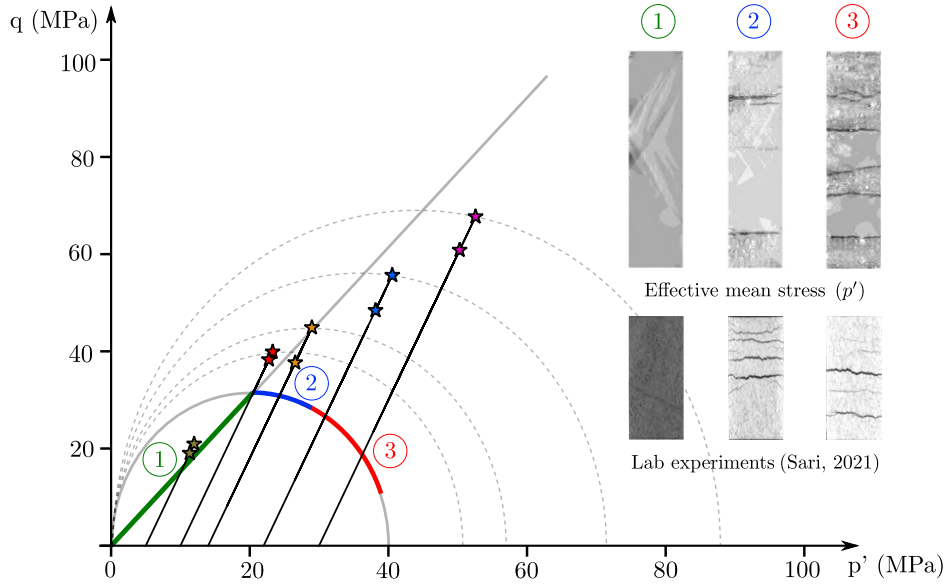


Figure 5.11: Stress paths of the triaxial compression tests under different confinement pressures.

Chapter 6

Conclusions and research perspectives

This work represented a theoretical and computational approach to studying strain localisation in the form of compaction bands in geomaterials, specifically in porous rocks. The first part of the work established a robust numerical framework for a wave-mechanics approach to the compaction band localisation process. It was based on an adaptive stabilised finite element method, extended successfully to nonlinear reaction–diffusion equations. With this tool, we were able to analyse the influence of the internal diffusivity interplay in the mechanical response of the specimen, showing the versatility of the proposed numerical method in finding stable solutions for a wide range of scenarios, including periodic conditions. The second part of the research dealt with the compaction band localisation as a bifurcation problem in rate-dependent critical-state-based materials. We studied the onset of compaction bands for well-known stress scenarios in geomechanical tests and established a series of statements for the localisation onset conditions, including compaction and dilation bands.

6.1 Conclusions

We can summarise the main conclusions of this work in the following items:

- We developed an efficient nonlinear extension for an adaptive stabilised finite element method, robust enough to find different branches of solutions in well-known problems such as Bratu's equation and also to recover a family

of solutions for the governing equation of the cnoidal theory in compaction bands, a differentiating advantage compared with standard finite element formulations.

- The numerical analysis of the governing equation of the cnoidal theory allowed us to recover the variety of conditions that can trigger different localisation configurations. Unlike the primal work of Veveakis & Regenauer-Lieb [2015], where uniquely the diffusivity ratio λ determines the behaviour, we were able to find different symmetry configurations in the solutions depending strongly on the initial condition.

- We extend the possible range of scenarios by adding the periodic conditions to the numerical analysis of the governing cnoidal equation. This state allowed a strong correlation between the diffusivity parameter λ and the number of localisation zones, which is consistent with the physics of the problem. In general, the presented numerical tool narrows the computational gap within the general framework of this wave-mechanics theory applied to localisation processes.

- The proposed consistent viscoplastic constitutive model, constructed from basic axiomatic statements, demonstrated to be an efficient framework for the bifurcation analysis from homogeneous deformation states in rate-dependent materials. Besides, the spectral analysis of the localisation indicator tensor $\mathbb{L}$ overcame the issues associated with the determinant of the classical acoustic tensor $\mathbb{Q}$ under isotropic stress states, allowing to give a more consistent explanation of the localisation phenomenon for these cases.

- The numerical experiments for the compaction banding phenomenon were understood as a particularisation of the analytical approach followed in the bifurcation analysis. The obtained results confirmed the transitional effect due to the confinement pressure on the onset of this type of localisation. Additionally, they agreed with experimental tests carried out under similar loading conditions [Sari et al., 2022], which in turn found a physical correlation with the original cnoidal wave theory.

6.2 Research outlook

The following research perspectives derive from this project:

- This project focused on developing a robust numerical implementation of the adimensional equation from the cnoidal wave theory. However, a proper extension of the cnoidal theory to higher dimensional scenarios is still pending, with a proper re-derivation of the balance equations. This modification represents a considerably complex task, not due to the presence of the hyperbolic stress equilibrium equations, but because it involves resolving the temporal numerical instabilities in an indirect manner, where the blow-up term is hidden in the plastic behaviour instead of appearing in the system of equations directly [Cier et al., 2021].

- All the numerical simulations in this project has been carried out to a sample level. Field-scale simulations may be part of future work, as they represent a more realistic condition in rock mechanics, implying improvements in both the numerics and the theory. From the numerical perspective, efficient three-dimensional frameworks would need to overcome the size differences between the phenomenon (cm thick) compared to the model's size (m or km), as well as time scales, as these processes involve thousands of years. From the theoretical aspects, enhancements will need to consider paleostresses and tectonics in the way of anisotropic properties in the materials, along with coupled processes that lead to the phenomenon's onset.

- The consistent failure criterion presented in Chapter 5 was evaluated through the use of a particularisation, the Vermeer & Neher [1999] model, already available in a finite element framework. Proper implementation of the proposed consistent viscoplastic constitutive model could motivate new developments in terms of the influence of the consistency condition in the performance of this type of constitutive model.

- This project dealt with a specific type of localisation in the form of compaction bands in geomaterials focused on porous rocks, but the theory may also study their occurrence in soils. Although not reporting compaction bands, notorious work in drained triaxial tests in uncemented clays have been developed by Hicher et al. [1994] and Wei & Wang [2022] that could be potentially reproduced by the proposed theoretical framework.

- Currently, there is a strong industrial interest in a better understanding of flow liquefaction in soils, which implies a complete strength loss and fluid behaviour. This phenomenon has produced economic, social, and environmental losses worldwide related to failure in tailings dams and natural slopes. Flow liquefaction can be studied from a bifurcation approach, and it can predict their instability under given conditions. In this context, applying the consistent failure criterion to the study of the localisation onset that triggers flow liquefaction represents a promising field of study.

Appendices

Appendix A

Cnoidal wave theory in solids

The cnoidal wave theory in solids [Veveakis & Regenauer-Lieb, 2015; Alevizos et al., 2017] represents a new approach of the localised deformation phenomenon from a wave mechanics framework.

