Advanced polyhedral discretization methods for poromechanical modelling

Michele Botti

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Mathématiques et Modélisation

École doctorale : Information, Structures, Systèmes

Unité de recherche : Institut Montpelliérain Alexander Grothendieck

Advanced polyhedral discretization methods for poromechanical modelling

Présentée par Michele BOTTI

Le 27 Novembre 2018

Sous la direction de Daniele DI PIETRO

Devant le jury composé de

BEIRÃO DA VEIGA Lourenço, Professeur des universités, Università di Milano-Bicocca	Rapporteur
BOFFI Daniele, Professeur des universités, Università di Pavia	Examinateur
DI PIETRO Daniele, Professeur des universités, Université de Montpellier	Directeur
GALLOUËT Thierry, Professeur des universités, Institut de Mathématiques de Marseille	Rapporteur
SOCHALA Pierre, Ingénieur, BRGM	Co-encadrant
VOHRALÍK Martin, Directeur de recherche, INRIA Paris	Président du jury

UNIVERSITÉ DE MONTPELLIER

Acknowledgements

Sono profondamente grato a Daniele A. Di Pietro per essere stato un eccezionale direttore di tesi durante l'intero svolgersi del dottorato. Ammiro la tua passione e dedizione per la ricerca e sono onorato di aver lavorato insieme a questa tesi. Ti ringrazio per avermi sempre manifestato fiducia ed avermi incoraggiato durante tutto il percorso. La tua disponibilità e il tuo supporto sono stati di grandissimo aiuto per raggiungere al meglio questo traguardo.

I would like to thank my second supervisor, Pierre Sochala, for welcoming me at BRGM within the best conditions. Thank you for your kindness, your listening, your passion and for all the time we have spent together. It has been a great pleasure to work with you.

I would also like to express my sincere gratitude to Lourenço Beirão da Veiga and Thierry Gallouët, who have accepted the heavy task of being referees, for the interest they have shown in my work, and their pertinent suggestions. I am also very grateful to Daniele Boffi and Martin Vohralik for having accepted to be members of the PhD committee.

I would like to thank all the members of the SPU research team that I met during the 9 months at the BRGM. I am particularly grateful to André Burnol, Farid Smaï, Fernanda M.L. Veloso, and Hideo Aochi for the fruitful discussions about poromechanics and reservoir modelling. I also express my special thanks to Olivier Le Maître for his collaboration in the results of the last chapter of this manuscript. I really appreciate your great intuitions and clear explanations.

At the IMAG, I would like to thank the researchers and professors with whom I had the opportunity to exchange and work with, especially Berardo Ruffini, Michele Bolognesi, Matthieu Alfaro, Philippe Roche and Vanessa Lleras. I am also grateful to the administrative staff, including Bernadette Lacan, Geneviève Carrière, and Carmela Madonia for their efficiency and patience. Many thanks also to the PhD and post-doctoral students met during the years in Montpellier, especially Abel, Alexandre, Francesco, Jocelyn, Joubine, Mario, Nestor, Paul, Paul-Marie, Rita, Robert, Rodrigo, and Tiffany. I am also particularly grateful to Alice, Fatima, Roberta for the good moments we shared in our office. Finally, I would like to express a huge thank to Florent for being a true friend during these three years, not only for the fantastic time we had while traveling around the world, but also for being close to me when I needed it most.

Questi tre anni di dottorato mi hanno permesso di incontrare tante altre persone stupende che desidero ringraziare perché, a modo loro, hanno fortemente contribuito alla realizzazione di questa tesi. Ringrazio Florent, Alex, Tim, Paul e Alec per avermi accolto al mio arrivo all'IMAG ed avermi fatto sentire sin da subito a casa. Un ringraziamento speciale va anche a tutti i coinquilini dell'appartamento di *Rue Rondelet*: Gregoire, Jean-Florent, Claire, Swan

e Pierre-Alexis. Condividere le cene, le serate, i weekend con voi ha riempito di gioia la mia esperienza a Montpellier.

Desidero ringraziare gli amici dello *Sborgomeo*: Rumi, Stiv, Max, Maffo, Zenk, Andre e Susi; che sono sempre al mio fianco. I momenti condivisi con voi li porterò nel cuore. Un grande grazie anche agli *Sugarcandy Mountains*: Cano, Jaya, Oscar, J, Zenk e Zoia. I concerti e le prove insieme sono per me un fondamentale momento di amicizia e divertimento. Sono felice che il nostro progetto musicale continui. Scusandomi con chi non ho menzionato, ringrazio anche tutti coloro che, nonostante la lontananza, mi sono stati vicini manifestandomi il loro affetto e la loro amicizia.

Sarò eternamente riconoscente ai miei genitori per avermi fatto sentire il loro amore anche a centinaia di chilometri di distanza. Vi sono grato per tutto ciò che avete fatto e continuate a fare per me, per gli insegnamenti ed i valori che mi avete trasmesso. Grazie mamma per il tuo sesto senso capace di sentire quando qualche nuvolone grigio fa capolino nel mio cielo anche quando cerco di nasconderlo. Grazie papà per il tuo cuore immenso e per avermi insegnato la gioia del dare agli altri. Grazie Mamie per esserti sempre preoccupata per me. Grazie Stefano per la tua vicinanza, la tua energia e la voglia di passare dei bei momenti insieme. Grazie Lorenzo per tutte le ore trascorse a confrontarci sulle nostre teorie e riflessioni. Sono davvero felice di condividere con te la passione per la matematica. Ringrazio immensamente Anna e Martine per il loro supporto e per esserci per celebrare questo importante traguardo.

Concludo ringraziando la persona con cui più di tutte ho condiviso le soddisfazioni e le fatiche di questo dottorato. Grazie Claudia per esserci sempre stata con il tuo prezioso sostegno e per tutta la bellezza e l'amore che mi hai regalato in questi anni.

Résumé

Dans cette thèse, on s'intéresse à de nouveaux schémas de discrétisation afin de résoudre les équations couplées de la poroélasticité et nous présentons des résultats analytiques et numériques concernant des problèmes issus de la poromécanique. Nous proposons de résoudre ces problèmes en utilisant les méthodes Hybrid High-Order (HHO), une nouvelle classe de méthodes de discrétisation polyédriques d'ordre arbitraire. Cette thèse a été conjointement financée par le Bureau de Recherches Géologiques et Minières (BRGM) et le LabEx NUMEV. Le couplage entre l'écoulement souterrain et la déformation géomécanique est un sujet de recherche crucial pour les deux institutions de cofinancement.

Contexte et motivations

On prend en considération des matériaux constitués d'un squelette solide et d'un réseau de pores, connectés entre eux et permettant le passage d'un fluide interstitiel. Le comportement de ce fluide peut influer sur celui du squelette et réciproquement. De nombreux matériaux présentent cette caractéristique : les matériaux minéraux comme les sols et les roches, des matériaux organiques comme le bois, les os ou le cerveau et les matériaux industriels comme certains joints d'étanchéité. Dans cette thèse, l'accent porte sur la modélisation des sols et des roches. Le milieu poreux est vu comme la superposition de deux milieux continus : le squelette solide déformable et la phase fluide. Le milieu poreux possède une cinématique déterminée par le squelette et les sollicitations considérées peuvent être de type mécanique, hydraulique et thermique. Les sollicitations mécaniques peuvent être causées par des contraintes imposées, telles que le creusement d'ouvrages, ou bien par des contraintes intrinsèques, telles que la consolidation du sol par gravité. Les sollicitations hydrauliques sont liées à l'écoulement du fluide à travers le milieu, tandis que les sollicitations thermiques peuvent être dues à la présence de sources de chaleur. Lorsque les trois phénomènes sont traités de manière couplée, le cadre est celui dit de la thermo-hydro-mécanique. Dans cette thèse, on s'intéresse avant tout au modèle hydro-mécanique saturé obtenu en supposant que les variations de température soient négligeables.

L'intérêt pour les mécanismes couplés de diffusion-déformation était initialement motivé par le problème de la consolidation [27, 189], à savoir, le tassement progressif du sol dû à l'extraction des fluides. Aujourd'hui la théorie de la poroélasticité présente un intérêt pour les scientifiques et les ingénieurs en raison de son potentiel d'applications dans la mécanique des sols, l'industrie pétrolière, la géomécanique environnementale et la biomécanique. En simulation de réservoir, le couplage mécanique-écoulement joue un rôle important pour l'étude des problèmes de compaction et subsidence induits par la mise en production de

réservoirs peu consolidés, pour la stabilité des puits, ou encore la fracturation hydraulique. La non prise en compte de ce couplage peut aussi conduire à de mauvaises prédictions de la production. Le couplage poroélastic est par ailleurs crucial pour l'étude des risques liés à l'injection et au stockage du CO_2 dans les aquifères salins, comme la fuite du CO_2 par le puits ou bien la réactivation mécanique des failles. La surpression entraînée par l'injection peut aussi modifier les contraintes appliquées sur la roche couverture et provoquer une fracturation qui remettrait en cause l'intégrité du stockage.

Dans cette thèse, nous considérons principalement trois problèmes liées à la modélisation de la géomécanique : le problème couplé de Biot linéaire dans le Chapitre 2, le comportement non-linéaire en mécanique et poromécanique dans les chapitres 3 et 4 et la poroélasticité avec coefficients aléatoires dans le Chapitre 5. Comme noté dans [68], de nombreux résultats expérimentaux suggèrent que la réponse volumétrique de la roche poreuse au changement de la pression totale est en réalité non-linéaire. Cette non-linéarité est généralement associée à la fermeture/ouverture de fissures ou, dans les roches très poreuses, à l'effondrement progressif des pores. Les études physiques du comportement mécanique non-linéaire des milieux poreux [22, 28] ont été motivées par la nécessité de quantifier l'effet de la diminution de la pression interstitielle lors de l'épuisement d'un gisement de pétrole ou de gaz sur le volume de la matrice rocheuse.

En simulation numérique, la prise en compte des incertitudes dans les données d'entrée (telles que les constantes physiques, les conditions limites et initiales, le forçage externe et la géométrie) est un problème crucial, en particulier dans l'analyse des risques. L'évaluation des risques dans les applications de poroélasticité reste un défi majeur pour un large éventail d'applications de gestion des ressources essentielles. Les préoccupations récentes de la population concernant la sismicité induite et la contamination des eaux souterraines soulignent la nécessité de quantifier la probabilité d'événements nocifs associés à l'écoulement et à la déformation de la subsurface. La prédiction des contraintes critiques qui peuvent compromettre l'intégrité des roches couvertures et la stabilité des puits dans les réservoirs de stockage présente un intérêt particulier. La clé de la réussite de l'évaluation des risques dans les applications poroélastiques est la capacité de prédire la probabilité d'événements critiques en se basant sur une approximation empirique des paramètres matériels et de la variabilité du terme source. La quantification d'incertitude fournit des barres d'erreur numériques qui facilitent la comparaison avec l'observation expérimentale et l'évaluation des modèles physiques. De plus, elle permet d'identifier les paramètres incertains qui doivent être mesurés avec plus de précision car ils ont un impact plus significatif sur la solution. Parmi les différentes techniques conçues pour la propagation et la quantification de l'incertitude dans les modèles numériques, dans cette thèse nous considérons les méthodes spectrales stochastiques. L'idée centrale de ces méthodes est la décomposition de quantités aléatoires sur des bases d'approximation appropriées telles que les expansions en Polynômes de Chaos [112, 137].

Discrétisation

Ces dernières années, un effort important a été consacré au développement et à l'analyse de méthodes numériques capables de traiter des maillages plus généraux que les maillages simpliciaux ou cartésiens. Dans le contexte de la poroélasticité, cette nécessité est motivée par

la présence des couches géologiques et des fractures dans le milieu poreux. La discrétisation du domaine peut également inclure des éléments dégénérés (comme dans la région proche du puits dans la modélisation des réservoirs) et des interfaces non-conformes apparaissant lors de la modélisation de l'érosion et de la formation de failles. Dans le contexte de la mécanique des structures, les méthodes supportant les maillages polyédriques généraux peuvent être utiles pour plusieurs raisons, notamment l'utilisation de nœuds suspendus pour les problèmes de contact et d'interface et la robustesse par rapport aux distorsions et fractures. De plus, les méthodes polyédriques permettent de simplifier les procédures de raffinement et de déraffinement du maillage utilisées pour l'adaptation.

Dans cette thèse, on s'intéresse à deux familles de méthodes de discrétisation polyédriques d'ordre arbitraire : les méthodes HHO et Galerkin discontinues (dG). L'utilisation de méthodes d'ordre élevé peut accélérer la convergence en présence de solutions régulières, ou lorsqu'elles sont associées au raffinement local du maillage. De plus, la construction de problèmes discrets à l'aide des discrétisations HHO et dG est valable en dimension d'espace arbitraire permettant d'envisager une mise en œuvre indépendante de la dimension. Les deux méthodes sont non-conformes dans le contexte de formulations primales de problèmes elliptiques, car aucune condition de continuité n'est imposée entre les éléments voisins en dG et entre les faces voisines en HHO. Nous nous concentrons sur la méthode HHO pour la discrètisation de la mécanique et sur la méthode dG pour la discrétisation de l'écoulement.

Les méthodes HHO, introduites dans [78] et [74], reposent sur des formulations primales de problèmes elliptiques et aboutissent à des systèmes symétriques et définis positifs. Les inconnues discrètes sont des polynômes du même ordre sur les éléments et les faces. Ces derniéres établissent des connexions inter-éléments aux interfaces et peuvent être utilisées pour imposer des conditions limites essentielles aux faces de bord. Les inconnues de maille sont des variables intermédiaires qui peuvent être éliminées du système global par condensation statique, comme détaillé dans le Chapitre 2. La conception des méthodes HHO se fait en deux étapes : (i) tout d'abord, on reconstruit des opérateurs différentiels discrèts basés sur la résolution de problèmes locaux peu coûteux à l'intérieur de chaque élément et ensuite, (ii) on introduit une stabilisation qui relie entre elles les inconnues discrètes d'élément et de face. Les définitions des opérateurs de reconstructions différentiels sont basées sur des contreparties discrètes des intégrations par parties. Même si les problèmes locaux liés aux opérateurs de reconstruction doivent être résolus, les résultats numériques de [74] indiquent que le coût associé devient marginal par rapport au coût de la résolution du système global. De plus, les méthodes HHO sont efficaces, car l'utilisation d'inconnues de face donne des stencils compacts. Les méthodes HHO offrent d'autres avantages, dont la possibilité d'établir des propriétés de conservation de quantités physiques et la robustesse par rapport à l'hétérogénéité des coefficients. La discrétisation HHO étudiée dans le Chapitre 3 s'inspire des travaux récents sur les opérateurs Leray–Lions [69, 70], où les auteurs montrent que la méthode est robuste en ce qui concerne les non-linéarités.

Les méthodes dG peuvent être considérées comme des méthodes éléments finis permettant des discontinuités dans les espaces discrets, ou bien comme des volumes finis dans lesquels la solution approchée est représentée dans chaque élément du maillage par un polynôme au lieu que par une constante. Permettre à la solution d'être continue par morceaux offre une grande flexibilité dans la conception du maillage. En plus d'être adaptées aux maillages

généraux y compris non-conformes, les méthodes dG ont l'avantage de pouvoir monter en ordre très facilement. Leur principal inconvénient est le grand nombre d'inconnues qu'elles engendrent, et donc le coût de résolution des systèmes. Une analyse unifiée des méthodes dG pour les problèmes elliptiques peut être trouvée dans [42] et [76]. La stratégie fondamentale pour approcher le problème de la diffusion hétérogène (tel que le flux darcéen dans les milieux poreux) en utilisant les méthodes dG consiste à pénaliser les sauts d'interface en utilisant une pondération dépendante de la moyenne harmonique des composantes normales du tenseur de diffusion. Cette technique permet de pénaliser de façon optimale. De plus, comme indiqué dans le chapitres 2 et 4 et dans [77], la méthode proposée est robuste en ce qui concerne les variations spatiales du coefficient de perméabilité, avec des constantes dans les estimations d'erreur ayant une dépendance faible du rapport d'hétérogénéité.

Structure du manuscrit

Par la suite, nous allons résumer brièvement le contenu du manuscrit en mettant l'accent sur les résultats marquants.

Dans le **Chapitre 1** nous présentons les modèles de poroélasticité que nous allons considérer successivement. Une fois ceux-ci introduits, nous faisons un inventaire des difficultés liées à leur approximation numérique, avant de présenter un état de l'art documenté nous permettant de définir les orientations des chapitres suivants.

Dans le **Chapitre 2**, publié dans *SIAM Journal on Scientific Computing* (voir [30]), nous introduisons un nouvel algorithme pour le problème de Biot, basé sur une discrétisation HHO de la mécanique et une discrétisation dG Symmetric Weighted Interior Penalty (SWIP) du flux. La méthode a plusieurs atouts, notamment la validité en deux et trois dimensions, la stabilité inf-sup, le support des maillages polyédriques généraux et l'utilisation de l'ordre d'approximation arbitraire en espace. De plus, le coût de résolution peut être réduit en condensant statiquement un grand sous-ensemble d'inconnues. Notre analyse fournit des estimations de stabilité et d'erreur qui se maintiennent même lorsque le coefficient de stockage spécifique est nul et montre que les constantes dépendent faiblement de l'hétérogénéité du coefficient de perméabilité. Nous discutons les détails de la mise en œuvre de la méthode et fournissons des tests numériques démontrant ses performances. Enfin, nous montrons que le schéma est localement conservateur sur le maillage primal, une propriété souhaitable pour les praticiens et fondamentale pour les estimations a postériori basées sur des flux équilibrés.

Dans le **Chapitre 3**, publié dans *SIAM Journal on Numerical Analysis* (voir [33]), nous proposons et analysons une nouvelle discrétisation HHO d'une classe de modèles d'élasticité (linéaire et) non-linéaire couramment utilisées en mécanique des solides. La méthode satisfait un principe local de travail virtuel à l'intérieur de chaque élément du maillage, avec des tractions numériques d'interface qui obéissent à la loi d'action et réaction. Une analyse complète couvrant des lois de comportement mécanique très générales est effectuée. En particulier, nous prouvons l'existence d'une solution discrète et nous identifions une hypothèse de monotonie stricte sur la loi contrainte-déformation qui assure l'unicité. La convergence aux solutions de régularité minimale est démontrée en utilisant un argument de compacité. Une estimation d'erreur optimale d'ordre h^{k+1} de la norme d'énergie est alors prouvée dans les conditions supplémentaires de continuité de Lipschitz et de forte monotonie sur la loi de

comportement. Les performances de la méthode sont largement étudiées sur un panel complet de tests numériques avec des lois contrainte-déformation correspondant à des matériaux réels.

Dans le **Chapitre 4**, tiré de [35], nous construisons et analysons une méthode couplée HHO-dG pour la poroélasticité non-linéaire. Il y a deux différences principales par rapport à la méthode conçue dans le Chapitre 2 pour la version linéaire du problème. Premièrement, le gradient symétrique discret se situe dans l'espace complet des polynômes à valeurs tensorielles, par opposition aux gradients symétriques des polynômes à valeurs vectorielles. Comme montré dans le Chapitre 3, cette modification est nécessaire pour l'analyse de convergence en présence de lois de comportement non-linéaires. Deuxièmement, le terme de droite du problème discret est obtenu en prenant la moyenne en temps de la force de chargement $\boldsymbol{f}$ et de la source de fluide g entre deux pas de temps consécutifs. Cette modification nous permet de prouver la stabilité et l'estimation d'erreur sans aucune hypothèse de régularité supplémentaire sur les données. Un autre résultat marquant de ce chapitre est une nouvelle preuve de l'inégalité de Korn discrète, ne nécessitant pas d'hypothèse géométrique particulière sur le maillage.

Le **Chapitre 5** contient des perspectives sur la solution numérique du problème de Biot avec coefficients poroélastiques aléatoires dans le contexte de la quantification des incertitudes. Il recueille une partie des travaux en cours réalisés lors du stage au BRGM qui s'est déroulé de janvier à septembre 2018. L'incertitude est modélisée par un ensemble fini de paramètres avec une distribution de probabilité prescrite. Une attention particulière est accordée pour assurer que la paramétrisation des coefficients soit physiquement admissible. Nous présentons la formulation faible du système d'équations différentielles stochastiques et établissons sa bonne position. Nous discutons ensuite de l'approximation du problème par des techniques non-intrusives basées sur le développement de solutions sur des bases de Polynômes de Chaos. La procédure de projection spectrale considérée permet de réduire le problème stochastique à un ensemble fini de simulations déterministes paramétriques discrétisées par le schéma HHO-dG du Chapitre 2. Nous étudions numériquement la convergence de l'erreur par rapport au niveau de la grille dans l'espace de probabilité. Enfin, nous effectuons une analyse de sensibilité pour évaluer la propagation de l'incertitude en entrée sur les champs de déplacement et pression dans un cas test d'injection et un problème de traction.

En **Annexe A**, publié dans les actes de la conférence *Finite Volumes for Complex Applications VIII* (voir [34]), nous présentons une variante du schéma HHO-dG pour le problème de la poroélasticité non-linéaire étudié dans le Chapitre 4. En particulier cette annexe fournit des tests numériques démontrant la convergence de la méthode en presence de lois de comportement non-linéaires de type Hencky-Mises.

L'**Annexe B** présente des travaux réalisés en marge de la ligne directrice de ce manuscrit. Elle est tirée de [36] publié dans *Computational Methods in Applied Mathematics*. Nous considérons les problèmes hyperélastiques et leurs solutions numériques en utilisant des algorithmes de discrétisation par éléments finis conformes. Pour ces problèmes, nous présentons des reconstructions du tenseur de contraintes conformes en $H(\mathrm{div})$, équilibrées et faiblement symétriques, obtenues à partir de problèmes locaux à l'aide des espaces d'éléments finis Arnold–Falk–Winther. Les reconstructions sont indépendantes de la loi de comportement

mécanique. Sur la base de ces reconstructions du tenseur des contraintes, nous obtenons une estimation d'erreur à postériori en distinguant les estimations d'erreur de discrétisation, de linéarisation et de quadrature, et proposons un algorithme adaptatif équilibrant ces différentes sources d'erreur. Nous confirmons l'efficacité de l'estimation par un test numérique avec une solution analytique. Nous appliquons ensuite l'algorithme adaptatif à un test davantage axé sur l'application.

Contents

Chapter 1

Introduction

Contents

In this manuscript we focus on novel discretization schemes for solving the coupled equations of poroelasticity and we present analytical and numerical results for poromechanics problems relevant to geoscience applications. We propose to solve these problems using Hybrid High-Order (HHO) methods, a new class of nonconforming high-order methods supporting general polyhedral meshes.

1.1 Poroelasticity

Porous media are solid materials comprising a great number of interconnected pores allowing for fluid flow through the medium. The presence of a moving fluid influences the mechanical response of the solid matrix. Two mechanisms play a central role in this interaction between the mechanical behavior and the fluid dynamics: (i) an increase of the pore pressure induces a dilatation of the rock in response to the added stress, and (ii) a compression of the porous skeleton leads to a rise of pore pressure, if the change of the mechanical state is fast relative to the fluid flow rate. These two coupled deformation-diffusion phenomena lie at the heart

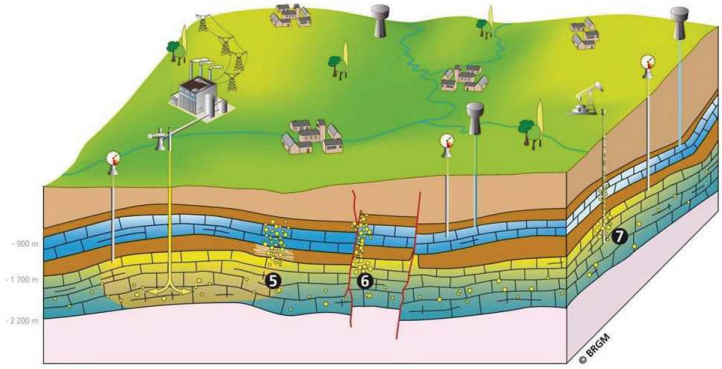


Figure 1.1: Geologic storage of carbon dioxide in saline aquifers

of the theory of poroelasticity. In accordance with these key phenomena, the fluid-filled porous medium acts in a time-dependent manner. Indeed, if the fluid pressure caused by the deformation is allowed to dissipate through a Darcean mass transport, further deformation of the rock progressively appear. At the same time, the induced poroelastic stresses will, in turn, respond back to the fluid pressure field.

1.1.1 Context

This Ph.D. thesis was conjointly founded by the Bureau de recherches géologiques et minières (BRGM) and LabEx NUMEV. The coupling between subsurface flow and geomechanical deformation is a crucial research topic for both cofunding institutions. The main goals of BRGM are to implement decision-support tools designed to anticipate and prevent subsurface risks and to establish safety criteria for human geological activities. There is, therefore, a particular interest in poromechanical modelling, that is critical in the assessment of the environmental impacts of groundwater use and exploitation of shale gas reserves. In particular, seismicity induced by fluid injection and withdrawal has emerged as a central element of the scientific discussion around subsurface technologies that tap into water and energy resources.

Interest in the coupled diffusion-deformation mechanisms was initially motivated by the problem of consolidation, namely the progressive settlement of a soil due to fluid extraction. The earliest theory modeling the effects of the pore fluid on the deformation of soils was developed in the pioneering work of Terzaghi [189], who proposed a model for consolidation accounting for the fluid-to-solid coupling only. In this case, the problem can be decoupled and solved in two stages. This kind of theory can successfully model some of the poroelastic processes in the case of highly compressible fluids such as air. However, when one deals with slightly compressible (or incompressible) fluids, the solid-to-fluid coupling cannot be neglected since the changes in the stress can significantly influence the pore pressure. The first detailed mathematical theory of poroelasticity incorporating both the basic phenomena outlined above, was formulated by Biot [27]. The model proposed by Biot was subsequently re-derived via homogenization [9, 45] and mixture theory [146, 147], what placed the Biot theory on a rigorous basis.

Poroelasticity has since been explored in a large number of geomechanical applications,

such as: subsidence due to fluid withdrawal, reservoir impoundment, tensile failure induced by pressurization of a borehole, waste disposal, earthquake triggering due to pressure induced faults slip, and injection-production cycles in geothermal fields. Recently, coupled flow and geomechanics has also gained attention due to its role in the long-term geologic storage of carbon dioxide in saline aquifers (cf. Figure 1.1), which is widely regarded as a promising technology to help mitigate the climate change by reducing anthropogenic CO_2 emissions into the atmosphere. Injection of CO_2 requires a compression of the ambient groundwater and an overpressurization of the aquifer, which could fracture the caprock, trigger seismicity, or activate faults. BRGM plays a key role in the research on CO_2 storage in geological layers, especially in deep aquifers, and develops tools to understand and prevent the onset of the previously mentioned side effects. Other examples of poroelastic structures include cartilage, skin, bone, the myocardium, the brain, and the lungs. Consequently, notable contributions have also been made in a diverse range of biomechanics applications [6, 15, 44, 187, 193].

1.1.2 Governing equations

We use a classical continuum representation in which the fluid and the solid skeleton are viewed as overlapping continua [18, 61]. The poroelasticity system, describing the fluid flow in a deformable saturated porous medium, consists of one equation expressing the momentum balance and one expressing the fluid mass conservation law. At the macroscopic scale, they are derived in the work of Biot [27] and Terzaghi [190]. In what follows, the elastic structure $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, forms a porous and permeable matrix saturated by a slightly compressible and viscous fluid which diffuses through it. We assume that the material is isotropic and the conditions are isothermal. In many important applications, such as geothermal energy extraction, nuclear waste disposal, and carbon storage, temperature plays a significant role and must therefore be included in the model. Thermo-poroelastic models are derived in [43, 61, 107] and consists in an additional coupling between deformation, heat diffusion, and fluid flow. Since the pressure and temperature play a similar role in the deformation of the body, in what follows we only focus on the hydromechanical coupling. The displacement of the structure is denoted by $\boldsymbol{u} := \{u_1, ..., u_d\}$ and the fluid pore pressure by p.

The momentum conservation equation is similar to that found in the context of elasticity, the exception being the addition of a fluid pressure term. Letting V be a fixed, arbitrary open subset of Ω, the momentum of the corresponding portion of the matrix is given by $\int_V \frac{\partial \boldsymbol{u}}{\partial t} \, \mathrm{d}V$. The forces acting on V consist in the traction forces applied by the complement of V across its boundary ∂V and the volume-distributed external forces. Then, for a volumetric load $\boldsymbol{f}$ and total stress tensor $\tilde{\boldsymbol{\sigma}}$, the momentum balance equation reads

$$\frac{\partial}{\partial t} \int_V \frac{\partial \boldsymbol{u}}{\partial t} \, \mathrm{d}V = \int_{\partial V} \tilde{\boldsymbol{\sigma}} \boldsymbol{n} \, \mathrm{d}S + \int_V \boldsymbol{f} \, \mathrm{d}V,$$

where $\boldsymbol{n}$ is the outward normal of ∂V. Owing to the divergence theorem, this gives

$$\int_V \frac{\partial^2 \boldsymbol{u}}{\partial t^2} \, \mathrm{d}V = \int_V \boldsymbol{\nabla} \cdot \tilde{\boldsymbol{\sigma}} \, \mathrm{d}V + \int_V \boldsymbol{f} \, \mathrm{d}V.$$

In the classical poroelasticity model, since the deformation of the material is usually much slower than the flow rate, the inertial effects are considered negligible. This quasi-static assumption consists in ignoring the second time-derivative for displacements and, since V was chosen arbitrarily, it follows that

$$-\boldsymbol{\nabla}\cdot\tilde{\boldsymbol{\sigma}} = \boldsymbol{f} \qquad \text{in } \Omega. \tag{1.1}$$

Turning to the mass conservation equation, the variables of interest are the fluid content η (fluid mass per unit bulk volume of porous medium), the flux $\boldsymbol{w}$ (fluid mass flow rate per unit area and time), and the volumetric fluid source g. Taking $V \subset \Omega$ as before, the rate at which fluid moves across the boundary ∂V is given by $\int_{\partial V} \boldsymbol{w}\cdot\boldsymbol{n}\, \mathrm{d}S$. Then the conservation of mass for isothermal single-phase flow of a slightly compressible fluid takes the integral form

$$\frac{\partial}{\partial t}\int_V \eta \,\mathrm{d}V + \int_{\partial V} \boldsymbol{w}\cdot\boldsymbol{n}\,\mathrm{d}S = \int_V g\,\mathrm{d}V.$$

The divergence theorem applied to the second term of the left-hand side of the above equation and the fact that V was chosen arbitrarily lead to

$$\frac{\partial \eta}{\partial t} + \boldsymbol{\nabla}\cdot\boldsymbol{w} = g. \tag{1.2}$$

The coupling between the mechanical behavior of the matrix and the fluid dynamics is realized by constitutive relations relating the total stress $\tilde{\boldsymbol{\sigma}}$, the flux $\boldsymbol{w}$, and the fluid content η to the primary variables $\boldsymbol{u}$ and p. The total stress must account for the usual material stress, as in solid mechanics, and for the fluid pressure. Owing to the Terzaghi decomposition of the total stress [190], one has

$$\tilde{\boldsymbol{\sigma}} = \boldsymbol{\sigma} - \alpha p \boldsymbol{I}_d, \tag{1.3}$$

where $\boldsymbol{\sigma}$ is referred to as the *effective stress tensor* and measures the material properties of the medium. The Biot–Willis coefficient $0 < \alpha \leq 1$, defined as the ratio of the volume change of the fluid over the volume change of the medium, accounts for the pressure-deformation interaction, and $\boldsymbol{I}_d$ denotes the d-dimensional identity matrix. The effective stress tensor depends on the deformation according to the stress-strain relation $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})$, where the symmetric gradient operator measures the strain in the case of small deformations and is defined as

$$\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} := \frac{\boldsymbol{\nabla}\boldsymbol{u} + (\boldsymbol{\nabla}\boldsymbol{u})^{\mathrm{T}}}{2}, \quad \text{with } (\boldsymbol{\nabla}\boldsymbol{u})_{ij} := \frac{\partial u_i}{\partial x_j}, \quad \text{for } 1 \leq i, j \leq d.$$

The standard assumption of Darcy's law for porous media holds for the flux:

$$\boldsymbol{w} = -\frac{\mathbb{K}\boldsymbol{\nabla}p}{\mu_{\mathrm{f}}}, \tag{1.4}$$

where $\mathbb{K}$ is the tensor-valued intrinsic permeability field and μ_{f} is the fluid viscosity. The third constitutive assumption links the change in the fluid content η with the changes in the

fluid pressure p and and the material volume, which is locally measured by $\boldsymbol{\nabla}\cdot\boldsymbol{u}$. More specifically, following [61, 68], we set

$$\eta = c_0 p + \alpha \boldsymbol{\nabla}\cdot\boldsymbol{u}, \tag{1.5}$$

where the Biot–Willis coefficient α quantifies the amount of fluid that can be forced into the medium by a variation of pore volume for a constant fluid pressure, while the constrained specific storage coefficient $c_0 \geq 0$ measures the amount of fluid that can be forced into the medium by pressure increments due to the compressibility of the structure. The case of a solid matrix with incompressible grains corresponds to the limit value $c_0 = 0$.

1.1.3 The Biot model

The Biot model of linear poroelasticity is valid under the following assumptions: (i) small relative variations of porosity with respect to the equilibrium value $\varphi \in [0, 1]$, (ii) small relative variations of the fluid density, and (iii) infinitesimal strain theory. Moreover, the material is assumed to be isotropic, linearly elastic, and homogeneous and, as a consequence, its mechanical behavior is described by Hooke's stress-strain law

$$\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) \coloneqq 2\mu\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} + \lambda(\boldsymbol{\nabla}\cdot\boldsymbol{u})\boldsymbol{I}_d, \tag{1.6}$$

the Lamé's parameters $\mu > 0$ and $\lambda > 0$ correspond to the dilatation and shear moduli, respectively. We recall that the shear modulus μ remains bounded and bounded away from 0, namely $0 < \underline{\mu} \leq \mu \leq \overline{\mu} \in \mathbb{R}$, whereas λ can take unboundedly large values in the case of a quasi-incompressible material ($\lambda \to \infty$). We note that the condition $\lambda > 0$ is stronger than the one required to have the coercivity of the elasticity operator in Section 2.2.2. Indeed, $\lambda \geq -\frac{2}{d}\mu$ is sufficient to ensure the positivity of the bulk modulus and the well-posedness of the weak formulation (5.18) (see also Lemma 5.4 and [178]).

The Biot model is derived by inserting the constitutive relations (1.3), (1.4), and (1.5) into (1.1) and (1.2). Additional details on the model derivation and investigations on the poromechanical coefficient are given in Section 5.2. Hence, for a given bounded connected domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, with boundary $\partial\Omega$ and outward normal $\boldsymbol{n}$, a finite time $t_{\mathrm{F}} > 0$, a volumetric load $\boldsymbol{f}$, and a fluid source g, the corresponding problem consists in finding a vector-valued displacement field $\boldsymbol{u} : \Omega \times (0, t_{\mathrm{F}}] \to \mathbb{R}^d$ and a scalar-valued pore pressure field $p : \Omega \times (0, t_{\mathrm{F}}] \to \mathbb{R}$ solution of

$$-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) + \alpha\boldsymbol{\nabla}p = \boldsymbol{f} \qquad \text{in } \Omega \times (0, t_{\mathrm{F}}], \tag{1.7a}$$

$$c_0 \mathrm{d}_t p + \alpha \mathrm{d}_t(\boldsymbol{\nabla}\cdot\boldsymbol{u}) - \boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}p) = g \qquad \text{in } \Omega \times (0, t_{\mathrm{F}}], \tag{1.7b}$$

where d_t denotes the time-derivative operator and $\boldsymbol{\kappa} \coloneqq \frac{\mathbb{K}}{\mu_{\mathrm{f}}}$ the tensor-valued fluid mobility field. There are two distinct sets of boundary conditions, one corresponding to the deformation and one corresponding to the flow, and an initial condition on the fluid content to be added

to the above equations to close the model for the Biot problem:

$$\boldsymbol{u} = \boldsymbol{u}_D \qquad \text{on } \Gamma_D \times (0, t_{\mathrm{F}}], \tag{1.7c}$$

$$(\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) + \alpha p \boldsymbol{I}_d)\boldsymbol{n} = \boldsymbol{t}_N \qquad \text{on } \Gamma_N \times (0, t_{\mathrm{F}}], \tag{1.7d}$$

$$p = p_d \qquad \text{on } \Gamma_d \times (0, t_{\mathrm{F}}], \tag{1.7e}$$

$$\boldsymbol{\kappa}\boldsymbol{\nabla} p \cdot \boldsymbol{n} = w_n \qquad \text{on } \Gamma_n \times (0, t_{\mathrm{F}}], \tag{1.7f}$$

$$(c_0 p + \alpha \boldsymbol{\nabla}\cdot\boldsymbol{u})(\cdot, 0) = \eta_0 \qquad \text{in } \Omega, \tag{1.7g}$$

where $\Gamma_D, \Gamma_N, \Gamma_d$, and Γ_n are subsets of the boundary $\partial\Omega$ such that $\Gamma_D \cup \Gamma_N = \Gamma_d \cup \Gamma_n = \partial\Omega$, $\Gamma_D \cap \Gamma_N \neq \emptyset$ and $\Gamma_d \cap \Gamma_n \neq \emptyset$. If $\Gamma_D = \emptyset$, owing to the Neumann condition (1.7d), we need to impose a compatibility condition on the average of the forcing term $\boldsymbol{f}$ and the boundary traction $\boldsymbol{t}_N$, as well as to prescribe the rigid-body motions of the medium. Thus, in this case, we have $\Gamma_N = \partial\Omega$ and we set

$$\int_\Omega \boldsymbol{f}(\cdot, t) = \int_{\partial\Omega} \boldsymbol{t}_N(\cdot, t), \quad \int_\Omega \boldsymbol{u}(\cdot, t) = \boldsymbol{u}_{\mathrm{av}}, \quad \text{and} \quad \int_\Omega \boldsymbol{\nabla}\times\boldsymbol{u}(\cdot, t) = \boldsymbol{u}_{\mathrm{rot}} \quad \forall t \in (0, t_F], \tag{1.7h}$$

where $\boldsymbol{\nabla}\times$ denotes the curl operator. For the sake of simplicity, we exclude the case of $\Gamma_D \neq \emptyset$ with zero $(d-1)$-dimensional Hausdorff measure. We only mention that, if the domain is clamped on a single point $\boldsymbol{x}_D \in \partial\Omega$ there is no need to fix the translations (by imposing the zero-average constraint on $\boldsymbol{u}$), but it is still necessary to prescribe the rotations. Similarly, in the case $\Gamma_d = \Gamma_N = \emptyset$ and $c_0 = 0$, we observe that, owing to (1.7b) and the inhomogeneous boundary conditions (1.7c) and (1.7f), we need the following compatibility condition relating the fluid source g, the normal flux w_n, and the Dirichlet datum $\boldsymbol{u}_D$, and the following constraint on the average of p:

$$\int_\Omega g(\cdot, t) = \int_{\partial\Omega} w_n(\cdot, t) + \alpha \mathrm{d}_t(\boldsymbol{u}_D \cdot \boldsymbol{n})(\cdot, t), \quad \text{and} \quad \int_\Omega p(\cdot, t) = p_{\mathrm{av}} \quad \forall t \in (0, t_F]. \tag{1.7i}$$

In the context of geomechanics, a reasonable choice for the rigid-body motions of the medium and the pressure average in (1.7h) and (1.7i) can be $\boldsymbol{u}_{\mathrm{av}} = \boldsymbol{0}$, $\boldsymbol{u}_{\mathrm{rot}} = \boldsymbol{0}$, and p_{av} equal to the hydrostatic pressure measured at the barycenter of Ω, respectively. Finally, in the case $\Gamma_N = 0$ and $c_0 = 0$, if looking for a strong (space-time continuous up to the boundary) solution of problem (1.7), one should also: (i) assume that the boundary data in (1.7c) and (1.7e) admit continuous traces at $\partial\Omega \times \{t = 0\}$, and (ii) require the average of the initial fluid content to be compatible with the Dirichlet datum $\boldsymbol{u}_D$, namely

$$\int_\Omega \eta_0 = \alpha \int_\Omega \boldsymbol{\nabla}\cdot\boldsymbol{u}(\cdot, 0) = \int_{\partial\Omega} (\boldsymbol{u}_D \cdot \boldsymbol{n})(\cdot, 0). \tag{1.7j}$$

The symmetric mobility tensor $\boldsymbol{\kappa} : \Omega \to \mathbb{R}^{d\times d}_{\mathrm{sym}}$ is assumed to be uniformly bounded and uniformly elliptic; that means there exist positive constants $\underline{\kappa}$ and $\overline{\kappa}$ such that, for all $\boldsymbol{x} \in \Omega$ and all vector $\boldsymbol{\xi} \in \mathbb{R}^d$,

$$\underline{\kappa}\|\boldsymbol{\xi}\|^2 \leq \boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\kappa}(\boldsymbol{x})\boldsymbol{\xi} \leq \overline{\kappa}\|\boldsymbol{\xi}\|^2.$$

Parameter	Description	Unit
$\mu \in [\underline{\mu}, \overline{\mu}]$	Shear modulus	Pa
$\lambda \in (0, +\infty)$	Dilatation modulus	Pa
$\boldsymbol{\kappa} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$	Uniformly elliptic mobility tensor	$\mathrm{m}^2\mathrm{s}^{-1}\mathrm{Pa}^{-1}$
$\varphi \in [0, 1]$	Reference porosity value	–
$\alpha \in (\varphi, 1]$	Biot–Willis coefficient	–
$c_0 \geq 0$	Constrained specific storage coefficient	Pa^{-1}

Table 1.1: Summary of physical parameters

The above hypothesis allows the permeability tensor $\boldsymbol{\kappa}$ to be discontinuous in Ω even if, in practice, it has often more regularity than just being uniformly bounded. Henceforth, it is assumed that there is a partition $P_\Omega := \{\Omega_i\}_{1\leq i\leq N_\Omega}$ of Ω such that the restriction of $\boldsymbol{\kappa}$ to each Ω_i is constant. In geomechanics applications, the partition P_Ω typically results from the partitioning of the porous medium into various geological layers.

In some applications, such as consolidation processes, the fluid is considered to be incompressible and the solid matrix to have very low sensitiveness with respect to pressure increments. Hence, the constrained specific storage coefficient c_0 can be very small and sometimes vanishing. Since, in this case, the volume change of the solid grains composing the matrix is neglected, the volume of fluid only depends on the variations of pore volume. This model is referred to as *Biot's consolidation model*. From a numerical point of view, as we will detail further, the correct approximation of the consolidation problem is more involved than the one of the linear poroelasticity problem with c_0 bounded away from 0. The well-posedness of the canonical two-field weak formulation of problem (1.7) with displacement and pressure as variables was carried out in [181, 205]. Three-field and four-field formulations, obtained by taking the Darcy flux and the total stress as independent variables, have also been analyzed and can be found in several studies, *e.g.* [131, 167, 204].

1.1.4 Nonlinear poroelasticity

As noted in [68, Section 3.2], many experimental results suggest that the volumetric response of porous rock to the change of total pressure is actually nonlinear. The nonlinear behavior is generally associated with the closing/opening of crack-like pores, but in very porous and weak rocks it is mainly caused by progressive pore collapse. Investigations of the nonlinear deformation of porous media have been motivated by the need to quantify the effect of the pore pressure decline during depletion of an oil or gas reservoir on the volume of the rock matrix. For a physical investigation of the nonlinear mechanical behavior of porous solids we refer the reader to [22, 28]. Another possible source of nonlinearity in the poroelasticity system is due to the dependence of the hydraulic mobility on the stress and fluid pressure. In fact, some porous media exhibit a significant difference in conductivity once the material deformations start to occur. The analysis of the nonlinear problem obtained by considering $\boldsymbol{\kappa}$ depending on $\boldsymbol{\nabla}\cdot\boldsymbol{u}$ is provided in [48], while the case of a negative exponential dependence on the volumetric part of total stress is examined in [145]. A model for finite-strain poroelasticity

with deformation dependent permeability field has been proposed and analyzed in [23, 90].

In this manuscript we focus on nonlinearities of the stress-strain relation $\boldsymbol{\sigma} : \Omega \times \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}^{d\times d}_{\mathrm{sym}}$ in the small deformation regime. In particular, in Chapter 4 we study the poroelasticity problem obtained by replacing (1.6) with the hyperelastic nonlinear laws considered in Chapter 3. *Hyperelasticity* is a type of constitutive model for ideally elastic materials in which the stress is determined by the current state of deformation by deriving a stored energy density function $\Psi : \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}$, namely

$$\boldsymbol{\sigma}(\boldsymbol{\tau}) := \frac{\partial \Psi(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}}.$$

We next discuss a number of meaningful examples.