For completeness, this section briefly recapitulates the formulation of the physical model behind (1.2), presented more in detail in Alevizos et al. [2017]. We consider a one-dimensional representative elementary volume (REV) of porous material under compression in the z direction. In this approach, the material is taken as homogeneous and all material properties are therefore constant. The sample of height H, under constant loading p'_n at its boundaries, is considered already past its limit of elasticity and we track its mean effective stress p' using the framework of overstress viscoplasticity by Perzyna [1966]. Using Terzaghi's definition of effective stress $p = p' + p_f$, with p the mean stress, taken positive in compression, and p_f the pore pressure, we can express the momentum balance in the z direction as

$$\frac{\partial p'}{\partial z} = -\frac{\partial p_f}{\partial z}. \tag{A.1}$$

Internal mass transfer is allowed between the solid and fluid phases through chemical reactions of dissolution/precipitation, which can be homogenized as a single effective reaction (see Alevizos et al. [2017]; Law [2006]) written generically as

$$AB_{(solid)} \rightleftharpoons A_{(solid)} + B_{(fluid)}. \tag{A.2}$$

Defining the solid and fluid phase densities as

$$\rho_1 = (1-\phi)\rho_s, \tag{A.3a}$$

$$\rho_2 = \phi\rho_f, \tag{A.3b}$$

where ϕ denotes the porosity, and ρ_s (resp. ρ_f) the solid (resp. fluid) density, the mass balance equations of the solid and fluid phases can be written as

$$\frac{\partial \rho_1}{\partial t} + \frac{\partial \left(\rho_1 v_z^{(1)}\right)}{\partial z} = -j, \tag{A.4a}$$

$$\frac{\partial \rho_2}{\partial t} + \frac{\partial \left(\rho_2 v_z^{(2)}\right)}{\partial z} = j, \tag{A.4b}$$

with j the mass rate of fluid produced by the chemical reaction (A.2) and $v_z^{(1)}$ and $v_z^{(2)}$ the velocities of phases 1 and 2 respectively. Combining (A.4) with Darcy's law for the filter velocity $\phi\left(v_k^{(2)} - v_k^{(1)}\right) = -\frac{k}{\mu}\frac{\partial p_f}{\partial z}$ (with constant permeability k and fluid viscosity μ) leads to the mass balance equation for the solid-fluid mixture [Veveakis et al., 2015]

$$-\frac{k}{\mu}\frac{\partial^2 p_f}{\partial z^2} + \dot{\varepsilon}_v = j\left(\frac{1}{\rho_f} - \frac{1}{\rho_s}\right), \tag{A.5}$$

where $\dot{\epsilon}_V$ denotes the volumetric strain rate. Combining (A.1) and (A.5), we obtain

$$\frac{k}{\mu}\frac{\partial^2 p'}{\partial z^2} + \dot{\varepsilon}_v = j\left(\frac{1}{\rho_f} - \frac{1}{\rho_s}\right). \tag{A.6}$$

The volumetric strain rate is then decomposed into its elastic and (visco)plastic components, ε_v^e and ε_v^{vp}, with the latter expressed through a typical power law rheology [Kohlstedt et al., 1995], under isothermal and overstress assumptions

$$\dot{\epsilon}_V = \dot{\epsilon}_V^e + \dot{\epsilon}_V^{vp} = -\frac{\dot{p'}}{K} - \dot{\epsilon}_n\left[\frac{p' - p'_Y}{p'_n - p'_Y}\right]^m, \tag{A.7}$$

where K is the bulk modulus, m the stress exponent, p_Y the yield value, p'_n the loading boundary conditions for $z \in \{0, H\}$ and $\dot{\epsilon}_n$ the corresponding loading strain rate. The negative signs match the sign convention of positive stresses in compression. Using the overstress definition $\bar{p} = p' - p'_Y$, with p'_Y constant, along

with (A.7), (A.6) becomes

$$\frac{k}{\mu}\frac{\partial^2 \bar{p}}{\partial z^2} - \frac{1}{K}\frac{\partial \bar{p}}{\partial t} - \dot{\varepsilon}_n \left[\frac{\bar{p}}{\bar{p}_n}\right]^m = j\left(\frac{1}{\rho_f} - \frac{1}{\rho_s}\right). \tag{A.8}$$

All variables can be normalized

$$\sigma = \frac{\bar{p}}{\bar{p}_n}, \quad \tau = \frac{kK}{\mu H^2}t, \quad z^* = \frac{z}{H}, \tag{A.9}$$

and following Alevizos et al. [2017], the rate of fluid production j follows an Arrhenius relationship with a dependence on mean pressure of the activation enthalpy. Assuming pressure-enhanced precipitation, it can be expressed as

$$j = -A\mathrm{e}^{\beta\sigma}, \tag{A.10}$$

where A is a coefficient and β is a chemo-mechanical parameter determining the interplay between the external loading and the mass exchange rate [Alevizos et al., 2017]. Equation (A.8) then gets rewritten in dimensionless form as

$$\frac{\partial \sigma}{\partial \tau} = \frac{\partial^2 \sigma}{\partial z^{*2}} - \lambda\sigma^m + \eta\mathrm{e}^{\beta\sigma}, \tag{A.11}$$

with $\lambda = \dfrac{\mu\dot{\varepsilon}_n}{k\bar{p}_n}H^2$ and $\eta = \dfrac{A\mu H^2}{k\bar{p}_n}\left(\dfrac{1}{\rho_f} - \dfrac{1}{\rho_s}\right)$. Dropping the asterisk and considering the stationary case $\partial/\partial t = 0$, we recover (1.2).

Appendix B

Stability in finite elements

In abstract setting, we look for a solution for a PDE of the following way:

$$\begin{cases} \text{Find } u \text{ in } \Omega, \text{ such that:} \\ \mathcal{L}u = f \qquad \text{in } \Omega, \\ \quad u = g \qquad \text{on } \partial\Omega, \end{cases} \tag{B.1}$$

where $\mathcal{L}$ represents a differential operator. Problem (B.1) is called the *strong form* of the PDE. Similarly, the *weak form* (also known as *variational formulation*) can be derived from (B.1) through using a proper test function v and applying integration by parts:

$$\begin{cases} \text{Find } u \in X, \text{ such that:} \\ b(u, v) = \ell(v), \quad \forall v \in Y, \end{cases} \tag{B.2}$$

In general, the variational formulation is well-posed if enjoys

1. Existence: There exists a solution satisfying the equation and each boundary condition.

2. Uniqueness: There is at most one solution.

3. Stability: The solution is a continuous function of the data, that is, small changes in the given data produces small changes in the solution.