Example 1.1 (Hencky–Mises model)**.** The nonlinear Hencky–Mises model of [105, 157] corresponds to the stored energy density function

$$\Psi_{\mathrm{hm}}(\boldsymbol{\tau}) := \frac{\alpha}{2}\operatorname{tr}(\boldsymbol{\tau})^2 + \Phi(\operatorname{dev}(\boldsymbol{\tau})), \tag{1.8}$$

where $\operatorname{dev} : \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}$ defined by $\operatorname{dev}(\boldsymbol{\tau}) = \operatorname{tr}(\boldsymbol{\tau}^2) - \frac{1}{d}\operatorname{tr}(\boldsymbol{\tau})^2$ is the deviatoric operator. Here, $\alpha \in (0, +\infty)$ and $\Phi : [0, +\infty) \to \mathbb{R}$ is a function of class C^2 satisfying, for some positive constants C_1, C_2, and C_3,

$$C_1 \le \Phi'(\rho) < \alpha, \quad |\rho\Phi''(\rho)| \le C_2 \quad \text{and} \quad \Phi'(\rho) + 2\rho\Phi''(\rho) \ge C_3 \quad \forall \rho \in [0, +\infty). \tag{1.9}$$

We observe that taking $\alpha = \lambda + \frac{2}{d}\mu$ and $\Phi(\rho) = \mu\rho$ in (1.8) leads to the linear case (1.6). Deriving the energy density function (1.8) yields

$$\boldsymbol{\sigma}(\boldsymbol{\tau}) = \tilde{\lambda}(\operatorname{dev}(\boldsymbol{\tau}))\operatorname{tr}(\boldsymbol{\tau})\boldsymbol{I}_d + 2\tilde{\mu}(\operatorname{dev}(\boldsymbol{\tau}))\boldsymbol{\tau}, \tag{1.10}$$

with nonlinear Lamé functions $\tilde{\mu}(\rho) := \Phi'(\rho)$ and $\tilde{\lambda}(\rho) := \alpha - \Phi'(\rho)$.

In the previous example the nonlinearity of the model only depends on the deviatoric part of the strain. In the following model it depends on the term $\boldsymbol{\tau} : \mathfrak{C}\boldsymbol{\tau}$.

Example 1.2 (An isotropic reversible damage model)**.** The isotropic reversible damage model of [50] can also be interpreted in the framework of hyperelasticity by setting up the energy density function as

$$\Psi_{\mathrm{dam}}(\boldsymbol{\tau}) := \frac{(1 - D(\boldsymbol{\tau}))}{2}\boldsymbol{\tau} : \mathfrak{C}\boldsymbol{\tau} + \Phi(D(\boldsymbol{\tau})), \tag{1.11}$$

where $D : \mathbb{R}^{d\times d}_{\mathrm{sym}} \to [0, 1]$ is the scalar damage function and $\mathfrak{C}$ is a fourth-order symmetric and uniformly elliptic tensor, namely, for some positive constants C_* and C^*, it holds

$$C_*\|\boldsymbol{\tau}\|^2_{d\times d} \le \mathfrak{C}\boldsymbol{\tau} : \boldsymbol{\tau} \le C^*\|\boldsymbol{\tau}\|^2_{d\times d}, \quad \forall \boldsymbol{\tau} \in \mathbb{R}^{d\times d}. \tag{1.12}$$

The function $\Phi : [0, 1] \to \mathbb{R}$ defines the relation between $\boldsymbol{\tau}$ and D by $\frac{\partial \phi}{\partial D} = \frac{1}{2}\boldsymbol{\tau} : \mathfrak{C}\boldsymbol{\tau}$. The resulting stress-strain relation reads

$$\boldsymbol{\sigma}(\boldsymbol{\tau}) = (1 - D(\boldsymbol{\tau}))\mathfrak{C}\boldsymbol{\tau}. \tag{1.13}$$

Another generalization of the linear stress-strain relation is obtained by adding second-order terms to (1.6).

Example 1.3 (A second-order elasticity model)**.** In the second-order isotropic elasticity model of [66, 124] the stress-strain relation is

$$\boldsymbol{\sigma}(\boldsymbol{\tau}) \;=\; \lambda\,\mathrm{tr}(\boldsymbol{\tau})\boldsymbol{I}_d \;+\; 2\mu\boldsymbol{\tau} \;+\; B\,\mathrm{tr}((\boldsymbol{\tau})^2)\boldsymbol{I}_d \;+\; 2B\,\mathrm{tr}(\boldsymbol{\tau})\boldsymbol{\tau} \;+\; C\,\mathrm{tr}(\boldsymbol{\tau})^2\boldsymbol{I}_d \;+\; A(\boldsymbol{\tau})^2, \quad (1.14)$$

where λ and μ are the standard Lamé parameter, and $A, B, C \in \mathbb{R}$ are the second-order moduli. We note that the hyperelastic stored energy function associated to the the stress-strain relation (1.14) has third-order terms and in general is not polyconvex, *i.e.* convex with respect to each of the strain invariants.

1.2 Discretization

In order to discretize the Biot and nonlinear poroelasticity problems presented in the previous sections, we consider a coupling of the Hybrid High-Order (HHO) and discontinuous Galerkin (dG) method. Concerning the time discretization, we consider backward differentiation formulas (especially the implicit Euler method), which are the simplest and most widely used methods in the literature. We show that this construction yields an inf-sup stable hydro-mechanical coupling, which is a crucial property to ensure robustness for small time steps combined with small permeabilities. Additionally, the analysis presented in Chapter 2 and 4 allows to derive error estimates that are robust in the limit $c_0 \to 0$ which, as noted in Section 1.2.1 and in [164, Section 5.2], is expected to mitigate the problem of the nonphysical pressure oscillations which can sometimes arise in the numerical simulation of poroelastic systems. Before discussing the main features of the discretization methods considered in the next chapters, we give an overview of the numerical issues related to the discretization of poroelasticity problems.

1.2.1 Numerical issues

Several difficulties have to be accounted for in the design of a discretization methods for problem (1.7). These issues have three origins: the discretization of the elasticity operator, the (possibly saddle-point) coupling between the flow and the mechanics, and the rough variations of the permeability coefficient that the Darcy operator has to accomodate.

First, we remark that the elasticity operator has to be carefully engineered in order to ensure stability expressed by a discrete version of Korn's inequality. Conforming finite elements naturally yields coercive discretizations, but this is not necessarily the case when considering nonconforming approximations. For instance, it is well known that the Crouzeix–Raviart space (spanned by piecewise affine functions that are continuous at the midpoint of mesh interfaces) does not fulfill a discrete Korn's inequality. Another numerical issue concerning the discretization of elasticity problems is the so-called locking phenomenon. When the dilatation modulus λ in (1.6) is very large, which corresponds to a quasi-incompressible material, results of poor quality can be obtained. More specifically, it can be observed that

the material deforms as if it were much stiffer. In other words, it appears to lock. The key point to ensure that a numerical method is locking-free is to establish an error estimate with a multiplicative constant not blowing up in the limit $\lambda \to \infty$, *i.e.* being able to prove uniform convergence with respect to λ.

The stability of the saddle-point mechanics-flow coupling is closely related to the elasticity locking phenomenon. Indeed, for both of them, the difficulty lies in the approximation of the divergence operator. In the linear elasticity context, locking can be handled by projecting the divergence operator onto a discrete (pressure) space which satisfies an inf-sup condition when coupled with the displacement approximation space. In the context of a poroelastic displacement-pressure coupling, stability can thus be obtained by considering a discrete pressure belonging to that space, which is in fact equivalent to projecting the discrete divergence operator onto the pressure space in the coupling term. From a mathematical point of view, the inf-sup condition yields an estimate in the $L^\infty((0,t_\mathrm{F});L^2(\Omega))$-norm on the discrete pore pressure which is independent of $\underline{\kappa}^{-1}$. It is of some importance to note that inf-sup stability is not strictly needed in the compressible case (i.e. with $c_0 > 0$). As a matter of fact, the presence of the term $c_0 \mathrm{d}_t p$ in the left-hand side of the fluid balance Equation (1.7b) directly yields a discrete $L^\infty((0,t_\mathrm{F});L^2(\Omega))$ estimate on the pore pressure which does not depend neither on $\underline{\kappa}^{-1}$ nor on λ. An investigation of the role of the inf-sup condition in the context of finite element discretizations of linear poroelasticity can be found, *e.g.* , in [154, 155] and in [165, 166].

The problem of spurious spatial oscillations of the pore pressure is actually more involved than a simple saddle-point coupling issue. The difficulty comes from the fact that, in very early times (or when the permeability is low), the pressure is quasi-$L^2(\Omega)$ as the diffusion term gives an almost vanishing contribution. However, as soon as $t > 0$, boundary conditions are imposed on the pore pressure hence giving necessarily to this latter a $H^1(\Omega)$ regularity. When $c_0 = 0$, if no discrete inf-sup condition holds, the only control of pressure is given by the diffusion term, which is almost inexistent in early times. Spurious spatial oscillations then arise. If an inf-sup condition holds, then it yields a control of the pressure approximation, hence reducing the oscillations. However, it has been recently pointed out in [176] that, even for discretization methods leading to an inf-sup stable discretization of the Stokes problem in the steady case, pressure oscillations can arise owing to a lack of monotonicity of the discrete operator.

1.2.2 Polyhedral element methods

In recent years, a large effort has been devoted to the development and analysis of numerical methods that apply to more general meshes than simplicial or Cartesian ones. In the context of poroelaticity, this necessity arises from the presence of various layers and fractures in the porous medium. The discretization of the domain may also include degenerate elements (as in the near wellbore region in reservoir modelling, e.g. Figure 1.2) and nonconforming interfaces (accounting for the presence of cracks or resulting from local mesh refinement). In the context of structural mechanics, discretization methods supporting polyhedral meshes and nonconforming interfaces can be useful for several reasons including, e.g., the use of hanging nodes for contact and interface elasticity problems, and the greater robustness to mesh

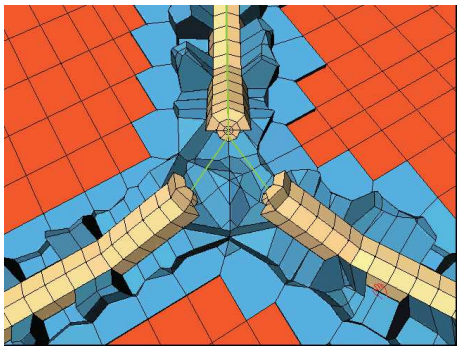


Figure 1.2: Example of polyhedral mesh with degenerate elements in reservoir applications

distorsion and fracture. Moreover, polyhedral methods allow for simple mesh refinement and coarsening procedures for adaptivity.

In this dissertation, we focus on two families of polyhedral discretization methods: the HHO and dG methods. Both are nonconforming in the context of primal formulations of elliptic problems, since no continuity condition is imposed between neighboring mesh elements. They are also high-order methods, because they allow to increase the space approximation order. The use of high-order methods can classically accelerate the convergence in the presence of regular exact solutions or when combined with local mesh refinement. Additionally, the construction of discrete problems using the HHO and dG discretizations is valid for arbitrary space dimension and this enables dimension-independent implementations. In what follows, we consider the HHO method of [33, 74] for the discretization of the (possibly nonlinear) elasticity operator in (1.7a) and the Symmetric Weighted Interior Penalty (SWIP) dG method of [77, 93] for the Darcy operator in (1.7b).

Discontinuous Galerkin methods can be viewed as finite element methods allowing for discontinuities in the discrete trial and test spaces. Localizing test functions to single mesh elements, they can also be viewed as finite volume methods in which the approximate solution is represented on each mesh element by a polynomial function and not only by a constant function. Allowing the approximate solution to be only piecewise continuous offers a substantial amount of flexibility in the mesh design. A unified analysis of dG methods for elliptic problems can be found in [42] and in [76, Chapter 4]. The fundamental strategy to approximate heterogeneous diffusion (such us Darcean flow in porous media) problem using dG methods is to penalize interface and boundary jumps using a diffusion-dependent parameter scaling as the harmonic mean of the normal component of the diffusion tensor. Indeed, using the harmonic mean to penalize will turn out to tune automatically the amount of penalty. Moreover, as pointed out in Chapter 4 and in [77] the proposed method is robust with respect to the spatial variations of the permeability coefficient, with constants in the error estimates having a mild dependendance on the heterogeneity ratio.

Hybrid High-Order methods, introduced in [78] and [74], rely on primal formulations of elliptic problems and lead to a symmetric, positive definite system matrices. The methods can be deployed on general polyhedral meshes. The degrees of freedom are polynomials of the same order at mesh elements and faces: face-based discrete unknowns establish inter-elements connections at interfaces and can be used to strongly enforce essential boundary conditions

at boundary faces, element-based discrete unknowns are intermediate variables which can be eliminated from the global system by static condensation, as detailed in [57, Section 2.4] and in Section 2.5. The design of these methods proceeds in two step: (i) the discrete reconstructions of differential operators hinging on the solution of inexpensive local problems inside each element and (ii) the definition of a least-squares stabilization that weakly enforces the matching of element- and face-based discrete unknowns. The definitions of the differential reconstructions operators are based on discrete counterparts of integrations by parts. Even if local problems related to the reconstruction operators have to be solved, the numerical results of [74] indicate that the associated cost becomes marginal with respect to the cost of solving the global systems as the number of discrete unknowns is increased. Moreover, HHO methods are computationally effective, since the use of face unknowns yields to compact stencils, and they can be efficiently implemented thanks to the possibility of statically condensing a large subset of the unknowns. The HHO discretization studied in Chapter 3 for nonlinear elasticity problems is inspired by recent works on Leray–Lions operators [69, 70], where the authors show that the method is robust with respect to nonlinearities. The robustness of the method with respect to locking has been proved in [74, Section 5].

Other locking-free polyhedral schemes have been proposed for the discretization of linear elasticity problems. A non-exhaustive list includes: (i) the Hybridizable discontinuous Galerkin (HDG) method of [185], (ii) the Mimetic Finite Difference (MFD) scheme of [19], (iii) the Virtual Element Method (VEM) of [20], (iv) the generalized Crouzeix–Raviart scheme introduced in [80], and (v) the Weak Galerkin (WG) method of [197]. Concerning the nonlinear version of the problem, discretization methods supporting general meshes have also been considered in [21] and [54], where the authors propose a low-order VEM scheme for small and finite deformations, respectively. The HHO method has been applied for the discretization of finite deformations elasticity problems in [1]. To the best of our knowledge, the existing literature on the approximation of poroelasticity problem on general polyhedral meshes is more scarce. We can cite the vectorial Multi-Point Flux Approximation methods of [160], the Hybrid Finite Volume discretization considered in [138], and the very recent HDG method of [101] .

1.3 Uncertainty quantification

In numerical simulation, accounting for uncertainties in input data (such as physical constants, boundary and initial conditions, external forcing, and geometry) is a crucial issue, especially in risk analysis, safety, and design. Risk assessment in poroelasticity is a critical challenge for a wide range of vital resource management applications. Recent public concerns over induced seismicity [63, 113] and groundwater contamination [182] underscore the need to quantify the probability of harmful events associated with subsurface flow and deformation. Of particular interest is the prediction of critical stresses which can compromise the caprock integrity [162] and wellbore stability [153] in storage reservoirs. The key to successful risk assessment in these contexts is the ability to predict the probability of critical events based on some empirical approximation of the material parameters and source term variability. For this reason there is a particular need to combine numerical methods for poromechanical

simulation with efficient techniques for quantifying uncertainties.

Uncertainty Quantification (UQ) methods have been developed in the last decades to take into account the effect of random input parameters on the quantity of interest. They enable to obtain information on the model output that is richer than in a deterministic context, since they provide the statistical moments and the probability distribution. This information makes the comparison with experimental observation easier and facilitate the evaluation of the validity of the physical models. Indeed, UQ methods provide confidence intervals in computed predictions and identify the uncertain parameters that should be measured or controlled with more accuracy because they have the most significant impact on the solution. Additionally, they allow to assess the limits of predictability and the level of reliability that can be attached to numerical simulations.

Among the techniques designed for UQ in numerical models, the stochastic spectral methods have received a considerable attention. The core idea of these methods is the decomposition of random quantities on suitable approximation bases such as the Karhunen–Loéve [142, 180] or the Polynomial Chaos [112, 137] expansions. The former represents the random fields as a linear combination of an infinite number of uncorrelated random variables, while the latter uses polynomial expansions in terms of independent random variables. Their main interest is that they provide a complete probabilistic description of the uncertain solution.

In Chapter 5, in order to investigate the effect of uncertainty in poroelasticity problems, we rely on a Polynomial Chaos (PC) approach. The fundamental concept on which PC decompositions are based is to regard uncertainty as generating a new dimension and the solution as being dependent on this dimension. A convergent expansion along the new dimension is then sought in terms of a set of orthogonal basis functions, whose coefficients can be used to characterize the uncertainty. The motivation behind PC approaches includes: (i) the suitability to model expressed in terms of partial differential equations, (ii) the ability to deal with situations exhibiting steep nonlinear dependence of the solution on random model data, and (iii) the promise of obtaining efficient and accurate estimates of uncertainty. In addition, the provided information is given in a format that can be readily exploited to probe the dependence of specific observables on particular components of the input data.

1.4 Plan of the manuscript

The rest manuscript is organized as follows.

In **Chapter 2**, published in *SIAM Journal on Scientific Computing* (cf. [30]), we introduce a novel algorithm for the Biot problem based on a HHO discretization of the mechanics and a Symmetric Weighted Interior Penalty dG discretization of the flow. The method has several assets, including, in particular, the validity in two and three space dimensions, inf-sup stability, and the support of general polyhedral meshes, nonmatching interfaces, and arbitrary space approximation order. Additionally, the resolution cost can be reduced by statically condensing a large subset of the unknowns. Our analysis delivers stability and error estimates that hold also when the constrained specific storage coefficient vanishes, and shows that the constants have only a mild dependence on the heterogeneity of the permeability coefficient. We discuss implementation details and provide numerical tests demonstrating

the performance of the method. In particular, we numerically check the robustness of the method with respect to pressure spurious oscillations. Finally, we show that the scheme is locally conservative on the primal mesh, a desirable property for practitioners and key for a posteriori estimates based on equilibrated fluxes.

In **Chapter 3**, published in *SIAM Journal on Numerical Analysis* (cf. [33]), we propose and analyze a novel HHO discretization of a class of (linear and) nonlinear elasticity models in the small deformation regime which are of common use in solid mechanics. The method satisfies a local principle of virtual work inside each mesh element, with interface tractions that obey the law of action and reaction. A complete analysis covering very general stress-strain laws is carried out. In particular, we prove the existence of a discrete solution and we identify a strict monotonicity assumption on the stress-strain law which ensures uniqueness. Convergence to minimal regularity solutions $\boldsymbol{u} \in H_0^1(\Omega;\mathbb{R}^d)$ is proved using a compactness argument. An optimal energy-norm error estimate in h^{k+1} is then proved under the additional conditions of Lipschitz continuity and strong monotonicity on the stress-strain law. The performance of the method is extensively investigated on a complete panel of model problems using stress-strain laws corresponding to real materials. The numerical tests show that the method is robust with respect to strong nonlinearities.

In **Chapter 4**, building on the material of the previous chapters, we construct and analyze a coupled HHO-dG discretization method for the nonlinear poroelasticity problem. The chapter is submitted for publication (see [35] for the preprint version). Compared to the method proposed in Chapter 2 for the linear poroelasticity problem, there are two main differences in the design. First, the discrete symmetric gradient sits in the full space of tensor-valued polynomials, as opposed to symmetric gradients of vector-valued polynomials. As shown in Chapter 3, this modification is required for the convergence analysis in the presence of nonlinear stress-strain laws. Second, the right-hand side of the discrete problem is obtained by taking the average in time of the loading force $\boldsymbol{f}$ and fluid source g between two consecutive time steps. This modification allows us to prove stability and optimal error estimates without any additional time regularity assumptions on data. Moreover, the results holds for both nonzero and vanishing storage coefficients. In this chapter we also give a new simple proof of discrete Korn's inequality, not requiring particular geometrical assumptions on the mesh.

In **Chapter 5**, we consider the numerical solution of the Biot problem with random poroelastic coefficients in the context of uncertainty quantification. The chapter collects part of the ongoing work carried out during the internship at the BRGM that took place from January to September 2018. The uncertainty is modelled with a finite set of parameters with prescribed probability distribution. We present the variational formulation of the stochastic partial differential system and establish its well-posedness. We then discuss the approximation of the parameter-dependent problem by non-intrusive techniques based on on Polynomial Chaos decompositions. We specifically focus on sparse spectral projection methods which essentially amount to performing an ensemble of deterministic model simulations to estimate the expansion coefficients. We numerically investigate the convergence of the probability error of the PC approximation with respect to the level of the sparse grid. Finally, we perform a sensitivity analysis to asses the propagation of the input uncertainty on the solutions

considering an injection test and a traction problem.

In **Appendix A**, published in the conference book *Finite Volumes for Complex Applications VIII* (cf. [34]), we present a variant of the HHO-dG algorithm for the quasi-static nonlinear poroelasticity problem studied in Chapter 4. In particular, this appendix provides numerical tests demonstrating the convergence of the method considering the Hencky-Mises non-linear behavior law.

Appendix B is an excerpt of a complementary work that has been published in *Computational Methods in Applied Mathematics* (cf. [36]). I decided not to include the full version to give a more consistent structure to the manuscript. Therein we consider hyperelastic problems and their numerical solution using a conforming finite element discretization and iterative linearization algorithms. For these problems, we present equilibrated, weakly symmetric, $H(\mathrm{div})$-conforming stress tensor reconstructions, obtained from local problems on patches around vertices using the Arnold–Falk–Winther finite element spaces. We distinguish two stress reconstructions, one for the discrete stress and one representing the error of the iterative linearization algorithm. The reconstructions are independent of the mechanical behavior law. Based on these stress tensor reconstructions, we derive an a posteriori error estimate distinguishing the discretization, linearization, and quadrature error estimates, and propose an adaptive algorithm balancing these different error sources. We prove the efficiency of the estimate, and confirm it on a numerical test with analytical solution. We then apply the adaptive algorithm to a more application-oriented test.

Chapter 2

The Biot problem

This chapter has been published in the following peer-reviewed journal (see [30]):

SIAM Journal on Scientific Computing,
Volume 38, Issue 3, 2016, Pages *A*1508–*A*1537.

Contents

2.1 Introduction

We consider in this chapter the quasi-static Biot's consolidation problem describing Darcian flow in a deformable saturated porous medium. Our original motivation comes from applications in geosciences, where the support of general polyhedral meshes is crucial, e.g., to handle nonconforming interfaces arising from local mesh adaptation or Voronoi elements in the near wellbore region when modelling petroleum extraction. Let $\Omega \subset \mathbb{R}^d$, $1 \le d \le 3$, denote a bounded connected polyhedral domain with boundary $\partial\Omega$ and outward normal $\boldsymbol{n}$. For a given finite time $t_F > 0$, volumetric load $\boldsymbol{f}$, fluid source g, the Biot problem consists in finding a vector-valued displacement field $\boldsymbol{u}$ and a scalar-valued pore pressure field p solution of

$$-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{u}) + \alpha\boldsymbol{\nabla}p = \boldsymbol{f} \qquad \text{in } \Omega\times(0,t_F), \tag{2.1a}$$

$$c_0 \mathrm{d}_t p + \boldsymbol{\nabla}\cdot(\alpha \mathrm{d}_t \boldsymbol{u}) - \boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}p) = g \qquad \text{in } \Omega\times(0,t_F), \tag{2.1b}$$

where $c_0 \ge 0$ and $\alpha > 0$ are real numbers corresponding to the constrained specific storage and Biot–Willis coefficients, respectively, $\boldsymbol{\kappa}$ is a real-valued permeability field such that $\underline{\kappa} \le \boldsymbol{\kappa} \le \overline{\kappa}$ a.e. in Ω for given real numbers $0 < \underline{\kappa} \le \overline{\kappa}$, and the Cauchy stress tensor is given by

$$\boldsymbol{\sigma}(\boldsymbol{u}) \coloneqq 2\mu\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} + \lambda \boldsymbol{I}_d \boldsymbol{\nabla}\cdot\boldsymbol{u},$$

with real numbers $\lambda \ge 0$ and $\mu > 0$ corresponding to Lamé's parameters, $\boldsymbol{\nabla}_{\mathrm{s}}$ denoting the symmetric part of the gradient operator applied to vector-valued fields, and $\boldsymbol{I}_d$ denoting the identity matrix of $\mathbb{R}^{d\times d}$. Equations (2.1a) and (2.1b) express, respectively, the mechanical equilibrium and the fluid mass balance. We consider, for the sake of simplicity, the following homogeneous boundary conditions:

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \partial\Omega\times(0,t_F), \tag{2.1c}$$

$$\boldsymbol{\kappa}\boldsymbol{\nabla}p\cdot\boldsymbol{n} = 0 \qquad \text{on } \partial\Omega\times(0,t_F). \tag{2.1d}$$

Initial conditions are set prescribing $\boldsymbol{u}(\cdot,0) = \boldsymbol{u}^0$ and, if $c_0 > 0$, $p(\cdot,0) = p^0$. In the incompressible case $c_0 = 0$, we also need the following compatibility condition on g:

$$\int_\Omega g(\cdot,t) = 0 \qquad \forall t \in (0,t_F), \tag{2.1e}$$

as well as the following zero-average constraint on p:

$$\int_\Omega p(\cdot,t) = 0 \qquad \forall t \in (0,t_F). \tag{2.1f}$$

For the derivation of the Biot model we refer to the seminal work of Terzaghi [190] and Biot [27, 29]. A theoretical study of problem (2.1) can be found in [181]. For the precise regularity assumptions on the data and on the solution under which our a priori bounds and convergence estimates are derived, we refer to Lemma 2.7 and Theorem 2.12, respectively.

A few simplifications are made to keep the exposition as simple as possible while still retaining all the principal difficulties. For the Biot–Willis coefficient we take

$$\alpha = 1,$$

an assumption often made in practice. For the scalar-valued permeability κ, we assume that it is piecewise constant on a partition P_Ω of Ω into bounded open polyhedra. The treatment of more general permeability coefficients can be done following the ideas of [77]. Also, more general boundary conditions than (2.1c)–(2.1d) can be considered up to minor modifications.

Our focus is here on a novel space discretization for the Biot problem (standard choices are made for the time discretization). Several difficulties have to be accounted for in the design of the space discretization of problem (2.1): in the context of nonconforming methods, the linear elasticity operator has to be carefully engineered to ensure stability expressed by a discrete counterpart of the Korn's inequality; the Darcy operator has to accomodate rough variations of the permeability coefficient; the choice of discrete spaces for the displacement and the pressure must satisfy an inf-sup condition to contribute reducing spurious pressure oscillations for small time steps combined with small permeabilities when $c_0 = 0$. An investigation of the role of the inf-sup condition in the context of finite element discretizations can be found, e.g., in Murad and Loula [154, 155]. A recent work of Rodrigo, Gaspar, Hu, and Zikatanov [176] has pointed out that, even for discretization methods leading to an inf-sup stable discretization of the Stokes problem in the steady case, pressure oscillations can arise owing to a lack of monotonicity. Therein, the authors suggest that stabilizing is possible by adding to the mass balance equation an artificial diffusion term with coefficient proportional to h^2/τ (with h and τ denoting, respectively, the spatial and temporal meshsizes). However, computing the exact amount of stabilization required is in general feasible only in 1 space dimension.

Several space discretization methods for the Biot problem have been considered in the literature. Finite element discretizations are discussed, e.g., in the monograph of Lewis and Schrefler [140]; cf. also references therein. A finite volume discretization for the three-dimensional Biot problem with discontinuous physical coefficients is considered by Naumovich [156]. In [165, 166], Phillips and Wheeler propose and analyze an algorithm that models displacements with continuous elements and the flow with a mixed method. In [167], the same authors also propose a different method where displacements are instead approximated using discontinuous Galerkin methods. In [200], Wheeler, Xue and Yotov study the coupling of multipoint flux discretization for the flow with a discontinuous Galerkin discretization of the displacements. While certainly effective on matching simplicial meshes, discontinuous Galerkin discretizations of the displacements usually do not allow to prove inf-sup stability on general polyhedral meshes.

In this chapter, we propose a novel space discretization of problem (2.1) where the linear elasticity operator is discretized using the Hybrid High-Order (HHO) method of [74] (c.f. also [71, 73, 78]), while the flow relies on the Symmetric Weighted Interior Penalty (SWIP) discontinuous Galerkin method of [77], see also [76, Chapter 4]. The proposed method has several assets: (i) It delivers an inf-sup stable discretization on general meshes including, e.g., polyhedral elements and nonmatching interfaces; (ii) it allows to increase the space approximation order to accelerate convergence in the presence of (locally) regular solutions;

(iii) it is locally conservative on the primal mesh, a desirable property for practitioners and key for a posteriori estimates based on equilibrated fluxes; (iv) it is robust with respect to the spatial variations of the permeability coefficient, with constants in the error estimates that depend on the square root of the heterogeneity ratio; (v) it is (relatively) inexpensive: at the lowest order, after static condensation of element unknowns for the displacement, we have 4 (resp. 9) unknowns per face for the displacements + 3 (resp. 4) unknowns per element for the pore pressure in 2d (resp. 3d). Finally, the proposed construction is valid for arbitrary space dimension, a feature which can be exploited in practice to conceive dimension-independent implementations.

The material is organized as follows. In Section 2.2, we introduce the discrete setting and formulate the method. In Section 2.3, we derive a priori bounds on the exact solution for regular-in-time volumetric load and mass source. The convergence analysis of the method is carried out in Section 2.4. Implementation details are discussed in Section 2.5, while numerical tests proposed in Section 2.6. Finally, in Appendix 2.7, we investigate the local conservation properties of the method by identifying computable conservative normal tractions and mass fluxes.

2.2 Discretization

In this section we introduce the assumptions on the mesh, define the discrete counterparts of the elasticity and Darcy operators and of the hydro-mechanical coupling terms, and formulate the discretization method.

2.2.1 Mesh and notation

Denote by $\mathcal{H} \subset \mathbb{R}_*^+$ a countable set of meshsizes having 0 as its unique accumulation point. Following [76, Chapter 1], we consider h-refined spatial mesh sequences $(\mathcal{T}_h)_{h\in\mathcal{H}}$ where, for all $h \in \mathcal{H}$, $\mathcal{T}_h$ is a finite collection of nonempty disjoint open polyhedral elements T such that $\overline{\Omega} = \bigcup_{T\in\mathcal{T}_h} \overline{T}$ and $h = \max_{T\in\mathcal{T}_h} h_T$ with h_T standing for the diameter of the element T. We assume that mesh regularity holds in the following sense: For all $h \in \mathcal{H}$, $\mathcal{T}_h$ admits a matching simplicial submesh $\mathfrak{T}_h$ and there exists a real number $\varrho > 0$ independent of h such that, for all $h \in \mathcal{H}$, (i) for all simplex $S \in \mathfrak{T}_h$ of diameter h_S and inradius r_S, $\varrho h_S \le r_S$ and (ii) for all $T \in \mathcal{T}_h$, and all $S \in \mathfrak{T}_h$ such that $S \subset T$, $\varrho h_T \le h_S$. A mesh sequence with this property is called regular. It is worth emphasizing that the simplicial submesh $\mathfrak{T}_h$ is just an analysis tool, and it is not used in the actual construction of the discretization method. These assumptions are essentially analogous to those made in the context of other recent methods supporting general meshes; cf., e.g., [20, Section 2.2] for the Virtual Element method. For a collection of useful geometric and functional inequalities that hold on regular mesh sequences we refer to [76, Chapter 1] and [69].

Remark 2.1 (Face degeneration). The above regularity assumptions on the mesh imply that the diameter of the mesh faces is uniformly comparable to that of the cell(s) they belong to, i.e., face degeneration is not allowed. Face degeneration has been considered, on the other hand, in [47] in the context of interior penalty discontinuous Galerkin methods. One

could expect that this framework could be used herein while adapting accordingly the penalty strategy (2.11) and (2.19). This point lies out of the scope of the present work and will be inspected in the future.

To avoid dealing with jumps of the permeability inside elements, we additionally assume that, for all $h \in \mathcal{H}$, $\mathcal{T}_h$ is compatible with the known partition P_Ω on which the diffusion coefficient $\boldsymbol{\kappa}$ is piecewise constant, so that jumps can only occur at interfaces.

We define a face F as a hyperplanar closed connected subset of $\overline{\Omega}$ with positive $(d-1)$-dimensional Hausdorff measure and such that (i) either there exist $T_1, T_2 \in \mathcal{T}_h$ such that $F \subset \partial T_1 \cap \partial T_2$ (with ∂T_i denoting the boundary of T_i) and F is called an interface or (ii) there exists $T \in \mathcal{T}_h$ such that $F \subset \partial T \cap \partial\Omega$ and F is called a boundary face. Interfaces are collected in the set $\mathcal{F}_h^{\mathrm{i}}$, boundary faces in $\mathcal{F}_h^{\mathrm{b}}$, and we let $\mathcal{F}_h \coloneqq \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^{\mathrm{b}}$. The diameter of a face $F \in \mathcal{F}_h$ is denoted by h_F. For all $T \in \mathcal{T}_h$, $\mathcal{F}_T \coloneqq \{F \in \mathcal{F}_h \,:\, F \subset \partial T\}$ denotes the set of faces contained in ∂T and, for all $F \in \mathcal{F}_T$, $\boldsymbol{n}_{TF}$ is the unit normal to F pointing out of T. For a regular mesh sequence, the maximum number of faces in $\mathcal{F}_T$ can be bounded by an integer N_∂ uniformly in h. For each interface $F \in \mathcal{F}_h^{\mathrm{i}}$, we fix once and for all the ordering for the elements $T_1, T_2 \in \mathcal{T}_h$ such that $F \subset \partial T_1 \cap \partial T_2$ and we let $\boldsymbol{n}_F \coloneqq \boldsymbol{n}_{T_1,F}$. For a boundary face, we simply take $\boldsymbol{n}_F = \boldsymbol{n}$, the outward unit normal to Ω.

For integers $l \geq 0$ and $s \geq 1$, we denote by $\mathbb{P}_d^l(\mathcal{T}_h)$ the space of fully discontinuous piecewise polynomial functions of total degree $\leq l$ on $\mathcal{T}_h$ and by $H^s(\mathcal{T}_h)$ the space of functions in $L^2(\Omega)$ that lie in $H^s(T)$ for all $T \in \mathcal{T}_h$. The notation $H^s(P_\Omega)$ will also be used with obvious meaning. Under the mesh regularity assumptions detailed above, using [76, Lemma 1.40] together with the results of [88], one can prove that there exists a real number C_{app} depending on ϱ and l, but independent of h, such that, denoting by π_h^l the L^2-orthogonal projector on $\mathbb{P}_d^l(\mathcal{T}_h)$, the following holds: For all $s \in \{1, \ldots, l+1\}$ and all $v \in H^s(\mathcal{T}_h)$,

$$|v - \pi_h^l v|_{H^m(\mathcal{T}_h)} \leq C_{\mathrm{app}} h^{s-m} |v|_{H^s(\mathcal{T}_h)} \qquad \forall m \in \{0, \ldots, s-1\}. \tag{2.2}$$

For an integer $l \geq 0$, we consider the space

$$C^l(V) \coloneqq C^l([0, t_{\mathrm{F}}]; V),$$

spanned by V-valued functions that are l times continuously differentiable in the time interval $[0, t_{\mathrm{F}}]$. The space $C^0(V)$ is a Banach space when equipped with the norm $\|\varphi\|_{C^0(V)} \coloneqq \max_{t\in[0,t_{\mathrm{F}}]} \|\varphi(t)\|_V$, and the space $C^l(V)$ is a Banach space when equipped with the norm $\|\varphi\|_{C^l(V)} \coloneqq \max_{0\leq m\leq l} \|\mathrm{d}_t^m \varphi\|_{C^0(V)}$. For the time discretization, we consider a uniform mesh of the time interval $(0, t_{\mathrm{F}})$ of step $\tau \coloneqq t_{\mathrm{F}}/N$ with $N \in \mathbb{N}^*$, and introduce the discrete times $t^n \coloneqq n\tau$ for all $0 \leq n \leq N$. For any $\varphi \in C^l(V)$, we set $\varphi^n \coloneqq \varphi(t^n) \in V$, and we introduce the backward differencing operator δ_t such that, for all $1 \leq n \leq N$,

$$\delta_t \varphi^n \coloneqq \frac{\varphi^n - \varphi^{n-1}}{\tau} \in V. \tag{2.3}$$

In what follows, for $X \subset \overline{\Omega}$, we respectively denote by $(\cdot,\cdot)_X$ and $\|\cdot\|_X$ the standard inner product and norm in $L^2(X)$, with the convention that the subscript is omitted whenever $X = \Omega$. The same notation is used in the vector- and tensor-valued cases. For the sake of

brevity, throughout the chapter, we will often use the notation $a \lesssim b$ for the inequality $a \le Cb$ with generic constant $C > 0$ independent of h, τ, c_0, λ, μ, and $\boldsymbol{\kappa}$, but possibly depending on ϱ and the polynomial degree k. We will name generic constants only in statements or when this helps to follow the proofs.

2.2.2 Linear elasticity operator

The discretization of the linear elasticity operator is based on the Hybrid High-Order method of [74]. Let a polynomial degree $k \ge 1$ be fixed. The degrees of freedom (DOFs) for the displacement are collected in the space

$$\underline{\boldsymbol{U}}_h^k := \left\{ \bigtimes_{T\in\mathcal{T}_h} \mathbb{P}_d^k(T)^d \right\} \times \left\{ \bigtimes_{F\in\mathcal{F}_h} \mathbb{P}_{d-1}^k(F)^d \right\}.$$

For a generic collection of DOFs in $\underline{\boldsymbol{U}}_h^k$ we use the notation $\underline{\boldsymbol{v}}_h := ((\boldsymbol{v}_T)_{T\in\mathcal{T}_h}, (\boldsymbol{v}_F)_{F\in\mathcal{F}_h})$. We also denote by $\boldsymbol{v}_h$ (not underlined) the function of $\mathbb{P}_d^k(\mathcal{T}_h)^d$ such that $\boldsymbol{v}_{h|T} = \boldsymbol{v}_T$ for all $T \in \mathcal{T}_h$. The restrictions of $\underline{\boldsymbol{U}}_h^k$ and $\underline{\boldsymbol{v}}_h$ to an element T are denoted by $\underline{\boldsymbol{U}}_T^k$ and $\underline{\boldsymbol{v}}_T = (\boldsymbol{v}_T, (\boldsymbol{v}_F)_{F\in\mathcal{F}_T})$, respectively. For further use, we define the reduction map $\underline{\boldsymbol{I}}_h^k : H^1(\Omega)^d \to \underline{\boldsymbol{U}}_h^k$ such that, for all $\boldsymbol{v} \in H^1(\Omega)^d$,

$$\underline{\boldsymbol{I}}_h^k \boldsymbol{v} = ((\pi_T^k \boldsymbol{v})_{T\in\mathcal{T}_h}, (\pi_F^k \boldsymbol{v})_{F\in\mathcal{F}_h}), \tag{2.4}$$

where π_T^k and π_F^k denote the L^2-orthogonal projectors on $\mathbb{P}_d^k(T)$ and $\mathbb{P}_{d-1}^k(F)$, respectively. For all $T \in \mathcal{T}_h$, the reduction map on $\underline{\boldsymbol{U}}_T^k$ obtained by a restriction of $\underline{\boldsymbol{I}}_h^k$ is denoted by $\underline{\boldsymbol{I}}_T^k$.

For all $T \in \mathcal{T}_h$, we obtain a high-order polynomial reconstruction $\boldsymbol{r}_T^{k+1} : \underline{\boldsymbol{U}}_T^k \to \mathbb{P}_d^{k+1}(T)^d$ of the displacement field by solving the following local pure traction problem:c2: For a given local collection of DOFs $\underline{\boldsymbol{v}}_T = (\boldsymbol{v}_T, (\boldsymbol{v}_F)_{F\in\mathcal{F}_T}) \in \underline{\boldsymbol{U}}_T^k$, find $\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T \in \mathbb{P}_d^{k+1}(T)^d$ such that

$$(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w})_T = (\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_T, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w})_T + \sum_{F\in\mathcal{F}_T} (\boldsymbol{v}_F - \boldsymbol{v}_T, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w}\,\boldsymbol{n}_{TF})_F \quad \forall \boldsymbol{w} \in \mathbb{P}_d^{k+1}(T)^d. \tag{2.5}$$

In order to uniquely define the solution to (2.5), we prescribe the conditions $\int_T \boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T = \int_T \boldsymbol{v}_T$ and $\int_T \boldsymbol{\nabla}_{\mathrm{ss}}\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T = \sum_{F\in\mathcal{F}_T} \int_F \frac{1}{2}(\boldsymbol{v}_F \otimes \boldsymbol{n}_{TF} - \boldsymbol{n}_{TF} \otimes \boldsymbol{v}_F)$, where $\boldsymbol{\nabla}_{\mathrm{ss}}$ denotes the skew-symmetric part of the gradient operator. We also define the global reconstruction of the displacement $\boldsymbol{r}_h^{k+1} : \underline{\boldsymbol{U}}_h^k \to \mathbb{P}_d^{k+1}(\mathcal{T}_h)^d$ such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$,

$$(\boldsymbol{r}_h^{k+1}\underline{\boldsymbol{v}}_h)_{|T} = \boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T \qquad \forall T \in \mathcal{T}_h.$$

The following approximation property is proved in [74, Lemma 2]: For all $\boldsymbol{v} \in H^1(\Omega)^d \cap H^{k+2}(P_\Omega)^d$,

$$\|\boldsymbol{\nabla}_{\mathrm{s}}(\boldsymbol{r}_h^{k+1}\underline{\boldsymbol{I}}_h^k\boldsymbol{v} - \boldsymbol{v})\| \lesssim h^{k+1}\|\boldsymbol{v}\|_{H^{k+2}(P_\Omega)^d}. \tag{2.6}$$

We next introduce the discrete divergence operator $D_T^k : \underline{\boldsymbol{U}}_T^k \to \mathbb{P}_d^k(T)$ such that, for all

$q \in \mathbb{P}_d^k(T)$

$$(D_T^k \underline{\boldsymbol{v}}_T, q)_T = (\boldsymbol{\nabla}\cdot\boldsymbol{v}_T, q)_T + \sum_{F\in\mathcal{F}_T} (\boldsymbol{v}_F - \boldsymbol{v}_T, q\boldsymbol{n}_{TF})_F \tag{2.7a}$$

$$= -(\boldsymbol{v}_T, \boldsymbol{\nabla} q)_T + \sum_{F\in\mathcal{F}_T} (\boldsymbol{v}_F, q\boldsymbol{n}_{TF})_F, \tag{2.7b}$$

where we have used integration by parts to pass to the second line. The divergence operator satisfies the following commuting property: For all $T \in \mathcal{T}_h$ and all $\boldsymbol{v} \in H^1(T)^d$,

$$D_T^k \underline{\boldsymbol{I}}_T^k \boldsymbol{v} = \pi_T^k(\boldsymbol{\nabla}\cdot\boldsymbol{v}). \tag{2.8}$$

The local contribution to the discrete linear elasticity operator is expressed by the bilinear form a_T on $\underline{\boldsymbol{U}}_T^k \times \underline{\boldsymbol{U}}_T^k$ such that, for all $\underline{\boldsymbol{w}}_T, \underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}_T^k$,

$$a_T(\underline{\boldsymbol{w}}_T, \underline{\boldsymbol{v}}_T) \coloneqq 2\mu \left\{ (\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{r}_T^{k+1} \underline{\boldsymbol{w}}_T, \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{r}_T^{k+1} \underline{\boldsymbol{v}}_T)_T + s_T(\underline{\boldsymbol{w}}_T, \underline{\boldsymbol{v}}_T) \right\} + \lambda (D_T^k \underline{\boldsymbol{w}}_T, D_T^k \underline{\boldsymbol{v}}_T)_T, \tag{2.9}$$

where the stabilization bilinear form s_T is such that

$$s_T(\underline{\boldsymbol{w}}_T, \underline{\boldsymbol{v}}_T) \coloneqq \sum_{F\in\mathcal{F}_T} h_F^{-1} (\boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{w}}_T, \boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{v}}_T)_F, \tag{2.10}$$

with face-based residual such that, for all $\underline{\boldsymbol{w}}_T \in \underline{\boldsymbol{U}}_T^k$,

$$\boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{w}}_T \coloneqq (\pi_F^k \boldsymbol{r}_T^{k+1} \underline{\boldsymbol{w}}_T - \boldsymbol{w}_F) - (\pi_T^k \boldsymbol{r}_T^{k+1} \underline{\boldsymbol{w}}_T - \boldsymbol{w}_T).$$

The global bilinear form a_h on $\underline{\boldsymbol{U}}_h^k \times \underline{\boldsymbol{U}}_h^k$ is assembled element-wise from local contributions:

$$a_h(\underline{\boldsymbol{w}}_h, \underline{\boldsymbol{v}}_h) \coloneqq \sum_{T\in\mathcal{T}_h} a_T(\underline{\boldsymbol{w}}_T, \underline{\boldsymbol{v}}_T). \tag{2.11}$$

To account for the zero-displacement boundary condition (2.1c), we consider the subspace

$$\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \coloneqq \left\{ \underline{\boldsymbol{v}}_h = ((\boldsymbol{v}_T)_{T\in\mathcal{T}_h}, (\boldsymbol{v}_F)_{F\in\mathcal{F}_h}) \in \underline{\boldsymbol{U}}_h^k \;:\; \boldsymbol{v}_F \equiv \boldsymbol{0} \quad \forall F \in \mathcal{F}_h^{\mathrm{b}} \right\}. \tag{2.12}$$

Define on $\underline{\boldsymbol{U}}_h^k$ the discrete strain seminorm

$$\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2 \coloneqq \sum_{T\in\mathcal{T}_h} \|\underline{\boldsymbol{v}}_h\|_{\epsilon,T}^2, \qquad \|\underline{\boldsymbol{v}}_h\|_{\epsilon,T}^2 \coloneqq \|\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1} \|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F^2. \tag{2.13}$$

It can be proved that $\|\cdot\|_{\epsilon,h}$ defines a norm on $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$. Moreover, using [74, Corollary 6], one has the following coercivity and boundedness result for a_h:

$$\eta^{-1}(2\mu)\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2 \le \|\underline{\boldsymbol{v}}_h\|_{a,h}^2 \coloneqq a_h(\underline{\boldsymbol{v}}_h, \underline{\boldsymbol{v}}_h) \le \eta(2\mu + d\lambda)\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2, \tag{2.14}$$

where $\eta > 0$ is a real number independent of h, τ and the physical coefficients. Additionally, we know from [74, Theorem 8] that, for all $\boldsymbol{w} \in H_0^1(\Omega)^d \cap H^{k+2}(P_\Omega)^d$ such that $\boldsymbol{\nabla}\cdot\boldsymbol{w} \in H^{k+1}(P_\Omega)$ and all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, the following consistency result holds:

$$\left| a_h(\underline{\boldsymbol{I}}_h^k \boldsymbol{w}, \underline{\boldsymbol{v}}_h) + (\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{w}), \boldsymbol{v}_h) \right| \lesssim h^{k+1} \left(2\mu \|\boldsymbol{w}\|_{H^{k+2}(P_\Omega)^d} + \lambda \|\boldsymbol{\nabla}\cdot\boldsymbol{w}\|_{H^{k+1}(P_\Omega)} \right) \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}. \tag{2.15}$$

To close this section, we give the following discrete counterpart of Korn's inequality (see 4.3 for the proof).