Let $\mathscr{P}_h$ the partition of the domain Ω, such that $\cup_{T\in\mathscr{P}_h} = \Omega$. We can rewrite (B.2) in a discrete way as:

$$\begin{cases} \text{Find } u_h \in U_h, \text{ such that:} \\ b_h(u_h, v_h) = \ell_h(v_h), \quad \forall v_h \in U_h, \end{cases} \tag{B.3}$$

Similar to the abstract setting, the well-posedness of the FEM discretisation relies on the following property (cf. Strang & Fix [1973]):

1. Coercivity and consistency means convergence

Coercivity of a variational formulation is lost for extreme values of material parameters. Following Ern & Guermond [2013], let us consider a continuous, coercive, bilinear form $b_\eta(u, v)$ on $V \times V$, that satisfies (B.2). Besides, $b_\eta(u, v)$ is dependant on a parameter η which will take small values further. We set $\|b_\eta\| := \|b_\eta\|_{V,V}$. Coercivity reads:

$$b_\eta(u, u) \geq \alpha_\eta \|u\|_V^2 \tag{B.4}$$

where α_η is the coercivity constant of b_η. By definition, coercivity loss occurs in (B.2) if:

$$\lim_{\eta\to 0} \frac{\|b_\eta\|}{\alpha_\eta} = \infty. \tag{B.5}$$

For the disrete setting, let V_h be a V-conforming approximation space and assume that accomplishes the optimal interpolation property, i.e.

$$\forall u \in W, \quad \inf_{v_h\in V_h} \|u - u_h\|_V \leq c_i h^k \|u\|_W, \tag{B.6}$$

where W is a dense subspace of V and c_i is an interpolation constant. Morevover, let u_h solve the discrete problem:

$$\begin{cases} \text{Find } u_h \in V_h, \text{ such that:} \\ b_\eta(u_h, v_h) = \ell_h(v_h), \quad \forall v_h \in V_h, \end{cases} \tag{B.7}$$

Assuming that the exact solution u lives in W yields the error estimate:

$$\forall u \in W, \quad \|u - u_h\|_V \leq \frac{\|b_\eta\|}{\alpha_\eta} c_i h^k \|u\|_W. \tag{B.8}$$

If (B.2) suffers from coercivity loss, the error estimate does not ensure any practical control of the error, Unless that η remains constant and $h \to 0$ (very fine mesh), which is expensive and impractical. Advection-diffusion-reaction problems arise a wide range of phenomena relevant to many areas of applied physics and engineering, and their accurate and stable numerical solution has been the focus of intense research for several decades. The advection-diffusion-reaction problem reads:

$$\begin{cases} \text{Find } u \text{ in } \Omega, \text{ such that:} \\ \mathcal{L}u = -\kappa\Delta u + \beta \cdot \nabla u + \sigma u = f \qquad \text{in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad u = 0 \qquad \text{on } \partial\Omega, \end{cases} \tag{B.9}$$

where $\kappa \in L^\infty(\Omega)$ represents the diffusion term, $\beta \in [L^\infty(\Omega)]^d$ denotes an advection coefficient, $\sigma \in L^\infty(\Omega)$ is a reactive coefficient, and $f \in L^2(\Omega)$ denotes a spatial source.

Similar to (B.2), we consider the bilinear form $b_\eta(u, v)$ and linear form $\ell(v)$ as:

$$\begin{aligned} b_\eta(u, v) &:= (\kappa\nabla u, \nabla v) + (\beta \cdot \nabla u, \nabla v) + (\sigma u, v) \\ \ell_h(v) &:= (v, f) \end{aligned} \tag{B.10}$$

The paremeter η in this kind of problem reads:

$$\eta = \frac{\min(\|\kappa\|_\infty, \|\sigma\|_\infty)}{\|\beta\|_\ell}, \tag{B.11}$$

relating the advective, diffusive and reactive effects. For the advection-dominated regime, occurring when $\|\sigma\|_\infty, \|\kappa\|_\infty \ll \|\beta\|_\ell$, the parameter $\eta \ll 1$, which implies:

$$\frac{\|\beta_\eta\|}{\alpha_\eta} = \mathcal{O}\left(\frac{\|\beta\|_\ell}{\min(\|\kappa\|_\infty, \|\sigma\|_\infty)}\right) = \mathcal{O}\left(\frac{1}{\eta}\right) \ll 1 \tag{B.12}$$

leading to loss of coercivity in (B.2). Thus, in this regime, this equation develops non-physical oscillatory solutions on coarse meshes under a classical discrete formulation.

Appendix C

Acoustic tensor expansion

Considering index notation, we can express the acoustic tensor in a three-dimensional state as follows:

$$\mathbb{Q}_{jk} = \boldsymbol{n}_i \, \mathbb{C}_{ijkl} \, \boldsymbol{n}_l \tag{C.1}$$

$$\mathbb{Q}_{jk} = (\boldsymbol{n}_1 \, \mathbb{C}_{1jk1} + \boldsymbol{n}_2 \, \mathbb{C}_{2jk1} + \boldsymbol{n}_3 \, \mathbb{C}_{3jk1}) \cdot \boldsymbol{n}_1 + (\boldsymbol{n}_1 \, \mathbb{C}_{1jk2} + \boldsymbol{n}_2 \, \mathbb{C}_{2jk2} + \boldsymbol{n}_3 \, \mathbb{C}_{3jk2}) \cdot \boldsymbol{n}_2 \tag{C.2}$$

$$+ (\boldsymbol{n}_1 \, \mathbb{C}_{1jk3} + \boldsymbol{n}_2 \, \mathbb{C}_{2jk3} + \boldsymbol{n}_3 \, \mathbb{C}_{3jk3}) \cdot \boldsymbol{n}_3, \tag{C.3}$$

where each component is expressed by:

$$\mathbb{Q}_{11} = \boldsymbol{n}_1\mathbb{C}_{1111}\boldsymbol{n}_1 + \boldsymbol{n}_2\mathbb{C}_{2112}\boldsymbol{n}_2 + \boldsymbol{n}_3\mathbb{C}_{3113}\boldsymbol{n}_3 = \boldsymbol{n}_1\mathbb{C}_{11}\boldsymbol{n}_1 + \boldsymbol{n}_2\mathbb{C}_{44}\boldsymbol{n}_2 + \boldsymbol{n}_3\mathbb{C}_{55}\boldsymbol{n}_3 \ , \tag{C.4}$$

$$\mathbb{Q}_{12} = \boldsymbol{n}_2\mathbb{C}_{2121}\boldsymbol{n}_1 + \boldsymbol{n}_1\mathbb{C}_{1122}\boldsymbol{n}_2 = \boldsymbol{n}_2\mathbb{C}_{44}\boldsymbol{n}_1 + \boldsymbol{n}_1\mathbb{C}_{12}\boldsymbol{n}_2 \ , \tag{C.5}$$

$$\mathbb{Q}_{13} = \boldsymbol{n}_3\mathbb{C}_{3131}\boldsymbol{n}_1 + \boldsymbol{n}_1\mathbb{C}_{1133}\boldsymbol{n}_3 = \boldsymbol{n}_3\mathbb{C}_{55}\boldsymbol{n}_1 + \boldsymbol{n}_1\mathbb{C}_{13}\boldsymbol{n}_3, \tag{C.6}$$