Proposition 2.2 (Discrete Korn's inequality)**.** *There is a real number $C_{\mathrm{K}} > 0$ depending on ϱ and on k but independent of h such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, recalling that $\boldsymbol{v}_h \in \mathbb{P}_d^k(\mathcal{T}_h)^d$ denotes the broken polynomial function such that $\boldsymbol{v}_{h|T} = \boldsymbol{v}_T$ for all $T \in \mathcal{T}_h$,*

$$\|\boldsymbol{v}_h\| \le C_{\mathrm{K}} d_\Omega \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}, \tag{2.16}$$

where d_Ω denotes the diameter of Ω.

2.2.3 Darcy operator

The discretization of the Darcy operator is based on the Symmetric Weighted Interior Penalty method of [77], cf. also [76, Section 4.5]. At each time step, the discrete pore pressure is sought in the broken polynomial space

$$P_h^k := \begin{cases} \mathbb{P}_d^k(\mathcal{T}_h) & \text{if } c_0 > 0, \\ \mathbb{P}_{d,0}^k(\mathcal{T}_h) & \text{if } c_0 = 0, \end{cases} \tag{2.17}$$

where we have introduced the zero-average subspace $\mathbb{P}_{d,0}^k(\mathcal{T}_h) := \left\{ q_h \in \mathbb{P}_d^k(\mathcal{T}_h) \,:\, (q_h, 1) = 0 \right\}$. For all $F \in \mathcal{F}_h^{\mathrm{i}}$, we define the jump and (weighted) average operators such that, for all $\varphi \in H^1(\mathcal{T}_h)$, denoting by φ_T and $\boldsymbol{\kappa}_T$ the restrictions of φ and $\boldsymbol{\kappa}$ to $T \in \mathcal{T}_h$, respectively,

$$[\varphi]_F := \varphi_{T_1} - \varphi_{T_2}, \qquad \{\varphi\}_F := \omega_{T_1}\varphi_{T_1} + \omega_{T_2}\varphi_{T_2}, \tag{2.18}$$

where $\omega_{T_1} = 1 - \omega_{T_2} := \frac{\boldsymbol{\kappa}_{T_2}}{(\boldsymbol{\kappa}_{T_1} + \boldsymbol{\kappa}_{T_2})}$. Denoting by $\boldsymbol{\nabla}_h$ the broken gradient on $H^1(\mathcal{T}_h)$ and letting, for all $F \in \mathcal{F}_h^{\mathrm{i}}$, $\lambda_{\boldsymbol{\kappa},F} := \frac{2\boldsymbol{\kappa}_{T_1}\boldsymbol{\kappa}_{T_2}}{(\boldsymbol{\kappa}_{T_1} + \boldsymbol{\kappa}_{T_2})}$, we define the bilinear form c_h on $P_h^k \times P_h^k$ such that, for all $q_h, r_h \in P_h^k$,

$$\begin{aligned} c_h(r_h, q_h) := {} & (\boldsymbol{\kappa}\boldsymbol{\nabla}_h r_h, \boldsymbol{\nabla}_h q_h) + \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} \frac{\varsigma \lambda_{\boldsymbol{\kappa},F}}{h_F} ([r_h]_F, [q_h]_F)_F \\ & - \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} ((\{\boldsymbol{\kappa}\boldsymbol{\nabla}_h r_h\}_F \cdot \boldsymbol{n}_F, [q_h]_F)_F + ([r_h]_F, \{\boldsymbol{\kappa}\boldsymbol{\nabla}_h q_h\}_F \cdot \boldsymbol{n}_F)_F), \end{aligned} \tag{2.19}$$

where $\varsigma > 0$ is a user-defined penalty parameter. The fact that the boundary terms only appear on internal faces in (2.19) reflects the Neumann boundary condition (2.1d). From this point on, we will assume that $\varsigma > C_{\mathrm{tr}}^2 N_\partial$ with C_{tr} denoting the constant from the discrete trace inequality [76, Eq. (1.37)], which ensures that the bilinear form c_h is coercive (in the numerical tests of Section 2.6, we took $\varsigma = (N_\partial + 0.1)k^2$). Since the bilinear form c_h is also symmetric, it defines a seminorm on P_h^k, denoted hereafter by $\|\cdot\|_{c,h}$ (the map $\|\cdot\|_{c,h}$ is in fact a norm on $\mathbb{P}_{d,0}^k(\mathcal{T}_h)$).

Remark 2.3 (Alternative stabilization). To get rid of the dependence of the lower threshold of ς on C_{tr}, one can resort to the BR2 stabilization; c.f. [17] and also [76, Section 5.3.2]. In passing, this stabilization could also contribute to handle face degeneration since the penalty parameter no longer depends on the inverse of the face diameter (cf. Remark 2.1). This topic will make the object of future investigations.

The following known results will be needed in the analysis. Let

$$P_* \coloneqq \left\{ r \in H^1(\Omega) \cap H^2(P_\Omega) \ : \ \boldsymbol{\kappa}\boldsymbol{\nabla} r \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega \right\}, \qquad P_{*h}^k \coloneqq P_* + P_h^k.$$

Extending the bilinear form c_h to $P_{*h}^k \times P_{*h}^k$, the following consistency result can be proved adapting the arguments of [76, Chapter 4] to account for the homogeneous Neumann boundary condition (2.1d):

$$\forall r \in P_*, \qquad -(\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla} r), q) = c_h(r, q) \qquad \forall q \in P_{*h}. \tag{2.20}$$

Assuming, additionally, that $r \in H^{k+2}(P_\Omega)$, as a consequence of [76, Lemma 5.52] together with the optimal approximation properties (2.2) of π_h^k on regular mesh sequences one has,

$$\sup_{q_h \in \mathbb{P}_{d,0}^k(\mathcal{T}_h)\setminus\{0\}} \frac{c_h(r - \pi_h^k r, q_h)}{\|q_h\|_{c,h}} \lesssim \overline{\kappa}^{1/2} h^k \|r\|_{H^{k+1}(P_\Omega)}.$$

2.2.4 Hydro-mechanical coupling

The hydro-mechanical coupling is realized by means of the bilinear form b_h on $\underline{\boldsymbol{U}}_h^k \times \mathbb{P}_d^k(\mathcal{T}_h)$ such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$ and all $q_h \in \mathbb{P}_d^k(\mathcal{T}_h)$,

$$b_h(\underline{\boldsymbol{v}}_h, q_h) \coloneqq \sum_{T\in\mathcal{T}_h} b_T(\underline{\boldsymbol{v}}_T, q_{h|T}), \qquad b_T(\underline{\boldsymbol{v}}_T, q_{h|T}) \coloneqq -(D_T^k \underline{\boldsymbol{v}}_T, q_{h|T})_T, \tag{2.21}$$

where D_T^k is the discrete divergence operator defined by (2.7a). A simple verification shows that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$ and all $q_h \in \mathbb{P}_d^k(\mathcal{T}_h)$,

$$b_h(\underline{\boldsymbol{v}}_h, q_h) \lesssim \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h} \|q_h\|. \tag{2.22}$$

Additionally, using the definition (2.7a) of D_T^k and (2.12) of $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, it can be proved that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, it holds ($\chi_\Omega$ denotes here the characteristic function of Ω),

$$b_h(\underline{\boldsymbol{v}}_h, \chi_\Omega) = 0. \tag{2.23}$$

The following inf-sup condition expresses the stability of the hydro-mechanical coupling:

Lemma 2.4 (Inf-sup condition for b_h)**.** *There is a real number β depending on Ω, ϱ and k but independent of h such that, for all $q_h \in \mathbb{P}_{d,0}^k(\mathcal{T}_h)$,*

$$\|q_h\| \le \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k\setminus\{\underline{\boldsymbol{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, q_h)}{\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}}. \tag{2.24}$$

Proof. Let $q_h \in \mathbb{P}_{d,0}^k(\mathcal{T}_h)$. Classically [31], there is $\boldsymbol{v}_{q_h} \in H_0^1(\Omega)^d$ such that $\boldsymbol{\nabla}\cdot\boldsymbol{v}_{q_h} = q_h$ and $\|\boldsymbol{v}_{q_h}\|_{H^1(\Omega)^d} \lesssim \|q_h\|$. Let $T \in \mathcal{T}_h$. Using the H^1-stability of the L^2-orthogonal projector (cf., e.g., [69, Corollary 3.7]), it is inferred that

$$\|\boldsymbol{\nabla}_{\mathrm{s}} \pi_T^k \boldsymbol{v}_{q_h}\|_T \le \|\boldsymbol{\nabla}\boldsymbol{v}_{q_h}\|_T.$$

Moreover, for all $F \in \mathcal{F}_T$, using the boundedness of π_F^k and the continuous trace inequality of [76, Lemma 1.49] followed by a local Poincaré's inequality for the zero-average function $(\pi_T^k \boldsymbol{v}_{q_h} - \boldsymbol{v}_{q_h})$, we have

$$h_F^{-1/2} \|\pi_F^k(\pi_T^k \boldsymbol{v}_{q_h} - \boldsymbol{v}_{q_h})\|_F \le h_F^{-1/2} \|\pi_T^k \boldsymbol{v}_{q_h} - \boldsymbol{v}_{q_h}\|_F \lesssim \|\boldsymbol{\nabla} \boldsymbol{v}_{q_h}\|_T.$$

As a result, recalling the definition (2.4) of the local reduction map $\underline{I}_T^k$ and (2.13) of the strain norm $\|\cdot\|_{\epsilon,T}$, it follows that $\|\underline{I}_T^k \boldsymbol{v}_{q_h}\|_{\epsilon,T} \lesssim \|\boldsymbol{v}_{q_h}\|_{H^1(T)^d}$. Squaring and summing over $T \in \mathcal{T}_h$ the latter inequality, we get

$$\|\underline{I}_h^k \boldsymbol{v}_{q_h}\|_{\epsilon,h} \lesssim \|\boldsymbol{v}_{q_h}\|_{H^1(\Omega)^d} \lesssim \|q_h\|. \tag{2.25}$$

Using (2.25), the commuting property (2.8), and denoting by S the supremum in (2.24), one has

$$\|q_h\|^2 = (\boldsymbol{\nabla}\cdot \boldsymbol{v}_{q_h}, q_h) = \sum_{T\in\mathcal{T}_h} (D_T^k \underline{I}_T^k \boldsymbol{v}_{q_h}, q_h)_T = -b_h(\underline{I}_h^k \boldsymbol{v}_{q_h}, q_h) \le \mathsf{S}\|\underline{I}_h^k \boldsymbol{v}_{q_h}\|_{\epsilon,h} \lesssim \mathsf{S}\|q_h\|. \quad \square$$

2.2.5 Formulation of the method

For all $1 \le n \le N$, the discrete solution $(\underline{\boldsymbol{u}}_h^n, p_h^n) \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times P_h^k$ at time t^n is such that, for all $(\underline{\boldsymbol{v}}_h, q_h) \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times \mathbb{P}_d^k(\mathcal{T}_h)$,

$$a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{v}}_h) + b_h(\underline{\boldsymbol{v}}_h, p_h^n) = l_h^n(\underline{\boldsymbol{v}}_h), \tag{2.26a}$$

$$(c_0 \delta_t p_h^n, q_h) - b_h(\delta_t \underline{\boldsymbol{u}}_h^n, q_h) + c_h(p_h^n, q_h) = (g^n, q_h), \tag{2.26b}$$

where the linear form l_h^n on $\underline{\boldsymbol{U}}_h^k$ is defined as

$$l_h^n(\underline{\boldsymbol{v}}_h) \coloneqq (\boldsymbol{f}^n, \boldsymbol{v}_h) = \sum_{T\in\mathcal{T}_h} (\boldsymbol{f}^n, \boldsymbol{v}_T)_T. \tag{2.27}$$

In petroleum engineering, the usual way to enforce the initial condition is to compute a displacement from an initial (usually hydrostatic) pressure distribution. For a given scalar-valued initial pressure field $p^0 \in L^2(\Omega)$, we let $\widehat{p}_h^0 \coloneqq \pi_h^k p^0$ and set $\underline{\boldsymbol{u}}_h^0 = \underline{\widehat{\boldsymbol{u}}}_h^0$ with $\underline{\widehat{\boldsymbol{u}}}_h^0 \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ unique solution of

$$a_h(\underline{\widehat{\boldsymbol{u}}}_h^0, \underline{\boldsymbol{v}}_h) = l_h^0(\underline{\boldsymbol{v}}_h) - b_h(\underline{\boldsymbol{v}}_h, \widehat{p}_h^0) \qquad \forall \underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k. \tag{2.28}$$

If $c_0 = 0$, the value of $\widehat{p}_h^0$ is only needed to enforce the initial condition on the displacement while, if $c_0 > 0$, we also set $p_h^0 = \widehat{p}_h^0$ to initialize the discrete pressure.

Remark 2.5 (Discrete compatibility condition for $c_0 = 0$). Also when $c_0 = 0$ it is possible to take the test function q_h in (2.26b) in the full space $\mathbb{P}_d^k(\mathcal{T}_h)$ instead of the zero-average subspace $\mathbb{P}_{d,0}^k(\mathcal{T}_h)$, since the compatibility condition is verified at the discrete level. To check it, it suffices to let $q_h = \chi_\Omega$ in (2.26b), observe that the right-hand side is equal to zero since g^n has zero average on Ω (cf. (2.1e)), and use the definition (2.19) of c_h together with (2.23) to prove that the left-hand side also vanishes. This remark is crucial to ensure the local conservation properties of the method detailed in Section 2.7.

2.3 Stability analysis

In this section we study the stability of problem (2.26) and prove its well-posedness. We recall the following discrete Gronwall's inequality, which is a minor variation of [121, Lemma 5.1].

Lemma 2.6 (Discrete Gronwall's inequality)**.** *Let an integer N and reals $\delta, G > 0$, and $K \geq 0$ be given, and let $(a^n)_{0\leq n\leq N}$, $(b^n)_{0\leq n\leq N}$, and $(\gamma^n)_{0\leq n\leq N}$ denote three sequences of nonnegative real numbers such that, for all $0 \leq n \leq N$*

$$a^n + \delta \sum_{m=0}^{n} b^m + K \leq \delta \sum_{m=0}^{n} \gamma^m a^m + G.$$

Then, if $\gamma^m \delta < 1$ for all $0 \leq m \leq N$, letting $\varsigma^m := (1-\gamma^m\delta)^{-1}$, it holds, for all $0 \leq n \leq N$,

$$a^n + \delta \sum_{m=0}^{n} b^m + K \leq \exp\left(\delta \sum_{m=0}^{n} \varsigma^m \gamma^m\right) \times G. \tag{2.29}$$

Lemma 2.7 (A priori bounds)**.** *Assume $\boldsymbol{f} \in C^1(L^2(\Omega)^d)$ and $g \in C^0(L^2(\Omega))$, and let $(\underline{\boldsymbol{u}}_h^0, p_h^0) = (\widehat{\underline{\boldsymbol{u}}}_h^0, \widehat{p}_h^0)$ with $(\widehat{\underline{\boldsymbol{u}}}_h^0, \widehat{p}_h^0)$ defined as in Section 2.2.5. For all $1 \leq n \leq N$, denote by $(\underline{\boldsymbol{u}}_h^n, p_h^n)$ the solution to (2.26). Then, for τ small enough, it holds that*

$$\begin{aligned}
&\|\underline{\boldsymbol{u}}_h^N\|_{a,h}^2 + \|c_0^{1/2} p_h^N\|^2 + \frac{1}{2\mu + d\lambda}\|p_h^N - \overline{p}_h^N\|^2 + \sum_{n=1}^{N} \tau \|p_h^n\|_{c,h}^2 \lesssim \\
&\left(\frac{1}{2\mu} + c_0\right)\|p^0\|^2 + \frac{d_\Omega^2}{2\mu}\|\boldsymbol{f}\|_{C^1(L^2(\Omega)^d)}^2 + (2\mu + d\lambda) t_{\mathrm{F}}^2 \|g\|_{C^0(L^2(\Omega))}^2 + \frac{t_{\mathrm{F}}^2}{c_0}\|\overline{g}\|_{C^0(L^2(\Omega))}^2,
\end{aligned} \tag{2.30}$$

with the convention that $c_0^{-1}\|\overline{g}\|_{C^0(L^2(\Omega))}^2 = 0$ if $c_0 = 0$ and, for $0 \leq n \leq N$, $\overline{p}_h^n := (p_h^n, 1)$.

Remark 2.8 (Well-posedness). Owing to linearity, the well-posedness of (2.26) is an immediate consequence of Lemma 2.7.

Remark 2.9 (A priori bound for $c_0 = 0$). When $c_0 = 0$, the choice (2.17) of the discrete space for the pressure ensures that $\overline{p}_h^n = 0$ for all $0 \leq n \leq N$. Thus, the third term in the left-hand side of (2.30) yields an estimate on $\|p_h^N\|^2$, and the a priori bound reads

$$\begin{aligned}
\|\underline{\boldsymbol{u}}_h^N\|_{a,h}^2 + \frac{1}{2\mu + d\lambda}\|p_h^N\|^2 + \sum_{n=1}^{N} \tau \|p_h^n\|_{c,h}^2 \lesssim \\
(2\mu)^{-1}\left(d_\Omega^2 \|\boldsymbol{f}\|_{C^1(L^2(\Omega)^d)}^2 + \|p^0\|^2\right) + (2\mu + d\lambda) t_{\mathrm{F}}^2 \|g\|_{C^0(L^2(\Omega))}^2.
\end{aligned}$$

The convention $c_0^{-1}\|\overline{g}\|_{C^0(L^2(\Omega))}^2 = 0$ if $c_0 = 0$ is justified since the term $\mathfrak{T}_2$ in point (4) of the following proof vanishes in this case thanks to the compatibility condition (2.1e).

Proof of Lemma 2.7. Throughout the proof, C_i with $i \in \mathbb{N}^*$ will denote a generic positive constant independent of h, τ, and of the physical parameters c_0, λ, μ, and $\boldsymbol{\kappa}$.

(1) *Estimate of* $\|p_h^n - \overline{p}_h^n\|$. Using the inf-sup condition (2.24) followed by (2.23) to infer that $b_h(\underline{\boldsymbol{v}}_h, \overline{p}_h^n) = 0$, the mechanical equilibrium equation (2.26a), and the second inequality in (2.14), for all $1 \le n \le N$ we get

$$\begin{aligned}\|p_h^n - \overline{p}_h^n\| &\le \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\underline{\mathbf{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, p_h^n - \overline{p}_h^n)}{\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}} = \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\underline{\mathbf{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, p_h^n)}{\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}} \\ &= \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\underline{\mathbf{0}}\}} \frac{l_h^n(\underline{\boldsymbol{v}}_h) - a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{v}}_h)}{\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}} \le C_1^{1/2} \left(d_\Omega \|\boldsymbol{f}^n\| + (2\mu + d\lambda)^{1/2} \|\underline{\boldsymbol{u}}_h^n\|_{a,h} \right),\end{aligned}$$

where we have set, for the sake of brevity, $C_1^{1/2} \coloneqq \beta \max(C_{\mathrm{K}}, \eta)$. This implies, in particular,

$$\|p_h^n - \overline{p}_h^n\|^2 \le 2C_1 \left(d_\Omega^2 \|\boldsymbol{f}^n\|^2 + (2\mu + d\lambda) \|\underline{\boldsymbol{u}}_h^n\|_{a,h}^2 \right) \tag{2.31}$$

(2) *Energy balance.* Adding (2.26a) with $\underline{\boldsymbol{v}}_h = \tau \delta_t \underline{\boldsymbol{u}}_h^n$ to (2.26b) with $q_h = \tau p_h^n$, and summing the resulting equation over $1 \le n \le N$, it is inferred

$$\sum_{n=1}^N \tau a_h(\underline{\boldsymbol{u}}_h^n, \delta_t \underline{\boldsymbol{u}}_h^n) + \sum_{n=1}^N \tau (c_0 \delta_t p_h^n, p_h^n) + \sum_{n=1}^N \tau \|p_h^n\|_{c,h}^2 = \sum_{n=1}^N \tau l_h^n(\delta_t \underline{\boldsymbol{u}}_h^n) + \sum_{n=1}^N \tau (g^n, p_h^n). \tag{2.32}$$

We denote by $\mathcal{L}$ and $\mathcal{R}$ the left- and right-hand side of (2.32) and proceed to find suitable lower and upper bounds, respectively.

(3) *Lower bound for* $\mathcal{L}$. Using twice the formula

$$2x(x - y) = x^2 + (x - y)^2 - y^2, \tag{2.33}$$

and telescoping out the appropriate summands, the first two terms in the left-hand side of (2.32) can be rewritten as, respectively,

$$\begin{aligned}\sum_{n=1}^N \tau a_h(\underline{\boldsymbol{u}}_h^n, \delta_t \underline{\boldsymbol{u}}_h^n) &= \frac{1}{2} \|\underline{\boldsymbol{u}}_h^N\|_{a,h}^2 + \frac{1}{2} \sum_{n=1}^N \tau^2 \|\delta_t \underline{\boldsymbol{u}}_h^n\|_{a,h}^2 - \frac{1}{2} \|\underline{\boldsymbol{u}}_h^0\|_{a,h}^2, \\ \sum_{n=1}^N \tau (c_0 \delta_t p_h^n, p_h^n) &= \frac{1}{2} \|c_0^{1/2} p_h^N\|^2 + \frac{1}{2} \sum_{n=1}^N \tau^2 \|c_0^{1/2} \delta_t p_h^n\|^2 - \frac{1}{2} \|c_0^{1/2} p_h^0\|^2.\end{aligned}$$

Using the above relation together with (2.31) and $\|\boldsymbol{f}^N\| \le \|\boldsymbol{f}\|_{C^1(L^2(\Omega)^d)}$, it is inferred that

$$\begin{aligned}&\frac{1}{4} \|\underline{\boldsymbol{u}}_h^N\|_{a,h}^2 - \frac{1}{2} \|\underline{\boldsymbol{u}}_h^0\|_{a,h}^2 + \frac{1}{2} \|c_0^{1/2} p_h^N\|^2 - \frac{1}{2} \|c_0^{1/2} p_h^0\|^2 \\ &\quad + \frac{1}{8C_1(2\mu + d\lambda)} \|p_h^N - \overline{p}_h^N\|^2 + \sum_{n=1}^N \tau \|p_h^n\|_{c,h}^2 \le \mathcal{L} + \frac{d_\Omega^2}{4(2\mu + d\lambda)} \|\boldsymbol{f}\|_{C^1(L^2(\Omega)^d)}^2.\end{aligned} \tag{2.34}$$

(4) *Upper bound for* $\mathcal{R}$. For the first term in the right-hand side of (2.32), discrete integration by parts in time yields

$$\sum_{n=1}^{N} \tau l_h^n(\delta_t \underline{\boldsymbol{u}}_h^n) = (\boldsymbol{f}^N, \boldsymbol{u}_h^N) - (\boldsymbol{f}^0, \boldsymbol{u}_h^0) - \sum_{n=1}^{N} \tau(\delta_t \boldsymbol{f}^n, \boldsymbol{u}_h^{n-1}),$$

hence, using the Cauchy–Schwarz inequality, the discrete Korn's inequality followed by (2.14) to estimate $\|\boldsymbol{u}_h^n\|^2 \le \frac{C_2 d_\Omega^2}{\mu} \|\underline{\boldsymbol{u}}_h^n\|_{a,h}^2$ for all $1 \le n \le N$ (with $C_2 := C_{\mathrm{K}}^2 \eta/2$), and Young's inequality, one has

$$\begin{aligned}
\left| \sum_{n=1}^{N} \tau l_h^n(\delta_t \underline{\boldsymbol{u}}_h^n) \right| &\le \frac{1}{8} \left(\|\underline{\boldsymbol{u}}_h^N\|_{a,h}^2 + \|\underline{\boldsymbol{u}}_h^0\|_{a,h}^2 + \frac{1}{2t_{\mathrm{F}}} \sum_{n=1}^{N} \tau \|\underline{\boldsymbol{u}}_h^{n-1}\|_{a,h}^2 \right) \\
&\quad + \frac{C_2 d_\Omega^2}{\mu} \left(\|\boldsymbol{f}^N\|^2 + \|\boldsymbol{f}^0\|^2 + 2t_{\mathrm{F}} \sum_{n=1}^{N} \tau \|\delta_t \boldsymbol{f}^n\|^2 \right) \\
&\le \frac{1}{8} \left(\|\underline{\boldsymbol{u}}_h^N\|_{a,h}^2 + \|\underline{\boldsymbol{u}}_h^0\|_{a,h}^2 + \frac{1}{2t_{\mathrm{F}}} \sum_{n=0}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{a,h}^2 \right) + \frac{C_2 C_3 d_\Omega^2}{\mu} \|\boldsymbol{f}\|_{C^1(L^2(\Omega)^d)}^2,
\end{aligned} \tag{2.35}$$

where we have used the classical bound $\|\boldsymbol{f}^N\|^2 + \|\boldsymbol{f}^0\|^2 + 2t_{\mathrm{F}} \sum_{n=1}^{N} \tau \|\delta_t \boldsymbol{f}^n\|^2 \le C_3 \|\boldsymbol{f}\|_{C^1(L^2(\Omega)^d)}^2$ to conclude. We proceed to estimate the second term in the right-hand side of (2.32) by splitting it into two contributions as follows (here, $\overline{g}^n := (g^n, 1)$):

$$\sum_{n=1}^{N} \tau(g^n, p_h^n) = \sum_{n=1}^{N} \tau(g^n, p_h^n - \overline{p}_h^n) + \sum_{n=1}^{N} \tau(\overline{g}^n, p_h^n) := \mathfrak{T}_1 + \mathfrak{T}_2.$$

Using the Cauchy–Schwarz inequality, the bound $\sum_{n=1}^{N} \tau \|g^n\|^2 \le t_{\mathrm{F}} \|g\|_{C^0(L^2(\Omega))}^2$ together with (2.31) and Young's inequality, it is inferred that

$$\begin{aligned}
|\mathfrak{T}_1| &\le \left\{ \sum_{n=1}^{N} \tau \|g^n\|^2 \right\}^{1/2} \times \left\{ \sum_{n=1}^{N} \tau \|p_h^n - \overline{p}_h^n\|^2 \right\}^{1/2} \\
&\le t_{\mathrm{F}} \|g\|_{C^0(L^2(\Omega))} \times \left\{ \frac{2C_1}{t_{\mathrm{F}}} \sum_{n=1}^{N} \tau \left(d_\Omega^2 \|\boldsymbol{f}^n\|^2 + (2\mu + d\lambda) \|\underline{\boldsymbol{u}}_h^n\|_{a,h}^2 \right) \right\}^{1/2} \\
&\le 8C_1 t_{\mathrm{F}}^2 (2\mu + d\lambda) \|g\|_{C^0(L^2(\Omega))}^2 + \frac{d_\Omega^2}{16(2\mu + d\lambda)} \|\boldsymbol{f}\|_{C^1(L^2(\Omega)^d)}^2 + \frac{1}{16t_{\mathrm{F}}} \sum_{n=1}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{a,h}^2.
\end{aligned} \tag{2.36}$$

Owing the compatibility condition (2.1e), $\mathfrak{T}_2 = 0$ if $c_0 = 0$. If $c_0 > 0$, using the Cauchy–

Schwarz and Young's inequalities, we have

$$\begin{aligned}|\mathfrak{T}_2| &\le \left\{ t_{\mathrm{F}} \sum_{n=1}^{N} \tau c_0^{-1} \|\overline{g}^n\|^2 \right\}^{1/2} \times \left\{ t_{\mathrm{F}}^{-1} \sum_{n=1}^{N} \tau \|c_0^{1/2} p_h^n\|^2 \right\}^{1/2} \\ &\le \frac{t_{\mathrm{F}}^2}{2c_0} \|\overline{g}\|^2_{C^0(L^2(\Omega))} + \frac{1}{2t_{\mathrm{F}}} \sum_{n=1}^{N} \tau \|c_0^{1/2} p_h^n\|^2. \end{aligned} \tag{2.37}$$

Using (2.35), (2.36), and (2.37), we infer

$$\begin{aligned}\mathcal{R} \le \frac{1}{8} \left(\|\underline{\boldsymbol{u}}_h^N\|_{a,h}^2 + t_{\mathrm{F}}^{-1} \sum_{n=0}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{a,h}^2 + \|\underline{\boldsymbol{u}}_h^0\|_{a,h}^2 \right) + \frac{1}{2t_{\mathrm{F}}} \sum_{n=1}^{N} \tau \|c_0^{1/2} p_h^n\|^2 + \frac{t_{\mathrm{F}}^2}{2c_0} \|\overline{g}\|^2_{C^0(L^2(\Omega))} \\ + 8C_1 t_{\mathrm{F}}^2 (2\mu + d\lambda) \|g\|^2_{C^0(L^2(\Omega))} + \left(\frac{1}{16(2\mu + d\lambda)} + \frac{C_2 C_3}{\mu} \right) d_\Omega^2 \|\boldsymbol{f}\|^2_{C^1(L^2(\Omega)^d)}. \end{aligned} \tag{2.38}$$

(5) *Conclusion.* Using (2.34), the fact that $\mathcal{L} = \mathcal{R}$ owing to (2.32), and (2.38), it is inferred that

$$\begin{aligned}\|\underline{\boldsymbol{u}}_h^N\|_{a,h}^2 + 4\|c_0^{1/2} p_h^N\|^2 + \frac{1}{(2\mu + d\lambda)} \|p_h^N - \overline{p}_h^N\|^2 + 8 \sum_{n=1}^{N} \tau \|p_h^n\|_{c,h}^2 \le \\ \frac{C_4}{t_{\mathrm{F}}} \sum_{n=0}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{a,h}^2 + \frac{C_4}{t_{\mathrm{F}}} \sum_{n=1}^{N} \tau 4 \|c_0^{1/2} p_h^n\|^2 + G, \end{aligned}$$

where $C_4 := \max(1, C_1)$ while, observing that $\|c_0^{1/2} p_h^0\| \le \|c_0^{1/2} p^0\|$ since π_h^k is a bounded operator, and that it follows from (2.39) below that $\|\underline{\boldsymbol{u}}_h^0\|_{a,h}^2 \le C_5 (2\mu)^{-1} \left(d_\Omega^2 \|\boldsymbol{f}^0\|^2 + \|p^0\|^2 \right)$,

$$\begin{aligned}C_4^{-1} G := \frac{5C_5}{2\mu} \left(d_\Omega^2 \|\boldsymbol{f}^0\|^2 + \|p^0\|^2 \right) + 4\|c_0^{1/2} p^0\|^2 + \frac{4t_{\mathrm{F}}^2}{c_0} \|\overline{g}\|^2_{C^0(L^2(\Omega))} \\ + 64 C_1 t_{\mathrm{F}}^2 (2\mu + d\lambda) \|g\|^2_{C^0(L^2(\Omega))} + \left(\frac{5}{2(2\mu + d\lambda)} + \frac{8C_2 C_3}{\mu} \right) d_\Omega^2 \|\boldsymbol{f}\|^2_{C^1(L^2(\Omega)^d)}. \end{aligned}$$

Using Gronwall's Lemma 2.6 with $a^0 := \|\underline{\boldsymbol{u}}_h^0\|_{a,h}^2$ and $a^n := \|\underline{\boldsymbol{u}}_h^n\|_{a,h}^2 + 4\|c_0^{1/2} p_h^n\|^2$ for $1 \le n \le N$, $\delta := \tau$, $b^0 := 0$ and $b^n := \|p_h^n\|_{c,h}^2$ for $1 \le n \le N$, $K = \frac{1}{(2\mu + d\lambda)} \|p_h^N - \overline{p}_h^N\|^2$, and $\gamma^n = \frac{C_4}{t_{\mathrm{F}}}$, the desired result follows. □

Proposition 2.10 (Stability and approximation properties for $\widehat{\underline{\boldsymbol{u}}}_h^0$)**.** *The initial displacement* (2.28) *satisfies the following stability condition:*

$$\|\widehat{\underline{\boldsymbol{u}}}_h^0\|_{a,h} \lesssim (2\mu)^{-1/2} \left(d_\Omega \|\boldsymbol{f}^0\| + \|p^0\| \right). \tag{2.39}$$

Additionally, recalling the global reduction map $\underline{I}_h^k$ defined by (2.4)*, and assuming the additional regularity $p_0 \in H^{k+1}(P_\Omega)$, $\boldsymbol{u}^0 \in H^{k+2}(P_\Omega)^d$, and $\boldsymbol{\nabla}\cdot\boldsymbol{u}^0 \in H^{k+1}(P_\Omega)$, it holds*

$$(2\mu)^{1/2} \|\widehat{\underline{\boldsymbol{u}}}_h^0 - \underline{I}_h^k \boldsymbol{u}^0\|_{a,h} \lesssim h^{k+1} \left(2\mu \|\boldsymbol{u}^0\|_{H^{k+2}(P_\Omega)^d} + \lambda \|\boldsymbol{\nabla}\cdot\boldsymbol{u}^0\|_{H^{k+1}(P_\Omega)} + \rho_{\boldsymbol{\kappa}}^{1/2} \|p^0\|_{H^{k+1}(P_\Omega)} \right), \tag{2.40}$$

with global heterogeneity ratio $\rho_{\boldsymbol{\kappa}} := \overline{\kappa}/\underline{\kappa}$.

Proof. (1) *Proof of* (2.39). Using the first inequality in (2.14) followed by the definition (2.28) of $\widehat{\underline{\boldsymbol{u}}}_h^0$, we have

$$\|\widehat{\underline{\boldsymbol{u}}}_h^0\|_{a,h} \lesssim \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\mathbf{0}\}} \frac{a_h(\widehat{\underline{\boldsymbol{u}}}_h^0, \underline{\boldsymbol{v}}_h)}{(2\mu)^{1/2}\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}}$$
$$= (2\mu)^{-1/2} \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\mathbf{0}\}} \frac{l_h^0(\underline{\boldsymbol{v}}_h) - b_h(\underline{\boldsymbol{v}}_h, \pi_h^k p^0)}{\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}} \lesssim (2\mu)^{-1/2} \left(d_\Omega \|\boldsymbol{f}^0\| + \|p^0\| \right),$$

where to conclude we have used the Cauchy–Schwarz and discrete Korn's (2.16) inequalities for the first term in the numerator and the continuity (2.22) of b_h together with the $L^2(\Omega)$-stability of π_h^k for the second. (2) *Proof of* (2.40). The proof is analogous to that of point (3) in Lemma 2.11 except that we use the approximation properties (2.2) of π_h^k instead of (2.45). For this reason, elliptic regularity is not needed. □

2.4 Error analysis

In this section we carry out the error analysis of the method.

2.4.1 Projection

We consider the error with respect to the sequence of projections $(\widehat{\underline{\boldsymbol{u}}}_h^n, \widehat{p}_h^n)_{1\le n\le N}$, of the exact solution defined as follows: For $1 \le n \le N$, $\widehat{p}_h^n \in P_h^k$ solves

$$c_h(\widehat{p}_h^n, q_h) = c_h(p^n, q_h) \qquad \forall q_h \in \mathbb{P}_d^k(\mathcal{T}_h), \tag{2.41a}$$

with the closure condition $\int_\Omega \widehat{p}_h^n = \int_\Omega p^n$. Once $\widehat{p}_h^n$ has been computed, $\widehat{\underline{\boldsymbol{u}}}_h^n \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ solves

$$a_h(\widehat{\underline{\boldsymbol{u}}}_h^n, \underline{\boldsymbol{v}}_h) = l_h^n(\underline{\boldsymbol{v}}_h) - b_h(\underline{\boldsymbol{v}}_h, \widehat{p}_h^n) \qquad \forall \underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k. \tag{2.41b}$$

The well-posedness of problems (2.41a) and (2.41b) follow, respectively, from the coercivity of c_h on $\mathbb{P}_{d,0}^k(\mathcal{T}_h)$ and of a_h on $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$. The projection $(\widehat{\underline{\boldsymbol{u}}}_h^n, \widehat{p}_h^n)$ is chosen so that a convergence rate of $(k+1)$ in space analogous to the one derived in [74] can be proved for the $\|\cdot\|_{a,h}$-norm of the displacement at final time t_F. To this purpose, we also need in what follows the following elliptic regularity, which holds, e.g., when Ω is convex: There is a real number $C_{\mathrm{ell}} > 0$ only depending on Ω such that, for all $\psi \in L_0^2(\Omega)$, with $L_0^2(\Omega) := \left\{ q \in L^2(\Omega) \ : \ (q,1) = 0 \right\}$, the unique function $\zeta \in H^1(\Omega) \cap L_0^2(\Omega)$ solution of the homogeneous Neumann problem

$$-\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}\zeta) = \psi \quad \text{in } \Omega, \qquad \boldsymbol{\kappa}\boldsymbol{\nabla}\zeta \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega, \tag{2.42}$$

is such that

$$\|\zeta\|_{H^2(P_\Omega)} \le C_{\mathrm{ell}} \underline{\kappa}^{-1/2} \|\psi\|. \tag{2.43}$$

For further insight on the role of the choice (2.41) and of the elliptic regularity assumption we refer to Remark 2.14.

Lemma 2.11 (Approximation properties for $(\widehat{\underline{\boldsymbol{u}}}_h^n, \widehat{p}_h^n)$)**.** *Let a time step $1 \le n \le N$ be fixed. Assuming the regularity $p^n \in H^{k+1}(P_\Omega)$, it holds*

$$\|\widehat{p}_h^n - p^n\|_{c,h} \lesssim h^k \overline{\kappa}^{1/2} \|p^n\|_{H^{k+1}(P_\Omega)}. \tag{2.44}$$

Moreover, recalling the global reduction map $\underline{I}_h^k$ defined by (2.4), further assuming the regularity $\boldsymbol{u}^n \in H^{k+2}(P_\Omega)^d$, $\boldsymbol{\nabla}\cdot\boldsymbol{u}^n \in H^{k+1}(P_\Omega)$, and provided that the elliptic regularity (2.43) holds, one has

$$\|\widehat{p}_h^n - p^n\| \lesssim h^{k+1} \rho_{\boldsymbol{\kappa}}^{1/2} \|p^n\|_{H^{k+1}(P_\Omega)}, \tag{2.45}$$

$$(2\mu)^{1/2} \|\widehat{\underline{\boldsymbol{u}}}_h^n - \underline{I}_h^k \boldsymbol{u}^n\|_{a,h} \lesssim h^{k+1} \left(2\mu \|\boldsymbol{u}^n\|_{H^{k+2}(P_\Omega)^d} + \lambda \|\boldsymbol{\nabla}\cdot\boldsymbol{u}^n\|_{H^{k+1}(P_\Omega)} + \rho_{\boldsymbol{\kappa}}^{1/2} \|p^n\|_{H^{k+1}(P_\Omega)}\right). \tag{2.46}$$

Proof. (1) *Proof of* (2.44). By definition, we have that $\|\widehat{p}_h^n - p^n\|_{c,h} = \inf_{q_h \in \mathbb{P}_d^k(\mathcal{T}_h)} \|q_h - p^n\|_{c,h}$. To prove (2.44), it suffices to take $q_h = \pi_h^k p^n$ in the right-hand side of the previous expression and use the approximation properties (2.2) of π_h^k.
(2) *Proof of* (2.45). Let $\zeta \in H^1(\Omega)$ solve (2.42) with $\psi = p^n - \widehat{p}_h^n$. From the consistency property (2.20), it follows that

$$\|\widehat{p}_h^n - p^n\|^2 = -(\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}\zeta), \widehat{p}_h^n - p^n) = c_h(\zeta, \widehat{p}_h^n - p^n) = c_h(\zeta - \pi_h^1\zeta, \widehat{p}_h^n - p^n).$$

Then, using the Cauchy–Schwarz inequality, the estimate (2.44) together with the approximation properties (2.2) of π_h^1, and elliptic regularity, it is inferred that

$$\begin{aligned}\|\widehat{p}_h^n - p^n\|^2 &= c_h(\zeta - \pi_h^1\zeta, \widehat{p}_h^n - p^n) \le \|\zeta - \pi_h^1\zeta\|_{c,h} \|\widehat{p}_h^n - p^n\|_{c,h} \\ &\lesssim h^{k+1} \overline{\kappa}^{1/2} \|\zeta\|_{H^2(P_\Omega)} \|p^n\|_{H^{k+1}(P_\Omega)} \lesssim h^{k+1} \rho_{\boldsymbol{\kappa}}^{1/2} \|\widehat{p}_h^n - p^n\| \|p^n\|_{H^{k+1}(P_\Omega)},\end{aligned}$$

and (2.45) follows.
(3) *Proof of* (2.46). We start by observing that

$$\|\widehat{\underline{\boldsymbol{u}}}_h^n - \underline{I}_h^k \boldsymbol{u}^n\|_{a,h} = \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k \setminus \{\underline{\boldsymbol{0}}\}} \frac{a_h(\widehat{\underline{\boldsymbol{u}}}_h^n - \underline{I}_h^k \boldsymbol{u}^n, \underline{\boldsymbol{v}}_h)}{\|\underline{\boldsymbol{v}}_h\|_{a,h}} \lesssim \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k \setminus \{\underline{\boldsymbol{0}}\}} \frac{a_h(\widehat{\underline{\boldsymbol{u}}}_h^n - \underline{I}_h^k \boldsymbol{u}^n, \underline{\boldsymbol{v}}_h)}{(2\mu)^{1/2} \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}},$$

where we have used the first inequality in (2.14). Recalling the definition (2.27) of the linear form l_h^n, the fact that $\boldsymbol{f}^n = -\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{u}) + \boldsymbol{\nabla}p$, and using (2.41a), it is inferred that

$$\begin{aligned}a_h(\widehat{\underline{\boldsymbol{u}}}_h^n - \underline{I}_h^k \boldsymbol{u}^n, \underline{\boldsymbol{v}}_h) &= l_h^n(\underline{\boldsymbol{v}}_h^n) - a_h(\underline{I}_h^k \boldsymbol{u}^n, \underline{\boldsymbol{v}}_h) - b_h(\underline{\boldsymbol{v}}_h, \widehat{p}_h n) \\ &= \{ - a_h(\underline{I}_h^k \boldsymbol{u}^n, \underline{\boldsymbol{v}}_h) - (\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{u}^n), \boldsymbol{v}_h)\} + \{(\boldsymbol{\nabla}p^n, \boldsymbol{v}_h) - b_h(\underline{\boldsymbol{v}}_h, \widehat{p}_h^n)\}.\end{aligned} \tag{2.47}$$

Denote by $\mathfrak{T}_1$ and $\mathfrak{T}_2$ the terms in braces. Using (2.15), it is readily inferred that

$$|\mathfrak{T}_1| \lesssim h^{k+1} \left(2\mu \|\boldsymbol{u}^n\|_{H^{k+2}(P_\Omega)^d} + \lambda \|\boldsymbol{\nabla}\cdot\boldsymbol{u}^n\|_{H^{k+1}(P_\Omega)}\right) \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}. \tag{2.48}$$

For the second term, performing an element-wise integration by parts on $(\boldsymbol{\nabla} p, \boldsymbol{v}_h)$ and recalling the definition (2.21) of b_h and (2.7a) of D_T^k with $q = \widehat{p}_h^n$, it is inferred that

$$\begin{aligned} |\mathfrak{T}_2| &= \left| \sum_{T\in\mathcal{T}_h} \left\{ (\widehat{p}_h^n - p^n, \boldsymbol{\nabla}\cdot\boldsymbol{v}_T)_T + \sum_{F\in\mathcal{F}_T} (\widehat{p}_h^n - p^n, (\boldsymbol{v}_F - \boldsymbol{v}_T)\boldsymbol{n}_{TF})_F \right\} \right| \\ &\lesssim h^{k+1} \rho_{\boldsymbol{\kappa}}^{1/2} \|p^n\|_{H^{k+1}(P_\Omega)} \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}, \end{aligned} \tag{2.49}$$

where the conclusion follows from the Cauchy–Schwarz inequality together with (2.45). Plugging (2.48)–(2.49) into (2.47) we obtain (2.46). □

2.4.2 Error equations

We define the discrete error components as follows: For all $1 \le n \le N$,

$$\underline{\boldsymbol{e}}_h^n \coloneqq \underline{\boldsymbol{u}}_h^n - \underline{\widehat{\boldsymbol{u}}}_h^n, \qquad \rho_h^n \coloneqq p_h^n - \widehat{p}_h^n. \tag{2.50}$$

Owing to the choice of the initial condition detailed in Section 2.2.5, the inital error $(\underline{\boldsymbol{e}}_h^0, \rho_h^0) \coloneqq (\underline{\boldsymbol{u}}_h^0 - \underline{\widehat{\boldsymbol{u}}}_h^0, p_h^0 - \widehat{p}_h^0)$ is the null element in the product space $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times P_h^k$. On the other hand, for all $1 \le n \le N$, $(\underline{\boldsymbol{e}}_h^n, \rho_h^n)$ solves

$$a_h(\underline{\boldsymbol{e}}_h^n, \underline{\boldsymbol{v}}_h) + b_h(\underline{\boldsymbol{v}}_h, \rho_h^n) = 0 \qquad \forall \underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k, \tag{2.51a}$$

$$(c_0 \delta_t \rho_h^n, q_h) - b_h(\delta_t \underline{\boldsymbol{e}}_h^n, q_h) + c_h(\rho_h^n, q_h) = \mathcal{E}_h^n(q_h), \qquad \forall q_h \in P_h^k, \tag{2.51b}$$

with consistency error

$$\mathcal{E}_h^n(q_h) \coloneqq (g^n, q_h) - (c_0 \delta_t \widehat{p}_h^n, q_h) - c_h(\widehat{p}_h^n, q_h) + b_h(\delta_t \underline{\widehat{\boldsymbol{u}}}_h^n, q_h).$$

2.4.3 Convergence

Theorem 2.12 (Estimate for the discrete errors)**.** *Let $(\boldsymbol{u}, p)$ denote the unique solution to (2.1), for which we assume the regularity*

$$\boldsymbol{u} \in C^2(H^1(P_\Omega)^d) \cap C^1(H^{k+2}(P_\Omega)^d), \qquad p \in C^1(H^{k+1}(P_\Omega)).$$

If $c_0 > 0$, we further assume $p \in C^2(L^2(\Omega))$. Define, for the sake of brevity, the bounded quantities

$$\begin{aligned} \mathcal{N}_1 &\coloneqq (2\mu + d\lambda)^{1/2} \|\boldsymbol{u}\|_{C^2(H^1(P_\Omega)^d)} + \|c_0^{1/2} p\|_{C^2(L^2(\Omega)^d)}, \\ \mathcal{N}_2 &\coloneqq \frac{(2\mu + d\lambda)^{1/2}}{2\mu} \left(2\mu \|\boldsymbol{u}\|_{C^1(H^{k+2}(P_\Omega)^d)} + \lambda \|\boldsymbol{\nabla}\cdot\boldsymbol{u}\|_{C^1(H^{k+1}(P_\Omega))} + \rho_{\boldsymbol{\kappa}}^{1/2} \|p\|_{C^1(H^{k+1}(P_\Omega))} \right) \\ &\quad + \|c_0^{1/2} p\|_{C^0(H^{k+1}(P_\Omega))}. \end{aligned}$$

Then, assuming the elliptic regularity (2.43), it holds, letting $\overline{\rho}_h^n \coloneqq (\rho_h^n, 1)$,

$$\|\underline{\boldsymbol{e}}_h^N\|_{a,h}^2 + \|c_0^{1/2} \rho_h^N\|^2 + \frac{1}{2\mu + d\lambda} \|\rho_h^N - \overline{\rho}_h^N\|^2 + \sum_{n=1}^N \tau \|\rho_h^n\|_{c,h}^2 \lesssim \left(\tau \mathcal{N}_1 + h^{k+1} \mathcal{N}_2 \right)^2. \tag{2.52}$$

Remark 2.13 (Pressure estimate for $c_0 = 0$). In the incompressible case $c_0 = 0$, the third term in the left-hand side of (2.52) delivers an estimate on the L^2-norm of the pressure since $\overline{\rho}_h^N = 0$ (cf. (2.1f)).

Proof of Theorem 2.12. Throughout the proof, C_i with $i \in \mathbb{N}^*$ will denote a generic positive constant independent of h, τ, and of the physical parameters c_0, λ, μ, and $\boldsymbol{\kappa}$.
(1) *Basic error estimate.* Using the inf-sup condition (2.24), equation (2.23) followed by (2.51a), and the second inequality in (2.14), it is readily seen that

$$\begin{aligned}\|\rho_h^n - \overline{\rho}_h^n\| &\le \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\underline{\boldsymbol{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, \rho_h^n - \overline{\rho}_h^n)}{\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}} \\ &= \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\underline{\boldsymbol{0}}\}} \frac{-a_h(\underline{\boldsymbol{e}}_h^n, \underline{\boldsymbol{v}}_h)}{\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}} \le C_1^{1/2}(2\mu + d\lambda)^{1/2}\|\underline{\boldsymbol{e}}_h^n\|_{a,h},\end{aligned}$$

with $C_1^{1/2} = \beta\eta^{1/2}$. Adding (2.51a) with $\underline{\boldsymbol{v}}_h = \tau\delta_t\underline{\boldsymbol{e}}_h$ to (2.51b) with $q_h = \tau\rho_h^n$ and summing the resulting equation over $1 \le n \le N$, it is inferred that

$$\sum_{n=1}^N \tau a_h(\underline{\boldsymbol{e}}_h^n, \delta_t\underline{\boldsymbol{e}}_h^n) + \sum_{n=1}^N \tau(c_0\delta_t\rho_h^n, \rho_h^n) + \sum_{n=1}^N \tau\|\rho_h^n\|_{c,h}^2 = \sum_{n=1}^N \tau\mathcal{E}_h^n(\rho_h^n). \tag{2.53}$$

Proceeding as in point (3) of the proof of Lemma 2.7, and recalling that $(\underline{\boldsymbol{e}}_h^0, \rho_h^0) = (\underline{\boldsymbol{0}}, 0)$, we arrive at the following error estimate:

$$\frac{1}{4}\|\underline{\boldsymbol{e}}_h^N\|_{a,h}^2 + \frac{1}{4C_1(2\mu + d\lambda)}\|\rho_h^N - \overline{\rho}_h^N\|^2 + \frac{1}{2}\|c_0^{1/2}\rho_h^N\|^2 + \sum_{n=1}^N \tau\|\rho_h^n\|_{c,h}^2 \le \sum_{n=1}^N \tau\mathcal{E}_h^n(\rho_h^n). \tag{2.54}$$

(2) *Bound of the consistency error.* Using $g^n = c_0\mathrm{d}_t p^n + \boldsymbol{\nabla}\cdot(\mathrm{d}_t\boldsymbol{u}^n - \boldsymbol{\kappa}\boldsymbol{\nabla}p^n)$, the consistency property (2.20), and observing that, using the definition (2.19) of c_h, integration by parts together with the homogeneous displacement boundary condition (2.1c), and (2.23),

$$c_h(p^n - \widehat{p}_h^n, \overline{\rho}_h^n) + (\boldsymbol{\nabla}\cdot(\mathrm{d}_t\boldsymbol{u}^n), \overline{\rho}_h^n) + b_h(\delta_t\widehat{\underline{\boldsymbol{u}}}_h^n, \overline{\rho}_h^n) = 0,$$

we can decompose the right-hand side of (2.54) as follows:

$$\begin{aligned}\sum_{n=1}^N \tau\mathcal{E}_h^n(\rho_h^n) &= \sum_{n=1}^N \tau(c_0(\mathrm{d}_t p^n - \delta_t\widehat{p}_h^n), \rho_h^n) + \sum_{n=1}^N \tau c_h(p^n - \widehat{p}_h^n, \rho_h^n - \overline{\rho}_h^n) \\ &\quad + \sum_{n=1}^N \tau\left\{(\boldsymbol{\nabla}\cdot(\mathrm{d}_t\boldsymbol{u}^n), \rho_h^n - \overline{\rho}_h^n) + b_h(\delta_t\widehat{\underline{\boldsymbol{u}}}_h^n, \rho_h^n - \overline{\rho}_h^n)\right\} := \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3.\end{aligned} \tag{2.55}$$

For the first term, inserting $\pm\delta_t p^n$ into the first factor and using the Cauchy-Schwarz inequality followed by the approximation properties of $\widehat{p}_h^0$ (a consequence of (2.2)) and (2.45) of $\widehat{p}_h^n$, it

is inferred that

$$\begin{aligned}|\mathfrak{T}_1| &\lesssim \left\{c_0 \sum_{n=1}^{N} \tau \left[\|\mathrm{d}_t p^n - \delta_t p^n\|^2 + \|\delta_t(p^n - \widehat{p}_h^n)\|^2\right]\right\}^{1/2} \times \left\{\sum_{n=1}^{N} \tau \|c_0^{1/2}\rho_h^n\|^2\right\}^{1/2} \\ &\le C_2\left(\tau\mathcal{N}_1 + h^{k+1}\mathcal{N}_2\right) + \frac{1}{2}\sum_{n=1}^{N}\tau\|c_0^{1/2}\rho_h^n\|^2.\end{aligned} \tag{2.56}$$

For the second term, the choice (2.41a) of the pressure projection readily yields

$$\mathfrak{T}_2 = 0.$$

For the last term, inserting $\pm\underline{I}_h^k \boldsymbol{u}^n$ into the first argument of b_h, and using the commuting property (2.8) of D_T^k, it is inferred that

$$\mathfrak{T}_3 = \sum_{n=1}^{N}\tau\left\{\sum_{T\in\mathcal{T}_h}\left[(\boldsymbol{\nabla}\cdot(\mathrm{d}_t\boldsymbol{u}^n - \delta_t\boldsymbol{u}^n), \rho_h^n - \overline{\rho}_h^n)_T + (D_T^k\delta_t(\underline{I}_T^k\boldsymbol{u}^n - \widehat{\underline{\boldsymbol{u}}}_T^n), \rho_h^n - \overline{\rho}_h^n)_T\right]\right\}.$$

Using the Cauchy–Schwarz inequality, the bound $\|D_T^k\delta_t(\underline{I}_T^k\boldsymbol{u}^n - \widehat{\underline{\boldsymbol{u}}}_T^n)\|_T \lesssim \|\delta_t(\underline{I}_T^k\boldsymbol{u}^n - \widehat{\underline{\boldsymbol{u}}}_T^n)\|_{\epsilon,T}$ valid for all $T \in \mathcal{T}_h$, and the approximation properties (2.40) and (2.46) of $\widehat{\underline{\boldsymbol{u}}}_h^0$ and $\widehat{\underline{\boldsymbol{u}}}_h^n$, respectively, we obtain

$$\begin{aligned}|\mathfrak{T}_3| &\lesssim \left\{\sum_{n=1}^{N}\tau\left[\|\mathrm{d}_t\boldsymbol{u}^n - \delta_t\boldsymbol{u}^n\|_{H^1(\Omega)^d}^2 + \|\delta_t(\underline{I}_h^k\boldsymbol{u}^n - \widehat{\underline{\boldsymbol{u}}}_h^n)\|_{\epsilon,h}^2\right]\right\}^{\frac{1}{2}} \times \left\{\sum_{n=1}^{N}\tau\|\rho_h^n - \overline{\rho}_h^n\|^2\right\}^{\frac{1}{2}} \\ &\le C_3C_1\left(\tau\mathcal{N}_1 + h^{k+1}\mathcal{N}_2\right)^2 + \frac{1}{4C_1(2\mu + d\lambda)}\sum_{n=1}^{N}\tau\|\rho_h^n - \overline{\rho}_h^n\|^2.\end{aligned} \tag{2.57}$$

Using (2.56)–(2.57) to bound the right-hand side of (2.55), it is inferred

$$\begin{aligned}\|\underline{\boldsymbol{e}}_h^N\|_{a,h}^2 + \frac{1}{C_1(2\mu+d\lambda)}\|\rho_h^N - \overline{\rho}_h^N\|^2 + 2\|c_0^{1/2}\rho_h^N\|^2 + 4\sum_{n=1}^{N}\tau\|\rho_h^n\|_{c,h}^2 \\ \le \frac{1}{C_1(2\mu+d\lambda)}\sum_{n=1}^{N}\tau\|\rho_h^n - \overline{\rho}_h^n\|^2 + 2\sum_{n=1}^{N}\tau\|c_0^{1/2}\rho_h^n\|^2 + G,\end{aligned}$$

with $G := 4(C_1C_3 + C_2)\left(\tau\mathcal{N}_1 + h^{k+1}\mathcal{N}_2\right)^2$. The conclusion follows using the discrete Gronwall's inequality (2.29) with $\delta = \tau$, $K = \|\underline{\boldsymbol{e}}_h^N\|_{a,h}^2$, $a^0 = 0$ and $a^n = \frac{1}{C_1(2\mu+d\lambda)}\|\rho_h^n - \overline{\rho}_h^n\|^2 + 2\|c_0^{1/2}\rho_h^n\|^2$ for $1 \le n \le N$, $b^n = 4\|\rho_h^n\|_{c,h}^2$, and $\gamma^n = 1$. □

Remark 2.14 (Role of the choice (2.41) and of elliptic regularity). The choice (2.41) for the projection ensures that the term $\mathfrak{T}_2$ in step (2) of the proof of Theorem 2.12 vanishes. This is a key point to obtain an order of convergence of $(k+1)$ in space. For a different choice, say $\widehat{p}_h^n = \pi_h^k p^n$, this term would be of order k, and therefore yield a suboptimal estimate for the terms in the left-hand side of (2.58) below (the estimate (2.59) would not change and remain optimal). This would also be the case if we removed the elliptic regularity assumption (2.43).

Remark 2.15 (BDF2 time discretization). In some of the numerical test cases of Section 2.6, we have used a BDF2 time discretization, which corresponds to the backward differencing operator

$$\delta_t^{(2)}\varphi^{n+2} := \frac{3\varphi^{n+2} - 4\varphi^{n+1} + \varphi^n}{2\tau},$$

used in place of (2.3). As BDF2 requires two starting values, we perform a first march in time using the backward Euler scheme (another possibility would have been to resort to the second-order Crank–Nicolson scheme). For the BDF2 time discretization, stability estimates similar to those of Lemma 2.7 can be proved with this initialization, while the error can be shown to scale as $\tau^2 + h^{k+1}$ (compare with (2.52)). The main difference with respect to the present analysis focused on the backward Euler scheme is that formula (2.33) is replaced in the proofs by

$$2x(3x - 4y + z) = x^2 - y^2 + (2x - y)^2 - (2y - z)^2 + (x - 2y + z)^2.$$

The modifications of the proofs are quite classical and are not detailed here for the sake of conciseness (for a pedagogic exposition, one can consult, e.g., [92, Chapter 6]).

Corollary 2.16 (Convergence)**.** *Under the assumptions of Theorem 2.12, it holds that*

$$(2\mu)^{1/2}\|\boldsymbol{\nabla}_{\mathrm{s},h}(\boldsymbol{r}_h^{k+1}\underline{\boldsymbol{u}}_h^N - \boldsymbol{u}^N)\| + \|c_0^{1/2}(p_h^N - p^N)\| + \frac{1}{2\mu + d\lambda}\|(p_h^N - p^N) - (\overline{p}_h^N - \overline{p}^N)\|$$
$$\lesssim \tau\mathcal{N}_1 + h^{k+1}\mathcal{N}_2 + c_0^{1/2}h^{k+1}\|p^N\|_{H^{k+1}(P_\Omega)}, \quad (2.58)$$

$$\left\{\sum_{n=1}^N \tau\|p_h^n - p^n\|_{c,h}^2\right\}^{1/2} \lesssim \tau\mathcal{N}_1 + h^{k+1}\mathcal{N}_2 + h^k\overline{\kappa}^{1/2}t_{\mathrm{F}}^{1/2}\|p\|_{C^0(H^{k+1}(P_\Omega))}. \quad (2.59)$$

Proof. Using the triangle inequality, recalling the definition (2.50) of $\underline{\boldsymbol{e}}_h^N$ and $\widehat{p}_h^N$ and (2.14) of $\|\cdot\|_{a,h}$-norm, it is inferred that

$$\begin{aligned}
(2\mu)^{1/2}\|\boldsymbol{\nabla}_{\mathrm{s},h}(\boldsymbol{r}_h^{k+1}\underline{\boldsymbol{u}}_h^N - \boldsymbol{u}^N)\| &\le \|\underline{\boldsymbol{e}}_h^N\|_{a,h} + (2\mu)^{1/2}\|\boldsymbol{\nabla}_{\mathrm{s},h}(\boldsymbol{r}_h^{k+1}\widehat{\boldsymbol{u}}_h - \boldsymbol{r}_h^{k+1}\underline{\boldsymbol{I}}_h^k\boldsymbol{u}^N)\| \\
&\quad + (2\mu)^{1/2}\|\boldsymbol{\nabla}_{\mathrm{s}}(\boldsymbol{r}_h^{k+1}\underline{\boldsymbol{I}}_h^k\boldsymbol{u}^N - \boldsymbol{u}^N)\|, \\
\|p_h^N - p^N - (\overline{p}_h^N - \overline{p}^N)\| &\le \|\rho_h^N - \overline{\rho}_h^N\| + \|\widehat{p}_h^N - p^N\|, \\
\|c_0^{1/2}(p_h^N - p^N)\| &\le \|c_0^{1/2}\rho_h^N\| + \|c_0^{1/2}(\widehat{p}_h^N - p^N)\|.
\end{aligned}$$

To conclude, use (2.52) to estimate the left-most terms in the right-hand sides of the above equations. Use (2.46) and (2.45), the approximation properties (2.6) of $\boldsymbol{r}_h^{k+1}\underline{\boldsymbol{I}}_h^k$, respectively, for the right-most terms. This proves (2.58). A similar decomposition of the error yields (2.59). □

2.5 Implementation

In this section we discuss practical aspects including, in particular, static condensation. The implementation is based on the hho platform[1], which relies on the linear algebra facilities provided by the Eigen3 library [116].

The starting point consists in selecting a basis for each of the polynomial spaces appearing in the construction. Let $\boldsymbol{s} = (s_1, ..., s_d)$ be a d-dimensional multi-index with the usual notation $|\boldsymbol{s}|_1 = \sum_{i=1}^d s_i$, and let $\boldsymbol{x} = (x_1, ..., x_d) \in \mathbb{R}^d$. Given $k \geq 0$ and $T \in \mathcal{T}_h$, we denote by $\mathcal{B}_T^k$ a basis for the polynomial space $\mathbb{P}_d^k(T)$. In the numerical experiments of Section 2.6, we have used the set of locally scaled monomials:

$$\mathcal{B}_T^k := \left\{ \left(\frac{\boldsymbol{x} - \boldsymbol{x}_T}{h_T} \right)^{\boldsymbol{s}}, \ |\boldsymbol{s}|_1 \leq k \right\}, \tag{2.60}$$

with $\boldsymbol{x}_T$ denoting the barycenter of T. Similarly, for all $F \in \mathcal{F}_h$, we denote by $\mathcal{B}_F^k$ a basis for the polynomial space $\mathbb{P}_{d-1}^k(F)$ which, in the proposed implementation, is again a set of locally scaled monomials similar to (2.60).

Remark 2.17 (Choice of the polynomial bases). The choice of the polynomial bases can have a sizeable impact on the conditioning of both the local problems defining the displacement reconstruction $\boldsymbol{r}_T^{k+1}$ (cf. (2.5)) and the global problem. This is particularly the case when using high polynomial orders (typically, $k \geq 7$). The scaled monomial basis (2.60) is appropriate when dealing with isotropic elements. In the presence of anisotropic elements, a better choice is to use for each element a local frame aligned with its principal axes of rotation together with normalization factors tailored for each direction. A further improvement, originally investigated in [16] in the context of dG methods, consists in performing a Gram–Schmidt orthonormalization with respect to a suitably selected inner product. In the numerical test cases of Section 2.6, which focus on isotropic meshes and moderate polynomial degrees ($k \leq 3$), the basis (2.60) proved fully satisfactory.

Introducing the vector bases $\boldsymbol{\mathcal{B}}_T^k := (\mathcal{B}_T^k)^d$, $T \in \mathcal{T}_h$, and $\boldsymbol{\mathcal{B}}_F^k := (\mathcal{B}_F^k)^d$, $F \in \mathcal{F}_h^{\mathrm{i}}$, a basis $\underline{\boldsymbol{\mathcal{U}}}_{h,0}^k$ for the space $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ (cf. (2.12)) is given by

$$\underline{\boldsymbol{\mathcal{U}}}_{h,0}^k := \boldsymbol{\mathcal{U}}_{\mathcal{T}}^k \times \boldsymbol{\mathcal{U}}_{\mathcal{F}}^k, \qquad \boldsymbol{\mathcal{U}}_{\mathcal{T}}^k := \bigtimes_{T \in \mathcal{T}_h} \boldsymbol{\mathcal{B}}_T^k, \qquad \boldsymbol{\mathcal{U}}_{\mathcal{F}}^k := \bigtimes_{F \in \mathcal{F}_h^{\mathrm{i}}} \boldsymbol{\mathcal{B}}_F^k,$$

while a basis $\mathcal{P}_h^k$ for the space P_h^k (cf. (2.17)) is obtained setting

$$\mathcal{P}_h^k := \bigtimes_{T \in \mathcal{T}_h} \mathcal{B}_T^k.$$

When $c_0 = 0$, the zero average constraint in P_h^k can be accounted for using as a Lagrange multiplier the characteristic function of Ω. Notice also that boundary faces have been excluded from the Cartesian product in the definition of $\boldsymbol{\mathcal{U}}_{\mathcal{F}}^k$ to strongly account for boundary

[1] DL15105 Université de Montpellier

conditions. Letting, for the sake of brevity, $N_n^k := \binom{k+n}{k}$, $n \in \mathbb{N}$, a simple computation shows that

$$\dim(\boldsymbol{\mathcal{U}}_{\mathcal{T}}^k) = d\,\mathrm{card}(\mathcal{T}_h)N_d^k, \qquad \dim(\boldsymbol{\mathcal{U}}_{\mathcal{F}}^k) = d\,\mathrm{card}(\mathcal{F}_h^{\mathrm{i}})N_{d-1}^k, \qquad \dim(\mathcal{P}_h^k) = \mathrm{card}(\mathcal{T}_h)N_d^k.$$

The total DOF count thus yields

$$d\,\mathrm{card}(\mathcal{T}_h)N_d^k + d\,\mathrm{card}(\mathcal{F}_h^{\mathrm{i}})N_{d-1}^k + \mathrm{card}(\mathcal{T}_h)N_d^k. \tag{2.61}$$

In what follows, for a given time step $0 \le n \le N$, we denote by $\mathsf{U}_{\mathcal{T}}^n$ and $\mathsf{U}_{\mathcal{F}}^n$ the vectors collecting element-based and face-based displacement DOFs, respectively, and by P^n the vector collecting pressure DOFs.

Denote now by A and B, respectively, the matrices that represent the bilinear forms a_h (cf. (2.11)) and b_h (cf. (2.21)) in the selected basis. Distinguishing element-based and face-based displacement DOFs, the matrices A and B display the following block structure:

$$\mathsf{A} = \left[\begin{array}{c:c} \mathsf{A}_{\mathcal{T}\mathcal{T}} & \mathsf{A}_{\mathcal{T}\mathcal{F}} \\ \hdashline \mathsf{A}_{\mathcal{T}\mathcal{F}}^{\mathrm{T}} & \mathsf{A}_{\mathcal{F}\mathcal{F}} \end{array}\right], \qquad \mathsf{B} = \left[\begin{array}{c} \mathsf{B}_{\mathcal{T}} \\ \hdashline \mathsf{B}_{\mathcal{F}} \end{array}\right].$$

For every mesh element $T \in \mathcal{T}_h$, the element-based displacement DOFs are only coupled with those face-based displacement DOFs that lie on the boundary of T and with the (element-based) pressure DOFs in T. This translates into the fact that the submatrix $\mathsf{A}_{\mathcal{T}\mathcal{T}}$ is block-diagonal, i.e.,

$$\mathsf{A}_{\mathcal{T}\mathcal{T}} = \mathrm{diag}(\mathsf{A}_{TT})_{T\in\mathcal{T}_h},$$

with each elementary block A_{TT} of size $\dim(\boldsymbol{\mathcal{B}}_T^k)^2$. Additionally, it can be proved that the blocks A_{TT}, $T \in \mathcal{T}_h$, are invertible, so that the inverse of $\mathsf{A}_{\mathcal{T}\mathcal{T}}$ can be efficiently computed setting

$$\mathsf{A}_{\mathcal{T}\mathcal{T}}^{-1} = \mathrm{diag}(\mathsf{A}_{TT}^{-1})_{T\in\mathcal{T}_h}. \tag{2.62}$$

The above remark can be exploited in practice to efficiently eliminate the element-based displacement DOFs from the global system. This process, usually referred to as "static condensation", is detailed in what follows.

For a given time step $1 \le n \le N$, the linear system corresponding to the discrete problem (2.26) is of the form

$$\left[\begin{array}{c:c:c} \mathsf{A}_{\mathcal{T}\mathcal{T}} & \mathsf{A}_{\mathcal{T}\mathcal{F}} & \mathsf{B}_{\mathcal{T}} \\ \hdashline \mathsf{A}_{\mathcal{T}\mathcal{F}}^{\mathrm{T}} & \mathsf{A}_{\mathcal{F}\mathcal{F}} & \mathsf{B}_{\mathcal{F}} \\ \hdashline -\mathsf{B}_{\mathcal{T}}^{\mathrm{T}} & -\mathsf{B}_{\mathcal{F}}^{\mathrm{T}} & \frac{\tau}{\theta}\mathsf{C} + c_0\mathsf{M} \end{array}\right] \left[\begin{array}{c} \mathsf{U}_{\mathcal{T}}^n \\ \hdashline \mathsf{U}_{\mathcal{F}}^n \\ \hdashline \mathsf{P}^n \end{array}\right] = \left[\begin{array}{c} \mathsf{F}_{\mathcal{T}}^n \\ \hdashline 0_{\mathcal{F}} \\ \hdashline \widetilde{\mathsf{G}}^n \end{array}\right], \tag{2.63}$$

where C denotes the matrix that represents the bilinear form c_h in the selected basis, M is the (block diagonal) pressure mass matrix, $\mathsf{F}_{\mathcal{T}}^n$ is the vector corresponding to the discretization of the volumetric load $\boldsymbol{f}^n$, while $0_{\mathcal{F}}$ is the zero vector of length $\dim(\boldsymbol{\mathcal{U}}_{\mathcal{F}}^k)$. Denoting by G^n the vector corresponding to the discretization of the fluid source g^n, when the backward Euler method is used to march in time, we let $\theta = 1$ and set

$$\widetilde{\mathsf{G}}^n := \tau\mathsf{G}^n - \mathsf{B}^{\mathrm{T}}\mathsf{U}^{n-1} + c_0\mathsf{M}\mathsf{P}^{n-1}.$$

For the BDF2 method (and $n \geq 2$) , we let $\theta = 3/2$ and set

$$\widetilde{\mathsf{G}}^n := \frac{2}{3}\tau \mathsf{G}^n - \frac{4}{3}\mathsf{B}^{\mathrm{T}}\mathsf{U}^{n-1} + \frac{1}{3}\mathsf{B}^{\mathrm{T}}\mathsf{U}^{n-2} + \frac{4}{3}c_0\mathsf{M}\mathsf{P}^{n-1} - \frac{1}{3}c_0\mathsf{M}\mathsf{P}^{n-2}.$$

Recalling (2.62), instead of assemblying the full system, we can effectively compute the Schur complement of $\mathsf{A}_{\mathcal{T}\mathcal{T}}$ and code, instead, the following reduced version, where the element-based displacement DOFs collected in the subvector $\mathsf{U}^n_{\mathcal{T}}$ no longer appear:

$$\left[\begin{array}{c|c} \mathsf{A}_{\mathcal{F}\mathcal{F}} - \mathsf{A}^{\mathrm{T}}_{\mathcal{T}\mathcal{F}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{A}_{\mathcal{T}\mathcal{F}} & \mathsf{B}_{\mathcal{F}} - \mathsf{A}^{\mathrm{T}}_{\mathcal{T}\mathcal{F}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{B}_{\mathcal{T}} \\ \hline -\mathsf{B}^{\mathrm{T}}_{\mathcal{F}} + \mathsf{B}^{\mathrm{T}}_{\mathcal{T}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{A}_{\mathcal{T}\mathcal{F}} & \frac{\tau}{\theta}\mathsf{C} + c_0\mathsf{M} + \mathsf{B}^{\mathrm{T}}_{\mathcal{T}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{B}_{\mathcal{T}} \end{array}\right] \left[\begin{array}{c} \mathsf{U}^n_{\mathcal{F}} \\ \hline \mathsf{P}^n \end{array}\right] = \left[\begin{array}{c} -\mathsf{A}^{\mathrm{T}}_{\mathcal{T}\mathcal{F}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{F}^n_{\mathcal{T}} \\ \hline \widetilde{\mathsf{G}}^n + \mathsf{B}^{\mathrm{T}}_{\mathcal{T}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{F}^n_{\mathcal{T}} \end{array}\right]. \tag{2.64}$$

All matrix products appearing in (2.64) are directly assembled from their local counterparts (i.e., the factors need not be constructed separately). Specifically, introducing, for all $T \in \mathcal{T}_h$, the following local matrices $\mathsf{A}(T)$ and $\mathsf{B}(T)$ representing the local bilinear forms a_T (cf. (2.9)) and b_T (cf. (2.21)), respectively:

$$\mathsf{A}(T) = \left[\begin{array}{c|c} \mathsf{A}_{TT} & \mathsf{A}_{T\mathcal{F}_T} \\ \hline \mathsf{A}^{\mathrm{T}}_{T\mathcal{F}_T} & \mathsf{A}_{\mathcal{F}_T\mathcal{F}_T} \end{array}\right], \qquad \mathsf{B}(T) = \left[\begin{array}{c} \mathsf{B}_T \\ \hline \mathsf{B}_{\mathcal{F}_T} \end{array}\right],$$

one has for the left-hand side matrix, denoting by $\underset{T\in\mathcal{T}_h}{\longleftarrow}$ the usual assembly procedure based on a global DOF map,

$$\mathsf{A}^{\mathrm{T}}_{\mathcal{T}\mathcal{F}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{A}_{\mathcal{T}\mathcal{F}} \underset{T\in\mathcal{T}_h}{\longleftarrow} \mathsf{A}^{\mathrm{T}}_{T\mathcal{F}_T}\mathsf{A}^{-1}_{TT}\mathsf{A}_{T\mathcal{F}_T}, \qquad \mathsf{A}^{\mathrm{T}}_{\mathcal{T}\mathcal{F}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{B}_{\mathcal{T}} \underset{T\in\mathcal{T}_h}{\longleftarrow} \mathsf{A}^{\mathrm{T}}_{T\mathcal{F}_T}\mathsf{A}^{-1}_{TT}\mathsf{B}_T,$$
$$\mathsf{B}^{\mathrm{T}}_{\mathcal{T}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{A}_{\mathcal{T}\mathcal{F}} \underset{T\in\mathcal{T}_h}{\longleftarrow} \mathsf{B}^{\mathrm{T}}_{T}\mathsf{A}^{-1}_{TT}\mathsf{A}_{T\mathcal{F}_T}, \qquad \mathsf{B}^{\mathrm{T}}_{\mathcal{T}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{B}_{\mathcal{T}} \underset{T\in\mathcal{T}_h}{\longleftarrow} \mathsf{B}^{\mathrm{T}}_{T}\mathsf{A}^{-1}_{TT}\mathsf{B}_T,$$

and, similarly, for the right-hand side vector

$$\mathsf{A}^{\mathrm{T}}_{\mathcal{T}\mathcal{F}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{F}^n_{\mathcal{T}} \underset{T\in\mathcal{T}_h}{\longleftarrow} \mathsf{A}^{\mathrm{T}}_{T\mathcal{F}_T}\mathsf{A}^{-1}_{TT}\mathsf{F}^n_{T}, \qquad \mathsf{B}^{\mathrm{T}}_{\mathcal{T}}\mathsf{A}^{-1}_{\mathcal{T}\mathcal{T}}\mathsf{F}^n_{\mathcal{T}} \underset{T\in\mathcal{T}_h}{\longleftarrow} \mathsf{B}^{\mathrm{T}}_{T}\mathsf{A}^{-1}_{TT}\mathsf{F}^n_{T}.$$

The advantage of implementing (2.64) over (2.63) is that the number of DOFs appearing in the linear system reduces to (compare with (2.61))

$$d\,\mathrm{card}(\mathcal{F}_h^{\mathrm{i}})N^k_{d-1} + \mathrm{card}(\mathcal{T}_h)N^k_d.$$

Additionally, since the reduced left-hand side matrix in (2.64) does not depend on the time step n, it can be assembled (and, possibly, factored) once and for all in a preliminary stage, thus leading to a further reduction in the computational cost. Finally, for all $T \in \mathcal{T}_h$, the local vector U^n_T of element-based displacement DOFs can be recovered from the local right-hand side vector F^n_T and the local vector of face-based displacement DOFs and (element-based) pressure DOFs $(\mathsf{U}^n_{\mathcal{F}_T}, \mathsf{P}^n_T)$ by the following element-by-element post-processing:

$$\mathsf{U}^n_T = \mathsf{A}^{-1}_{TT}\left(\mathsf{F}^n_T - \mathsf{A}_{T\mathcal{F}_T}\mathsf{U}^n_{\mathcal{F}_T} - \mathsf{B}_T\mathsf{P}^n_T\right).$$

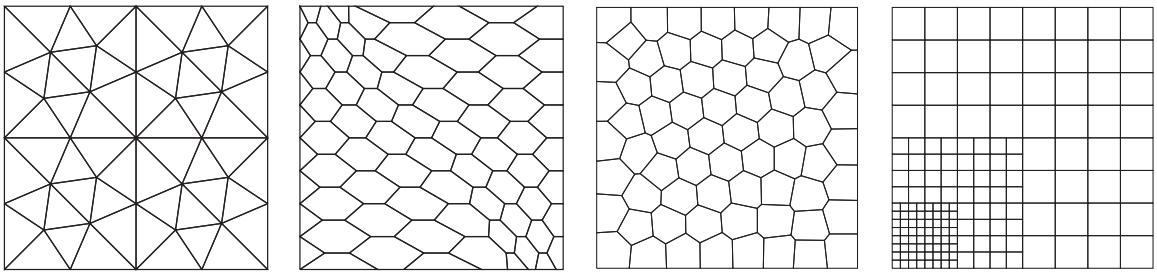


Figure 2.1: Triangular, hexagonal-dominant, Voronoi, and nonmatching quadrangular meshes for the numerical tests. The triangular and nonmatching quadrangular meshes were originally proposed for the FVCA5 benchmark [120], whereas the hexagonal-dominant mesh is the same used in [80, Section 4.2.3].

2.6 Numerical tests

In this section we present a comprehensive set of numerical tests to assess the properties of our method.

2.6.1 Convergence

We first consider a manufactured regular exact solution to confirm the convergence rates predicted in (2.52). Specifically, we solve the two-dimensional incompressible Biot problem ($c_0 = 0$) in the unit square domain $\Omega = (0,1)^2$ with $t_F = 1$ and physical parameters $\mu = 1$, $\lambda = 1$, and $\boldsymbol{\kappa} = 1$. The exact displacement $\boldsymbol{u}$ and exact pressure p are given by, respectively

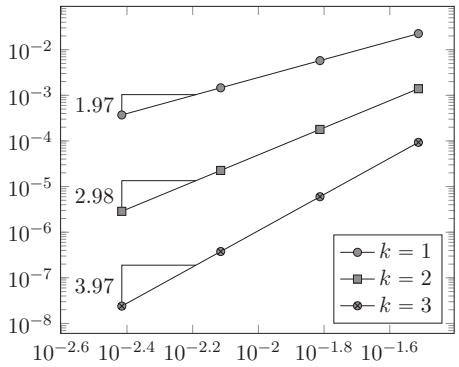
$$\begin{aligned}\boldsymbol{u}(\boldsymbol{x},t) &= (-\sin(\pi t)\cos(\pi x_1)\cos(\pi x_2), \sin(\pi t)\sin(\pi x_1)\sin(\pi x_2)),\\ p(\boldsymbol{x},t) &= -\cos(\pi t)\sin(\pi x_1)\cos(\pi x_2).\end{aligned}$$

The volumetric load is given by

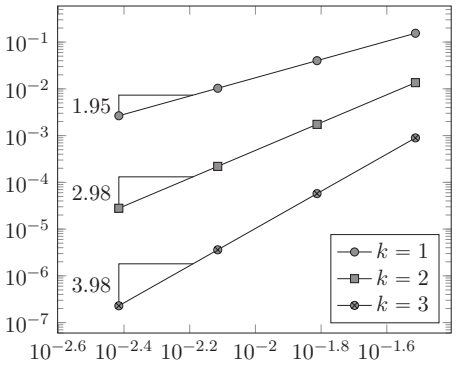
$$\boldsymbol{f}(\boldsymbol{x},t) = 6\pi^2(\sin(\pi t) + \pi\cos(\pi t)) \times (-\cos(\pi x_1)\cos(\pi x_2), \sin(\pi x_1)\sin(\pi x_2)),$$

while $g(\boldsymbol{x},t) \equiv 0$. Dirichlet boundary conditions for the displacement and Neumann boundary conditions for the pressure are inferred from exact solutions to $\partial\Omega$.

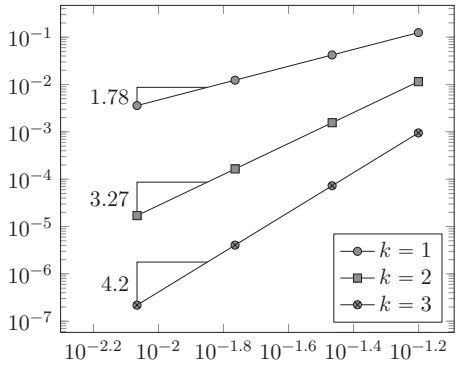
We consider the triangular, (predominantly) hexagonal, Voronoi, and nonmatching quadrangular mesh families depicted in Figure 2.1. The Voronoi mesh family was obtained using the PolyMesher algorithm of [188]. The nonmatching mesh is simply meant to show that the method supports nonconforming interfaces: refining in the corner has no particular meaning for the selected solution. The time discretization is based on the second order Backward Differentiation Formula (BDF2); cf. Remark 2.15. The time step τ on the coarsest mesh is taken to be $0.1/2^{\frac{(k+1)}{2}}$ for every choice of the spatial degree k, and it decreases with the mesh size h according to the theoretical convergence rates, thus, if $h_2 = h_1/2$, then $\tau_2 = \tau_1/2^{\frac{(k+1)}{2}}$. Figure 2.2 displays convergence results for the various mesh families and polynomial degrees up to 3. The error measures are $\|p_h^N - \pi_h^k p^N\|$ for the pressure and $\|\underline{\boldsymbol{u}}_h^N - \underline{\boldsymbol{I}}_h^k \boldsymbol{u}^N\|_{a,h}$ for the displacement. Using the triangle inequality together with (2.52) and the approximation properties (2.2) of π_h^k and (2.6) of $(\boldsymbol{r}_h^{k+1} \circ \underline{\boldsymbol{I}}_h^k)$, it is a simple matter to prove that these quantities



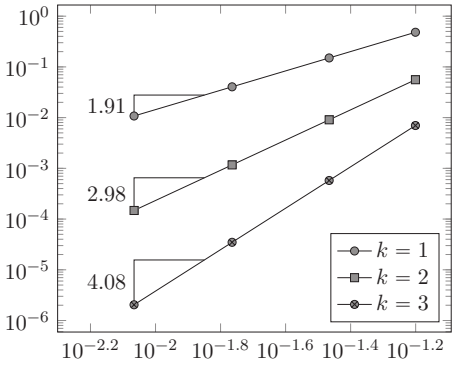
(a) $\|p_h^N - \pi_h^k p^N\|$, triangular



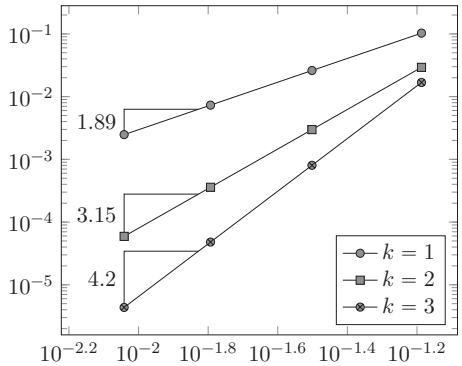
(b) $\|\underline{\boldsymbol{u}}_h^N - \underline{I}_h^k \boldsymbol{u}^N\|_{a,h}$, triangular



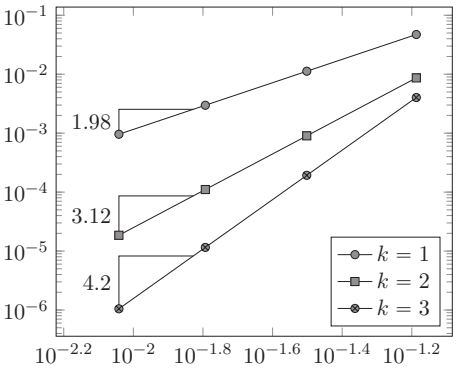
(c) $\|p_h^N - \pi_h^k p^N\|$, hexagonal



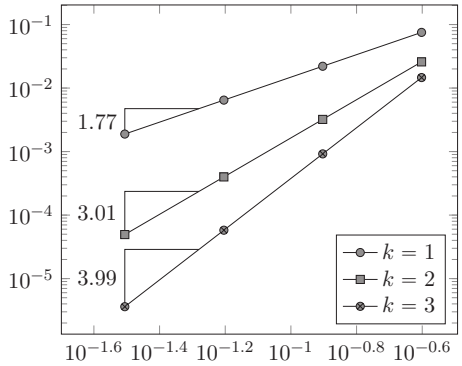
(d) $\|\underline{\boldsymbol{u}}_h^N - \underline{I}_h^k \boldsymbol{u}^N\|_{a,h}$, hexagonal



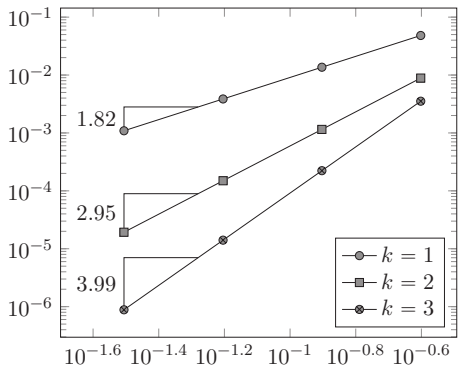
(e) $\|p_h^N - \pi_h^k p^N\|$, Voronoi



(f) $\|\underline{\boldsymbol{u}}_h^N - \underline{I}_h^k \boldsymbol{u}^N\|_{a,h}$, Voronoi



(g) $\|p_h^N - \pi_h^k p^N\|$, nonmatching



(h) $\|\underline{\boldsymbol{u}}_h^N - \underline{I}_h^k \boldsymbol{u}^N\|_{a,h}$, nonmatching

Figure 2.2: Errors vs. h

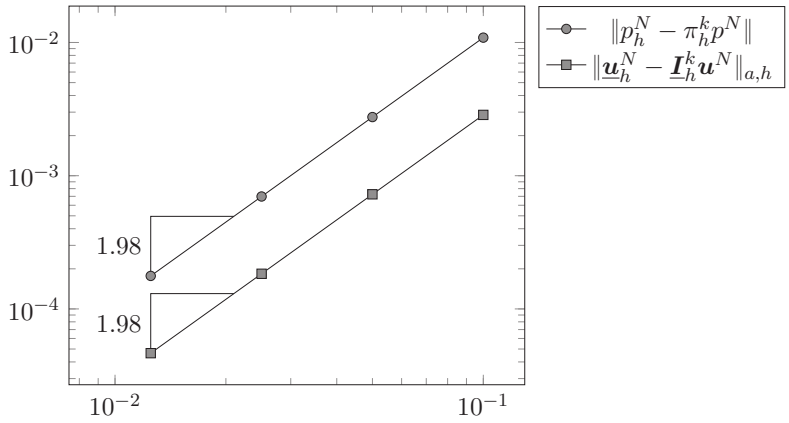


Figure 2.3: Time convergence rate with BDF2, hexagonal mesh

have the same convergence behaviour as the terms in the left-hand side of (2.52). In all the cases, the numerical results show asymptotic convergence rates that are in agreement with theoretical predictions. This test was also used to numerically check that the mechanical equilibrium and mass conservation relations of Lemma 2.18 hold up to machine precision.

The convergence in time was also separately checked considering the method with spatial degree $k = 3$ on the hexagonal mesh with mesh size $h = 0.0172$ and time step decreasing from $\tau = 0.1$ to $\tau = 0.0125$. With this choice, the time-component of the error is dominant, and Figure 2.3 confirms the second order convergence of the BDF2 scheme.