$$\mathbb{Q}_{21} = \boldsymbol{n}_2\mathbb{C}_{2211}\boldsymbol{n}_1 + \boldsymbol{n}_1\mathbb{C}_{1212}\boldsymbol{n}_2 = \boldsymbol{n}_2\mathbb{C}_{21}\boldsymbol{n}_1 + \boldsymbol{n}_1\mathbb{C}_{44}\boldsymbol{n}_2, \tag{C.7}$$

$$\mathbb{Q}_{22} = \boldsymbol{n}_1\mathbb{C}_{1221}\boldsymbol{n}_1 + \boldsymbol{n}_2\mathbb{C}_{2222}\boldsymbol{n}_2 + \boldsymbol{n}_3\mathbb{C}_{3223}\boldsymbol{n}_3 = \boldsymbol{n}_1\mathbb{C}_{44}\boldsymbol{n}_1 + \boldsymbol{n}_2\mathbb{C}_{22}\boldsymbol{n}_2 + \boldsymbol{n}_3\mathbb{C}_{66}\boldsymbol{n}_3, \tag{C.8}$$

$$\mathbb{Q}_{23} = \boldsymbol{n}_3\mathbb{C}_{3232}\boldsymbol{n}_2 + \boldsymbol{n}_2\mathbb{C}_{2233}\boldsymbol{n}_3 = \boldsymbol{n}_3\mathbb{C}_{66}\boldsymbol{n}_2 + \boldsymbol{n}_2\mathbb{C}_{23}\boldsymbol{n}_3, \tag{C.9}$$

$$\mathbb{Q}_{31} = \boldsymbol{n}_3\mathbb{C}_{3311}\boldsymbol{n}_1 + \boldsymbol{n}_1\mathbb{C}_{1313}\boldsymbol{n}_3 = \boldsymbol{n}_3\mathbb{C}_{31}\boldsymbol{n}_1 + \boldsymbol{n}_1\mathbb{C}_{55}\boldsymbol{n}_3, \tag{C.10}$$

$$\mathbb{Q}_{32} = \boldsymbol{n}_3\mathbb{C}_{3322}\boldsymbol{n}_2 + \boldsymbol{n}_2\mathbb{C}_{2323}\boldsymbol{n}_3 = \boldsymbol{n}_3\mathbb{C}_{32}\boldsymbol{n}_2 + \boldsymbol{n}_2\mathbb{C}_{66}\boldsymbol{n}_3, \tag{C.11}$$

$$\mathbb{Q}_{33} = \boldsymbol{n}_1\mathbb{C}_{1331}\boldsymbol{n}_1 + \boldsymbol{n}_2\mathbb{C}_{2332}\boldsymbol{n}_2 + \boldsymbol{n}_3\mathbb{C}_{3333}\boldsymbol{n}_3 = \boldsymbol{n}_1\mathbb{C}_{55}\boldsymbol{n}_1 + \boldsymbol{n}_2\mathbb{C}_{66}\boldsymbol{n}_2 + \boldsymbol{n}_3\mathbb{C}_{33}\boldsymbol{n}_3. \tag{C.12}$$

Then, we derive the components of the $\mathbb{L}_{ij}$ tensor in the following way:

$$\mathbb{L}_{ij} = \mathbb{C}_{ijkl}\ \mathbf{N}_{kl} \tag{C.13}$$

$$\mathbb{L}_{ij} = \mathbb{C}_{ij11}\ \mathbf{N}_{11} + \mathbb{C}_{ij22}\ \mathbf{N}_{22} + \mathbb{C}_{ij33}\ \mathbf{N}_{33} + \mathbb{C}_{ij12}\ \mathbf{N}_{12} + \tag{C.14}$$

$$\mathbb{C}_{ij13}\ \mathbf{N}_{13} + \mathbb{C}_{ij23}\ \mathbf{N}_{23} + \mathbb{C}_{ij21}\ \mathbf{N}_{21} + \mathbb{C}_{ij31}\ \mathbf{N}_{31} + \mathbb{C}_{ij32}\ \mathbf{N}_{32}, \tag{C.15}$$

thus, the components read:

$$\mathbb{L}_{11} = \mathbb{C}_{11}\mathbf{N}_{11} + \mathbb{C}_{12}\mathbf{N}_{22} + \mathbb{C}_{13}\mathbf{N}_{33}\ , \tag{C.16}$$

$$\mathbb{L}_{12} = 2\mathbb{C}_{44}\mathbf{N}_{12}\ , \tag{C.17}$$

$$\mathbb{L}_{13} = 2\mathbb{C}_{55}\mathbf{N}_{13}, \tag{C.18}$$

$$\mathbb{L}_{21} = 2\mathbb{C}_{44}\mathbf{N}_{12}, \tag{C.19}$$

$$\mathbb{L}_{22} = \mathbb{C}_{12}\mathbf{N}_{11} + \mathbb{C}_{22}\mathbf{N}_{22} + \mathbb{C}_{13}\mathbf{N}_{33}, \tag{C.20}$$

$$\mathbb{L}_{23} = 2\mathbb{C}_{66}\mathbf{N}_{23}, \tag{C.21}$$

$$\mathbb{L}_{31} = 2\mathbb{C}_{55}\mathbf{N}_{13}, \tag{C.22}$$

$$\mathbb{L}_{32} = 2\mathbb{C}_{66}\mathbf{N}_{23}, \tag{C.23}$$

$$\mathbb{L}_{33} = \mathbb{C}_{13}\mathbf{N}_{11} + \mathbb{C}_{23}\mathbf{N}_{22} + \mathbb{C}_{33}\mathbf{N}_{33}. \tag{C.24}$$

Bibliography

Y. Abdallah, et al. (2021). 'Compaction Banding in High-Porosity Carbonate Rocks: 1. Experimental Observations'. *Journal of Geophysical Research: Solid Earth* **126**(1):e2020JB020538.

Y. Abdallah, et al. (2020). 'Compaction Banding in High-Porosity Carbonate Rocks: 2. A Gradient-Dependent Plasticity Model'. *Journal of Geophysical Research: Solid Earth* **125**(12):e2020JB020610.

S. Alevizos, et al. (2017). 'A Framework for Fracture Network Formation in Overpressurised Impermeable Shale: Deformability Versus Diagenesis'. *Rock Mechanics and Rock Engineering* **50**(3):689–703.

M. S. Alnæs, et al. (2015). 'The FEniCS project version 1.5'. *Archive of Numerical Software* **3**(100):9–23.

M. Anguiano, et al. (2020). 'Chemo-mechanical coupling and material evolution in finitely deforming solids with advancing fronts of reactive fluids'. *Acta Mechanica* **231**(5):1933–1961.