2.6.2 Barry and Mercer's test case

A test case more representative of actual physical configurations is that of Barry and Mercer [14], for which an exact solution is available (we refer to the cited paper and also to [164, Section 4.2.1] for the its expression). We let $\Omega = (0,1)^2$ and consider the following time-independent boundary conditions on $\partial\Omega$

$$\boldsymbol{u} \cdot \boldsymbol{\tau} = 0, \qquad \boldsymbol{n}^{\mathrm{T}} \boldsymbol{\nabla} \boldsymbol{u}\, \boldsymbol{n} = 0, \qquad p = 0,$$

where $\boldsymbol{\tau}$ denotes the tangent vector on $\partial\Omega$. The evolution of the displacement and pressure fields is driven by a periodic pointwise source (mimicking a well) located at $\boldsymbol{x}_0 = (0.25, 0.25)$:

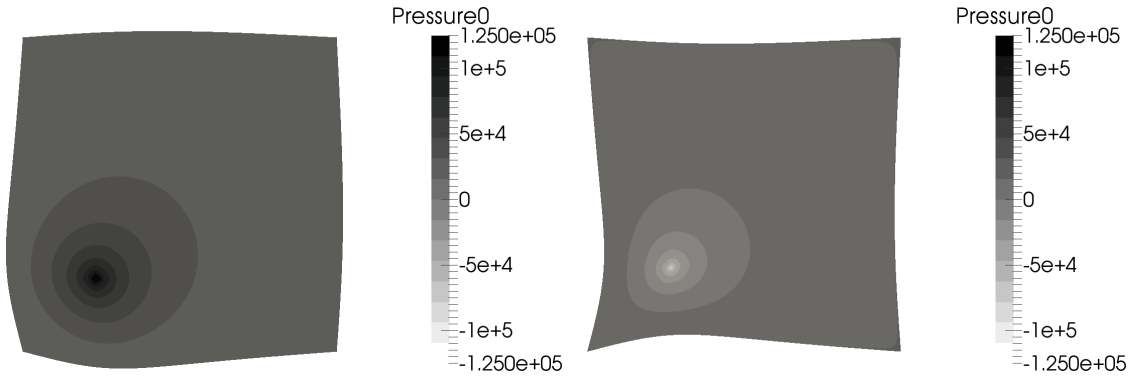
$$g = \delta(\boldsymbol{x} - \boldsymbol{x}_0) \sin(\hat{t}),$$

with normalized time $\hat{t} := \beta t$ for $\beta := (\lambda + 2\mu)\boldsymbol{\kappa}$. As in [166, 176], we use the following values for the physical parameters:

$$c_0 = 0, \qquad E = 1 \cdot 10^5, \qquad \nu = 0.1, \qquad \boldsymbol{\kappa} = 1 \cdot 10^{-2},$$

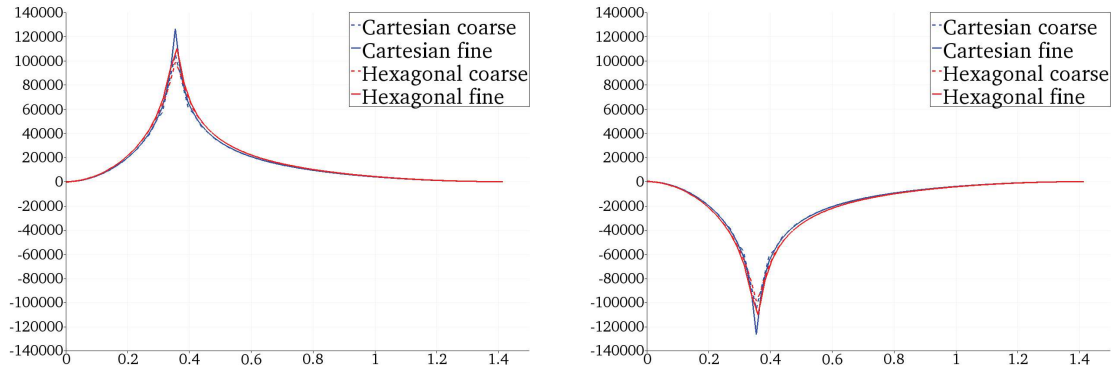
where E and ν denote Young's modulus and Poisson ratio, respectively, and

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{E}{2(1+\nu)}.$$



(a) $\hat{t} = \pi/2$ (b) $\hat{t} = 3\pi/2$

Figure 2.4: Pressure field on the deformed domain at different times for the finest Cartesian mesh containing 4,192 elements



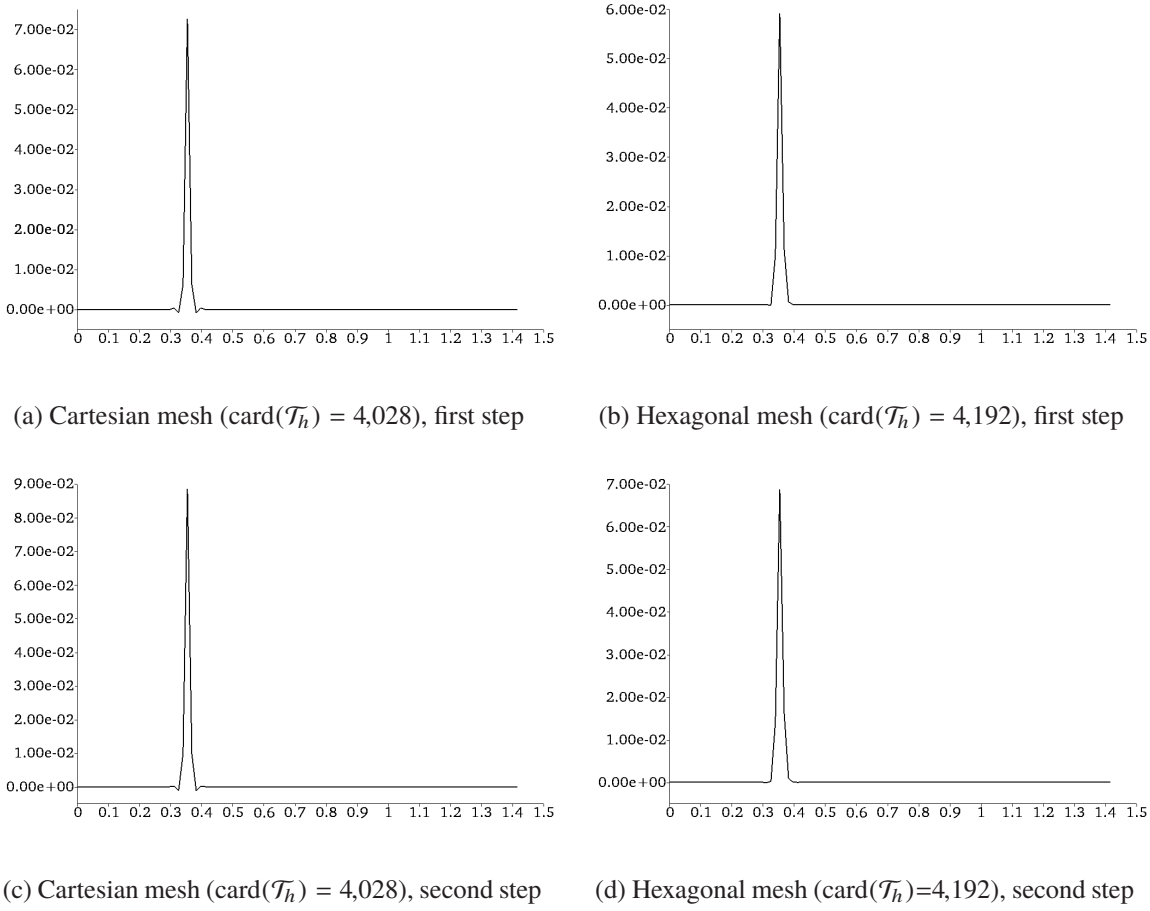
(a) $\hat{t} = \pi/2$, $\kappa = 1 \cdot 10^{-2}$ (b) $\hat{t} = 3\pi/2$, $\kappa = 1 \cdot 10^{-2}$

Figure 2.5: Pressure profiles along the diagonal $(0,0)$–$(1,1)$ of the domain for different normalized times $\hat{t}$ and meshes ($k = 1$). The time step is here $\tau = (2\pi/\beta) \cdot 10^{-2}$.

In the injection phase $\hat{t} \in (0, \pi)$, we observe an inflation of the domain, which reaches its maximum at $\hat{t} = \pi/2$; cf. Figure 2.4a. In the extraction phase $\hat{t} \in (\pi, 2\pi)$, on the other hand, we have a contraction of the domain which reaches its maximum at $\hat{t} = 3\pi/2$; cf. Figure 2.4b.

The following results have been obtained with the lowest-order version of the method corresponding to $k = 1$ (taking advantage of higher orders would require local mesh refinement, which is out of the scope of the present work). In Figure 2.5 we plot the pressure profile at normalized times $\hat{t} = \pi/2$ and $\hat{t} = 3\pi/2$ along the diagonal $(0,0)$–$(1,1)$ of the domain. We consider two Cartesian meshes containing 1,024 and 4,096 elements, respectively, as well as two (predominantly) hexagonal meshes containing 1,072 and 4,192 elements, respectively. In all the cases, a time step $\tau = (2\pi/\beta) \cdot 10^{-2}$ is used. We note that the behaviour of the pressure is well-captured even on the coarsest meshes. For the finest hexagonal mesh, the relative error on the pressure in the L^2-norm at times $\hat{t} = \pi/2$ and $\hat{t} = 3\pi/2$ is 2.85%.

To check the robustness of the method with respect to pressure oscillations for small



(a) Cartesian mesh (card($\mathcal{T}_h$) = 4,028), first step

(b) Hexagonal mesh (card($\mathcal{T}_h$) = 4,192), first step

(c) Cartesian mesh (card($\mathcal{T}_h$) = 4,028), second step

(d) Hexagonal mesh (card($\mathcal{T}_h$)=4,192), second step

Figure 2.6: Pressure profiles along the diagonal $(0,0)$–$(1,1)$ of the domain for $\kappa = 1 \cdot 10^{-6}$ and time step $\tau = 1 \cdot 10^{-4}$. Small oscillations are present on the Cartesian mesh (left), whereas no sign of oscillations is present on the hexagonal mesh (right).

permeabilities combined with small time steps, we also show in Figure 2.6 the pressure profile after one and two step with $\kappa = 1 \cdot 10^{-6}$ and $\tau = 1 \cdot 10^{-4}$ on the Cartesian and hexagonal meshes with 4,096 and 4,192 elements, respectively. We remark that the first time step is performed using the backward Euler scheme, while the second with the second order BDF2 scheme. This situation corresponds to the one considered in [176, Figure 5.10] to highlight the onset of spurious oscillations. In our case, small oscillations can be observed for the Cartesian mesh (cf. Figure 2.6a and Figure 2.6c), whereas no sign of oscillations in present for the hexagonal mesh (cf. Figure 2.6b and Figure 2.6d). One possible conjecture is that increasing the number of element faces contributes to the monotonicity of the scheme.

2.7 Appendix: Flux formulation

In this section, we reformulate the discrete problem (2.26) to unveil the local conservation properties of the method. Before doing so, we need to introduce a few operators and some

notation to treat the boundary terms.

We start from the mechanical equilibrium. Let an element $T \in \mathcal{T}_h$ be fixed and denote by $\boldsymbol{U}_{\partial T} \coloneqq \mathbb{P}^k_{d-1}(\mathcal{F}_T)^d$ the broken polynomial space of degree $\le k$ on the boundary ∂T of T. We define the boundary operator $\boldsymbol{L}^k_T : \boldsymbol{U}_{\partial T} \to \boldsymbol{U}_{\partial T}$ such that, for all $\boldsymbol{\varphi} \in \boldsymbol{U}_{\partial T}$,

$$\boldsymbol{L}^k_T\boldsymbol{\varphi}_{|F} \coloneqq \pi^k_F\left(\boldsymbol{\varphi}_{|F} - \boldsymbol{r}^{k+1}_T(\boldsymbol{0}, (\boldsymbol{\varphi}_{|F})_{F\in\mathcal{F}_T}) + \pi^k_T\boldsymbol{r}^{k+1}_T(\boldsymbol{0}, (\boldsymbol{\varphi}_{|F})_{F\in\mathcal{F}_T})\right) \qquad \forall F \in \mathcal{F}_T.$$

We also need the adjoint $\boldsymbol{L}^{k,*}_T$ of $\boldsymbol{L}^k_T$ such that

$$\forall \boldsymbol{\varphi} \in \boldsymbol{U}_{\partial T}, \qquad (\boldsymbol{L}^k_T\boldsymbol{\varphi}, \boldsymbol{\psi})_{\partial T} = (\boldsymbol{\varphi}, \boldsymbol{L}^{k,*}_T\boldsymbol{\psi})_{\partial T} \qquad \forall \boldsymbol{\psi} \in \boldsymbol{U}_{\partial T}.$$

For a collection of DOFs $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}^k_T$, we denote in what follows by $\boldsymbol{v}_{\partial T} \in \boldsymbol{U}_{\partial T}$ the function in $\boldsymbol{U}_{\partial T}$ such that $\boldsymbol{v}_{\partial T|F} = \boldsymbol{v}_F$ for all $F \in \mathcal{F}_T$. Finally, it is convenient to define the discrete stress operator $\boldsymbol{S}^k_T : \underline{\boldsymbol{U}}^k_T \to \mathbb{P}^k_d(T)^{d\times d}$ such that, for all $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}^k_T$,

$$\boldsymbol{S}^k_T\underline{\boldsymbol{v}}_T \coloneqq 2\mu\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{r}^{k+1}_T\underline{\boldsymbol{v}}_T + \lambda\boldsymbol{I}_d D^k_T\underline{\boldsymbol{v}}_T. \tag{2.65}$$

To reformulate the mass conservation equation, we need to introduce the classical lifting operator $R^k_{\boldsymbol{\kappa},h} : P^k_h \to \mathbb{P}^{k-1}_d(\mathcal{T}_h)^d$ such that, for all $q_h \in P^k_h$, it holds

$$(R^k_{\boldsymbol{\kappa},h}q_h, \boldsymbol{\xi}_h) = \sum_{F\in\mathcal{F}^{\mathrm{i}}_h}([q_h]_F, \{\boldsymbol{\kappa}\boldsymbol{\xi}_h\}_F\cdot\boldsymbol{n}_F)_F \qquad \forall \boldsymbol{\xi}_h \in \mathbb{P}^{k-1}_d(\mathcal{T}_h)^d.$$

Lemma 2.18 (Flux formulation of problem (2.26))**.** *Problem (2.26) can be reformulated as follows: Find $(\underline{\boldsymbol{u}}^n_h, p^n_h) \in \underline{\boldsymbol{U}}^k_{h,\mathrm{D}} \times P^k_h$ such that it holds, for all $(\underline{\boldsymbol{v}}_h, q_h) \in \underline{\boldsymbol{U}}^k_{h,\mathrm{D}} \times \mathbb{P}^k_d(\mathcal{T}_h)$ and all $T \in \mathcal{T}_h$,*

$$(\boldsymbol{S}^k_T\underline{\boldsymbol{u}}^n_T - p^n_h\boldsymbol{I}_d, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_T)_T + \sum_{F\in\mathcal{F}_T}(\boldsymbol{\Phi}^k_{TF}(\underline{\boldsymbol{u}}^n_T, p^n_{h|T}), \boldsymbol{v}_F - \boldsymbol{v}_T)_F = (\boldsymbol{f}^n, \boldsymbol{v}_T)_T, \tag{2.66a}$$

$$(c_0\delta_t p^n_h, q_h)_T - (\delta_t\boldsymbol{u}^n_T - \boldsymbol{\kappa}(\boldsymbol{\nabla}_h p^n_h - R^k_{\boldsymbol{\kappa},h}p^n_h), \boldsymbol{\nabla}_h q_h)_T - \sum_{F\in\mathcal{F}_T}(\phi^k_{TF}(\delta_t\boldsymbol{u}^n_F, p^n_h), q_{h|T})_F = (g^n, q_h), \tag{2.66b}$$

where, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, the numerical traction $\boldsymbol{\Phi}^k_{TF} : \underline{\boldsymbol{U}}^k_T \times \mathbb{P}^k_d(T) \to \mathbb{P}^k_{d-1}(F)^d$ and mass flux $\phi^k_{TF} : \mathbb{P}^k_{d-1}(F)^d \times \mathbb{P}^k_d(\mathcal{T}_h) \to \mathbb{P}^k_{d-1}(F)$ are such that

$$\begin{aligned}
\boldsymbol{\Phi}^k_{TF}(\underline{\boldsymbol{v}}_T, q) &\coloneqq (\boldsymbol{S}^k_T\underline{\boldsymbol{v}}_T - q\boldsymbol{I}_d)\boldsymbol{n}_{TF} + (2\mu)\boldsymbol{L}^{k,*}_T(\mathfrak{h}^{-1}_{\partial T}\boldsymbol{L}^k_T(\boldsymbol{v}_{\partial T} - \boldsymbol{v}_T)),\\
\phi^k_{TF}(\boldsymbol{v}_F, q_h) &\coloneqq \begin{cases}(-\boldsymbol{v}^n_F + \{\boldsymbol{\kappa}\boldsymbol{\nabla}_h q_h\}_F)\cdot\boldsymbol{n}_{TF} - \frac{\varsigma\lambda_{\boldsymbol{\kappa},F}}{h_F}[q_h]_F(\boldsymbol{n}_{TF}\cdot\boldsymbol{n}_F) & \text{if } F \in \mathcal{F}^{\mathrm{i}}_h,\\ 0 & \text{otherwise,}\end{cases}
\end{aligned} \tag{2.67}$$

with $\mathfrak{h}_{\partial T} \in \mathbb{P}^0_d(\mathcal{F}_T)$ such that $\mathfrak{h}_{\partial T|F} = h_F$ for all $F \in \mathcal{F}_T$, and it holds, for all $F \in \mathcal{F}^{\mathrm{i}}_h$ such that $F \in \mathcal{F}_{T_1} \cap \mathcal{F}_{T_2}$,

$$\boldsymbol{\Phi}^k_{T_1F}(\underline{\boldsymbol{u}}^n_{T_1}, p^n_{h|T_1}) + \boldsymbol{\Phi}^k_{T_2F}(\underline{\boldsymbol{u}}^n_{T_2}, p^n_{h|T_2}) = \boldsymbol{0} \tag{2.68a}$$

$$\phi^k_{T_1F}(\delta_t\boldsymbol{u}^n_F, p^n_h) + \phi^k_{T_2F}(\delta_t\boldsymbol{u}^n_F, p^n_h) = 0. \tag{2.68b}$$

Proof. (1) *Proof of* (2.66a). Proceeding as in [57, Section 3.1], the stabilization bilinear form s_T defined by (2.10) can be rewritten as

$$s_T(\underline{\boldsymbol{w}}_T, \underline{\boldsymbol{v}}_T) = \sum_{F\in\mathcal{F}_T} (\boldsymbol{L}_T^{k,*}(\mathfrak{h}_{\partial T}^{-1}\boldsymbol{L}_T^k(\boldsymbol{w}_{\partial T} - \boldsymbol{w}_T)), \boldsymbol{v}_F - \boldsymbol{v}_T)_F.$$

Therefore, using the definitions (2.5) of $\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T$ with $\boldsymbol{w} = \boldsymbol{r}_T^{k+1}\underline{\boldsymbol{u}}_T^n$ and (2.7a) of $D_T^k\underline{\boldsymbol{v}}_T$ with $q = p_{h|T}^n$, and recalling the definition (2.65) of $\boldsymbol{S}_T^k$, one has

$$a_T(\underline{\boldsymbol{u}}_T^n, \underline{\boldsymbol{v}}_T) = (\boldsymbol{S}_T^k\underline{\boldsymbol{u}}_T^n, \boldsymbol{\nabla}_s\boldsymbol{v}_T)_T + \sum_{F\in\mathcal{F}_T} (\boldsymbol{S}_T^k\underline{\boldsymbol{u}}_T^n\boldsymbol{n}_{TF} + (2\mu)\boldsymbol{L}_T^{k,*}(\mathfrak{h}_{\partial T}^{-1}\boldsymbol{L}_T^k(\boldsymbol{u}_{\partial T}^n - \boldsymbol{u}_T^n)), \boldsymbol{v}_F - \boldsymbol{v}_T)_F. \tag{2.69}$$

On the other hand, using again the definition (2.7a) of $D_T^k\underline{\boldsymbol{v}}_T$ with $q = p_{h|T}^n$, one has

$$b_T(\underline{\boldsymbol{v}}_T, p_{h|T}^n) = -(p_h^n\boldsymbol{I}_d, \boldsymbol{\nabla}_s\boldsymbol{v}_T)_T - \sum_{F\in\mathcal{F}_T} (p_{h|T}^n\boldsymbol{n}_{TF}, \boldsymbol{v}_F - \boldsymbol{v}_T)_F. \tag{2.70}$$

Equation (2.66a) follows summing (2.69) and (2.70).
(2) *Proof of* (2.66b). Using the definition (2.7b) of D_T^k with $\underline{\boldsymbol{v}}_T = \delta_t\underline{\boldsymbol{u}}_T^n$ and $q = q_{h|T}$, it is inferred that

$$b_T(\delta_t\underline{\boldsymbol{u}}_T^n, q_h) = -(\delta_t\boldsymbol{u}_T^n, \boldsymbol{\nabla}_h q_h)_T + \sum_{F\in\mathcal{F}_T} (\delta_t\boldsymbol{u}_F^n\cdot\boldsymbol{n}_{TF}, q_{h|T})_F. \tag{2.71}$$

On the other hand, adapting the results [76, Section 4.5.5] to the homogeneous Neumann boundary condition (2.1d), it is inferred

$$c_h(p_h^n, q_h) = \sum_{T\in\mathcal{T}_h} \Bigg\{ (\boldsymbol{\kappa}(\boldsymbol{\nabla}_h p_h^n - R_{\boldsymbol{\kappa},h}^k p_h^n)\cdot\boldsymbol{\nabla}_h q_h)_T - \sum_{F\in\mathcal{F}_T\cap\mathcal{F}_h^{\mathrm{i}}} (\{\boldsymbol{\kappa}\boldsymbol{\nabla}_h p_h^n\}_F\cdot\boldsymbol{n}_{TF} - \frac{\varsigma\lambda_{\boldsymbol{\kappa},F}}{h_F}[p_h^n]_F(\boldsymbol{n}_{TF}\cdot\boldsymbol{n}_F), q_{h|T})_F \Bigg\}. \tag{2.72}$$

Equation (2.66b) follows summing (2.71) and (2.72).
(3) *Proof of* (2.68). To prove (2.68a), let an internal face $F\in\mathcal{F}_h^{\mathrm{i}}$ be fixed, and make $\underline{\boldsymbol{v}}_h$ in (2.68a) such that $\boldsymbol{v}_T\equiv\boldsymbol{0}$ for all $T\in\mathcal{T}_h$, $\boldsymbol{v}_{F'}\equiv\boldsymbol{0}$ for all $F'\in\mathcal{F}_h\setminus\{F\}$, let $\boldsymbol{v}_F$ span $\mathbb{P}_{d-1}^k(F)$ and rearrange the sums. The mass flux conservation (2.68b) follows immediately from the expression of ϕ_{TF}^k observing that, for all $(\underline{\boldsymbol{v}}_h, q_h)\in\underline{\boldsymbol{U}}_h^k\times P_h^k$ and all $F\in\mathcal{F}_h^{\mathrm{i}}$, the quantity

$$(-\boldsymbol{v}_F + \{\boldsymbol{\kappa}\boldsymbol{\nabla}_h q_h\}_F)\cdot\boldsymbol{n}_F - \frac{\varsigma\lambda_{\boldsymbol{\kappa},F}}{h_F}[q_h]_F$$

is single-valued on F. □

Let now an element $T\in\mathcal{T}_h$ be fixed. Choosing as test functions in (2.66a) $\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ such that $\boldsymbol{v}_F\equiv\boldsymbol{0}$ for all $F\in\mathcal{F}_h$, $\boldsymbol{v}_{T'}\equiv\boldsymbol{0}$ for all $T'\in\mathcal{T}_h\setminus\{T\}$, and $\boldsymbol{v}_T$ spans $\mathbb{P}_d^k(T)^d$, we infer the following local mechanical equilibrium relation: For all $\boldsymbol{v}_T\in\mathbb{P}_d^k(T)^d$,

$$(\boldsymbol{S}_T^k\underline{\boldsymbol{u}}_T^n - p_h^n\boldsymbol{I}_d, \boldsymbol{\nabla}_s\boldsymbol{v}_T)_T - \sum_{F\in\mathcal{F}_T} (\boldsymbol{\Phi}_{TF}^k(\underline{\boldsymbol{u}}_T^n, p_{h|T}^n), \boldsymbol{v}_T)_F = (\boldsymbol{f}^n, \boldsymbol{v}_T)_T.$$

Similarly, selecting q_h in (2.66b) such that $q_{h|T'} \equiv 0$ for all $T' \in \mathcal{T}_h \setminus \{T\}$ and $q_T \coloneqq q_{h|T}$ spans $\mathbb{P}_d^k(T)$, we infer the following local mass conservation relation: For all $q_T \in \mathbb{P}_d^k(T)$,

$$(c_0\delta_t p_h^n, q_T)_T - (\delta_t \boldsymbol{u}_T^n - \boldsymbol{\kappa}(\boldsymbol{\nabla}_h p_h^n - R_{\boldsymbol{\kappa},h}^k p_h^n), \boldsymbol{\nabla} q_T)_T - \sum_{F\in\mathcal{F}_T} (\phi_{TF}^k(\delta_t \boldsymbol{u}_F^n, p_h^n), q_T)_F = (g^n, q_T).$$

To actually compute the numerical fluxes defined by (2.67), besides the operator $\boldsymbol{S}_T^k$ defined by (2.65) (which is readily available once $\boldsymbol{r}_T^{k+1}$ and D_T^k have been computed; cf. (2.5) and (2.7a), respectively), one also needs to compute the operators $\boldsymbol{L}_T^k$ and $\boldsymbol{L}_T^{k,*}$. The latter operation can be performed at marginal cost, since it only requires to invert the face mass matrices of $\mathbb{P}_{d-1}^k(F)$ for all $F \in \mathcal{F}_T$.

Chapter 3

Nonlinear elasticity

This chapter has been published in the following peer-reviewed journal (see [33]):

SIAM Journal on Numerical Analysis,
Volume 55, Issue 6, 2017, Pages 2687–2717.

Contents

3.1 Introduction

In this chapter we develop and analyze a novel Hybrid High-Order (HHO) method for a class of (linear and) nonlinear elasticity problems in the small deformation regime.

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, denote a bounded connected open polyhedral domain with Lipschitz boundary $\Gamma := \partial\Omega$ and outward normal $\boldsymbol{n}$. We consider a body that occupies the region Ω and is subjected to a volumetric force field $\boldsymbol{f} \in L^2(\Omega;\mathbb{R}^d)$. For the sake of simplicity, we assume the body fixed on Γ (extensions to other standard boundary conditions are possible). The nonlinear elasticity problem consists in finding a vector-valued displacement field $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ solution of

$$-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\cdot,\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) = \boldsymbol{f} \qquad \text{in } \Omega, \tag{3.1a}$$

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \Gamma, \tag{3.1b}$$

where $\boldsymbol{\nabla}_{\mathrm{s}}$ denotes the symmetric gradient. The stress-strain law $\boldsymbol{\sigma} : \Omega \times \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}^{d\times d}_{\mathrm{sym}}$ is assumed to satisfy regularity requirements closely inspired by [87], including conditions on its growth, coercivity, and monotonicity; cf. Assumption 3.1 below for a precise statement. Problem (3.1) is relevant, e.g., in modeling the mechanical behavior of soft materials [191] and metal alloys [169]. Examples of stress-strain laws of common use in the engineering practice are collected in Section 3.2.

The HHO discretization studied in this chapter is inspired by recent works on linear elasticity [74] (where HHO methods where originally introduced) and Leray–Lions operators [69, 70]. It hinges on degrees of freedom (DOFs) that are discontinuous polynomials of degree $k \geq 1$ on the mesh and on the mesh skeleton. Based on these DOFs, we reconstruct discrete counterparts of the symmetric gradient and of the displacement by solving local linear problems inside each mesh element. These reconstruction operators are used to formulate a local contribution composed of two terms: a consistency term inspired by the weak formulation of problem (3.1) with $\boldsymbol{\nabla}_{\mathrm{s}}$ replaced by its discrete counterpart, and a stabilization term penalizing cleverly designed face-based residuals. The resulting method has several advantageous features: (i) it is valid in arbitrary space dimension; (ii) it supports arbitrary polynomial orders ≥ 1 on fairly general meshes including, e.g., polyhedral elements and nonmatching interfaces; (iii) it satisfies inside each mesh element a local principle of virtual work with numerical tractions that obey the law of action and reaction; (iv) it can be efficiently implemented thanks to the possibility of statically condensing a large subset of the unknowns for linearized versions of the problem encountered, e.g., when solving the corresponding system of nonlinear algebraic equations by the Newton method (a numerical comparison between HHO methods and standard conforming finite element methods in the context of scalar diffusion problems can be found in [79]). Additionally, as shown by the numerical tests of Section 3.7, the method is robust with respect to strong nonlinearities.

In the context of structural mechanics, discretization methods supporting polyhedral meshes and nonconforming interfaces can be useful for several reasons including, e.g., the use of hanging nodes for contact [26, 202] and interface [106] elasticity problems, the simplicity in mesh refinement [186] and coarsening [16] for adaptivity, and the greater robustness to mesh distorsion [55] and fracture [139]. The use of high-order methods, on the other hand,

can classically accelerate the convergence in the presence of regular exact solutions or when combined with local mesh refinement. Over the last few years, several discretization schemes supporting polyhedral meshes and/or high-order have been proposed for the linear version of problem (3.1); a non-exhaustive list includes [20, 74, 80, 81, 185, 197, 199]. For the nonlinear version, the literature is more scarce. Conforming approximations on standard meshes have been considered in [104, 105], where the convergence analysis is carried out assuming regularity for the exact displacement field $\boldsymbol{u}$ and the constraint tensor $\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})$ beyond the minimal regularity required by the weak formulation. Discontinuous Galerkin methods on standard meshes have been considered in [163], where convergence is proved for $d = 2$ assuming $\boldsymbol{u} \in H^{m+1}(\Omega;\mathbb{R}^2)$ for some $m > 2$, and in [25], where convergence to minimal regularity solutions is proved for stress-strain functions similar to [21]. General meshes are considered, on the other hand, in [21] and [54], where the authors propose a low-order Virtual Element method, whose convergence analysis is carried out for nonlinear elasticity problems in the small deformation regime (more general problems are considered numerically). In [21], an energy-norm convergence estimate in h (with h denoting, as usual, the meshsize) is proved when $\boldsymbol{u} \in H^2(\Omega;\mathbb{R}^d)$ under the assumption that the function $\tau \mapsto \boldsymbol{\sigma}(\cdot,\tau)$ is piecewise C^1 with positive definite and bounded differential inside each element. A closer look at the proof reveals that properties essentially analogous to the ones considered in Assumption 3.14 below are in fact sufficient for the analysis, while C^1-regularity is used for the evaluation of the stability constant. Convergence to solutions that exhibit only the minimal regularity required by the weak formulation and for stress-strain functions as in Assumption 3.1 is proved in [87] for Gradient Schemes [85]. In this case, convergence rates are only proved for the linear case. We note, in passing, that the HHO method studied here fails to enter the Gradient Scheme framework essentially because the stabilization term is not embedded in the discrete symmetric gradient operator; see [72].

We carry out a complete analysis for the proposed HHO discretization of problem (3.1). Existence of a discrete solution is proved in Theorem 3.7, where we also identify a strict monotonicity assumption on the stress-strain law which ensures uniqueness. Convergence to minimal regularity solutions $\boldsymbol{u} \in H^1_0(\Omega;\mathbb{R}^d)$ is proved in Theorem 3.9 using a compactness argument inspired by [69, 87]. More precisely, we prove for monotone stress-strain laws that (i) the discrete displacement field strongly converges (up to a subsequence) to $\boldsymbol{u}$ in $L^q(\Omega;\mathbb{R}^d)$ with $1 \le q < +\infty$ if $d = 2$ and $1 \le q < 6$ if $d = 3$; (ii) the discrete strain tensor weakly converges (up to a subsequence) to $\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}$ in $L^2(\Omega,\mathbb{R}^{d\times d})$. Notice that our results are slightly stronger than [87, Theorem 3.5] (cf. also Remark 3.6 therein) because the HHO discretization is compact as proved in Lemma 3.12. If, additionally, strict monotonicity holds for $\boldsymbol{\sigma}$, the strain tensor strongly converges and convergence extends to the whole sequence. An optimal energy-norm error estimate in h^{k+1} is then proved in Theorem 3.16 under the additional conditions of Lipschitz continuity and strong monotonicity on the stress-strain law; cf. Assumption 3.14. The performance of the method is investigated in Section 3.7 on a complete panel of model problems using stress-strain laws corresponding to real materials.

The rest of the chapter is organized as follows. In Section 3.2 we formulate the assumptions on the stress-strain function $\boldsymbol{\sigma}$, provide several examples of models relevant in the engineering practice, and write the weak formulation of problem (3.1). In Section 3.3 we introduce the notation for the mesh and recall a few known results. In Section 3.4 we discuss the choice of

DOFs, formulate the local reconstructions and the discrete problem. The main results, stated and proved in Section 3.5, are collected in Theorems 3.7, 3.9, and 3.16 below. In Section 3.6 we show that the HHO method satisfies on each mesh element a discrete counterpart of the principle of virtual work, and that interface tractions obey the law of action and reaction. Section 3.7 contains numerical tests. Finally, Appendix 3.8 contains the proofs of intermediate technical results.

3.2 Setting and examples

For the stress-strain function, we make the following

Assumption 3.1 (Stress-strain function I)**.** The stress-strain function $\boldsymbol{\sigma} : \Omega \times \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}^{d\times d}_{\mathrm{sym}}$ is a Caratheodory function, namely

$$\boldsymbol{\sigma}(\boldsymbol{x}, \cdot) \text{ is continuous on } \mathbb{R}^{d\times d}_{\mathrm{sym}} \text{ for a.e. } \boldsymbol{x} \in \Omega, \tag{3.2a}$$

$$\boldsymbol{\sigma}(\cdot, \boldsymbol{\tau}) \text{ is measurable on } \Omega \text{ for all } \boldsymbol{\tau} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}, \tag{3.2b}$$

and it holds $\boldsymbol{\sigma}(\cdot, \mathbf{0}) \in L^2(\Omega; \mathbb{R}^{d\times d})$. Moreover, there exist real numbers $\overline{\sigma}$, $\underline{\sigma} \in (0, +\infty)$ such that, for a.e. $\boldsymbol{x} \in \Omega$, and all $\boldsymbol{\tau}, \boldsymbol{\eta} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$, the following conditions hold:

$$\|\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x}, \mathbf{0})\|_{d\times d} \le \overline{\sigma}\|\boldsymbol{\tau}\|_{d\times d}, \quad \text{(growth)} \tag{3.2c}$$

$$\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{\tau}) : \boldsymbol{\tau} \ge \underline{\sigma}\|\boldsymbol{\tau}\|^2_{d\times d}, \quad \text{(coercivity)} \tag{3.2d}$$

$$(\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{\eta})) : (\boldsymbol{\tau} - \boldsymbol{\eta}) \ge 0, \quad \text{(monotonicity)} \tag{3.2e}$$

where $\boldsymbol{\tau} : \boldsymbol{\eta} \coloneqq \sum_{i,j=1}^{d} \tau_{i,j}\eta_{i,j}$ and $\|\boldsymbol{\tau}\|^2_{d\times d} \coloneqq \boldsymbol{\tau} : \boldsymbol{\tau}$.

We next discuss a number of meaningful models that satisfy the above assumptions.

Example 3.2 (Linear elasticity)**.** The linear elasticity model corresponds to

$$\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) = \boldsymbol{C}(\cdot)\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u},$$

where $\boldsymbol{C}$ is a fourth order tensor. Being linear, the previous stress-strain relation clearly satisfies Assumption 3.1 provided that $\boldsymbol{C}$ is uniformly elliptic. A particular case of the previous stress-strain relation is the usual linear elasticity Cauchy stress tensor

$$\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) = \lambda \operatorname{tr}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})\boldsymbol{I}_d + 2\mu\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}, \tag{3.3}$$

where $\operatorname{tr}(\boldsymbol{\tau}) \coloneqq \boldsymbol{\tau} : \boldsymbol{I}_d$ and $\lambda \ge 0$, $\mu > 0$ are the Lamé's parameters.

Example 3.3 (Hencky–Mises model)**.** The nonlinear Hencky–Mises model of [105, 157] corresponds to the stress-strain relation

$$\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) = \tilde{\lambda}(\operatorname{dev}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})) \operatorname{tr}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})\boldsymbol{I}_d + 2\tilde{\mu}(\operatorname{dev}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}))\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}, \tag{3.4}$$

where $\tilde{\lambda}$ and $\tilde{\mu}$ are the nonlinear Lamé's scalar functions and $\operatorname{dev} : \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}$ defined by $\operatorname{dev}(\boldsymbol{\tau}) = \operatorname{tr}(\boldsymbol{\tau}^2) - \frac{1}{d}\operatorname{tr}(\boldsymbol{\tau})^2$ is the deviatoric operator. Conditions on $\tilde{\lambda}$ and $\tilde{\mu}$ such that $\boldsymbol{\sigma}$ satisfies Assumption 3.1 can be found in [13, 21].

Example 3.4 (An isotropic damage model)**.** The isotropic damage model of [50] corresponds to the stress-strain relation

$$\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_s \boldsymbol{u}) = (1 - D(\boldsymbol{\nabla}_s \boldsymbol{u}))\boldsymbol{C}(\cdot)\boldsymbol{\nabla}_s \boldsymbol{u}, \tag{3.5}$$

where $D : \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}$ is the scalar damage function and $\mathfrak{C}$ is a fourth-order symmetric and uniformly elliptic tensor. If there exists a continuous and bounded function $f : [0, +\infty) \to [a, b]$ for some $0 < a \le b$, such that $s \in [0, +\infty) \to sf(s)$ is non-decreasing and, for all $\boldsymbol{\tau} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$, $D(\boldsymbol{\tau}) = 1 - f(|\boldsymbol{\tau}|)$, the damage model constitutive relation satisfies Assumption 3.1.

In the numerical experiments of Section 3.7 we will also consider the following model, relevant in engineering applications, which however does not satisfy Assumption 3.1 in general.

Example 3.5 (The second-order elasticity model)**.** The nonlinear second-order isotropic elasticity model of [66, 124, 135] corresponds to the stress-strain relation

$$\begin{aligned}\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}) = {} & \lambda \operatorname{tr}(\boldsymbol{\nabla}_s \boldsymbol{u})\boldsymbol{I}_d + 2\mu \boldsymbol{\nabla}_s \boldsymbol{u} \\ & + B \operatorname{tr}((\boldsymbol{\nabla}_s \boldsymbol{u})^2)\boldsymbol{I}_d + 2B \operatorname{tr}(\boldsymbol{\nabla}_s \boldsymbol{u})\boldsymbol{\nabla}_s \boldsymbol{u} + C \operatorname{tr}(\boldsymbol{\nabla}_s \boldsymbol{u})^2 \boldsymbol{I}_d + A(\boldsymbol{\nabla}_s \boldsymbol{u})^2, \end{aligned} \tag{3.6}$$

where λ and μ are the standard Lamé's parameter, and $A, B, C \in \mathbb{R}$ are the second-order moduli.

Remark 3.6 (Energy density functions). Examples 3.2, 3.3, and 3.5, used in numerical tests of Section 3.7, can be interpreted in the framework of hyperelasticity. Hyperelasticity is a type of constitutive model for ideally elastic materials in which the stress-strain relation derives from a stored energy density function $\Psi : \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}$, namely

$$\boldsymbol{\sigma}(\boldsymbol{\tau}) := \frac{\partial \Psi(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}}.$$

The stored energy density function leading to the linear Cauchy stress tensor (3.3) is

$$\Psi_{\mathrm{lin}}(\boldsymbol{\tau}) := \frac{\lambda}{2} \operatorname{tr}(\boldsymbol{\tau})^2 + \mu \operatorname{tr}(\boldsymbol{\tau}^2), \tag{3.7}$$

while, in the Hencky–Mises model (3.4), it is defined such that

$$\Psi_{\mathrm{hm}}(\boldsymbol{\tau}) := \frac{\alpha}{2} \operatorname{tr}(\boldsymbol{\tau})^2 + \Phi(\operatorname{dev}(\boldsymbol{\tau})). \tag{3.8}$$

Here $\alpha \in (0, +\infty)$, while $\Phi : [0, +\infty) \to \mathbb{R}$ is a function of class C^2 satisfying, for some positive constants C_1, C_2, and C_3,

$$C_1 \le \Phi'(\rho) < \alpha, \quad |\rho\Phi''(\rho)| \le C_2, \text{ and } \quad \Phi'(\rho) + 2\rho\Phi''(\rho) \ge C_3 \quad \forall \rho \in [0, +\infty). \tag{3.9}$$

Deriving the energy density function (3.8) yields the stress-strain relation (3.4) with nonlinear Lamé's functions $\tilde{\mu}(\rho) := \Phi'(\rho)$ and $\tilde{\lambda}(\rho) := \alpha - \Phi'(\rho)$. Taking $\alpha = \lambda + \mu$ and $\Phi(\rho) = \mu\rho$ in (3.8) leads to the linear case. Finally, the second-order elasticity model (3.6) is obtained by adding third-order terms to the linear stored energy density function defined in (3.7):

$$\Psi_{\mathrm{snd}}(\boldsymbol{\tau}) := \frac{\lambda}{2} \operatorname{tr}(\boldsymbol{\tau})^2 + \mu \operatorname{tr}\left(\boldsymbol{\tau}^2\right) + \frac{C}{3} \operatorname{tr}(\boldsymbol{\tau})^3 + B \operatorname{tr}(\boldsymbol{\tau}) \operatorname{tr}(\boldsymbol{\tau}^2) + \frac{A}{3} \operatorname{tr}(\boldsymbol{\tau}^3). \tag{3.10}$$

The weak formulation of problem (3.1) that will serve as a starting point for the development and analysis of the HHO method reads

$$\text{Find } \boldsymbol{u} \in H^1_0(\Omega;\mathbb{R}^d) \text{ such that, for all } \boldsymbol{v} \in H^1_0(\Omega;\mathbb{R}^d),\ a(\boldsymbol{u},\boldsymbol{v}) = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}, \tag{3.11}$$

where $H^1_0(\Omega;\mathbb{R}^d)$ is the zero-trace subspace of $H^1(\Omega;\mathbb{R}^d)$ and the function $a : H^1_0(\Omega;\mathbb{R}^d)\times H^1_0(\Omega;\mathbb{R}^d)\to\mathbb{R}$ is such that

$$a(\boldsymbol{v},\boldsymbol{w}) \coloneqq \int_\Omega \boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}(\boldsymbol{x})) : \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Throughout the rest of the chapter, to alleviate the notation, we omit the dependence on the space variable $\boldsymbol{x}$ and the differential $\mathrm{d}\boldsymbol{x}$ from integrals.

3.3 Notation and basic results

We consider refined sequences of general polytopal meshes as in [85, Definition 7.2] matching the regularity requirements detailed in [82, Definition 3]. The main points are summarized hereafter. Denote by $\mathcal{H}\subset\mathbb{R}^+_*$ a countable set of meshsizes having 0 as its unique accumulation point, and let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a refined mesh sequence where each $\mathcal{T}_h$ is a finite collection of nonempty disjoint open polyhedral elements T with boundary ∂T such that $\overline{\Omega} = \bigcup_{T\in\mathcal{T}_h}\overline{T}$ and $h = \max_{T\in\mathcal{T}_h} h_T$ with h_T diameter of T.

For each $h\in\mathcal{H}$, let $\mathcal{F}_h$ be a set of faces with disjoint interiors which partitions the mesh skeleton, i.e., $\bigcup_{F\in\mathcal{F}_h} F = \bigcup_{T\in\mathcal{T}_h}\partial T$. A face F is defined here as a hyperplanar closed connected subset of $\overline{\Omega}$ with positive $(d-1)$-dimensional Hausdorff measure such that (i) either there exist distinct $T_1, T_2\in\mathcal{T}_h$ such that $F\subset\partial T_1\cap\partial T_2$ and F is called an interface or (ii) there exists $T\in\mathcal{T}_h$ such that $F\subset\partial T\cap\Gamma$ and F is called a boundary face. Interfaces are collected in the set $\mathcal{F}_h^{\mathrm{i}}$ and boundary faces in $\mathcal{F}_h^{\mathrm{b}}$, so that $\mathcal{F}_h \coloneqq \mathcal{F}_h^{\mathrm{i}}\cup\mathcal{F}_h^{\mathrm{b}}$. For all $T\in\mathcal{T}_h$, $\mathcal{F}_T \coloneqq \{F\in\mathcal{F}_h \,:\, F\subset\partial T\}$ denotes the set of faces contained in ∂T and, for all $F\in\mathcal{F}_T$, $\boldsymbol{n}_{TF}$ is the unit normal to F pointing out of T.

Mesh regularity holds in the sense that, for all $h\in\mathcal{H}$, $\mathcal{T}_h$ admits a matching simplicial submesh $\mathfrak{T}_h$ and there exists a real number $\varrho>0$ such that, for all $h\in\mathcal{H}$, (i) for any simplex $S\in\mathfrak{T}_h$ of diameter h_S and inradius r_S, $\varrho h_S\le r_S$ and (ii) for any $T\in\mathcal{T}_h$ and all $S\in\mathfrak{T}_h$ such that $S\subset T$, $\varrho h_T\le h_S$.