M. Anguiano, et al. (2022). 'Mixture model for thermo-chemo-mechanical processes in fluid-infused solids'. *International Journal of Engineering Science* **174**:103576.

D. N. Arnold (1982). 'An interior penalty finite element method with discontinuous elements'. *SIAM Journal on Numerical Analysis* **19**(4):742–760.

D. N. Arnold (2012). 'Lecture notes on numerical analysis of partial differential equations'.

D. N. Arnold, et al. (2002). 'Unified analysis of discontinuous Galerkin methods for elliptic problems'. *SIAM Journal on Numerical Analysis* **39**(5):1749–1779.

M. Arroyo, et al. (2005). 'Compaction bands and oedometric testing in cemented soils'. *Soils and foundations* **45**(2):181–194.

C. Bandle & H. Brunner (1998). 'Blowup in diffusion equations: a survey'. *Journal of Computational and Applied Mathematics* **97**(1-2):3–22.

R. E. Bank & D. J. Rose (1981). 'Global approximate Newton methods'. *Numerische Mathematik* **37**(2):279–295.

R. E. Bank, et al. (1983). 'Some refinement algorithms and data structures for regular local mesh refinement'. *Scientific Computing, Applications of Mathematics and Computing to the Physical Sciences* **1**:3–17.

R. E. Bank, et al. (1989). 'A class of iterative methods for solving saddle point problems'. *Numerische Mathematik* **56**(7):645–666.

P. Bastian, et al. (2012). 'Algebraic multigrid for discontinuous Galerkin discretizations of heterogeneous elliptic problems'. *Numerical Linear Algebra with Applications* **19**(2):367–388.

T. Baxevanis, et al. (2006). 'Compaction bands and induced permeability reduction in Tuffeau de Maastricht calcarenite'. *Acta Geotechnica* **1**(2):123–135.

Y. Bazilevs, et al. (2007). '$YZ\beta$ discontinuity capturing for advection-dominated processes with application to arterial drug delivery'. *International Journal for Numerical Methods in Fluids* **54**(6-8):593–608.

L. Beilina, et al. (2005). 'Nonobtuse tetrahedral partitions that refine locally towards Fichera-like corners'. *Applications of Mathematics* **50**(6):569–581.

M. Benzi, et al. (2005). 'Numerical solution of saddle point problems'. *Acta numerica* **14**:1.

P. Bésuelle (2001). 'Compacting and dilating shear bands in porous rock: Theoretical and experimental conditions'. *Journal of Geophysical Research: Solid Earth* **106**(B7):13435–13442.

L. Bjerrum (1967). 'Engineering geology of Norwegian normally-consolidated marine clays as related to settlements of buildings'. *Geotechnique* **17**(2):83–118.

P. B. Bochev & M. D. Gunzburger (2009). *Least-squares finite element methods*, vol. 166. Springer Science & Business Media.

R. I. Borja (2004). 'Computational modeling of deformation bands in granular media. II. Numerical simulations'. *Computer Methods in Applied Mechanics and Engineering* **193**(27-29):2699–2718.

R. I. Borja & A. Aydin (2004). 'Computational modeling of deformation bands in granular media. I. Geological and mathematical framework'. *Computer Methods in Applied Mechanics and Engineering* **193**(27-29):2667–2698.

R. I. Borja, et al. (2020). 'Cam-Clay plasticity. Part IX: On the anisotropy, heterogeneity, and viscoplasticity of shale'. *Computer Methods in Applied Mechanics and Engineering* **360**:112695.

F. Brezzi, et al. (2004). 'Discontinuous Galerkin methods for first-order hyperbolic problems'. *Mathematical Models and Methods in Applied Sciences* **14**(12):1893–1903.

A. N. Brooks & T. J. Hughes (1982). 'Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations'. *Computer Methods in Applied Mechanics and Engineering* **32**(1-3):199–259.

E. Burman & A. Ern (2005). 'Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence'. *Mathematics of Computation* **74**(252):1637–1652.

E. Burman & A. Ern (2007). 'Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations'. *Mathematics of computation* **76**(259):1119–1140.

E. Burman & A. Ern (2017). 'A nonlinear consistent penalty method weakly enforcing positivity in the finite element approximation of the transport equation'. *Computer Methods in Applied Mechanics and Engineering* **320**:122–132.

E. Burman & P. Hansbo (2004). 'Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems'. *Computer Methods in Applied Mechanics and Engineering* **193**(15-16):1437–1453.

E. Burman & P. Zunino (2006). 'A domain decomposition method based on weighted interior penalties for advection-diffusion-reaction problems'. *SIAM Journal on Numerical Analysis* **44**(4):1612–1638.

V. Calo, et al. (2008). 'Multiphysics model for blood flow and drug transport with application to patient-specific coronary artery flow'. *Computational Mechanics* **43**(1):161–177.

V. Calo, et al. (2011). 'A note on variational multiscale methods for high-contrast heterogeneous porous media flows with rough source terms'. *Advances in Water Resources* **34**(9):1177 – 1185. New Computational Methods and Software Tools.

V. M. Calo, et al. (2014a). 'Analysis of the discontinuous Petrov–Galerkin method with optimal test functions for the Reissner–Mindlin plate bending model'. *Computers & Mathematics with Applications* **66**(12):2570 – 2586.

V. M. Calo, et al. (2014b). 'Multiscale empirical interpolation for solving nonlinear PDEs'. *Journal of Computational Physics* **278**:204 – 220.

V. M. Calo, et al. (2016). 'Randomized Oversampling for Generalized Multiscale Finite Element Methods'. *Multiscale Modeling & Simulation* **14**(1):482–501.

V. M. Calo, et al. (2020). 'An adaptive stabilized conforming finite element method via residual minimization on dual discontinuous Galerkin norms'. *Computer Methods in Applied Mechanics and Engineering* **363**:112891.

V. M. Calo, et al. (2018). 'Automatic Variationally Stable Analysis for FE Computations: An Introduction'. *arXiv preprint arXiv:1808.01888* .

A. Carosio, et al. (2000). 'On the consistency of viscoplastic formulations'. *International Journal of Solids and Structures* **37**(48):7349–7369.

F. Cecinato & A. Gajo (2014). 'Dynamical effects during compaction band formation affecting their spatial periodicity'. *Journal of Geophysical Research: Solid Earth* **119**(10):7487–7502.

J. Chan & J. A. Evans (2013). 'A minimum-residual finite element method for the convection-diffusion equation'. Tech. rep., Texas Univ at Austin Institute for Computational Engineering and Sciences.

Y. Chen, et al. (2008). 'Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate'. *ACM Transactions on Mathematical Software (TOMS)* **35**(3):22.

R. J. Cier, et al. (2021). 'Automatically adaptive stabilized finite elements and continuation analysis for compaction banding in geomaterials'. *International Journal for Numerical Methods in Engineering* **122**(21):6234–6252.