Let X be a mesh element or face. For an integer $l\ge 0$, we denote by $\mathbb{P}^l_d(X;\mathbb{R})$ the space spanned by the restriction to X of scalar-valued, d-variate polynomials of total degree l. The L^2-projector $\pi^l_X : L^1(X;\mathbb{R})\to\mathbb{P}^l_d(X;\mathbb{R})$ is defined such that, for all $v\in L^1(X;\mathbb{R})$,

$$\int_X (\pi^l_X v - v)w = 0 \qquad \forall w\in\mathbb{P}^l_d(X;\mathbb{R}). \tag{3.12}$$

When dealing with the vector-valued polynomial space $\mathbb{P}^l_d(X;\mathbb{R}^d)$ or with the tensor-valued polynomial space $\mathbb{P}^l_d(X;\mathbb{R}^{d\times d})$, we use the boldface notation $\boldsymbol{\pi}^l_X$ for the corresponding L^2-orthogonal projectors acting component-wise.

On regular mesh sequences, we have the following optimal approximation properties for π_T^l (for a proof, cf. [76, Lemmas 1.58 and 1.59] and, in a more general framework, [70, Lemmas 3.4 and 3.6]): There exists a real number $C_{\mathrm{app}} > 0$ such that, for all $s \in \{0, \ldots, l+1\}$, all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$, and all $v \in H^s(T;\mathbb{R})$,

$$|v - \pi_T^l v|_{H^m(T;\mathbb{R})} \le C_{\mathrm{app}} h_T^{s-m} |v|_{H^s(T;\mathbb{R})} \qquad \forall m \in \{0, \ldots, s\}, \tag{3.13a}$$

and, if $s \ge 1$,

$$|v - \pi_T^l v|_{H^m(\mathcal{F}_T;\mathbb{R})} \le C_{\mathrm{app}} h_T^{s-m-\frac{1}{2}} |v|_{H^s(T;\mathbb{R})} \qquad \forall m \in \{0, \ldots, s-1\}, \tag{3.13b}$$

with $H^m(\mathcal{F}_T;\mathbb{R})$ denoting the broken hilbert space on the faces of T. Other useful geometric and functional analytic results on regular mesh sequences can be found in [76, Chapter 1] and [69, 70].

At the global level, we define broken versions of polynomial and Sobolev spaces. In particular, for an integer $l \ge 0$, we denote by $\mathbb{P}_d^l(\mathcal{T}_h;\mathbb{R})$, $\mathbb{P}_d^l(\mathcal{T}_h;\mathbb{R}^d)$, and $\mathbb{P}_d^l(\mathcal{T}_h;\mathbb{R}^{d\times d})$, respectively, the space of scalar-valued, vector-valued, and tensor-valued broken polynomial functions on $\mathcal{T}_h$ of total degree l. The space of broken vector-valued polynomial functions of total degree l on the trace of the mesh on the domain boundary Γ is denoted by $\mathbb{P}_d^l(\mathcal{F}_h^{\mathrm{b}};\mathbb{R}^d)$. Similarly, for an integer $s \ge 1$, $H^s(\mathcal{T}_h;\mathbb{R})$, $H^s(\mathcal{T}_h;\mathbb{R}^d)$, and $H^s(\mathcal{T}_h;\mathbb{R}^{d\times d})$ are the scalar-valued, vector-valued, and tensor-valued broken Sobolev spaces of index s.

Throughout the rest of the chapter, for $X \subset \overline{\Omega}$, we denote by $\|\cdot\|_X$ the standard norm in $L^2(X;\mathbb{R})$, with the convention that the subscript is omitted whenever $X = \Omega$. The same notation is used for the vector- and tensor-valued spaces $L^2(X;\mathbb{R}^d)$ and $L^2(X;\mathbb{R}^{d\times d})$.

3.4 The Hybrid High-Order method

In this section we define the space of DOFs and the local reconstructions, and we state the discrete problem.

3.4.1 Degrees of freedom

Let a polynomial degree $k \ge 1$ be fixed. The DOFs for the displacement are collected in the space

$$\underline{\boldsymbol{U}}_h^k := \left(\bigtimes_{T\in\mathcal{T}_h} \mathbb{P}_d^k(T;\mathbb{R}^d)\right) \times \left(\bigtimes_{F\in\mathcal{F}_h} \mathbb{P}_d^k(F;\mathbb{R}^d)\right),$$

see Figure 3.1. Observe that naming $\underline{\boldsymbol{U}}_h^k$ the space of DOFs involves a shortcut: the actual DOFs can be chosen in several equivalent ways (polynomial moments, point values, etc.), and the specific choice does not affect the following discussion. Details concerning the actual choice made in the implementation are given in Section 3.7 below.

For a generic collection of DOFs in $\underline{\boldsymbol{U}}_h^k$, we use the classical HHO underlined notation $\underline{\boldsymbol{v}}_h := ((\boldsymbol{v}_T)_{T\in\mathcal{T}_h}, (\boldsymbol{v}_F)_{F\in\mathcal{F}_h})$. We also denote by $\boldsymbol{v}_h \in \mathbb{P}_d^k(\mathcal{T}_h;\mathbb{R}^d)$ and $\boldsymbol{v}_{\Gamma,h} \in \mathbb{P}_d^k(\mathcal{F}_h^{\mathrm{b}};\mathbb{R}^d)$ (not underlined) the broken polynomial functions such that

$$(\boldsymbol{v}_h)_{|T} = \boldsymbol{v}_T \quad \forall T \in \mathcal{T}_h \quad \text{and} \quad (\boldsymbol{v}_{\Gamma,h})_{|F} = \boldsymbol{v}_F \quad \forall F \in \mathcal{F}_h^{\mathrm{b}}.$$

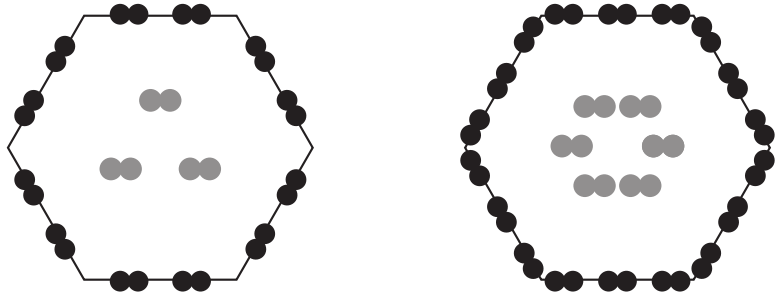


Figure 3.1: Local DOFs for $k = 1$ (left) and $k = 2$ (right). Shaded DOFs can be locally eliminated by static condensation when solving linearized versions of problem (3.21).

The restrictions of $\underline{\boldsymbol{U}}_h^k$ and $\underline{\boldsymbol{v}}_h$ to a mesh element T are denoted by $\underline{\boldsymbol{U}}_T^k$ and $\underline{\boldsymbol{v}}_T = (\boldsymbol{v}_T, (\boldsymbol{v}_F)_{F\in\mathcal{F}_T})$, respectively. The space $\underline{\boldsymbol{U}}_h^k$ is equipped with the following discrete strain semi-norm:

$$\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h} \coloneqq \left(\sum_{T\in\mathcal{T}_h} \|\underline{\boldsymbol{v}}_h\|_{\epsilon,T}^2\right)^{1/2}, \qquad \|\underline{\boldsymbol{v}}_h\|_{\epsilon,T}^2 \coloneqq \|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_T\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1}\|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F^2. \tag{3.14}$$

The DOFs corresponding to a given function $\boldsymbol{v} \in H^1(\Omega;\mathbb{R}^d)$ are obtained by means of the reduction map $\underline{I}_h^k : H^1(\Omega;\mathbb{R}^d) \to \underline{\boldsymbol{U}}_h^k$ such that

$$\underline{I}_h^k\boldsymbol{v} \coloneqq ((\boldsymbol{\pi}_T^k\boldsymbol{v})_{T\in\mathcal{T}_h}, (\boldsymbol{\pi}_F^k\boldsymbol{v})_{F\in\mathcal{F}_h}), \tag{3.15}$$

where we remind the reader that $\boldsymbol{\pi}_T^k$ and $\boldsymbol{\pi}_F^k$ denote the L^2-orthogonal projectors on $\mathbb{P}_d^k(T;\mathbb{R}^d)$ and $\mathbb{P}_d^k(F;\mathbb{R}^d)$, respectively. For all mesh elements $T \in \mathcal{T}_h$, the local reduction map $\underline{I}_T^k : H^1(T;\mathbb{R}^d) \to \underline{\boldsymbol{U}}_T^k$ is obtained by a restriction of $\underline{I}_h^k$, and is therefore such that for all $\boldsymbol{v} \in H^1(T;\mathbb{R}^d)$

$$\underline{I}_T^k\boldsymbol{v} = (\boldsymbol{\pi}_T^k\boldsymbol{v}, (\boldsymbol{\pi}_F^k\boldsymbol{v})_{F\in\mathcal{F}_T}). \tag{3.16}$$

3.4.2 Local reconstructions

We introduce symmetric gradient and displacement reconstruction operators devised at the element level that are instrumental to the formulation of the method.

Let a mesh element $T \in \mathcal{T}_h$ be fixed. The local symmetric gradient reconstruction operator

$$\boldsymbol{G}_{\mathrm{s},T}^k : \underline{\boldsymbol{U}}_T^k \to \mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$$

is obtained by solving the following pure traction problem:c3: For a given local collection of DOFs $\underline{\boldsymbol{v}}_T = (\boldsymbol{v}_T, (\boldsymbol{v}_F)_{F\in\mathcal{F}_T}) \in \underline{\boldsymbol{U}}_T^k$, find $\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T \in \mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$ such that, for all $\boldsymbol{\tau} \in \mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$,

$$\int_T \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T : \boldsymbol{\tau} = -\int_T \boldsymbol{v}_T\cdot(\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) + \sum_{F\in\mathcal{F}_T}\int_F \boldsymbol{v}_F\cdot(\boldsymbol{\tau}\boldsymbol{n}_{TF}) \tag{3.17a}$$

$$= \int_T \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_T : \boldsymbol{\tau} + \sum_{F\in\mathcal{F}_T}\int_F (\boldsymbol{v}_F - \boldsymbol{v}_T)\cdot(\boldsymbol{\tau}\boldsymbol{n}_{TF}). \tag{3.17b}$$

The right-hand side of (3.17a) is designed to resemble an integration by parts formula where the role of the function represented by the DOFs in $\underline{\boldsymbol{v}}_T$ is played by $\boldsymbol{v}_T$ inside the volumetric integral and by $(\boldsymbol{v}_F)_{F\in\mathcal{F}_T}$ inside boundary integrals. The reformulation (3.17b), obtained integrating by parts the first term in the right-hand side of (3.17a), highlights the fact that our method is nonconforming, as the second addend accounts for the difference between $\boldsymbol{v}_F$ and $\boldsymbol{v}_T$.

The definition of the symmetric gradient reconstruction is justified observing that, using the definitions (3.16) of the local reduction map $\underline{\boldsymbol{I}}_T^k$ and (3.12) of the L^2-orthogonal projectors $\boldsymbol{\pi}_T^k$ and $\boldsymbol{\pi}_F^k$ in (3.17a), one can prove the following commuting property: For all $T\in\mathcal{T}_h$ and all $\boldsymbol{v}\in H^1(T;\mathbb{R}^d)$,

$$\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{I}}_T^k\boldsymbol{v} = \boldsymbol{\pi}_T^k(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}). \tag{3.18}$$

As a result of (3.18) and (3.13), $\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{I}}_T^k$ has optimal approximation properties in $\mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$.

From $\boldsymbol{G}_{\mathrm{s},T}^k$, one can define the local displacement reconstruction operator

$$\boldsymbol{r}_T^{k+1}:\underline{\boldsymbol{U}}_T^k\to\mathbb{P}_d^{k+1}(T;\mathbb{R}^d)$$

such that, for all $\underline{\boldsymbol{v}}_T\in\underline{\boldsymbol{U}}_T^k$, $\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T$ is the orthogonal projection of $\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T$ on $\boldsymbol{\nabla}_{\mathrm{s}}\mathbb{P}_d^{k+1}(T;\mathbb{R}^d)\subset\mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$ and rigid-body motions are prescribed according to [74, Eq. (15)]. More precisely, we let $\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T$ be such that for all $\boldsymbol{w}\in\mathbb{P}_d^{k+1}(T;\mathbb{R}^d)$ it holds

$$\int_T(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T-\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T):\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w} = 0$$

and, denoting by $\boldsymbol{\nabla}_{\mathrm{ss}}$ the skew-symmetric part of the gradient operator, we have

$$\int_T\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T=\int_T\boldsymbol{v}_T,\qquad\int_T\boldsymbol{\nabla}_{\mathrm{ss}}\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T=\sum_{F\in\mathcal{F}_T}\int_F\frac{1}{2}\left(\boldsymbol{v}_F\otimes\boldsymbol{n}_{TF}-\boldsymbol{n}_{TF}\otimes\boldsymbol{v}_F\right).$$

Notice that, for a given $\underline{\boldsymbol{v}}_T\in\underline{\boldsymbol{U}}_T^k$, the displacement reconstruction $\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T$ is a vector-valued polynomial function one degree higher than the element-based DOFs $\boldsymbol{v}_T$. It was proved in [74, Lemma 2] that $\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{I}}_T^k$ has optimal approximation properties in $\mathbb{P}_d^{k+1}(T;\mathbb{R}^d)$.

In what follows, we will also need the global counterparts of the discrete gradient and displacement operators $\boldsymbol{G}_{\mathrm{s},h}^k:\underline{\boldsymbol{U}}_h^k\to\mathbb{P}_d^k(\mathcal{T}_h;\mathbb{R}_{\mathrm{sym}}^{d\times d})$ and $\boldsymbol{r}_h^{k+1}:\underline{\boldsymbol{U}}_h^k\to\mathbb{P}_d^{k+1}(\mathcal{T}_h;\mathbb{R}^d)$ defined setting, for all $\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_h^k$ and all $T\in\mathcal{T}_h$,

$$(\boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{v}}_h)_{|T}=\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T,\qquad(\boldsymbol{r}_h^{k+1}\underline{\boldsymbol{v}}_h)_{|T}=\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T. \tag{3.19}$$

3.4.3 Discrete problem

We define the following subspace of $\underline{\boldsymbol{U}}_h^k$ strongly accounting for the homogeneous Dirichlet boundary condition (3.1b):

$$\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k\coloneqq\left\{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_h^k\ :\ \boldsymbol{v}_F=\boldsymbol{0}\quad\forall F\in\mathcal{F}_h^{\mathrm{b}}\right\}, \tag{3.20}$$

and we notice that the map $\|\cdot\|_{\epsilon,h}$ defined by (3.14) is a norm on $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$. The HHO approximation of problem (3.11) reads:

Find $\underline{\boldsymbol{u}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$,

$$a_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{v}}_h) \coloneqq A_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{v}}_h) + s_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{v}}_h) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}_h, \quad (3.21)$$

where the consistency contribution $A_h : \underline{\boldsymbol{U}}_h^k \times \underline{\boldsymbol{U}}_h^k \to \mathbb{R}$ and the stability contribution $s_h : \underline{\boldsymbol{U}}_h^k \times \underline{\boldsymbol{U}}_h^k \to \mathbb{R}$ are respectively defined setting

$$A_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{v}}_h) \coloneqq \int_\Omega \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{u}}_h) : \boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{v}}_h,$$

$$s_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{v}}_h) \coloneqq \sum_{T\in\mathcal{T}_h} s_T(\underline{\boldsymbol{u}}_T, \underline{\boldsymbol{v}}_T), \quad s_T(\underline{\boldsymbol{u}}_T, \underline{\boldsymbol{v}}_T) \coloneqq \sum_{F\in\mathcal{F}_T} \frac{\gamma}{h_F} \int_F \boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{u}}_T \cdot \boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{v}}_T. \quad (3.22)$$

The scaling parameter $\gamma > 0$ in (3.22) can depend on $\overline{\sigma}$ and $\underline{\sigma}$ but is independent of the meshsize h. In the numerical tests of Section 3.7 we take $\gamma = 2\mu$ for the linear (3.3) and second-order (3.6) models and $\gamma = 2\tilde{\mu}(\boldsymbol{0})$ for the Hencky–Mises model (3.4). In s_T, we penalize in a least-square sense the face-based residual $\boldsymbol{\Delta}_{TF}^k : \underline{\boldsymbol{U}}_T^k \to \mathbb{P}_d^k(F;\mathbb{R}^d)$ such that, for all $T \in \mathcal{T}_h$, all $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}_T^k$, and all $F \in \mathcal{F}_T$,

$$\boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{v}}_T \coloneqq \boldsymbol{\pi}_F^k(\boldsymbol{r}_T^{k+1} \underline{\boldsymbol{v}}_T - \boldsymbol{v}_F) - \boldsymbol{\pi}_T^k(\boldsymbol{r}_T^{k+1} \underline{\boldsymbol{v}}_T - \boldsymbol{v}_T). \quad (3.23)$$

This particular choice ensures that $\boldsymbol{\Delta}_{TF}^k$ vanishes whenever its argument is of the form $\underline{\boldsymbol{I}}_T^k \boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{P}_d^{k+1}(T;\mathbb{R}^d)$, a crucial property to obtain an energy-norm error estimate in h^{k+1}; cf. Theorem 3.16 below. Additionally, s_h is stabilizing in the sense that the following uniform norm equivalence holds (the proof is a straightforward modification of [74, Lemma 4]; cf. also Corollary 6 therein): There exists a real number $\eta > 0$ independent of h such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$,

$$\eta^{-1} \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2 \le \|\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{v}}_h\|^2 + s_h(\underline{\boldsymbol{v}}_h, \underline{\boldsymbol{v}}_h) \le \eta \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2. \quad (3.24)$$

By (3.2d), this implies the coercivity of a_h.

3.5 Analysis

In this section we carry out a complete analysis of the method. To alleviate the notation, inside the proofs we abridge into $a \lesssim b$ the inequality $a \le Cb$ with real number $C > 0$ independent of h.

3.5.1 Existence and uniqueness

We start by discussing existence and uniqueness of the discrete solution.

Theorem 3.7 (Existence and uniqueness of a discrete solution)**.** *Let Assumption 3.1 hold and let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a regular mesh sequence. Then, for all $h \in \mathcal{H}$, there exists at least one solution $\underline{\boldsymbol{u}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ to problem (3.21). Additionally, if the stress-strain function $\boldsymbol{\sigma}$ is strictly monotone (i.e., if the inequality in (3.2e) is strict for $\boldsymbol{\tau} \neq \boldsymbol{\eta}$), the solution is unique.*

Proof. 1) *Existence.* We follow the argument of [64, Theorem 3.3]. If $(E, (\cdot,\cdot)_E, \|\cdot\|_E)$ is a Euclidean space and $\Phi : E \to E$ is a continuous map such that $\frac{(\Phi(x),x)_E}{\|x\|_E} \to +\infty$, as $\|x\|_E \to +\infty$, then Φ is surjective. We take $E = \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, endowed with the inner product

$$(\underline{\boldsymbol{v}}_h, \underline{\boldsymbol{w}}_h)_{\epsilon,h} \coloneqq \sum_{T\in\mathcal{T}_h}\left(\int_T \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_T : \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w}_T + \sum_{F\in\mathcal{F}_T}\frac{1}{h_F}\int_F (\boldsymbol{v}_F - \boldsymbol{v}_T)\cdot(\boldsymbol{w}_F - \boldsymbol{w}_T)\right),$$

and we define $\Phi : \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \to \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, $(\Phi(\underline{\boldsymbol{v}}_h), \underline{\boldsymbol{w}}_h)_{\epsilon,h} = a_h(\underline{\boldsymbol{v}}_h, \underline{\boldsymbol{w}}_h)$ for all $\underline{\boldsymbol{w}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$. The coercivity (3.2d) of $\boldsymbol{\sigma}$ together with the norm equivalence (3.24) yields $(\Phi(\underline{\boldsymbol{v}}_h), \underline{\boldsymbol{v}}_h)_{\epsilon,h} \geq \min\{1, \underline{\sigma}\}\eta^{-1}\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2$ for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, so that Φ is surjective. Let now $\underline{\boldsymbol{y}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ be such that $(\underline{\boldsymbol{y}}_h, \underline{\boldsymbol{w}}_h)_{\epsilon,h} = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{w}_h$ for all $\underline{\boldsymbol{w}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$. By the surjectivity of Φ, there exists $\underline{\boldsymbol{u}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ such that $\Phi(\underline{\boldsymbol{u}}_h) = \underline{\boldsymbol{y}}_h$. By definition of Φ and $\underline{\boldsymbol{y}}_h$, $\underline{\boldsymbol{u}}_h$ is a solution to the problem (3.21).

2) *Uniqueness.* Let $\underline{\boldsymbol{u}}_{h,1}, \underline{\boldsymbol{u}}_{h,2} \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ solve (3.21). We assume $\underline{\boldsymbol{u}}_{h,1} \neq \underline{\boldsymbol{u}}_{h,2}$ and proceed by contradiction. Subtracting (3.21) for $\underline{\boldsymbol{u}}_{h,2}$ from (3.21) for $\underline{\boldsymbol{u}}_{h,1}$, it is inferred that $a_h(\underline{\boldsymbol{u}}_{h,1}, \underline{\boldsymbol{v}}_h) - a(\underline{\boldsymbol{u}}_{h,2}, \underline{\boldsymbol{v}}_h) = 0$ for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$. Hence in particular, taking $\underline{\boldsymbol{v}}_h = \underline{\boldsymbol{u}}_{h,1} - \underline{\boldsymbol{u}}_{h,2}$ we obtain that

$$a_h(\underline{\boldsymbol{u}}_{h,1}, \underline{\boldsymbol{u}}_{h,1} - \underline{\boldsymbol{u}}_{h,2}) - a_h(\underline{\boldsymbol{u}}_{h,2}, \underline{\boldsymbol{u}}_{h,1} - \underline{\boldsymbol{u}}_{h,2}) = 0.$$

On the other hand, owing to the strict monotonicity of $\boldsymbol{\sigma}$ and to the fact that the bilinear form s_h is positive semidefinite, we have that

$$\begin{aligned} &a_h(\underline{\boldsymbol{u}}_{h,1}, \underline{\boldsymbol{u}}_{h,1} - \underline{\boldsymbol{u}}_{h,2}) - a_h(\underline{\boldsymbol{u}}_{h,2}, \underline{\boldsymbol{u}}_{h,1} - \underline{\boldsymbol{u}}_{h,2}) \\ &\quad= \int_\Omega (\boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_{h,1}) - \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_{h,2})) : \boldsymbol{G}_{\mathrm{s},h}^k(\underline{\boldsymbol{u}}_{h,1} - \underline{\boldsymbol{u}}_{h,2}) \\ &\qquad+ s_h(\underline{\boldsymbol{u}}_{h,1} - \underline{\boldsymbol{u}}_{h,2}, \underline{\boldsymbol{u}}_{h,1} - \underline{\boldsymbol{u}}_{h,2}) > 0. \end{aligned}$$

Hence, $\underline{\boldsymbol{u}}_{h,1} = \underline{\boldsymbol{u}}_{h,2}$ and the conclusion follows.

□

Remark 3.8 (Strict monotonicity of the stress-strain function). The strict monotonicity assumption is fulfilled, e.g., by the Hencky–Mises model (3.4) and by the damage model (3.5) when $D(\boldsymbol{\tau}) = 1 - f(|\boldsymbol{\tau}|)$, with f continuous, bounded, and such that $[0, +\infty) \ni s \mapsto sf(s)$ is strictly increasing. We observe, in passing, that the strict monotonicity is weaker than the strong monotonicity (3.40b) used in Theorem 3.16 below to prove error estimates.

3.5.2 Convergence

We now consider the convergence to solutions that only exhibit the minimal regularity required by the variational formulation (3.11). More specifically, in this section we prove the following **Theorem 3.9** (Convergence)**.** *Let Assumption 3.1 hold, let $k \geq 1$, and let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a regular mesh sequence. Further assume the existence of a real number $C_{\mathrm{K}} > 0$ independent of h but possibly depending on Ω, ϱ, and on k such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$,*

$$\|\boldsymbol{v}_h\| + \|\boldsymbol{\nabla}_h \boldsymbol{v}_h\| \leq C_{\mathrm{K}} \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}, \tag{3.25}$$

where $\boldsymbol{\nabla}_h$ denotes the broken gradient on $H^1(\mathcal{T}_h;\mathbb{R}^d)$. For all $h \in \mathcal{H}$, let $\underline{\boldsymbol{u}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ be a solution to the discrete problem (3.21) on $\mathcal{T}_h$. Then, for all q such that $1 \leq q < +\infty$ if $d = 2$ or $1 \leq q < 6$ if $d = 3$, as $h \to 0$ it holds, up to a subsequence,

- $\boldsymbol{u}_h \to \boldsymbol{u}$ *strongly in* $L^q(\Omega;\mathbb{R}^d)$,
- $\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{u}}_h \to \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{u}$ *weakly in* $L^2(\Omega;\mathbb{R}^{d\times d})$,

where $\boldsymbol{u} \in H_0^1(\Omega;\mathbb{R}^d)$ solves the weak formulation (3.11). Moreover, if we assume strict monotonicity for $\boldsymbol{\sigma}$ (i.e., the inequality in (3.2e) is strict for $\boldsymbol{\tau} \neq \boldsymbol{\eta}$), it holds that

- $\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{u}}_h \to \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{u}$ *strongly in* $L^2(\Omega;\mathbb{R}^{d\times d})$.

Finally, if the solution to (3.11) is unique, convergence extends to the whole sequence.

Some remarks are of order before proceeding with the proof of Theorem 3.9.

Remark 3.10 (Existence of a solution to the continuous problem). A side result of the existence of discrete solutions proved in Theorem 3.7 together with the convergence results of Theorem 3.9 is the existence of a solution to the weak formulation (3.11).

Remark 3.11 (Discrete Korn inequality). In Proposition 3.20 below we give a proof of the discrete Korn inequality (3.25) based on the results of [39], which require further assumptions on the mesh. While we have the feeling that these assumptions could probably be relaxed, we postpone this topic to a future work. Notice that inequality (3.25) is not required to prove the error estimate of Theorem 3.16 below.

The proof of Theorem 3.9 hinges on two technical results stated hereafter: a discrete Rellich–Kondrachov Lemma (cf. [40, Theorem 9.16]) and a proposition showing the approximation properties of the discrete symmetric gradient $\boldsymbol{G}_{\mathrm{s},h}^k$ defined by (3.19). Their proofs are given in Appendix 3.8.

Lemma 3.12 (Discrete compactness)**.** *Let the assumptions of Theorem 3.9 hold. Let $(\underline{\boldsymbol{v}}_h)_{h\in\mathcal{H}} \in (\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k)_{h\in\mathcal{H}}$, and assume that there is a real number $C \geq 0$ such that*

$$\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h} \leq C \qquad \forall h \in \mathcal{H}. \tag{3.26}$$

Then, for all q such that $1 \leq q < +\infty$ if $d = 2$ or $1 \leq q < 6$ if $d = 3$, the sequence $(\boldsymbol{v}_h)_{h\in\mathcal{H}} \in (\mathbb{P}_d^k(\mathcal{T}_h;\mathbb{R}^d))_{h\in\mathcal{H}}$ is relatively compact in $L^q(\Omega;\mathbb{R}^d)$. As a consequence, there is a function $\boldsymbol{v} \in L^q(\Omega;\mathbb{R}^d)$ such that as $h \to 0$, up to a subsequence, $\boldsymbol{v}_h \to \boldsymbol{v}$ strongly in $L^q(\Omega;\mathbb{R}^d)$.

Proof. See Appendix 3.8.2. □

Proposition 3.13 (Consistency of the discrete symmetric gradient operator)**.** *Let* $(\mathcal{T}_h)_{h\in\mathcal{H}}$ *be a regular mesh sequence, and let* $\boldsymbol{G}_{\mathrm{s},h}^k$ *be as in* (3.19) *with* $\boldsymbol{G}_{\mathrm{s},T}^k$ *defined by* (3.17) *for all* $T \in \mathcal{T}_h$.

1) Strong consistency. *For all* $\boldsymbol{v} \in H^1(\Omega;\mathbb{R}^d)$ *with* $\underline{I}_h^k$ *defined by* (3.15)*, it holds as* $h \to 0$

$$\boldsymbol{G}_{\mathrm{s},h}^k \underline{I}_h^k \boldsymbol{v} \to \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v} \text{ strongly in } L^2(\Omega;\mathbb{R}^{d\times d}). \tag{3.27}$$

2) Sequential consistency. *For all* $h \in \mathcal{H}$ *and all* $\boldsymbol{\tau} \in H^1(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})$*, denoting by* $\boldsymbol{\gamma_n}(\boldsymbol{\tau})$ *the normal trace of* $\boldsymbol{\tau}$ *on* Γ*, it holds*

$$\lim_{h\to 0}\left(\max_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_h^k,\,\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}=1}\left|\int_\Omega \boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{v}}_h : \boldsymbol{\tau} + \boldsymbol{v}_h\cdot(\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) - \int_\Gamma \boldsymbol{v}_{\Gamma,h}\cdot\boldsymbol{\gamma_n}(\boldsymbol{\tau})\right|\right) = 0, \tag{3.28}$$

Proof. See Appendix 3.8.3. □

We are now ready to prove convergence.

Proof of Theorem 3.9. The proof is subdivided into four steps: in **Step 1** we prove a uniform a priori bound on the solutions of the discrete problem (3.21); in **Step 2** we infer the existence of a limit for the sequence of discrete solutions and investigate its regularity; in **Step 3** we show that this limit solves the continuous problem (3.11); finally, in **Step 4** we prove strong convergence.

Step 1: ***A priori bound.*** We start by showing the following uniform a priori bound on the sequence of discrete solutions:

$$\|\underline{\boldsymbol{u}}_h\|_{\epsilon,h} \le C\|\boldsymbol{f}\|, \tag{3.29}$$

where the real number $C > 0$ only depends on Ω, $\underline{\sigma}$, γ, ϱ, and k. Making $\underline{\boldsymbol{v}}_h = \underline{\boldsymbol{u}}_h$ in (3.21) and using the coercivity property (3.2d) of $\boldsymbol{\sigma}$ in the left-hand side together with the Cauchy–Schwarz inequality in the right-hand side yields

$$\sum_{T\in\mathcal{T}_h}\left(\underline{\sigma}\|\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{u}}_T\|_T^2 + \sum_{F\in\mathcal{F}_h}\frac{\gamma}{h_F}\|\boldsymbol{\Delta}_{TF}^k\underline{\boldsymbol{u}}_T\|_F^2\right) \le \|\boldsymbol{f}\|\|\boldsymbol{u}_h\|.$$

Owing to the norm equivalence (3.24), and using the discrete Korn inequality (3.25) to estimate the right-hand side of the previous inequality, it is inferred that

$$\eta^{-1}\min(1,\underline{\sigma})\|\underline{\boldsymbol{u}}_h\|_{\epsilon,h}^2 \le \|\boldsymbol{f}\|\|\boldsymbol{u}_h\| \le C_{\mathrm{K}}\|\boldsymbol{f}\|\|\underline{\boldsymbol{u}}_h\|_{\epsilon,h}.$$

Dividing by $\|\underline{\boldsymbol{u}}_h\|_{\epsilon,h}$ yields (3.29) with $C = \eta\min(1,\underline{\sigma})^{-1}C_{\mathrm{K}}$.

Step 2: ***Existence of a limit and regularity.*** Let $1 \le q < +\infty$ if $d = 2$ or $1 \le q < 6$ if $d = 3$. Owing to the a priori bound (3.29) and the norm equivalence (3.24), the sequences $(\|\boldsymbol{u}_h\|_{\epsilon,h})_{h\in\mathcal{H}}$ and $(\|\boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_h\|)_{h\in\mathcal{H}}$ are uniformly bounded. Therefore, Lemma 3.12

and the Kakutani theorem [40, Theorem 3.17] yield the existence of $\boldsymbol{u} \in L^q(\Omega;\mathbb{R}^d)$ and $\boldsymbol{\mathcal{G}} \in L^2(\Omega;\mathbb{R}^{d\times d})$ such that as $h \to 0$, up to a subsequence,

$$\boldsymbol{u}_h \to \boldsymbol{u} \text{ strongly in } L^q(\Omega;\mathbb{R}^d) \text{ and } \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h \to \boldsymbol{\mathcal{G}} \text{ weakly in } L^2(\Omega;\mathbb{R}^{d\times d}).$$

This together with the fact that $\boldsymbol{u}_{h,\Gamma} = \boldsymbol{0}$ on Γ, shows that, for any $\boldsymbol{\tau} \in H^1(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})$,

$$\begin{aligned}
&\left|\int_\Omega \boldsymbol{\mathcal{G}} : \boldsymbol{\tau} + \boldsymbol{u}\cdot(\boldsymbol{\nabla}\cdot\boldsymbol{\tau})\right| \\
&= \lim_{h\to 0}\left|\int_\Omega \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h : \boldsymbol{\tau} + \boldsymbol{u}_h\cdot(\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) - \int_\Gamma \boldsymbol{u}_{h,\Gamma}\cdot\boldsymbol{\gamma}_{\boldsymbol{n}}(\boldsymbol{\tau})\right| \\
&\le \lim_{h\to 0}\left(\|\underline{\boldsymbol{u}}_h\|_{\epsilon,h}\max_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}^k_h,\,\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}=1}\left|\int_\Omega \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{v}}_h : \boldsymbol{\tau} + \boldsymbol{v}_h\cdot(\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) - \int_\Gamma \boldsymbol{v}_{h,\Gamma}\cdot\boldsymbol{\gamma}_{\boldsymbol{n}}(\boldsymbol{\tau})\right|\right) \\
&= 0.
\end{aligned} \tag{3.30}$$

To infer the previous limit we have used the uniform bound (3.29) on $\|\underline{\boldsymbol{u}}_h\|_{\epsilon,h}$ and the sequential consistency (3.28) of $\boldsymbol{G}^k_{\mathrm{s},h}$. Applying (3.30) with $\boldsymbol{\tau} \in C^\infty_c(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})$ leads to

$$\int_\Omega \boldsymbol{\mathcal{G}} : \boldsymbol{\tau} + \boldsymbol{u}\cdot(\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) = 0,$$

thus $\boldsymbol{\mathcal{G}} = \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}$ in the sense of distributions on Ω. As a result, owing to the isomorphism of Hilbert spaces between $H^1(\Omega;\mathbb{R}^d)$ and $\{\boldsymbol{v} \in L^2(\Omega;\mathbb{R}^d) \mid \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v} \in L^2(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})\}$ proved in [89, Theorem 3.1], we have $\boldsymbol{u} \in H^1(\Omega;\mathbb{R}^d)$. Using again (3.30) with $\boldsymbol{\tau} \in H^1(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})$ and integrating by parts, we obtain

$$\int_\Gamma \boldsymbol{\gamma}(\boldsymbol{u})\cdot\boldsymbol{\gamma}_{\boldsymbol{n}}(\boldsymbol{\tau}) = 0$$

with $\boldsymbol{\gamma}(\boldsymbol{u})$ denoting the trace of $\boldsymbol{u}$. As the set $\{\boldsymbol{\gamma}_{\boldsymbol{n}}(\boldsymbol{\tau}) : \boldsymbol{\tau} \in H^1(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})\}$ is dense in $L^2(\Gamma;\mathbb{R}^d)$, we deduce that $\boldsymbol{\gamma}(\boldsymbol{u}) = \boldsymbol{0}$ on Γ. In conclusion, with convergences up to a subsequence,

$$\begin{gathered}
\boldsymbol{u} \in H^1_0(\Omega;\mathbb{R}^d),\ \boldsymbol{u}_h \to \boldsymbol{u} \text{ strongly in } L^q(\Omega;\mathbb{R}^d) \\
\boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h \to \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} \text{ weakly in } L^2(\Omega;\mathbb{R}^{d\times d}).
\end{gathered}$$

Step 3: ***Identification of the limit.*** Let us now prove that $\boldsymbol{u}$ is a solution to (3.11). The growth property (3.2c) on $\boldsymbol{\sigma}$ and the bound on $(\|\boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h\|)_{h\in\mathcal{H}}$ ensure that the sequence $(\boldsymbol{\sigma}(\cdot,\boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h))_{h\in\mathcal{H}}$ is bounded in $L^2(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})$. Hence, there exists $\boldsymbol{\eta} \in L^2(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})$ such that, up to a subsequence as $h \to 0$,

$$\boldsymbol{\sigma}(\cdot,\boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h) \to \boldsymbol{\eta} \quad \text{weakly in } L^2(\Omega;\mathbb{R}^{d\times d}). \tag{3.31}$$

Plugging into (3.21) $\underline{\boldsymbol{v}}_h = \underline{I}^k_h\boldsymbol{\phi}$, with $\boldsymbol{\phi} \in C^\infty_c(\Omega;\mathbb{R}^d)$, gives

$$\int_\Omega \boldsymbol{\sigma}(\cdot,\boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h) : \boldsymbol{G}^k_{\mathrm{s},h}\underline{I}^k_h\boldsymbol{\phi} = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{\pi}^k_h\boldsymbol{\phi} - s_h(\underline{\boldsymbol{u}}_h,\underline{I}^k_h\boldsymbol{\phi}), \tag{3.32}$$

with $\boldsymbol{\pi}_h^k$ denoting the L^2-projector on the broken polynomial spaces $\mathbb{P}_d^k(\mathcal{T}_h;\mathbb{R}^d)$ and s_h defined by (3.22). Using the Cauchy–Schwarz inequality followed by the norm equivalence (3.24) to bound the first factor, we infer

$$|s_h(\underline{\boldsymbol{u}}_h, \underline{I}_h^k\boldsymbol{\phi})| \le s_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{u}}_h)^{1/2} s_h(\underline{I}_h^k\boldsymbol{\phi}, \underline{I}_h^k\boldsymbol{\phi})^{1/2} \le \|\underline{\boldsymbol{u}}_h\|_{\epsilon,h} s_h(\underline{I}_h^k\boldsymbol{\phi}, \underline{I}_h^k\boldsymbol{\phi})^{1/2}. \tag{3.33}$$

It was proved in [74, Eq. (35)] using the optimal approximation properties of $\boldsymbol{r}_T^{k+1}\underline{I}_T^k$ that it holds for all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$, all $\boldsymbol{v} \in H^{k+2}(T;\mathbb{R}^d)$, and all $F \in \mathcal{F}_T$ that

$$h_F^{-1/2}\|\boldsymbol{\Delta}_{TF}^k \underline{I}_T^k \boldsymbol{v}\|_F \lesssim h_T^{k+1}\|\boldsymbol{v}\|_{H^{k+2}(T;\mathbb{R}^d)}, \tag{3.34}$$

with $\boldsymbol{\Delta}_{TF}^k$ defined by (3.23). As a consequence, recalling the definition (3.22) of s_h, we have the following convergence result:

$$\forall \boldsymbol{v} \in H^1(\Omega;\mathbb{R}^d) \cap H^2(\mathcal{T}_h;\mathbb{R}^d), \quad \lim_{h\to 0} s_h(\underline{I}_h^k\boldsymbol{v}, \underline{I}_h^k\boldsymbol{v}) = 0. \tag{3.35}$$

Recalling the a priori bound (3.29) on the discrete solution and the convergence property (3.35), it follows from (3.33) that $|s_h(\underline{\boldsymbol{u}}_h, \underline{I}_h^k\boldsymbol{\phi})| \to 0$ as $h \to 0$. Additionally, by the approximation property (3.13a) of the L^2-projector, one has $\boldsymbol{\pi}_h^k\boldsymbol{\phi} \to \boldsymbol{\phi}$ strongly in $L^2(\Omega;\mathbb{R}^d)$ and, by virtue of Proposition 3.13, that $\boldsymbol{G}_{\mathrm{s},h}^k\underline{I}_h^k\boldsymbol{\phi} \to \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{\phi}$ strongly in $L^2(\Omega;\mathbb{R}^{d\times d})$. Thus, we can pass to the limit $h \to 0$ in (3.32) and obtain

$$\int_\Omega \boldsymbol{\eta} : \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{\phi} = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{\phi}. \tag{3.36}$$

By density of $C_c^\infty(\Omega;\mathbb{R}^d)$ in $H_0^1(\Omega;\mathbb{R}^d)$, this relation still holds if $\boldsymbol{\phi} \in H_0^1(\Omega;\mathbb{R}^d)$. On the other hand, plugging $\underline{\boldsymbol{v}}_h = \underline{\boldsymbol{u}}_h$ into (3.21) and using the fact that $s_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{u}}_h) \ge 0$, we obtain

$$\mathfrak{T}_h \coloneqq \int_\Omega \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_h) : \boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_h \le \int_\Omega \boldsymbol{f}\cdot\boldsymbol{u}_h.$$

Thus, using the previous bound, the strong convergence $\boldsymbol{u}_h \to \boldsymbol{u}$, and (3.36), it is inferred that

$$\lim_{h\to 0}\mathfrak{T}_h \le \int_\Omega \boldsymbol{f}\cdot\boldsymbol{u} = \int_\Omega \boldsymbol{\eta} : \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}. \tag{3.37}$$

We now use the monotonicity assumption on $\boldsymbol{\sigma}$ and the Minty trick [152] to prove that $\boldsymbol{\eta} = \boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})$. Let $\boldsymbol{\Lambda} \in L^2(\Omega;\mathbb{R}^{d\times d})$ and write, using the monotonicity (3.2e) of $\boldsymbol{\sigma}$, the convergence (3.31) of $\boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_h)$, and the bound (3.37),

$$0 \le \lim_{h\to 0}\left(\int_\Omega (\boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_h) - \boldsymbol{\sigma}(\cdot,\boldsymbol{\Lambda})) : (\boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_h - \boldsymbol{\Lambda})\right) \le \int_\Omega (\boldsymbol{\eta} - \boldsymbol{\sigma}(\cdot,\boldsymbol{\Lambda})) : (\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} - \boldsymbol{\Lambda}). \tag{3.38}$$

Applying the previous relation with $\boldsymbol{\Lambda} = \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} \pm t\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}$, for $t > 0$ and $\boldsymbol{v} \in H_0^1(\Omega;\mathbb{R}^d)$, and dividing by t, leads to

$$0 \le \pm \int_\Omega (\boldsymbol{\eta} - \boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} \mp t\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v})) : \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}.$$

Owing to the growth property (3.2c) and the Caratheodory property (3.2a) of $\boldsymbol{\sigma}$, we can let $t \to 0$ and pass the limit inside the integral and then inside the argument of $\boldsymbol{\sigma}$. In conclusion, for all $\boldsymbol{v} \in H^1_0(\Omega;\mathbb{R}^d)$, we infer

$$\int_\Omega \boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_s \boldsymbol{u}) : \boldsymbol{\nabla}_s \boldsymbol{v} = \int_\Omega \boldsymbol{\eta} : \boldsymbol{\nabla}_s \boldsymbol{v} = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v},$$

where we have used (3.36) with $\boldsymbol{\phi} = \boldsymbol{v}$ in order to obtain the second equality. The above equation shows that $\boldsymbol{\eta} = \boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_s \boldsymbol{u})$ and that $\boldsymbol{u}$ solves (3.11).