R. J. Cier, et al. (2020). 'A nonlinear weak constraint enforcement method for advection-dominated diffusion problems'. *Mechanics Research Communications* p. 103602.

B. Cockburn, et al. (2012). *Discontinuous Galerkin methods: theory, computation and applications*, vol. 11. Springer Science & Business Media.

R. Codina (1998). 'Comparison of some finite element methods for solving the diffusion-convection-reaction equation'. *Computer methods in applied mechanics and engineering* **156**(1-4):185–210.

R. Codina (2000). 'On Stabilized Finite Element Methods for Linear Systems of Convection-Diffusion-Reaction Equations'. *Computer Methods in Applied Mechanics and Engineering* **188**(1-3):61–82.

A. Cohen, et al. (2012). 'Adaptivity and variational stabilization for convection-diffusion equations'. *ESAIM: Mathematical Modelling and Numerical Analysis* **46**(5):1247–1273.

A. Das & G. Buscarnera (2014). 'Simulation of localized compaction in high-porosity calcarenite subjected to boundary constraints'. *International Journal of Rock Mechanics and Mining Sciences* **71**:91–104.

A. Das, et al. (2013). 'The propagation of compaction bands in porous rocks based on breakage mechanics'. *Journal of Geophysical Research: Solid Earth* **118**(5):2049–2066.

L. Demkowicz & J. Gopalakrishnan (2010). 'A class of discontinuous Petrov–Galerkin methods. Part I: The transport equation'. *Computer Methods in Applied Mechanics and Engineering* **199**(23-24):1558–1572.

L. Demkowicz & J. Gopalakrishnan (2011). 'A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions'. *Numerical Methods for Partial Differential Equations* **27**(1):70–105.

L. Demkowicz & J. Gopalakrishnan (2014). 'An overview of the Discontinuous Petrov Galerkin method'. In X. Feng, O. Karakashian, & Y. Xing (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations: 2012 John H Barrett Memorial Lectures*, vol. 157 of *The IMA Volumes in Mathematics and its Applications*, pp. 149–180. Springer, Cham.

L. Demkowicz, et al. (2012). 'A class of discontinuous Petrov–Galerkin methods. Part III: Adaptivity'. *Applied Numerical Mathematics* **62**(4):396 – 427. Third Chilean Workshop on Numerical Analysis of Partial Differential Equations (WONAPDE 2010).

D. A. Di Pietro & A. Ern (2012). *Mathematical aspects of discontinuous Galerkin methods*, vol. 69. Springer Science.

D. A. Di Pietro, et al. (2008). 'Discontinuous Galerkin methods for anisotropic semidefinite diffusion with advection'. *SIAM Journal on Numerical Analysis* **46**(2):805–831.

W. Dörfler (1996). 'A convergent adaptive algorithm for Poisson's equation'. *SIAM Journal on Numerical Analysis* **33**(3):1106–1124.

G. Duvaut & J. Lions (1982). *Les Inequations en Mécanique et en Physique*. Dunod, Paris.

Y. Epshteyn & B. Rivière (2007). 'Estimation of penalty parameters for symmetric interior penalty Galerkin methods'. *Journal of Computational and Applied Mathematics* **206**(2):843–872.

A. Ern & J.-L. Guermond (2006). 'Discontinuous Galerkin Methods for Friedrichs' Systems. I. General theory'. *SIAM Journal on Numerical Analysis* **44**(2):753–778.

A. Ern & J.-L. Guermond (2013). *Theory and practice of finite elements*, vol. 159. Springer Science & Business Media.

A. Ern & J.-L. Guermond (2017). 'Finite element quasi-interpolation and best approximation'. *ESAIM: Mathematical Modelling and Numerical Analysis* **51**(4):1367–1385.

A. Ern, et al. (2009). 'A discontinuous Galerkin method with weighted averages for advection–diffusion equations with locally small and anisotropic diffusivity'. *IMA Journal of Numerical Analysis* **29**(2):235–256.

R. E. Ewing & H. Wang (2001). 'A summary of numerical methods for time-dependent advection-dominated partial differential equations'. *Journal of Computational and Applied Mathematics* **128**(1-2):423–445.

J. Fortin, et al. (2006). 'Acoustic emission and velocities associated with the formation of compaction bands in sandstone'. *Journal of Geophysical Research: Solid Earth* **111**(B10).

H. Gajendran, et al. (2018). 'Chemo-mechanical coupling in curing and material-interphase evolution in multi-constituent materials'. *Acta Mechanica* **229**(8):3393–3414.

V. A. Galaktionov & J. L. Vazquez (1995). 'Necessary and sufficient conditions for complete blow-up and extinction for one-dimensional quasilinear heat equations'. *Archive for rational mechanics and analysis* **129**(3):225–244.

V. A. Galaktionov & J.-L. Vázquez (2002). 'The problem of blow-up in nonlinear parabolic equations'. *Discrete & Continuous Dynamical Systems* **8**(2):399.

J. Galvis, et al. (2018). 'On Overlapping Domain Decomposition Methods for High-Contrast Multiscale Problems'. In B. P. et al. (ed.), *Domain Decomposition Methods in Science and Engineering XXIV. DD 2017*, vol. 125 of *Lecture Notes in Computational Science and Engineering*. Springer.

A. Garavand, et al. (2020). 'Numerical modeling of plastic deformation and failure around a wellbore in compaction and dilation modes'. *International Journal for Numerical and Analytical Methods in Geomechanics* **44**(6):823–850.

A. Ghisi, et al. (2021). 'Consistent Implicit Time Integration for Viscoplastic Modelingof Subsidence above Hydrocarbon Reservoirs'. *Applied Sciences* **11**(8):3513.

J. F. Giraldo & V. Calo (2022). 'Residual minimisation as a variational multiscale method in singularly perturbated problems'. *In preparation* .

M. Gutierrez (2017). 'Rigorous comparison of the Rudnicki-Rice and Vermeer bifurcation criteria'. *Springer Series in Geomechanics and Geoengineering* **143**(191289):177–183.

M. Hajipour, et al. (2018). 'On the accurate discretization of a highly nonlinear boundary value problem'. *Numerical Algorithms* **79**(3):679–695.

R. Hartmann & P. Houston (2008). 'An optimal order interior penalty discontinuous Galerkin discretization of the compressible Navier–Stokes equations'. *Journal of Computational Physics* **227**(22):9670–9685.

J. Hasbani, et al. (2021). 'Adaptive stabilized finite elements via residual minimization onto bubble enrichments'. *In preparation* .

N. Hayward, et al. (2018). 'Spatial Periodicity in Self-Organized Ore Systems'. In *Metals, Minerals, and Society*, vol. 21 of *SEG Special Publications*, pp. 1–24. Society of Economic Geologists (SEG).

O. M. Heeres, et al. (2002). 'A comparison between the Perzyna viscoplastic model and the Consistency viscoplastic model'. *European Journal of Mechanics - A/Solids* **21**(1):1–12.

P.-Y. Hicher, et al. (1994). 'Microstructural analysis of strain localisation in clay'. *Computers and geotechnics* **16**(3):205–222.

R. Hill (1950). 'The mathematical theory of plasticity'. *Oxford: The Clarendon Press* **613**:614.

R. Hill (1958). 'A general theory of uniqueness and stability in elastic-plastic solids'. *Journal of the Mechanics and Physics of Solids* **6**(3):236–249.

D. J. Holcomb & W. A. Olsson (2003). 'Compaction localization and fluid flow'. *Journal of Geophysical Research: Solid Earth* **108**(B6).

S. S. Hossain, et al. (2012). 'Mathematical modeling of coupled drug and drug-encapsulated nanoparticle transport in patient-specific coronary artery walls'. *Computational Mechanics* **49**(2):213–242.

P. Houston, et al. (2020). 'Eliminating Gibbs phenomena: A non-linear Petrov–Galerkin method for the convection–diffusion–reaction equation'. *Computers & Mathematics with Applications* **80**(5):851–873.

B. Hu (2011). *Blow-up theories for semilinear parabolic equations*. Springer.

T. J. Hughes, et al. (1989). 'A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations'. *Computer Methods in Applied Mechanics and Engineering* **73**(2):173–189.

T. J. Hughes & G. Sangalli (2007). 'Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods'. *SIAM Journal on Numerical Analysis* **45**(2):539–557.

T. J. Hughes, et al. (2018). 'Multiscale and stabilized methods'. *Encyclopedia of Computational Mechanics Second Edition* pp. 1–64.

A. Hüpers & A. J. Kopf (2012). 'Data report: Consolidation properties of silty claystones and sandstones sampled seaward of the Nankai Trough subduction zone, IODP Sites C0011 and C0012'. In *Proc. IODP— Volume*, vol. 322, p. 2.

M. A. Iophis, et al. (2007). 'Experimental investigation of spatial periodicity of induced deformations in a rock mass'. *Journal of Mining Science* **43**(2):125–131.

K. Issen & J. W. Rudnicki (2001). 'Theory of compaction bands in porous rock'. *Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy* **26**(1-2):95–100.

C. Johnson & J. Pitkäranta (1986). 'An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation'. *Mathematics of Computation* **46**(173):1–26.

O. A. Karakashian & F. Pascal (2003). 'A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems'. *SIAM Journal on Numerical Analysis* **41**(6):2374–2399.

U. Kelka, et al. (2017). 'Zebra rocks: compaction waves create ore deposits'. *Scientific reports* **7**(1):14260.

M. K. Kim & P. V. Lade (1988). 'Single hardening constitutive model for frictional materials: I. Plastic potential function'. *Computers and Geotechnics* **5**(4):307–324.

D. L. Kohlstedt, et al. (1995). 'Strength of the lithosphere: Constraints imposed by laboratory experiments'. *Journal of Geophysical Research: Solid Earth* **100**(B9):17587–17602.

D. J. Korteweg & G. De Vries (1895). 'XLI. On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves'. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **39**(240):422–443.

D. Kuzmin & M. Möller (2010). 'Goal-oriented mesh adaptation for flux-limited approximations to steady hyperbolic problems'. *Journal of Computational and Applied Mathematics* **233**(12):3113–3120.

D. Kuzmin & S. Turek (2002). 'Flux correction tools for finite elements'. *Journal of Computational Physics* **175**(2):525–558.

N. A. Labanda, et al. (2022). 'A spatio-temporal adaptive phase-field fracture method'. *Computer Methods in Applied Mechanics and Engineering* **392**:114675.

C. K. Law (2006). *Combustion Physics*. Cambridge University Press, Cambridge.

P. Lesaint & P.-A. Raviart (1974). 'On a finite element method for solving the neutron transport equation'. *Publications mathématiques et informatique de Rennes* (S4):1–40.

J. Leuthold, et al. (2021). 'Effect of Compaction Banding on the Hydraulic Properties of Porous Rock: Part I—Experimental Investigation'. *Rock Mechanics and Rock Engineering* pp. 1–13.

R. Löhner, et al. (1987). 'Finite element flux-corrected transport (FEM–FCT) for the euler and Navier–Stokes equations'. *International Journal for Numerical Methods in Fluids* **7**(10):1093–1109.

A. Masud & A. A. Al-Naseem (2018). 'Variationally derived discontinuity capturing methods: Fine scale models with embedded weak and strong discontinuities'. *Computer Methods in Applied Mechanics and Engineering* **340**:1102–1134.

A. Masud, et al. (2020). 'Modeling of steep layers in singularly perturbed diffusion–reaction equation via flexible fine-scale basis'. *Computer Methods in Applied Mechanics and Engineering* **372**:113343.

A. Masud & T. J. Truster (2013). 'A framework for residual-based stabilization of incompressible finite elasticity: Stabilized formulations and F methods for linear triangles and tetrahedra'. *Computer Methods in Applied Mechanics and Engineering* **267**:359–399.

A. Masud & K. Xia (2006). 'A variational multiscale method for inelasticity: Application to superelasticity in shape memory alloys'. *Computer methods in applied mechanics and engineering* **195**(33-36):4512–4531.

W. F. Mitchell (2013). 'A collection of 2D elliptic problems for testing adaptive grid refinement algorithms'. *Applied Mathematics and Computation* **220**:350–364.

A. Mizukami & T. J. Hughes (1985). 'A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle'. *Computer Methods in Applied Mechanics and Engineering* **50**(2):181–193.

P. N. Mollema & M. A. Antonellini (1996). 'Compaction bands: A structural analog for anti-mode i cracks in aeolian sandstone'. *Tectonophysics* **267**(1-4):209–228.

G. Moore & A. Spence (1980). 'The calculation of turning points of nonlinear equations'. *SIAM Journal on Numerical Analysis* **17**(4):567–576.

I. Muga, et al. (2019). 'The discrete-dual minimal-residual method (DDMRes) for weak advection-reaction problems in Banach spaces'. *Computational Methods in Applied Mathematics* **19**(3):557–579.

A. Niemi, et al. (2011a). 'Discontinuous Petrov-Galerkin method with optimal test functions for thin-body problems in solid mechanics'. *Computer Methods in Applied Mechanics and Engineering* **200**(9-12):1291–1300.

A. H. Niemi, et al. (2011b). 'Discontinuous Petrov-Galerkin method based on the optimal test space norm for one-dimensional transport problems'. *Procedia Computer Science* **4**:1862 – 1869. Proceedings of the International Conference on Computational Science, ICCS 2011.

A. H. Niemi, et al. (2013). 'Automatically stable discontinuous Petrov–Galerkin methods for stationary transport problems: Quasi-optimal test space norm'. *Computers & Mathematics with Applications* **66**(10):2096 – 2113. ICNC-FSKD 2012.