Step 4: ***Strong convergence.*** We prove that if $\boldsymbol{\sigma}$ is strictly monotone then $\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h \to \boldsymbol{\nabla}_s \boldsymbol{u}$ strongly in $L^2(\Omega;\mathbb{R}^{d\times d})$. We define the function $\mathcal{D}_h : \Omega \to \mathbb{R}$ such that

$$\mathcal{D}_h \coloneqq (\boldsymbol{\sigma}(\cdot, \boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h) - \boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_s \boldsymbol{u})) : (\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h - \boldsymbol{\nabla}_s \boldsymbol{u}).$$

For all $h \in \mathcal{H}$, the function $\mathcal{D}_h$ is non-negative as a result of the monotonicity property (3.2e) and, by (3.38) with $\boldsymbol{\Lambda} = \boldsymbol{\nabla}_s \boldsymbol{u}$, it is inferred that $\lim_{h\to 0} \int_\Omega \mathcal{D}_h = 0$. Hence, $(\mathcal{D}_h)_{h\in\mathcal{H}}$ converges to 0 in $L^1(\Omega)$ and, therefore, also almost everywhere on Ω up to a subsequence. Let us take $\overline{\boldsymbol{x}} \in \Omega$ such that the above mentioned convergence hold at $\overline{\boldsymbol{x}}$. Developing the products in $\mathcal{D}_h$ and using the coercivity and growth properties (3.2d) and (3.2c) of $\boldsymbol{\sigma}$, one has

$$\mathcal{D}_h(\overline{\boldsymbol{x}}) \geq \underline{\sigma}\|\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h(\overline{\boldsymbol{x}})\|^2_{d\times d} - 2\overline{\sigma}\|\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h(\overline{\boldsymbol{x}})\|_{d\times d}\|\boldsymbol{\nabla}_s \boldsymbol{u}(\overline{\boldsymbol{x}})\|_{d\times d} + \underline{\sigma}\|\boldsymbol{\nabla}_s \boldsymbol{u}(\overline{\boldsymbol{x}})\|^2_{d\times d}.$$

Since the right hand side is quadratic in $\|\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h(\overline{\boldsymbol{x}})\|_{d\times d}$ and $(\mathcal{D}_h(\overline{\boldsymbol{x}}))_{h\in\mathcal{H}}$ is bounded, we deduce that also $(\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h(\overline{\boldsymbol{x}}))_{h\in\mathcal{H}}$ is bounded. Passing to the limit in the definition of $\mathcal{D}_h(\overline{\boldsymbol{x}})$ yields

$$(\boldsymbol{\sigma}(\overline{\boldsymbol{x}}, \boldsymbol{L}_{\overline{\boldsymbol{x}}}) - \boldsymbol{\sigma}(\overline{\boldsymbol{x}}, \boldsymbol{\nabla}_s \boldsymbol{u}(\overline{\boldsymbol{x}}))) : (\boldsymbol{L}_{\overline{\boldsymbol{x}}} - \boldsymbol{\nabla}_s \boldsymbol{u}(\overline{\boldsymbol{x}})) = 0,$$

where $\boldsymbol{L}_{\overline{\boldsymbol{x}}}$ is an adherence value of $(\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h(\overline{\boldsymbol{x}}))_{h\in\mathcal{H}}$. The strict monotonicity assumption forces $\boldsymbol{L}_{\overline{\boldsymbol{x}}} = \boldsymbol{\nabla}_s \boldsymbol{u}(\overline{\boldsymbol{x}})$ to be the unique adherence value of $(\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h(\overline{\boldsymbol{x}}))_{h\in\mathcal{H}}$, and therefore the sequence converges to this value. As a result,

$$\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h \to \boldsymbol{\nabla}_s \boldsymbol{u} \text{ a.e. on } \Omega. \tag{3.39}$$

Using (3.37) together with Fatou's Lemma, we see that

$$\lim_{h\to 0} \int_\Omega \boldsymbol{\sigma}(\cdot, \boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h) : \boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h = \int_\Omega \boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_s \boldsymbol{u}) : \boldsymbol{\nabla}_s \boldsymbol{u}.$$

Moreover, owing to (3.39), $(\boldsymbol{\sigma}(\cdot, \boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h) : \boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h)_{h\in\mathcal{H}}$ is a non-negative sequence converging almost everywhere on Ω. Using [84, Lemma 8.4] we see that this sequence also converges in $L^1(\Omega)$ and, therefore, it is equi-integrable in $L^1(\Omega)$. Thus, the coercivity (3.2d) of $\boldsymbol{\sigma}$ ensures that $(\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h)_{h\in\mathcal{H}}$ is equi-integrable in $L^2(\Omega;\mathbb{R}^{d\times d})$ and Vitali's theorem shows that

$$\boldsymbol{G}^k_{s,h}\underline{\boldsymbol{u}}_h \to \boldsymbol{\nabla}_s \boldsymbol{u} \text{ strongly in } L^2(\Omega;\mathbb{R}^{d\times d}).$$

This concludes the proof. □

3.5.3 Error estimate

We next estimate the order of convergence of the method for solutions that display further regularity beyond that required by the variational formulation. We stipulate the following additional assumptions on the stress-strain function $\boldsymbol{\sigma}$.

Assumption 3.14 (Stress-strain relation II)**.** There exist real numbers $\sigma^*, \sigma_* \in (0, +\infty)$ such that, for a.e. $\boldsymbol{x} \in \Omega$, and all $\boldsymbol{\tau}, \boldsymbol{\eta} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$,

$$\|\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\eta})\|_{d\times d} \le \sigma^* \|\boldsymbol{\tau} - \boldsymbol{\eta}\|_{d\times d}, \qquad \text{(Lipschitz continuity)} \tag{3.40a}$$

$$(\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\eta})) : (\boldsymbol{\tau} - \boldsymbol{\eta}) \ge \sigma_* \|\boldsymbol{\tau} - \boldsymbol{\eta}\|^2_{d\times d}. \qquad \text{(strong monotonicity)} \tag{3.40b}$$

Remark 3.15 (Lipschitz continuity and strong monotonocity). It has been proved in [13, Lemma 4.1] that, under the assumptions (3.9), the stress-strain tensor function for the Hencky–Mises model is strongly monotone and Lipschitz-continuous, namely Assumption 3.14 holds. Also the isotropic damage model satisfies Assumption 3.14 if the damage function in (3.5) is, for instance, such that $D(|\boldsymbol{\tau}|) = 1 - (1 + |\boldsymbol{\tau}|)^{-\frac{1}{2}}$.

Theorem 3.16 (Error estimate)**.** *Let Assumptions 3.1 and 3.14 hold, and let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a regular mesh sequence. Let $\boldsymbol{u}$ be the unique solution to (3.1). Let a polynomial degree $k \ge 1$ be fixed, and, for all $h \in \mathcal{H}$, let $\underline{\boldsymbol{u}}_h$ be the unique solution to (3.21) on the mesh $\mathcal{T}_h$. Then, under the additional regularity $\boldsymbol{u} \in H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)$ and $\boldsymbol{\sigma}(\cdot,\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) \in H^{k+1}(\mathcal{T}_h;\mathbb{R}^{d\times d})$, it holds*

$$\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} - \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h\| + s_h(\underline{\boldsymbol{u}}_h,\underline{\boldsymbol{u}}_h)^{\frac{1}{2}} \le Ch^{k+1}\left(\|\boldsymbol{u}\|_{H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)} + \|\boldsymbol{\sigma}(\cdot,\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})\|_{H^{k+1}(\mathcal{T}_h;\mathbb{R}^{d\times d})}\right), \tag{3.41}$$

where C is a positive constant depending only on Ω, k, the mesh regularity parameter ϱ, the real numbers $\overline{\sigma}$, $\underline{\sigma}$, σ^*, σ_* appearing in (3.2) and in (3.40), and an upper bound of $\|\boldsymbol{f}\|$.

Remark 3.17 (Locking-free error estimate). The proposed scheme, although different from the one of [74], is robust in the quasi-incompressible limit. The reason is that, as a result of the commuting property (3.18), we have $\pi^k_T(\boldsymbol{\nabla}\cdot\boldsymbol{v}) = \mathrm{tr}(\boldsymbol{G}^k_{\mathrm{s},T}\underline{\boldsymbol{I}}^k_T\boldsymbol{v})$. Thus, considering, e.g., the linear elasticity stress-strain relation (3.3), we can proceed as in [74, Theorem 8] in order to prove that, when $\boldsymbol{u} \in H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)$ and $\boldsymbol{\nabla}\cdot\boldsymbol{u} \in H^{k+1}(\mathcal{T}_h;\mathbb{R})$, and choosing $\gamma = 2\mu$, it holds

$$(2\mu)^{1/2}\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} - \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h\| \le Ch^{k+1}\left(2\mu\|\boldsymbol{u}\|_{H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)} + \lambda\|\boldsymbol{\nabla}\cdot\boldsymbol{u}\|_{H^{k+1}(\mathcal{T}_h;\mathbb{R})}\right),$$

with real number $C > 0$ independent of h, μ and λ. The previous bound leads to a locking-free estimate; see [74, Remark 9]. Note that the locking-free nature of polyhedral element methods has also been observed in [197] for the Weak Galerkin method and in [20] for the Virtual Element method.

Proof of Theorem 3.16. For the sake of conciseness, throughout the proof we let $\widehat{\underline{\boldsymbol{u}}}_h \coloneqq \underline{\boldsymbol{I}}^k_h\boldsymbol{u}$ and use the following abridged notations for the constraint field and its approximations:

$$\boldsymbol{\varsigma} \coloneqq \boldsymbol{\sigma}(\cdot,\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) \text{ and, for all } T \in \mathcal{T}_h,\ \boldsymbol{\varsigma}_T \coloneqq \boldsymbol{\sigma}(\cdot,\boldsymbol{G}^k_{\mathrm{s},T}\underline{\boldsymbol{u}}_T) \text{ and } \widehat{\boldsymbol{\varsigma}}_T \coloneqq \boldsymbol{\sigma}(\cdot,\boldsymbol{G}^k_{\mathrm{s},T}\widehat{\underline{\boldsymbol{u}}}_T).$$

First we want to show that (3.41) holds assuming that

$$\|\underline{\boldsymbol{u}}_h - \widehat{\underline{\boldsymbol{u}}}_h\|_{\epsilon,h} \lesssim h^{k+1}\left(\|\boldsymbol{u}\|_{H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)} + \|\boldsymbol{\varsigma}\|_{H^{k+1}(\mathcal{T}_h;\mathbb{R}^{d\times d})}\right). \tag{3.42}$$

Using the triangle inequality, we obtain

$$\begin{aligned}\|\boldsymbol{G}_{\mathrm{s},h}^k\underline{\boldsymbol{u}}_h - \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}\| + s_h(\underline{\boldsymbol{u}}_h,\underline{\boldsymbol{u}}_h)^{1/2} \le{}& \|\boldsymbol{G}_{\mathrm{s},h}^k(\underline{\boldsymbol{u}}_h - \widehat{\underline{\boldsymbol{u}}}_h)\| + s_h(\underline{\boldsymbol{u}}_h - \widehat{\underline{\boldsymbol{u}}}_h,\underline{\boldsymbol{u}}_h - \widehat{\underline{\boldsymbol{u}}}_h)^{1/2}\\ &+ \|\boldsymbol{G}_{\mathrm{s},h}^k\widehat{\underline{\boldsymbol{u}}}_h - \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}\| + s_h(\widehat{\underline{\boldsymbol{u}}}_h,\widehat{\underline{\boldsymbol{u}}}_h)^{1/2}.\end{aligned} \tag{3.43}$$

Using the norm equivalence (3.24) followed by (3.42) we obtain for the terms in the first line of (3.43)

$$\|\boldsymbol{G}_{\mathrm{s},h}^k(\underline{\boldsymbol{u}}_h - \widehat{\underline{\boldsymbol{u}}}_h)\| + s_h(\underline{\boldsymbol{u}}_h - \widehat{\underline{\boldsymbol{u}}}_h,\underline{\boldsymbol{u}}_h - \widehat{\underline{\boldsymbol{u}}}_h)^{1/2} \lesssim h^{k+1}\left(\|\boldsymbol{u}\|_{H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)} + \|\boldsymbol{\varsigma}\|_{H^{k+1}(\mathcal{T}_h;\mathbb{R}^{d\times d})}\right).$$

For the terms in the second line, using the approximation properties of $\boldsymbol{G}_{\mathrm{s},h}^k$ resulting from (3.18) together with (3.13a) for the first addend and (3.34) for the second, we get

$$\|\boldsymbol{G}_{\mathrm{s},h}^k\widehat{\underline{\boldsymbol{u}}}_h - \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}\| + s_h(\widehat{\underline{\boldsymbol{u}}}_h,\widehat{\underline{\boldsymbol{u}}}_h)^{1/2} \lesssim h^{k+1}\|\boldsymbol{u}\|_{H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)}.$$

It only remains to prove (3.42), which we do in two steps: in **Step 1** we prove a basic estimate in terms of a conformity error, which is then bounded in **Step 2**.

Step 1: ***Basic error estimate.*** Using for all $T\in\mathcal{T}_h$ the strong monotonicity (3.40b) with $\boldsymbol{\tau} = \boldsymbol{G}_{\mathrm{s},T}^k\widehat{\underline{\boldsymbol{u}}}_T$ and $\boldsymbol{\eta} = \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{u}}_T$, we infer

$$\|\boldsymbol{G}_{\mathrm{s},h}^k(\widehat{\underline{\boldsymbol{u}}}_h - \underline{\boldsymbol{u}}_h)\|^2 \lesssim \sum_{T\in\mathcal{T}_h}\int_T(\widehat{\boldsymbol{\varsigma}}_T - \boldsymbol{\varsigma}_T) : \boldsymbol{G}_{\mathrm{s},T}^k(\widehat{\underline{\boldsymbol{u}}}_T - \underline{\boldsymbol{u}}_T).$$

Owing to the norm equivalence (3.24) and the previous bound, we get

$$\begin{aligned}\|\widehat{\underline{\boldsymbol{u}}}_h - \underline{\boldsymbol{u}}_h\|_{\epsilon,h}^2 &\lesssim \sum_{T\in\mathcal{T}_h}\int_T(\widehat{\boldsymbol{\varsigma}}_T - \boldsymbol{\varsigma}_T) : \boldsymbol{G}_{\mathrm{s},T}^k(\widehat{\underline{\boldsymbol{u}}}_T - \underline{\boldsymbol{u}}_T) + s_h(\widehat{\underline{\boldsymbol{u}}}_h - \underline{\boldsymbol{u}}_h,\widehat{\underline{\boldsymbol{u}}}_h - \underline{\boldsymbol{u}}_h)\\ &= a_h(\widehat{\underline{\boldsymbol{u}}}_h,\widehat{\underline{\boldsymbol{u}}}_h - \underline{\boldsymbol{u}}_h) - \int_\Omega \boldsymbol{f}\cdot(\widehat{\underline{\boldsymbol{u}}}_h - \underline{\boldsymbol{u}}_h).\end{aligned}$$

where we have used the discrete problem (3.21) to conclude. Hence, dividing by $\|\widehat{\underline{\boldsymbol{u}}}_h - \underline{\boldsymbol{u}}_h\|_{\epsilon,h}$ and passing to the supremum in the right-hand side, we arrive at the following error estimate:

$$\|\widehat{\underline{\boldsymbol{u}}}_h - \underline{\boldsymbol{u}}_h\|_{\epsilon,h} \lesssim \sup_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k,\,\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}=1}\mathcal{E}_h(\underline{\boldsymbol{v}}_h), \tag{3.44}$$

with conformity error such that, for all $\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$,

$$\mathcal{E}_h(\underline{\boldsymbol{v}}_h) \coloneqq \sum_{T\in\mathcal{T}_h}\int_T\widehat{\boldsymbol{\varsigma}}_T : \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T - \int_\Omega\boldsymbol{f}\cdot\boldsymbol{v}_h + s_h(\widehat{\underline{\boldsymbol{u}}}_h,\underline{\boldsymbol{v}}_h). \tag{3.45}$$

Step 2: ***Bound of the conformity error.*** We bound the quantity $\mathcal{E}_h(\underline{\boldsymbol{v}}_h)$ defined above for a generic $\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$. Denote by $\mathfrak{T}_1$, $\mathfrak{T}_2$, and $\mathfrak{T}_3$ the three addends in the right-hand side of (3.45).

Using for all $T \in \mathcal{T}_h$ the definition (3.17) of $\boldsymbol{G}_{\mathrm{s},T}^k$ with $\boldsymbol{\tau} = \boldsymbol{\pi}_T^k \widehat{\boldsymbol{\varsigma}}_T$, we have that

$$\mathfrak{T}_1 = \sum_{T\in\mathcal{T}_h} \left(\int_T \widehat{\boldsymbol{\varsigma}}_T : \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T + \sum_{F\in\mathcal{F}_T} \int_F \boldsymbol{\pi}_T^k \widehat{\boldsymbol{\varsigma}}_T \boldsymbol{n}_{TF} \cdot (\boldsymbol{v}_F - \boldsymbol{v}_T) \right), \tag{3.46}$$

where we have used the fact that $\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T \in \mathbb{P}_d^{k-1}(T;\mathbb{R}^{d\times d})$ together with the definition (3.12) of the orthogonal projector to cancel $\boldsymbol{\pi}_T^k$ in the first term.

On the other hand, using the fact that $\boldsymbol{f} = -\boldsymbol{\nabla}\cdot\boldsymbol{\varsigma}$ a.e. in Ω and integrating by parts element by element, we get that

$$\mathfrak{T}_2 = -\sum_{T\in\mathcal{T}_h} \left(\int_T \boldsymbol{\varsigma} : \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T + \sum_{F\in\mathcal{F}_T} \int_F \boldsymbol{\varsigma} \boldsymbol{n}_{TF} \cdot (\boldsymbol{v}_F - \boldsymbol{v}_T) \right), \tag{3.47}$$

where we have additionally used that $\boldsymbol{\varsigma}_{|T_1} \boldsymbol{n}_{T_1F} + \boldsymbol{\varsigma}_{|T_2} \boldsymbol{n}_{T_2F} = \boldsymbol{0}$ for all interfaces $F \subset \partial T_1 \cap \partial T_2$ and that $\boldsymbol{v}_F$ vanishes on Γ (cf. (3.20)) to insert $\boldsymbol{v}_F$ into the second term.

Summing (3.46) and (3.47), taking absolute values, and using the Cauchy–Schwarz inequality to bound the right-hand side, we infer that

$$|\mathfrak{T}_1 + \mathfrak{T}_2| \le \left(\sum_{T\in\mathcal{T}_h} \left(\|\boldsymbol{\varsigma} - \widehat{\boldsymbol{\varsigma}}_T\|_T^2 + h_T \|\boldsymbol{\varsigma} - \boldsymbol{\pi}_T^k \widehat{\boldsymbol{\varsigma}}_T\|_{\partial T}^2 \right) \right)^{1/2} \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}. \tag{3.48}$$

It only remains to bound the first factor. Let a mesh element $T \in \mathcal{T}_h$ be fixed. Using the Lipschitz continuity (3.40a) with $\boldsymbol{\tau} = \boldsymbol{G}_{\mathrm{s},T}^k \widehat{\underline{\boldsymbol{u}}}_T$ and $\boldsymbol{\eta} = \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{u}$ and the optimal approximation properties of $\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{I}}_T^k$ resulting from (3.18) together with (3.13a) with $m = 1$ and $s = k+2$, leads to

$$\|\boldsymbol{\varsigma} - \widehat{\boldsymbol{\varsigma}}_T\|_T \lesssim \|\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{u} - \boldsymbol{G}_{\mathrm{s},T}^k \widehat{\underline{\boldsymbol{u}}}_T\|_T \lesssim h^{k+1} \|\boldsymbol{u}\|_{H^{k+2}(T;\mathbb{R}^d)}, \tag{3.49}$$

which provides an estimate for the first term inside the summation in the right-hand side of (3.48). To estimate the second term, we use the triangle inequality, the discrete trace inequality of [76, Lemma 1.46], and the boundedness of $\boldsymbol{\pi}_T^k$ to write

$$h_T^{1/2} \|\boldsymbol{\varsigma} - \boldsymbol{\pi}_T^k \widehat{\boldsymbol{\varsigma}}_T\|_{\partial T} \lesssim \|\boldsymbol{\pi}_T^k (\boldsymbol{\varsigma} - \widehat{\boldsymbol{\varsigma}}_T)\|_T + h_T^{1/2} \|\boldsymbol{\varsigma} - \boldsymbol{\pi}_T^k \boldsymbol{\varsigma}\|_{\partial T} \le \|\boldsymbol{\varsigma} - \widehat{\boldsymbol{\varsigma}}_T\|_T + h_T^{1/2} \|\boldsymbol{\varsigma} - \boldsymbol{\pi}_T^k \boldsymbol{\varsigma}\|_{\partial T}.$$

The first term in the right-hand side is bounded by (3.49). For the second, using the approximation properties (3.13b) of $\boldsymbol{\pi}_T^k$ with $m = 0$ and $s = k+1$, we get $h_T^{1/2} \|\boldsymbol{\varsigma} - \boldsymbol{\pi}_T^k \boldsymbol{\varsigma}\|_{\partial T} \lesssim h^{k+1} \|\boldsymbol{\varsigma}\|_{H^{k+1}(T;\mathbb{R}^{d\times d})}$ so that, in conclusion,

$$h_T^{1/2} \|\boldsymbol{\varsigma} - \boldsymbol{\pi}_T^k \widehat{\boldsymbol{\varsigma}}_T\|_{\partial T} \lesssim h^{k+1} \left(\|\boldsymbol{u}\|_{H^{k+2}(T;\mathbb{R}^d)} + \|\boldsymbol{\varsigma}\|_{H^{k+1}(T;\mathbb{R}^{d\times d})} \right). \tag{3.50}$$

Plugging the estimates (3.49) and (3.50) into (3.48) finally yields

$$|\mathfrak{T}_1 + \mathfrak{T}_2| \lesssim h^{k+1} \left(\|\boldsymbol{u}\|_{H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)} + \|\boldsymbol{\varsigma}\|_{H^{k+1}(\mathcal{T}_h;\mathbb{R}^{d\times d})} \right) \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}. \tag{3.51}$$

It only remains to bound $\mathfrak{T}_3 = s_h(\widehat{\underline{\boldsymbol{u}}}_h, \underline{\boldsymbol{v}}_h)$. Using the Cauchy–Schwarz inequality, the definition (3.22) of s_h, the approximation property (3.34) of $\boldsymbol{\Delta}^k_{TF}$, and the norm equivalence (3.24), we infer

$$|\mathfrak{T}_3| \lesssim \left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\frac{\gamma}{h_F}\|\boldsymbol{\Delta}^k_{TF}\widehat{\underline{\boldsymbol{u}}}_T\|_F^2\right)^{1/2} s_h(\underline{\boldsymbol{v}}_h,\underline{\boldsymbol{v}}_h)^{1/2} \lesssim h^{k+1}\|\boldsymbol{u}\|_{H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)}\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}. \tag{3.52}$$

Using (3.51) and (3.52), we finally get that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}^k_{h,\mathrm{D}}$,

$$\mathcal{E}_h(\underline{\boldsymbol{v}}_h) \lesssim h^{k+1}\left(\|\boldsymbol{u}\|_{H^{k+2}(\mathcal{T}_h;\mathbb{R}^d)} + \|\boldsymbol{\varsigma}\|_{H^{k+1}(\mathcal{T}_h;\mathbb{R}^{d\times d})}\right)\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}. \tag{3.53}$$

Thus, using (3.53) to bound the right-hand side of (3.44), (3.42) follows. □

3.6 Local principle of virtual work and law of action and reaction

We show in this section that the solution of the discrete problem (3.21) satisfies inside each element a local principle of virtual work with numerical tractions that obey the law of action and reaction. This property is important from both the mathematical and engineering points of view, and it can simplify the derivation of a posteriori error estimators based on equilibrated tractions; see, e.g., [3, 158]. It is worth emphasizing that local equilibrium properties on the primal mesh are a distinguishing feature of hybrid (face-based) methods: the derivation of similar properties for vertex-based methods usually requires to perform reconstructions on a dual mesh.

Define, for all $T \in \mathcal{T}_h$, the space

$$\underline{\boldsymbol{D}}^k_{\partial T} \coloneqq \bigtimes_{F\in\mathcal{F}_T}\mathbb{P}^k_d(F;\mathbb{R}^d),$$

as well as the boundary difference operator $\underline{\boldsymbol{\delta}}^k_{\partial T} : \underline{\boldsymbol{U}}^k_T \to \underline{\boldsymbol{D}}^k_{\partial T}$ such that, for all $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}^k_T$,

$$\underline{\boldsymbol{\delta}}^k_{\partial T}\underline{\boldsymbol{v}}_T = (\boldsymbol{\delta}^k_F\underline{\boldsymbol{v}}_T)_{F\in\mathcal{F}_T} \coloneqq (\boldsymbol{v}_F - \boldsymbol{v}_T)_{F\in\mathcal{F}_T}.$$

The following proposition shows that the stabilization can be reformulated in terms of boundary differences.

Proposition 3.18 (Reformulation of the local stabilization bilinear form)**.** *For all mesh elements $T \in \mathcal{T}_h$, the local stabilization bilinear form s_T defined by (3.22) satisfies, for all $\underline{\boldsymbol{u}}_T, \underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}^k_T$,*

$$s_T(\underline{\boldsymbol{u}}_T,\underline{\boldsymbol{v}}_T) = s_T((\boldsymbol{0},\underline{\boldsymbol{\delta}}^k_{\partial T}\underline{\boldsymbol{u}}_T),(\boldsymbol{0},\underline{\boldsymbol{\delta}}^k_{\partial T}\underline{\boldsymbol{v}}_T)). \tag{3.54}$$

Proof. Let a mesh element $T \in \mathcal{T}_h$ be fixed. Using the fact that $\boldsymbol{r}^{k+1}_T\underline{\boldsymbol{I}}^k_T\boldsymbol{v}_T = \boldsymbol{v}_T$ for all $\boldsymbol{v}_T \in \mathbb{P}^k_d(T)^d$ (this because $\boldsymbol{r}^{k+1}_T\underline{\boldsymbol{I}}^k_T$ is a projector on $\mathbb{P}^{k+1}_d(T;\mathbb{R}^d)$, cf. [74, Eq. (20)]) together

with the linearity of $\boldsymbol{r}_T^{k+1}$, it is inferred that, for all $F \in \mathcal{F}_T$, the face-based residual defined by (3.23) satisfies

$$\boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{v}}_T = \boldsymbol{\pi}_F^k(\boldsymbol{r}_T^{k+1}(\boldsymbol{0}, \underline{\boldsymbol{\delta}}_{\partial T}^k \underline{\boldsymbol{v}}_T) - \boldsymbol{\delta}_F^k \underline{\boldsymbol{v}}_T) - \boldsymbol{\pi}_T^k \boldsymbol{r}_T^{k+1}(\boldsymbol{0}, \underline{\boldsymbol{\delta}}_{\partial T}^k \underline{\boldsymbol{v}}_T) = \boldsymbol{\Delta}_{TF}^k(\boldsymbol{0}, \underline{\boldsymbol{\delta}}_{\partial T}^k \underline{\boldsymbol{v}}_T)$$

for all $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}_T^k$. Plugging this expression into (3.22) yields the assertion. □

Define now the boundary residual operator $\underline{\boldsymbol{R}}_{\partial T}^k : \underline{\boldsymbol{U}}_T^k \to \underline{\boldsymbol{D}}_{\partial T}^k$ such that, for all $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}_T^k$, $\underline{\boldsymbol{R}}_{\partial T}^k \underline{\boldsymbol{v}}_T := (\boldsymbol{R}_{TF}^k \underline{\boldsymbol{v}}_T)_{F \in \mathcal{F}_T}$ satisfies

$$-\sum_{F \in \mathcal{F}_T} \int_F \boldsymbol{R}_{TF}^k \underline{\boldsymbol{v}}_T \cdot \boldsymbol{\alpha}_F = s_T((\boldsymbol{0}, \underline{\boldsymbol{\delta}}_{\partial T}^k \underline{\boldsymbol{v}}_T), (\boldsymbol{0}, \underline{\boldsymbol{\alpha}}_{\partial T})) \qquad \forall \underline{\boldsymbol{\alpha}}_{\partial T} \in \underline{\boldsymbol{D}}_{\partial T}^k. \tag{3.55}$$

Problem (3.55) is well-posed, and computing $\boldsymbol{R}_{TF}^k \underline{\boldsymbol{v}}_T$ requires to invert the boundary mass matrix.

Lemma 3.19 (Local principle of virtual work and law of action and reaction)**.** *Denote by $\underline{\boldsymbol{u}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ a solution of problem (3.21) and, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, define the numerical traction*

$$\boldsymbol{T}_{TF}(\underline{\boldsymbol{u}}_T) := -\boldsymbol{\pi}_T^k \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{u}}_T) \boldsymbol{n}_{TF} + \boldsymbol{R}_{TF}^k \underline{\boldsymbol{u}}_T.$$

Then, for all $T \in \mathcal{T}_h$ we have the following discrete principle of virtual work: For all $\boldsymbol{v}_T \in \mathbb{P}_d^k(T; \mathbb{R}^d)$,

$$\int_T \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{u}}_T) : \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T + \sum_{F \in \mathcal{F}_T} \int_F \boldsymbol{T}_{TF}(\underline{\boldsymbol{u}}_T) \cdot \boldsymbol{v}_T = \int_T \boldsymbol{f} \cdot \boldsymbol{v}_T, \tag{3.56}$$

and, for any interface $F \in \mathcal{F}_{T_1} \cap \mathcal{F}_{T_2}$, the numerical tractions satisfy the law of action and reaction:

$$\boldsymbol{T}_{T_1 F}(\underline{\boldsymbol{u}}_{T_1}) + \boldsymbol{T}_{T_2 F}(\underline{\boldsymbol{u}}_{T_2}) = \boldsymbol{0}. \tag{3.57}$$

Proof. For all $T \in \mathcal{T}_h$, use the definition (3.17) of $\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T$ with $\boldsymbol{\tau} = \boldsymbol{\pi}_T^k \boldsymbol{\sigma}(\cdot \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{u}}_T)$ in A_h and the rewriting (3.54) of s_T together with the definition (3.55) of $\boldsymbol{R}_{TF}^k$ to infer that it holds, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$,

$$\begin{aligned}
\int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}_h &= A_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{v}}_h) + s_h(\underline{\boldsymbol{u}}_h, \underline{\boldsymbol{v}}_h) \\
&= \sum_{T \in \mathcal{T}_h} \int_T \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{u}}_T) : \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T \\
&\quad + \sum_{T \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_T} \int_F (\boldsymbol{\pi}_T^k \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{u}}_T) \boldsymbol{n}_{TF} - \boldsymbol{R}_{TF}^k \underline{\boldsymbol{u}}_T) \cdot (\boldsymbol{v}_F - \boldsymbol{v}_T),
\end{aligned}$$

where to cancel $\boldsymbol{\pi}_T^k$ inside the first integral in the second line we have used the fact that $\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T \in \mathbb{P}_d^{k-1}(T; \mathbb{R}^{d \times d})$ for all $T \in \mathcal{T}_h$. Selecting $\underline{\boldsymbol{v}}_h$ such that $\boldsymbol{v}_T$ spans $\mathbb{P}_d^k(T; \mathbb{R}^d)$ for a selected mesh element $T \in \mathcal{T}_h$ while $\boldsymbol{v}_{T'} \equiv \boldsymbol{0}$ for all $T' \in \mathcal{T}_h \setminus \{T\}$ and $\boldsymbol{v}_F \equiv \boldsymbol{0}$ for all $F \in \mathcal{F}_h$, we obtain (3.56). On the other hand, selecting $\underline{\boldsymbol{v}}_h$ such that $\boldsymbol{v}_T \equiv \boldsymbol{0}$ for all $T \in \mathcal{T}_h$, $\boldsymbol{v}_F$ spans $\mathbb{P}_d^k(F; \mathbb{R}^d)$ for a selected interface $F \in \mathcal{F}_{T_1} \cap \mathcal{F}_{T_2}$, and $\boldsymbol{v}_{F'} \equiv \boldsymbol{0}$ for all $F' \in \mathcal{F}_h \setminus \{F\}$ yields (3.57). □

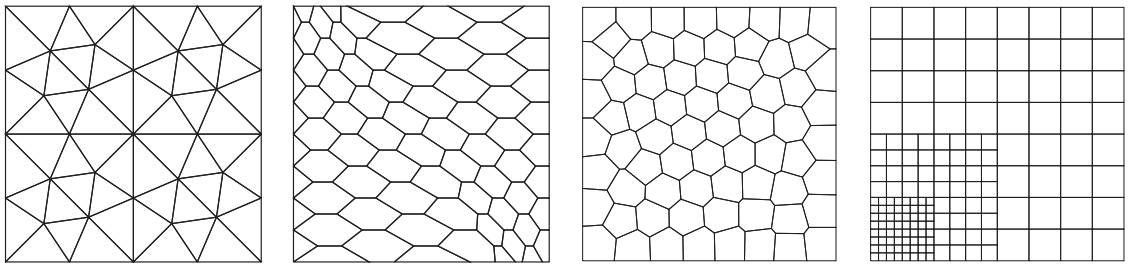


Figure 3.2: Triangular, hexagonal-dominant, Voronoi, and nonmatching quadrangular meshes for the numerical tests. The triangular and nonmatching quadrangular meshes were originally proposed for the FVCA5 benchmark [120]. The (predominantly) hexagonal was used in [80]. The Voronoi mesh family was obtained using the PolyMesher algorithm of [188].

3.7 Numerical results

In this section we present a comprehensive set of numerical tests to assess the properties of our method using the models of Examples 3.2, 3.3, and 3.5 (cf. also Remark 3.6). Note that an important step in the implementation of HHO methods consists in selecting a basis for each of the polynomial spaces that appear in the construction. In the numerical tests of the present section, for all $T \in \mathcal{T}_h$, we take as a basis for $\mathbb{P}_d^k(T;\mathbb{R}^d)$ the Cartesian product of the monomials in the translated and scaled coordinates $\left(h_T^{-1}(x_i - x_{T,i})\right)_{1\le i\le d}$, where $\boldsymbol{x}_T$ is the barycenter of T. Similarly, for all $F \in \mathcal{F}_h$ we define a basis for $\mathbb{P}_d^k(F;\mathbb{R}^d)$ by taking the monomials with respect to a local frame scaled using the face diameter h_F and the middle point of F. Further details on implementation aspects are given in [74, Section 6.1].

3.7.1 Convergence for the Hencky–Mises model

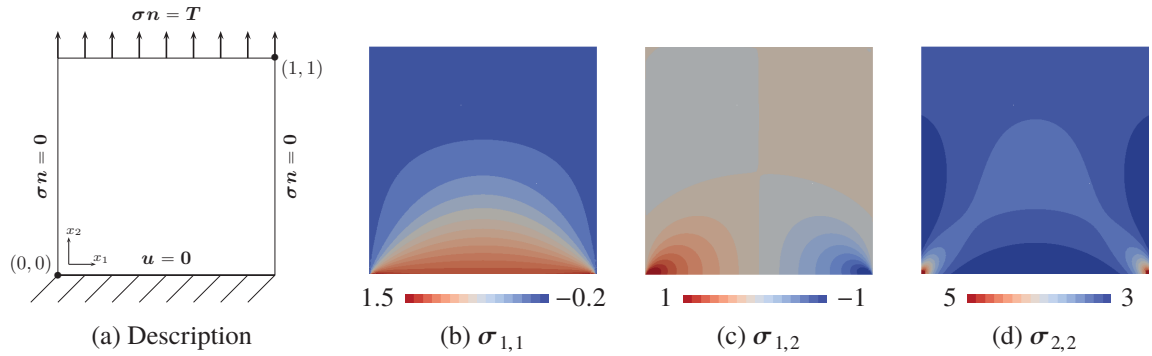
In order to check the error estimates stated in Theorem 3.16, we first solve a manufactured two-dimensional hyperelasticity problem. We consider the Hencky–Mises model with $\Phi(\rho) = \mu(e^{-\rho} + 2\rho)$ and $\alpha = \lambda + \mu$ in (3.8), so that the conditions (3.9) are satisfied. This choice leads to the following stress-strain relation:

$$\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) = ((\lambda-\mu) + \mu e^{-\operatorname{dev}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})})\operatorname{tr}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})\boldsymbol{I}_d + \mu(2 - e^{-\operatorname{dev}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})})\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}.$$

We consider the unit square domain $\Omega = (0,1)^2$ and take $\mu = 2$, $\lambda = 1$, and an exact displacement $\boldsymbol{u}$ given by

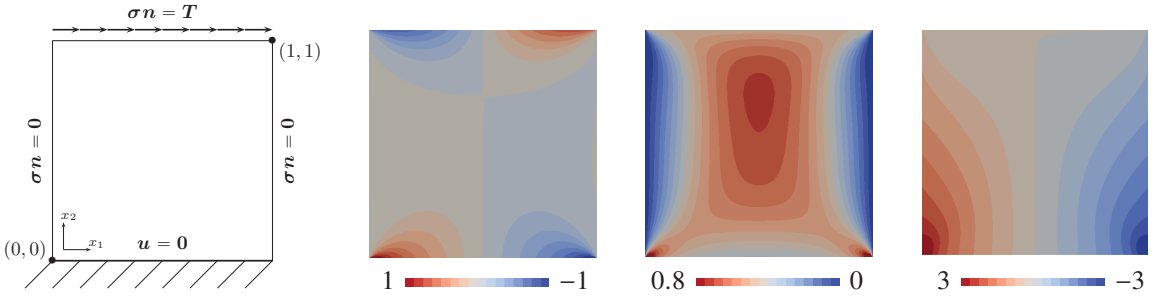
$$\boldsymbol{u}(\boldsymbol{x}) = (\sin(\pi x_1)\sin(\pi x_2), \sin(\pi x_1)\sin(\pi x_2)).$$

The volumetric load $\boldsymbol{f} = -\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})$ is inferred from the exact solution $\boldsymbol{u}$. In this case, since the selected exact displacement vanishes on Γ, we simply consider homogeneous Dirichlet conditions. We consider the triangular, hexagonal, Voronoi, and nonmatching quadrangular mesh families depicted in Figure 3.2 and polynomial degrees k ranging from 1 to 4. The nonmatching mesh is simply meant to show that the method supports nonconforming interfaces: refining in the corner has no particular meaning for the selected solution. The



(a) Description (b) $\boldsymbol{\sigma}_{1,1}$ (c) $\boldsymbol{\sigma}_{1,2}$ (d) $\boldsymbol{\sigma}_{2,2}$

Figure 3.3: Tensile test description and resulting stress components for the linear case. Values in 10^5Pa



(a) Description (b) $\boldsymbol{\sigma}_{1,1}$ (c) $\boldsymbol{\sigma}_{1,2}$ (d) $\boldsymbol{\sigma}_{2,2}$

Figure 3.4: Shear test description and resulting stress components for the linear case. Values in 10^5Pa

initialization of our iterative linearization procedure (Newton scheme) is obtained solving the linear elasticity model. This initial guess leads to a 40% reduction of the number of iterations with respect to a null initial guess. The energy-norm orders of convergence (OCV) displayed in the third column of Tables 3.1–3.4 at the end of the chapter are in agreement with the theoretical predictions. In particular, it is observed that the optimal convergence in h^{k+1} is reached for the triangular, nonmatching Cartesian, and hexagonal meshes for $1 \leq k \leq 3$, whereas for $k = 4$ the asymptotic convergence order does not appear to have been reached in the last mesh refinement. It can also be observed in Table 3.2 that the convergence rate exceeds the estimated one on the locally refined Cartesian mesh for $k = 1$ and $k = 2$. For the sake of completeness, we also display in the fourth column of Tables 3.1–3.4 the L^2-norm of the error defined as the difference between the L^2-projection $\boldsymbol{\pi}_h^k \boldsymbol{u}$ of the exact solution on $\mathbb{P}_d^k(\mathcal{T}_h; \mathbb{R}^d)$ and the broken polynomial function $\boldsymbol{u}_h$ obtained from element-based DOFs, while in the fifth column we display the corresponding observed convergence rates. In this case, orders of convergence up to h^{k+2} are observed.

3.7.2 Tensile and shear test cases

We next consider the two test cases schematically depicted in Figures 3.3 and 3.4. On the unit square domain Ω, we solve problem (3.1a) considering three different models of hyperelasticity (see Remark 3.6):

(i) *Linear.* The linear model corresponding to the stored energy density function (3.7) with Lamé's parameters

$$\lambda = 11 \times 10^5 \text{Pa}, \ \mu = 82 \times 10^4 \text{Pa}. \tag{3.58}$$

(ii) *Hencky–Mises.* The Hencky–Mises model (3.4) obtained by taking $\Phi(\rho) = \mu(\frac{\rho}{2} + (1+\rho)^{1/2})$ and $\alpha = \lambda + \mu$ in (3.8), with λ, μ as in (3.58) (also in this case the conditions (3.9) hold). This choice leads to

$$\begin{aligned} \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}) = ((\lambda + \frac{\mu}{2}) - \frac{\mu}{2}(1 + \text{dev}(\boldsymbol{\nabla}_s \boldsymbol{u}))^{-1/2})) \, \text{tr}(\boldsymbol{\nabla}_s \boldsymbol{u}) \boldsymbol{I}_d \\ + \mu(1 + (1 + \text{dev}(\boldsymbol{\nabla}_s \boldsymbol{u}))^{-1/2}) \boldsymbol{\nabla}_s \boldsymbol{u}. \end{aligned} \tag{3.59}$$

The Lamé's functions of the previous relation are inspired from those proposed in [21, Section 5.1]. In particular, the function $\tilde{\mu}(\rho) = \mu(1 + (1 + \text{dev}(\boldsymbol{\nabla}_s \boldsymbol{u}))^{-1/2})$ corresponds to the Carreau law for viscoplastic materials.

(iii) *Second-order.* The second-order model (3.6) with Lamé's parameter as in (3.58) and second-order moduli

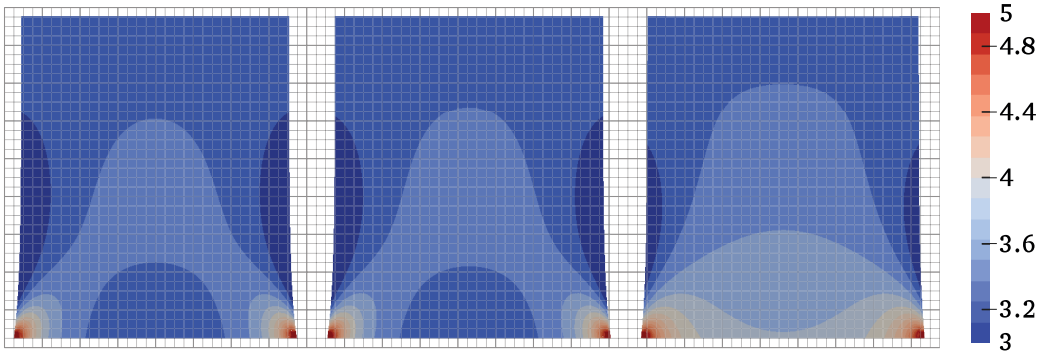
$$A = 11 \times 10^6 \text{Pa}, \ \text{B} = -48 \times 10^5 \text{Pa}, \ \text{C} = 13.2 \times 10^5 \text{Pa}.$$

These values correspond to the estimates provided in [124] for the Armco Iron. We recall that the second-order elasticity stress-strain relation does not satisfy in general the assumptions under which we are able to prove the convergence and error estimates. In particular, we observe that the stored energy density function defined in (3.10) is not convex.

The bottom part of the boundary of the domain is assumed to be fixed, the normal stress is equal to zero on the two lateral parts, and a traction is imposed at the top of the boundary. So, mixed boundary conditions are imposed as follows

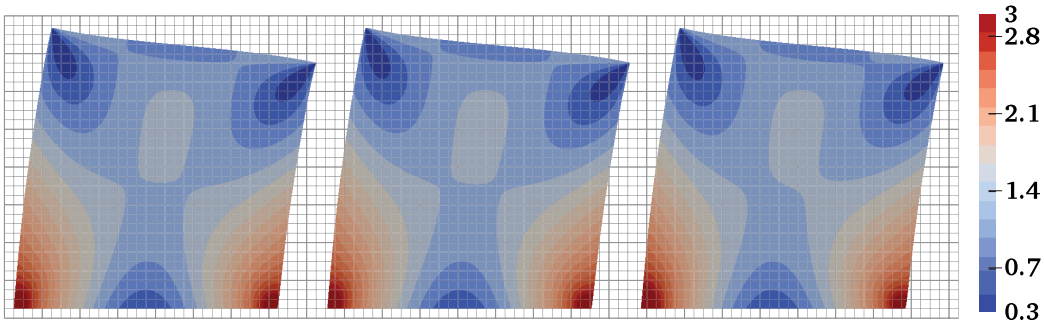
$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{0} \quad \text{on } \{\boldsymbol{x} \in \Gamma, x_2 = 0\}, \\ \boldsymbol{\sigma n} &= \boldsymbol{T} \ \text{on } \{\boldsymbol{x} \in \Gamma, x_2 = 1\}, \\ \boldsymbol{\sigma n} &= \boldsymbol{0} \ \text{on } \{\boldsymbol{x} \in \Gamma, x_1 = 0\}, \\ \boldsymbol{\sigma n} &= \boldsymbol{0} \ \text{on } \{\boldsymbol{x} \in \Gamma, x_1 = 1\}. \end{aligned}$$

For the tensile case, we impose a vertical traction at the top of the boundary equal to $\boldsymbol{T} = (0, 3.2 \times 10^5 \text{Pa})$. This type of boundary conditions produces large normal stresses (i.e., the diagonal components of $\boldsymbol{\sigma}$) and minor shear stresses (i.e., the off-diagonal components of $\boldsymbol{\sigma}$). It can be observed in Figure 3.3, where the components of the stress tensor are depicted for the linear case. In Figure 3.5 we plot the stress norm on the deformed domain obtained for the three hyperelasticity models. The results of Fig. 3.3, 3.4, 3.5, and 3.6 are obtained on a mesh with 3584 triangles (corresponding to a typical mesh-size of 3.84×10^{-3}) and with polynomial degree $k = 2$. Obviously, the symmetry of the results is visible, and we observe that the three displacement fields are very close. This is motivated by the fact that, with our choice of the parameters in (3.58) and in (3.59), the linear model exactly corresponds to the linear approximation at the origin of the nonlinear ones. The maximum value of the stress



(a) Linear (b) Hencky–Mises (c) Second order

Figure 3.5: Tensile test case: Stress norm on the deformed domain. Values in 10^5Pa



(a) Linear (b) Hencky–Mises (c) Second order

Figure 3.6: Shear test case: Stress norm on the deformed domain. Values in 10^4Pa

concentrates on the two bottom corners due to the homogeneous Dirichlet condition that totally locks the displacement when $x_1 = 0$. The repartition of the stress on the domain with the second-order model is visibly different from those obtained with the linear and Hencky–Mises models. At the energy level, we also have a higher difference between the second-order model and the linear one since $|\mathbb{E}_{\text{lin}} - \mathbb{E}_{\text{hm}}|/\mathbb{E}_{\text{lin}} = 0.44\%$ while $|\mathbb{E}_{\text{lin}} - \mathbb{E}_{\text{snd}}|/\mathbb{E}_{\text{lin}} = 4.45\%$, where $\mathbb{E}_\bullet$ is the total elastic energy obtained by integrating over the domain the strain energy density functions defined by (3.7), (3.8), and (3.10):

$$\mathbb{E}_\bullet := \int_\Omega \Psi_\bullet, \quad \text{with} \quad \bullet \in \{\text{lin}, \text{hm}, \text{snd}\}.$$

The reference values for the total energy, used in Figure 3.7 in order to assess convergence, are obtained on a fine Cartesian mesh having a mesh-size of 1.95×10^{-3} and $k = 3$. For the shear case, we consider an horizontal traction equal to $\boldsymbol{T} = (4.5 \times 10^4\text{Pa}, 0)$ which induces the stress pattern illustrated in Figure 3.4. The computed stress norm on the deformed domain is depicted in Figure 3.6, and we can see that the displacement fields associated with the three models are very close as for the tensile test case. Here, the maximum values of the stress are localized in the lower part of the domain near the lateral parts. Unlike the tensile test, the

difference between the three models is tiny as confirmed by the elastic energy equal to 3180 J, 3184 J, and 3190 J respectively. The decreasing of the energy difference in comparison with the previous test can be explained by the fact that the value of the Neumann boundary data on the top is divided by a factor 7 in order to obtain maximum displacements roughly equal to 15%.