R. Nova (1994). 'Controllability of the incremental response of soil specimens subjected to arbitrary loading programmes'. *Journal of the Mechanical behavior of Materials* **5**(2):193–202.

J. T. Oden & L. Demkowicz (2017). *Applied functional analysis*. CRC press.

J. T. Oden & A. Patra (1995). 'A parallel adaptive strategy for hp finite element computations'. *Computer Methods in Applied Mechanics and Engineering* **121**(1-4):449–470.

F. Oka, et al. (2011). 'An elasto-viscoplastic model for diatomaceous mudstone and numerical simulation of compaction bands'. *International Journal for Numerical and Analytical Methods in Geomechanics* **35**(2):244–263.

W. A. Olsson (1999). 'Theoretical and experimental investigation of compaction bands in porous rock'. *Journal of Geophysical Research* **104**(10):7219–7228.

P. Perzyna (1966). 'Fundamental Problems in Viscoplasticity'. *Advances in Applied Mechanics* **9**:244 – 377.

F. Pisano & C. d. Prisco (2016). 'A stability criterion for elasto-viscoplastic constitutive relationships'. *International Journal for Numerical and Analytical Methods in Geomechanics* **40**(1):141–156.

T. Poulet, et al. (2017). 'Multi-physics modelling of fault mechanics using REDBACK: a parallel open-source simulator for tightly coupled problems'. *Rock Mechanics and Rock Engineering* **50**(3):733–749.

M. Rabinowicz & J.-L. Vigneresse (2004). 'Melt segregation under compaction and shear channeling: Application to granitic magma segregation in a continental crust'. *Journal of Geophysical Research: Solid Earth* **109**(B4).

W. H. Reed & T. Hill (1973). 'Triangular mesh methods for the neutron transport equation'. Tech. rep., Los Alamos Scientific Lab., N. Mex.(USA).

K. Regenauer-Lieb, et al. (2016). 'A novel wave-mechanics approach for fluid flow in unconventional resources'. *The Leading Edge* **35**(1):90–97.

K. Regenauer-Lieb, et al. (2013). 'Multiscale coupling and multiphysics approaches in earth sciences: Applications'. *Journal of Coupled Systems and Multiscale Dynamics* **1**(3):281–323.

J. Rice (1976). 'The localization of plastic deformation'. *Theoretical and Applied Mechanics, 14th IUTAM Congress* p. 207–220.

J. Rice & J. Rudnicki (1980). 'A note on some features of the theory of localization of deformation' **16**:597–605.

M.-C. Rivara (1984). 'Mesh refinement processes based on the generalized bisection of simplices'. *SIAM Journal on Numerical Analysis* **21**(3):604–613.

K. Roscoe & J. Burland (1968). 'On the generalized stress-strain behaviour of wet clay' .

K. H. Roscoe, et al. (1958). 'On the yielding of soils'. *Geotechnique* **8**(1):22–53.

J. W. Rudnicki & J. Rice (1975). 'Conditions for the localization of deformation in pressure-sensitive dilatant materials'. *Journal of the Mechanics and Physics of Solids* **23**(6):371–394.

J. S. Russell (1844). 'Report on waves'. In *14th Meeting of the British Association for the Advancement of Science*, pp. 311–390.

M. Sari, et al. (2022). 'The Brittle–Ductile Transition and the Formation of Compaction Bands in the Savonnières Limestone: Impact of the Stress and Pore Fluid'. *Rock Mechanics and Rock Engineering* .

K. Shahbazi (2005). 'An explicit expression for the penalty parameter of the interior penalty method'. *Journal of Computational Physics* **205**(2):401–407.

G. Shahin, et al. (2019). 'Viscoplastic interpretation of localized compaction creep in porous rock'. *Journal of Geophysical Research: Solid Earth* **124**(10):10180–10196.

J. Simo & R. Taylor (1985). 'Consistent tangent operators for rate-independent elastoplasticity'. *Computer Methods in Applied Mechanics and Engineering* **48**(1):101–118.

G. Strang & G. J. Fix (1973). 'An analysis of the finite element method'. *Englewood Cliffs, N. J., Prentice-Hall, Inc., 1973. 318 p* .

R. Temam (2001). *Navier-Stokes equations: theory and numerical analysis*, vol. 343. American Mathematical Soc.

T. Thomas (1961). 'Plastic flow and fracture in solids'. *Academic Press,New York* p. 597–605.

I. Vardoulakis (2000). 'Regularization of Strain Softening Models for Geomaterials'. *Journal of the Mechanical Behavior of Materials* **11**(1-3):227–236.

K. Vemaganti (2007). 'Discontinuous Galerkin methods for periodic boundary value problems'. *Numerical Methods for Partial Differential Equations: An International Journal* **23**(3):587–596.

P. Vermeer (1982). 'A simple shear-band analysis using compliances' .

P. Vermeer & H. Neher (1999). 'A soft soil model that accounts for creep'. In *Beyond 2000 in computational geotechnics*, pp. 249–261. Routledge.

E. Veveakis & K. Regenauer-Lieb (2015). 'Cnoidal waves in solids'. *Journal of the Mechanics and Physics of Solids* **78**:231–248.

E. Veveakis, et al. (2015). 'Ductile compaction of partially molten rocks: The effect of non-linear viscous rheology on instability and segregation'. *Geophysical Journal International* **200**(1):519–523.

G. Volonté, et al. (2017). 'Advances in geomechanical subsidence modeling: effects of elasto-visco-plastic constitutive behavior.'. In *51st US Rock Mechanics/Geomechanics Symposium*. OnePetro.

W. M. Wang, et al. (1997). 'Viscoplasticity for instabilities due to strain softening and strain-rate softening'. *International Journal for Numerical Methods in Engineering* **40**(20):3839–3864.

L. Wei & G. Wang (2022). 'Deformation pattern and permeability change of compacted clay under triaxial compression'. *Bulletin of Engineering Geology and the Environment* **81**(5):1–10.

M. F. Wheeler (1978). 'An elliptic collocation-finite element method with interior penalties'. *SIAM Journal on Numerical Analysis* **15**(1):152–161.

S. Wolfram et al. (1999). *The MATHEMATICA® book, version 4*. Cambridge University Press.

H. Wu, et al. (2018). 'Multiscale modeling and analysis of compaction bands in high-porosity sandstones'. *Acta Geotechnica* **13**(3):575–599.

H. Wu, et al. (2020). 'Compaction bands in Tuffeau de Maastricht: insights from X-ray tomography and multiscale modeling'. *Acta Geotechnica* **15**(1):39–55.

K. Xia & A. Masud (2009). 'A stabilized finite element formulation for finite deformation elastoplasticity in geomechanics'. *Computers and Geotechnics* **36**(3):396–405.

T. Zeng, et al. (2020). 'A micromechanical-based elasto-viscoplastic model for the Callovo-Oxfordian argillite: Algorithms, validations, and applications'. *International Journal for Numerical and Analytical Methods in Geomechanics* **44**(2):183–207.