3.8 Appendix: Technical results

This appendix contains the proof of technical results.

3.8.1 Discrete Korn inequality

Proposition 3.20 (Discrete Korn inequality)**.** *Assume that the mesh further verifies the assumption of [39, Theorem 4.2] if $d = 2$ and [39, Theorem 5.2] if $d = 3$. Then, the discrete Korn inequality* (3.25) *holds.*

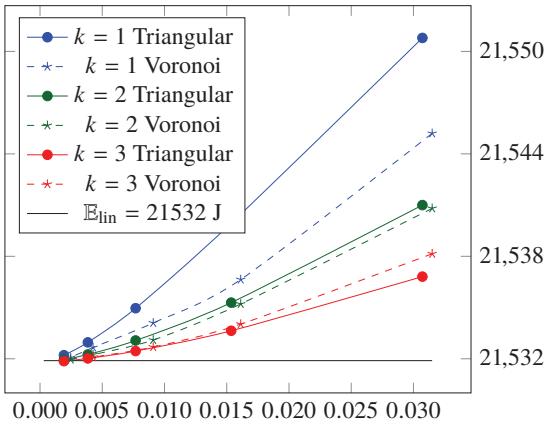
Proof. Using the broken Korn inequality [39, Eq. (1.22)] on $H^1(\mathcal{T}_h;\mathbb{R}^d)$ followed by the Cauchy–Schwarz inequality, one has

$$\begin{aligned}\|\boldsymbol{v}_h\|^2 + \|\boldsymbol{\nabla}_h \boldsymbol{v}_h\|^2 &\lesssim \|\boldsymbol{\nabla}_{\mathrm{s},h} \boldsymbol{v}_h\|^2 + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} h_F^{-1} \|[\boldsymbol{v}_h]_F\|_F^2 \\ &\quad + \sup_{\boldsymbol{m}\in\mathbb{P}_d^1(\mathcal{T}_h;\mathbb{R}^d),\, \|\boldsymbol{\gamma}_{\boldsymbol{n}}(\boldsymbol{m})\|_\Gamma = 1} \left(\int_\Gamma \boldsymbol{\gamma}(\boldsymbol{v}_h) \cdot \boldsymbol{\gamma}_{\boldsymbol{n}}(\boldsymbol{m}) \right)^2 \\ &\lesssim \|\boldsymbol{\nabla}_{\mathrm{s},h} \boldsymbol{v}_h\|^2 + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} h_F^{-1} \|[\boldsymbol{v}_h]_F\|_F^2 + \sum_{F\in\mathcal{F}_h^{\mathrm{b}}} \|\boldsymbol{v}_{h|F}\|_F^2. \end{aligned} \tag{3.60}$$

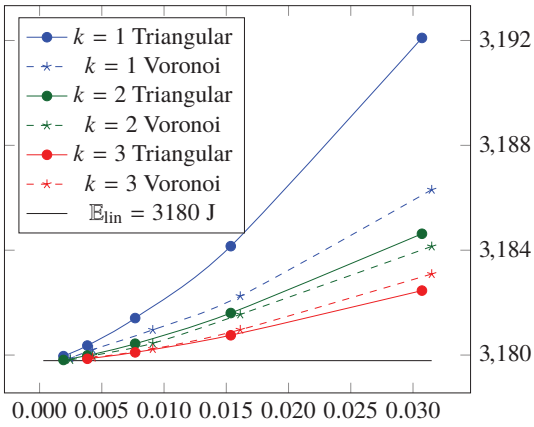
For an interface $F \in \mathcal{F}_{T_1} \cap \mathcal{F}_{T_2}$, we have introduced the jump $[\boldsymbol{v}_h]_F \coloneqq \boldsymbol{v}_{T_1} - \boldsymbol{v}_{T_2}$. Thus, using the triangle inequality, we get $\|[\boldsymbol{v}_h]_F\|_F \le \|\boldsymbol{v}_F - \boldsymbol{v}_{T_1}\|_F + \|\boldsymbol{v}_F - \boldsymbol{v}_{T_2}\|_F$. For a boundary face $F \in \mathcal{F}_h^{\mathrm{b}}$ such that $F \in \mathcal{F}_T \cap \mathcal{F}_h^{\mathrm{b}}$ for some $T \in \mathcal{T}_h$ we have, on the other hand, $\|\boldsymbol{v}_{h|F}\|_F = \|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F$ since $\boldsymbol{v}_F \equiv \boldsymbol{0}$ (cf. (3.20)). Using these relations in the right-hand side of (3.60) and rearranging the sums leads to

$$\begin{aligned}&\|\boldsymbol{v}_h\|^2 + \|\boldsymbol{\nabla}_h \boldsymbol{v}_h\|^2 \\ &\quad\lesssim \sum_{T\in\mathcal{T}_h} \left(\|\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T\|_T^2 + \sum_{F\in\mathcal{F}_T\cap\mathcal{F}_h^{\mathrm{i}}} h_F^{-1} \|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F^2 \right) + h \sum_{F\in\mathcal{F}_h^{\mathrm{b}}} h_F^{-1} \|\boldsymbol{v}_F - \boldsymbol{v}_{h|F}\|_F^2 \\ &\quad\lesssim \max\{1, d_\Omega\} \sum_{T\in\mathcal{T}_h} \left(\|\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1} \|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F^2 \right), \end{aligned}$$

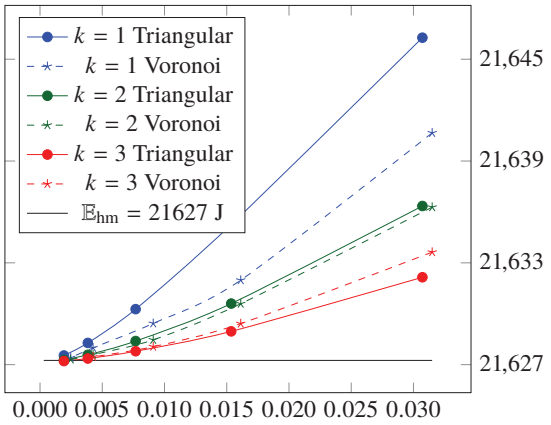
where d_Ω denotes the diameter of Ω. Owing to the definition (3.14) of the discrete strain seminorm, the latter yields the assertion. □



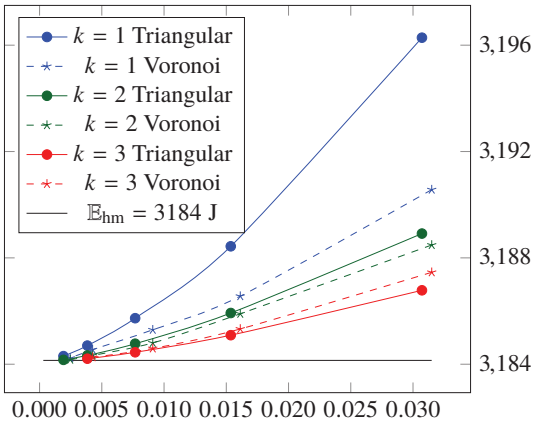
(a) Linear, tensile test



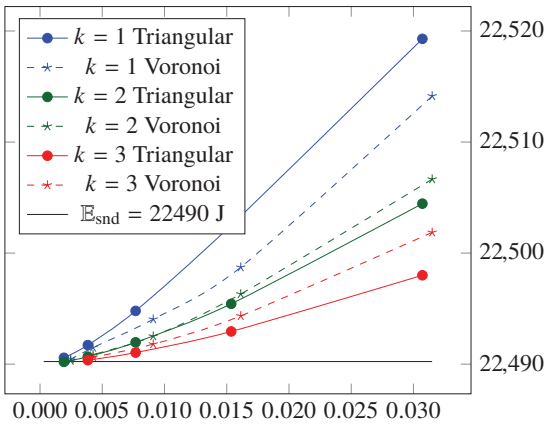
(b) Linear, shear test



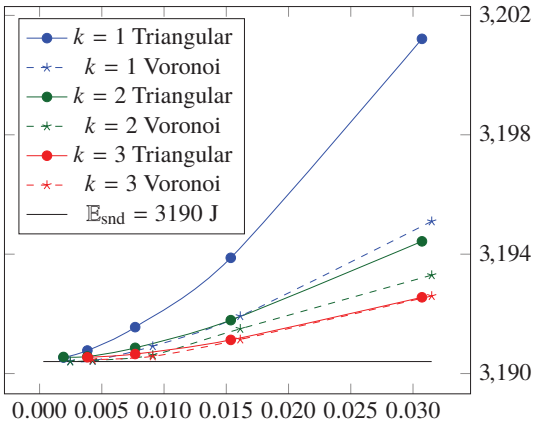
(c) Hencky–Mises, tensile test



(d) Hencky–Mises, shear test



(e) Second-order, tensile test



(f) Second-order, shear test

Figure 3.7: Energy vs h, tensile and shear test cases

3.8.2 Discrete compactness

Proof of Lemma 3.12. In the proof we use the same notation for functions in $L^2(\Omega;\mathbb{R}^d) \subset L^1(\Omega;\mathbb{R}^d)$ and for their extension by zero outside Ω. Let $(\underline{\boldsymbol{v}}_h)_{h\in\mathcal{H}} \in (\underline{\boldsymbol{U}}^k_{h,\mathrm{D}})_{h\in\mathcal{H}}$ be such that (3.26) holds. Define the space of integrable functions with bounded variation $\mathrm{BV}(\mathbb{R}^d) := \{\boldsymbol{v} \in L^1(\mathbb{R}^d;\mathbb{R}^d) \mid \|\boldsymbol{v}\|_{\mathrm{BV}} < +\infty\}$, where

$$\|\boldsymbol{v}\|_{\mathrm{BV}} := \sum_{i=1}^{d} \sup\left\{ \int_{\mathbb{R}^d} \boldsymbol{v}\cdot\partial_i\boldsymbol{\phi} \;\middle|\; \boldsymbol{\phi}\in C_c^\infty(\mathbb{R}^d;\mathbb{R}^d), \|\boldsymbol{\phi}\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} \le 1 \right\}.$$

Here, $\partial_i\boldsymbol{\phi}$ denotes the i-th column of $\boldsymbol{\nabla}\boldsymbol{\phi}$. Let $\boldsymbol{\phi} \in C_c^\infty(\mathbb{R}^d;\mathbb{R}^d)$ with $\|\boldsymbol{\phi}\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} \le 1$. Integrating by parts and using the fact that $\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\int_F(\boldsymbol{v}_F\cdot\boldsymbol{\phi})\boldsymbol{n}_{TF} = 0$, we have that

$$\begin{aligned}
\int_{\mathbb{R}^d} \boldsymbol{v}_h\cdot\partial_i\boldsymbol{\phi} &= \sum_{T\in\mathcal{T}_h}\int_T ((\boldsymbol{\nabla}\boldsymbol{\phi})^{\mathrm{T}}\boldsymbol{v}_T)_i \\
&= -\sum_{T\in\mathcal{T}_h}\left(\int_T((\boldsymbol{\nabla}\boldsymbol{v}_T)^{\mathrm{T}}\boldsymbol{\phi})_i + \sum_{F\in\mathcal{F}_T}\int_F(\boldsymbol{v}_F\cdot\boldsymbol{\phi}-\boldsymbol{v}_T\cdot\boldsymbol{\phi})(\boldsymbol{n}_{TF})_i\right) \\
&\le \sum_{T\in\mathcal{T}_h}\left(\int_T\sum_{j=1}^d|(\boldsymbol{\nabla}\boldsymbol{v}_T)_{ji}| + \sum_{F\in\mathcal{F}_T}\int_F\sum_{j=1}^d|(\boldsymbol{v}_F-\boldsymbol{v}_T)_j(\boldsymbol{n}_{TF})_i|\right),
\end{aligned}$$

where, in order to pass to the third line, we have used $\|\boldsymbol{\phi}\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)} \le 1$. Therefore, summing over $i \in \{1,\dots,d\}$, observing that, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, we have $\sum_{i=1}^d|(\boldsymbol{n}_{TF})_i| \le d^{1/2}$, and using the Lebesgue embeddings arising from the Hölder inequality on bounded domain, leads to

$$\|\boldsymbol{v}_h\|_{\mathrm{BV}} \lesssim \sum_{T\in\mathcal{T}_h}\left(|T|_d^{1/2}\|\boldsymbol{\nabla}\boldsymbol{v}_T\|_T + \sum_{F\in\mathcal{F}_T}|F|_{d-1}^{1/2}\|\boldsymbol{v}_F-\boldsymbol{v}_T\|_F\right),$$

where $|\cdot|_d$ denotes the d-dimensional Hausdorff measure. Moreover, using the Cauchy–Schwarz inequality together with the geometric bound $|F|_{d-1}h_F \lesssim |T|_d$, we obtain that

$$\|\boldsymbol{v}_h\|_{\mathrm{BV}} \lesssim |\Omega|_d^{1/2}\left(\sum_{T\in\mathcal{T}_h}\left[\|\boldsymbol{\nabla}\boldsymbol{v}_T\|_T^2 + \sum_{F\in\mathcal{F}_T}h_F^{-1}\|\boldsymbol{v}_F-\boldsymbol{v}_T\|_F^2\right]\right)^{1/2}.$$

Thus, using the discrete Korn inequality (3.25), it is readily inferred that

$$\|\boldsymbol{v}_h\|_{\mathrm{BV}} \lesssim \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h} \lesssim 1.$$

Owing to the Helly selection principle [99, Section 5.2.3], the sequence $(\boldsymbol{v}_h)_{h\in\mathcal{H}}$ is relatively compact in $L^1(\mathbb{R}^d;\mathbb{R}^d)$ and thus in $L^1(\Omega;\mathbb{R}^d)$. It only remains to prove that the sequence is also relatively compact in $L^q(\Omega;\mathbb{R}^d)$, with $1 < q < +\infty$ if $d = 2$ or $1 < q < 6$ if $d = 3$.

Owing to the discrete Sobolev embeddings [71, Proposition 5.4] together with the discrete Korn inequality (3.25), it holds, with $r = q + 1$ if $d = 2$ and $r = 6$ if $d = 3$, that

$$\|\boldsymbol{v}_h\|_{L^r(\Omega;\mathbb{R}^d)} \lesssim \left(\sum_{T\in\mathcal{T}_h}\left[\|\boldsymbol{\nabla}\boldsymbol{v}_T\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1}\|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F^2\right]\right)^{1/2} \lesssim 1,$$

Thus, we can complete the proof by means of the interpolation inequality [40, Remark 2 p. 93]. For all $h, h' \in \mathcal{H}$ we have with $\theta := \frac{r-q}{q(r-1)} \in (0,1)$,

$$\|\boldsymbol{v}_h - \boldsymbol{v}_{h'}\|_{L^q(\Omega;\mathbb{R}^d)} \le \|\boldsymbol{v}_h - \boldsymbol{v}_{h'}\|_{L^1(\Omega;\mathbb{R}^d)}^{\theta}\|\boldsymbol{v}_h - \boldsymbol{v}_{h'}\|_{L^r(\Omega;\mathbb{R}^d)}^{1-\theta} \lesssim \|\boldsymbol{v}_h - \boldsymbol{v}_{h'}\|_{L^1(\Omega;\mathbb{R}^d)}^{\theta}.$$

Therefore, up to a subsequence, $(\boldsymbol{v}_h)_{h\in\mathcal{H}}$ is a Cauchy sequence in $L^q(\Omega;\mathbb{R}^d)$, so it converges. □

3.8.3 Consistency of the discrete symmetric gradient operator

Proof of Proposition 3.13. 1) *Strong consistency.* We first assume that $\boldsymbol{v} \in H^2(\Omega;\mathbb{R}^d)$. Owing to the commuting property (3.18) and the approximation property (3.13a) with $m = 1$ and $s = 2$, it is inferred that $\|\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{I}}_T^k \boldsymbol{v} - \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}\|_T \lesssim h\|\boldsymbol{v}\|_{H^2(T;\mathbb{R}^d)}$. Squaring, summing over $T \in \mathcal{T}_h$, and taking the square root of the resulting inequality gives

$$\|\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{I}}_h^k \boldsymbol{v} - \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}\| \lesssim h\|\boldsymbol{v}\|_{H^2(\Omega;\mathbb{R}^d)}. \tag{3.61}$$

If $\boldsymbol{v} \in H^1(\Omega;\mathbb{R}^d)$ we reason by density, namely we take a sequence $(\boldsymbol{v}_\epsilon)_{\epsilon>0} \subset H^2(\Omega;\mathbb{R}^d)$ that converges to $\boldsymbol{v}$ in $H^1(\Omega;\mathbb{R}^d)$ as $\epsilon \to 0$ and, using twice the triangular inequality, we write

$$\|\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{I}}_h^k \boldsymbol{v} - \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}\| \le \|\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{I}}_h^k (\boldsymbol{v} - \boldsymbol{v}_\epsilon)\| + \|\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{I}}_h^k \boldsymbol{v}_\epsilon - \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_\epsilon\| + \|\boldsymbol{\nabla}_{\mathrm{s}}(\boldsymbol{v} - \boldsymbol{v}_\epsilon)\|.$$

By (3.61), the second term in the right-hand side tends to 0 as $h \to 0$. Moreover, owing to the commuting property (3.18) and the H^1-boundedness of $\boldsymbol{\pi}_T^k$, one has

$$\begin{aligned}\|\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{I}}_h^k (\boldsymbol{v} - \boldsymbol{v}_\epsilon)\| &= \left(\sum_{T\in\mathcal{T}_h}\|\boldsymbol{\pi}_T^k \boldsymbol{\nabla}_{\mathrm{s}}(\boldsymbol{v} - \boldsymbol{v}_\epsilon)\|_T^2\right)^{1/2} \le \left(\sum_{T\in\mathcal{T}_h}\|\boldsymbol{\nabla}_{\mathrm{s}}(\boldsymbol{v} - \boldsymbol{v}_\epsilon)\|_T^2\right)^{1/2} \\ &\le \|\boldsymbol{\nabla}_{\mathrm{s}}(\boldsymbol{v} - \boldsymbol{v}_\epsilon)\|.\end{aligned}$$

Therefore, taking the supremum limit as $h \to 0$ and then the supremum limit as $\epsilon \to 0$, concludes the proof of (3.27) (notice that the order in which the limits are taken is important).

2) *Sequential consistency.* In order to prove (3.28) we observe that, by the definitions (3.19) of $\boldsymbol{G}_{\mathrm{s},h}^k$ and (3.17b) of $\boldsymbol{G}_{\mathrm{s},T}^k$ one has, for all $\boldsymbol{\tau} \in H^1(\Omega;\mathbb{R}^{d\times d}_{\mathrm{sym}})$ and all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$,

$$
\begin{aligned}
\int_\Omega \boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{v}}_h : \boldsymbol{\tau} &= \sum_{T\in\mathcal{T}_h} \int_T \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T : \boldsymbol{\tau} \\
&= \sum_{T\in\mathcal{T}_h} \int_T (\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T - \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T) : (\boldsymbol{\tau} - \boldsymbol{\pi}_T^0 \boldsymbol{\tau}) + \sum_{T\in\mathcal{T}_h} \int_T (\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T - \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T) : \boldsymbol{\pi}_T^0 \boldsymbol{\tau} \\
&\quad + \sum_{T\in\mathcal{T}_h} \int_T \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T : \boldsymbol{\tau} \\
&= \mathfrak{T}_1 + \sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} \int_F (\boldsymbol{v}_F - \boldsymbol{v}_T) \cdot (\boldsymbol{\pi}_T^0 \boldsymbol{\tau}) \boldsymbol{n}_{TF} + \sum_{T\in\mathcal{T}_h} \int_T \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T : \boldsymbol{\tau} \qquad (3.62) \\
&= \mathfrak{T}_1 + \sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} \int_F (\boldsymbol{v}_F - \boldsymbol{v}_T) \cdot (\boldsymbol{\pi}_T^0 \boldsymbol{\tau} - \boldsymbol{\tau}) \boldsymbol{n}_{TF} - \sum_{T\in\mathcal{T}_h} \int_T \boldsymbol{v}_T \cdot (\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) \\
&\quad + \sum_{F\in\mathcal{F}_h^{\mathrm{b}}} \int_F \boldsymbol{v}_F \cdot (\boldsymbol{\tau} \boldsymbol{n}_{TF}) \\
&= \mathfrak{T}_1 + \mathfrak{T}_2 - \int_\Omega \boldsymbol{v}_h \cdot (\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) + \int_\Gamma \boldsymbol{v}_{\Gamma,h} \cdot \boldsymbol{\gamma}_{\boldsymbol{n}}(\boldsymbol{\tau}).
\end{aligned}
$$

To obtain the third equality, we used an element-wise integration by parts together with the relation

$$
\sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T\cap\mathcal{F}_h^{\mathrm{i}}} \int_F \boldsymbol{v}_F \cdot (\boldsymbol{\tau}\boldsymbol{n}_{TF}) = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F \boldsymbol{v}_F \cdot (\boldsymbol{\tau}\boldsymbol{n}_{T_1F} + \boldsymbol{\tau}\boldsymbol{n}_{T_2F}) = 0,
$$

where for all $F \in \mathcal{F}_h^{\mathrm{i}}$, $T_1, T_2 \in \mathcal{T}_h$ are such that $F \subset \partial T_1 \cap \partial T_2$. Owing to (3.62), the conclusion follows once we prove that $|\mathfrak{T}_1 + \mathfrak{T}_2| \lesssim h \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h} \|\boldsymbol{\tau}\|_{H^1(\Omega;\mathbb{R}^{d\times d})}$. By (3.13a) (with $m = 0$ and $s = 1$) we have $\|\boldsymbol{\tau} - \boldsymbol{\pi}_T^0\boldsymbol{\tau}\|_T \lesssim h_T \|\boldsymbol{\tau}\|_{H^1(T;\mathbb{R}^{d\times d})}$ and thus, using the Cauchy–Schwarz and triangle inequalities followed by the norm equivalence (3.24),

$$
\begin{aligned}
|\mathfrak{T}_1| &\le \left(\sum_{T\in\mathcal{T}_h} \|\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T - \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T\|_T^2 \right)^{1/2} \left(\sum_{T\in\mathcal{T}_h} \|\boldsymbol{\tau} - \boldsymbol{\pi}_T^0 \boldsymbol{\tau}\|_T^2 \right)^{1/2} \\
&\lesssim h \left(\|\boldsymbol{G}_{\mathrm{s},h}^k \underline{\boldsymbol{v}}_h\|^2 + \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2 \right)^{1/2} \|\boldsymbol{\tau}\|_{H^1(\Omega;\mathbb{R}^{d\times d})} \lesssim h \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h} \|\boldsymbol{\tau}\|_{H^1(\Omega;\mathbb{R}^{d\times d})}.
\end{aligned} \qquad (3.63)
$$

In a similar way, we obtain an upper bound for $\mathfrak{T}_2$. By (3.13b) (with $m = 0$ and $s = 1$), for all $F \in \mathcal{F}_T$, we have $\|\boldsymbol{\tau} - \boldsymbol{\pi}_T^0\boldsymbol{\tau}\|_F \lesssim h_T^{1/2} \|\boldsymbol{\tau}\|_{H^1(T;\mathbb{R}^{d\times d})} \lesssim h_F^{1/2} \|\boldsymbol{\tau}\|_{H^1(T;\mathbb{R}^{d\times d})}$ and thus, using the Cauchy–Schwarz inequality,

$$
|\mathfrak{T}_2| \lesssim \sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} h_F^{1/2} \|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F \|\boldsymbol{\tau}\|_{H^1(T;\mathbb{R}^{d\times d})} \lesssim h \|\underline{\boldsymbol{v}}_h\|_{\epsilon,h} \|\boldsymbol{\tau}\|_{H^1(\Omega;\mathbb{R}^{d\times d})}. \qquad (3.64)
$$

Owing to (3.63) and (3.64), the triangle inequality $|\mathfrak{T}_1 + \mathfrak{T}_2| \le |\mathfrak{T}_1| + |\mathfrak{T}_2|$ yields the conclusion.

□

Table 3.1: Convergence results on the triangular mesh family. OCV stands for order of convergence.

h	$\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} - \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h\|$	OCV	$\|\boldsymbol{\pi}^k_h\boldsymbol{u} - \boldsymbol{u}_h\|$	OCV
		$k = 1$		
$3.07 \cdot 10^{-2}$	$5.59 \cdot 10^{-2}$	—	$7.32 \cdot 10^{-3}$	—
$1.54 \cdot 10^{-2}$	$1.51 \cdot 10^{-2}$	1.9	$1.05 \cdot 10^{-3}$	2.81
$7.68 \cdot 10^{-3}$	$3.86 \cdot 10^{-3}$	1.96	$1.34 \cdot 10^{-4}$	2.96
$3.84 \cdot 10^{-3}$	$1.01 \cdot 10^{-3}$	1.93	$1.7 \cdot 10^{-5}$	2.98
$1.92 \cdot 10^{-3}$	$2.59 \cdot 10^{-4}$	1.96	$2.15 \cdot 10^{-6}$	2.98
		$k = 2$		
$3.07 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	—	$1.47 \cdot 10^{-3}$	—
$1.54 \cdot 10^{-2}$	$1.29 \cdot 10^{-3}$	3.35	$6.05 \cdot 10^{-5}$	4.62
$7.68 \cdot 10^{-3}$	$2.11 \cdot 10^{-4}$	2.6	$5.36 \cdot 10^{-6}$	3.48
$3.84 \cdot 10^{-3}$	$2.73 \cdot 10^{-5}$	2.95	$3.6 \cdot 10^{-7}$	3.9
$1.92 \cdot 10^{-3}$	$3.42 \cdot 10^{-6}$	3	$2.28 \cdot 10^{-8}$	3.98
		$k = 3$		
$3.07 \cdot 10^{-2}$	$2.81 \cdot 10^{-3}$	—	$2.39 \cdot 10^{-4}$	—
$1.54 \cdot 10^{-2}$	$3.72 \cdot 10^{-4}$	2.93	$1.95 \cdot 10^{-5}$	3.63
$7.68 \cdot 10^{-3}$	$2.16 \cdot 10^{-5}$	4.09	$5.47 \cdot 10^{-7}$	5.14
$3.84 \cdot 10^{-3}$	$1.43 \cdot 10^{-6}$	3.92	$1.66 \cdot 10^{-8}$	5.04
$1.92 \cdot 10^{-3}$	$9.51 \cdot 10^{-8}$	3.91	$5.34 \cdot 10^{-10}$	4.96
		$k = 4$		
$3.07 \cdot 10^{-2}$	$1.37 \cdot 10^{-3}$	—	$1.13 \cdot 10^{-4}$	—
$1.54 \cdot 10^{-2}$	$5.97 \cdot 10^{-5}$	4.54	$3.04 \cdot 10^{-6}$	5.24
$7.68 \cdot 10^{-3}$	$1.76 \cdot 10^{-6}$	5.07	$4.09 \cdot 10^{-8}$	6.19
$3.84 \cdot 10^{-3}$	$6.46 \cdot 10^{-8}$	4.77	$7.64 \cdot 10^{-10}$	5.74

Table 3.2: Convergence results on the locally refined mesh family. OCV stands for order of convergence.

h	$\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} - \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h\|$	OCV	$\|\boldsymbol{\pi}^k_h\boldsymbol{u} - \boldsymbol{u}_h\|$	OCV
		$k = 1$		
0.25	0.13	—	$1.9 \cdot 10^{-2}$	—
0.13	$2.64 \cdot 10^{-2}$	2.28	$2.54 \cdot 10^{-3}$	2.9
$6.25 \cdot 10^{-2}$	$4.97 \cdot 10^{-3}$	2.41	$3.22 \cdot 10^{-4}$	2.98
$3.12 \cdot 10^{-2}$	$9.14 \cdot 10^{-4}$	2.44	$4.12 \cdot 10^{-5}$	2.96
$1.56 \cdot 10^{-2}$	$1.67 \cdot 10^{-4}$	2.45	$5.21 \cdot 10^{-6}$	2.98
		$k = 2$		
0.25	$1.88 \cdot 10^{-2}$	—	$3.79 \cdot 10^{-3}$	—
0.13	$5.05 \cdot 10^{-3}$	1.9	$3.55 \cdot 10^{-4}$	3.42
$6.25 \cdot 10^{-2}$	$6.51 \cdot 10^{-4}$	2.96	$2.92 \cdot 10^{-5}$	3.6
$3.12 \cdot 10^{-2}$	$6.83 \cdot 10^{-5}$	3.25	$1.89 \cdot 10^{-6}$	3.94
$1.56 \cdot 10^{-2}$	$6.23 \cdot 10^{-6}$	3.45	$1.19 \cdot 10^{-7}$	3.99
		$k = 3$		
0.25	$7.84 \cdot 10^{-3}$	—	$1.41 \cdot 10^{-3}$	—
0.13	$1.09 \cdot 10^{-3}$	2.85	$7.5 \cdot 10^{-5}$	4.23
$6.25 \cdot 10^{-2}$	$8.22 \cdot 10^{-5}$	3.73	$3.93 \cdot 10^{-6}$	4.25
$3.12 \cdot 10^{-2}$	$5.64 \cdot 10^{-6}$	3.86	$1.45 \cdot 10^{-7}$	4.75
$1.56 \cdot 10^{-2}$	$3.44 \cdot 10^{-7}$	4.04	$5.23 \cdot 10^{-9}$	4.79
		$k = 4$		
0.25	$4.35 \cdot 10^{-3}$	—	$4.68 \cdot 10^{-4}$	—
0.13	$3.65 \cdot 10^{-4}$	3.58	$3.19 \cdot 10^{-5}$	3.87
$6.25 \cdot 10^{-2}$	$1.5 \cdot 10^{-5}$	4.6	$6.02 \cdot 10^{-7}$	5.73
$3.12 \cdot 10^{-2}$	$5.78 \cdot 10^{-7}$	4.69	$1.03 \cdot 10^{-8}$	5.86

Table 3.3: Convergence results on the hexagonal mesh family. OCV stands for order of convergence.

h	$\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} - \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h\|$	OCV	$\|\boldsymbol{\pi}^k_h\boldsymbol{u} - \boldsymbol{u}_h\|$	OCV
		$k = 1$		
$6.3 \cdot 10^{-2}$	0.22	—	$2.75 \cdot 10^{-2}$	—
$3.42 \cdot 10^{-2}$	$3.72 \cdot 10^{-2}$	2.89	$3.73 \cdot 10^{-3}$	3.27
$1.72 \cdot 10^{-2}$	$7.17 \cdot 10^{-3}$	2.4	$4.83 \cdot 10^{-4}$	2.97
$8.59 \cdot 10^{-3}$	$1.44 \cdot 10^{-3}$	2.31	$6.14 \cdot 10^{-5}$	2.97
$4.3 \cdot 10^{-3}$	$2.4 \cdot 10^{-4}$	2.59	$7.7 \cdot 10^{-6}$	3
		$k = 2$		
$6.3 \cdot 10^{-2}$	$2.68 \cdot 10^{-2}$	—	$3.04 \cdot 10^{-3}$	—
$3.42 \cdot 10^{-2}$	$7.01 \cdot 10^{-3}$	2.2	$3.56 \cdot 10^{-4}$	3.51
$1.72 \cdot 10^{-2}$	$1.09 \cdot 10^{-3}$	2.71	$3.31 \cdot 10^{-5}$	3.46
$8.59 \cdot 10^{-3}$	$1.41 \cdot 10^{-4}$	2.95	$2.53 \cdot 10^{-6}$	3.7
$4.3 \cdot 10^{-3}$	$1.96 \cdot 10^{-5}$	2.85	$1.72 \cdot 10^{-7}$	3.89
		$k = 3$		
$6.3 \cdot 10^{-2}$	$1.11 \cdot 10^{-2}$	—	$1.08 \cdot 10^{-3}$	—
$3.42 \cdot 10^{-2}$	$1.92 \cdot 10^{-3}$	2.87	$9.29 \cdot 10^{-5}$	4.02
$1.72 \cdot 10^{-2}$	$2.79 \cdot 10^{-4}$	2.81	$6.13 \cdot 10^{-6}$	3.95
$8.59 \cdot 10^{-3}$	$2.54 \cdot 10^{-5}$	3.45	$2.88 \cdot 10^{-7}$	4.4
$4.3 \cdot 10^{-3}$	$1.61 \cdot 10^{-6}$	3.99	$1.24 \cdot 10^{-8}$	4.55
		$k = 4$		
$6.3 \cdot 10^{-2}$	$5.53 \cdot 10^{-3}$	—	$4.49 \cdot 10^{-4}$	—
$3.42 \cdot 10^{-2}$	$5.76 \cdot 10^{-4}$	3.7	$3.07 \cdot 10^{-5}$	4.39
$1.72 \cdot 10^{-2}$	$6.29 \cdot 10^{-5}$	3.22	$1.21 \cdot 10^{-6}$	4.7
$8.59 \cdot 10^{-3}$	$2.21 \cdot 10^{-6}$	4.82	$2.69 \cdot 10^{-8}$	5.48

Table 3.4: Convergence results on the Voronoi mesh family. OCV stands for order of convergence.

h	$\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} - \boldsymbol{G}^k_{\mathrm{s},h}\underline{\boldsymbol{u}}_h\|$	OCV	$\|\boldsymbol{\pi}^k_h\boldsymbol{u} - \boldsymbol{u}_h\|$	OCV
		$k = 1$		
$6.5 \cdot 10^{-2}$	$8.82 \cdot 10^{-2}$	—	$1.55 \cdot 10^{-2}$	—
$3.15 \cdot 10^{-2}$	$1.49 \cdot 10^{-2}$	2.45	$2.29 \cdot 10^{-3}$	2.64
$1.61 \cdot 10^{-2}$	$3.63 \cdot 10^{-3}$	2.1	$3.01 \cdot 10^{-4}$	3.02
$9.09 \cdot 10^{-3}$	$8.68 \cdot 10^{-4}$	2.5	$3.95 \cdot 10^{-5}$	3.55
$4.26 \cdot 10^{-3}$	$2.04 \cdot 10^{-4}$	1.91	$4.97 \cdot 10^{-6}$	2.74
		$k = 2$		
$6.5 \cdot 10^{-2}$	$1.43 \cdot 10^{-2}$	—	$2.63 \cdot 10^{-3}$	—
$3.15 \cdot 10^{-2}$	$4.03 \cdot 10^{-3}$	1.75	$2.53 \cdot 10^{-4}$	3.23
$1.61 \cdot 10^{-2}$	$4.78 \cdot 10^{-4}$	3.18	$2.22 \cdot 10^{-5}$	3.63
$9.09 \cdot 10^{-3}$	$6.7 \cdot 10^{-5}$	3.44	$1.45 \cdot 10^{-6}$	4.77
$4.26 \cdot 10^{-3}$	$9.08 \cdot 10^{-6}$	2.64	$9.07 \cdot 10^{-8}$	3.66
		$k = 3$		
$6.5 \cdot 10^{-2}$	$7.12 \cdot 10^{-3}$	—	$9.08 \cdot 10^{-4}$	—
$3.15 \cdot 10^{-2}$	$8.34 \cdot 10^{-4}$	2.96	$6.78 \cdot 10^{-5}$	3.58
$1.61 \cdot 10^{-2}$	$7.03 \cdot 10^{-5}$	3.69	$3.18 \cdot 10^{-6}$	4.56
$9.09 \cdot 10^{-3}$	$4.17 \cdot 10^{-6}$	4.94	$9.67 \cdot 10^{-8}$	6.11
$4.26 \cdot 10^{-3}$	$2.42 \cdot 10^{-7}$	3.76	$3.15 \cdot 10^{-9}$	4.52
		$k = 4$		
$6.5 \cdot 10^{-2}$	$3.25 \cdot 10^{-3}$	—	$3.68 \cdot 10^{-4}$	—
$3.15 \cdot 10^{-2}$	$2.94 \cdot 10^{-4}$	3.32	$2.14 \cdot 10^{-5}$	3.93
$1.61 \cdot 10^{-2}$	$9.86 \cdot 10^{-6}$	5.06	$4.34 \cdot 10^{-7}$	5.81
$9.09 \cdot 10^{-3}$	$3.47 \cdot 10^{-7}$	5.85	$6.74 \cdot 10^{-9}$	7.29

Chapter 4

Nonlinear poroelasticity

This chapter has been submitted for publication. The preprint is available in HAL:

Analysis of a Hybrid High-Order–discontinuous Galerkin discretization method for nonlinear poroelasticity,
https://hal.archives-ouvertes.fr/hal-01785810, 2018.

Contents

4.1 Introduction

In this chapter, we formulate and analyze a coupled Hybrid High-Order–discontinuous Galerkin (HHO–dG) discretization method for nonlinear poroelastic problems. We overstep the work of Chapter 2 devoted to the Biot's model [27, 190] by considering more general nonlinear stress-strain constitutive laws as in Chapter 3. The model is valid under the assumptions of small deformations of the rock matrix, small variations of the porosity, and small relative variations of the fluid density. The interest of the poroelastic models considered here is particularly manifest in geosciences applications [123, 127, 151], where fluid flows in geological subsurface, modeled as a porous media, induce a deformation of the rock matrix. The challenge is then to design a discretization method able to (i) treat a complex geometry with polyhedral meshes and nonconforming interfaces, (ii) handle possible heterogeneities of the poromechanical parameters and nonlinearities of the stress-strain relation, and (iii) deal with the numerical instabilities encountered in this type of coupled problem. We focus here on the theoretical aspects while numerical tests assessing the convergence of the method have been presented in Section A.5; further numerical investigation will be carried out in a future work.

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, denote a bounded connected polyhedral domain with boundary $\partial\Omega$ and outward normal $\boldsymbol{n}$. Without loss of generality, we assume that the domain is scaled so that its diameter $\mathrm{diam}(\Omega)$ is equal to 1. For a given finite time $t_{\mathrm{F}} > 0$, volumetric load $\boldsymbol{f}$, fluid source g, the nonlinear poroelasticity problem considered here consists in finding a vector-valued displacement field $\boldsymbol{u}$ and a scalar-valued pore pressure field p solution of

$$-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\cdot,\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) + \alpha\boldsymbol{\nabla}p = \boldsymbol{f} \qquad \text{in } \Omega\times(0,t_{\mathrm{F}}), \tag{4.1a}$$

$$C_0\mathrm{d}_t p + \alpha\mathrm{d}_t\boldsymbol{\nabla}\cdot\boldsymbol{u} - \boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}(\cdot)\boldsymbol{\nabla}p) = g \qquad \text{in } \Omega\times(0,t_{\mathrm{F}}), \tag{4.1b}$$

where $\boldsymbol{\nabla}_{\mathrm{s}}$ denotes the symmetric gradient, d_t denotes the time derivative, α is the Biot–Willis coefficient, $C_0 \geq 0$ is the constrained specific storage coefficient, and $\boldsymbol{\kappa} : \Omega \to \mathbb{R}^{d\times d}_{\mathrm{s}}$ is the uniformly elliptic permeability tensor field which, for real numbers $0 < \underline{\kappa} \leq \overline{\kappa}$, satisfies for almost every (a.e.) $\boldsymbol{x} \in \Omega$ and all $\boldsymbol{\xi} \in \mathbb{R}^d$,

$$\underline{\kappa}|\boldsymbol{\xi}|^2 \leq \boldsymbol{\kappa}(\boldsymbol{x})\boldsymbol{\xi}\cdot\boldsymbol{\xi} \leq \overline{\kappa}|\boldsymbol{\xi}|^2.$$

For the sake of simplicity, we assume in the following discussion that $\boldsymbol{\kappa}$ is piecewise constant on a polyhedral partition P_Ω of Ω, an assumption typically verified in geoscience applications. In the poroelasticity theory [61], the medium is modeled as a continuous superposition of solid and fluid phases. The momentum equilibrium equation (4.1a) is based on the Terzaghi decomposition [190] of the total stress tensor into a mechanical contribution and a pore pressure contribution. Examples and assumptions for the constitutive stress-strain relation $\boldsymbol{\sigma} : \Omega\times\mathbb{R}^{d\times d}_{\mathrm{s}} \to \mathbb{R}^{d\times d}_{\mathrm{s}}$ are detailed in Section 4.2.2; we refer the reader to [22, 28] for a physical and experimental investigation of the nonlinear behavior of porous solids. On the other hand, the mass conservation equation (4.1b) is derived for fully saturated porous media assuming Darcean flow. The first two terms of this equation quantify the variation of fluid content in the pores. The dimensionless coupling coefficient α express the amount of fluid that can be forced into the medium by a variation of pore volume for a constant

fluid pressure, while C_0 measures the amount of fluid that can be forced into the medium by pressure increments due to compressibility of the structure. The case of a solid matrix with incompressible grains corresponds to the limit value $C_0 = 0$. Following [181, 205], for the sake of simplicity we take $\alpha = 1$ in what follows. To close the problem, we enforce homogeneous boundary conditions corresponding to a clamped, impermeable boundary, i.e.,

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \partial\Omega \times (0, t_{\mathrm{F}}), \tag{4.1c}$$

$$(\boldsymbol{\kappa}(\cdot)\boldsymbol{\nabla} p) \cdot \boldsymbol{n} = 0 \qquad \text{on } \partial\Omega \times (0, t_{\mathrm{F}}), \tag{4.1d}$$

as well as the following initial condition which prescribes the initial fluid content:

$$C_0 p(\cdot, 0) + \boldsymbol{\nabla}\cdot\boldsymbol{u}(\cdot, 0) = \phi^0(\cdot). \tag{4.1e}$$

In the case $C_0 = 0$, we also need the following compatibility conditions on g and ϕ^0 and zero-average constraint on p:

$$\int_\Omega \phi^0 = 0, \qquad \int_\Omega g(\cdot, t) = 0, \quad \text{and} \quad \int_\Omega p(\cdot, t) = 0 \qquad \forall t \in (0, t_{\mathrm{F}}). \tag{4.1f}$$

When discretizing the poroelasticity system (4.1), the main challenges are to ensure stability and convergence under mild assumptions on the nonlinear stress-strain relation and on the permeability field, and to prevent localized pressure oscillations arising in the case of low-permeable and low-compressible porous media. Since the latter issue is in part related to the saddle point structure in the coupled equations for $C_0 = 0$ and small $\underline{\kappa}$, the discrete spaces for the displacement and the pressure should satisfy an inf-sup condition. Indeed, as observed in [117, 155, 168] in the context of finite element discretizations of the linear poroelasticity problem, the inf-sup condition yields an L^2-estimate of the discrete pressure independent of $\underline{\kappa}^{-1}$, and allows one to prove the convergence of the approximate pressure towards the continuous pressure also in the incompressible case $C_0 = 0$. We notice, however, that the problem of spurious pressure oscillations is actually more involved than a simple saddle-point coupling issue. For instance, it has been recently pointed out in [176] that, even for discretization methods leading to an inf-sup stable discretization of the Stokes problem in the steady case, pressure oscillations can arise owing to a lack of monotonicity of the discrete operator. The robustness with respect to spurious oscillations has been numerically observed in Section 2.6.2 for a HHO–dG discretization of the linear poroelasticity model.

In this chapter, we present and analyze a nonconforming space discretization of problem (4.1) where the nonlinear elasticity operator is discretized using the HHO method of Chapter 3 (c.f. also [71, 74]), while the Darcy operator relies on the Symmetric Weighted Interior Penalty (SWIP) method of [77]. The proposed method has several assets: (i) it is valid in two and three space dimensions; (ii) it delivers an inf-sup stable discretization on general spatial meshes including, e.g., polyhedral elements and nonmatching interfaces; (iii) it allows one to increase the space approximation order to accelerate convergence in the presence of (locally) regular solutions. Compared to the method proposed in Chapter 2 for the linear poroelasticity problem, there are two main differences in the design. First, for a given polynomial degree $k \geq 1$, the symmetric gradient reconstruction sits in the full space

of tensor-valued polynomials of total degree $\leq k$, as opposed to symmetric gradients of vector-valued polynomials of total degree $\leq (k+1)$. Following [33, 72], this modification is required to obtain optimal convergence rates when considering nonlinear stress-strain laws. Second, the right-hand side of the discrete problem of Section 4.4.4 is obtained by taking the average in time of the loading force $\boldsymbol{f}$ and fluid source g over a time step instead of their value at the end of the time step. This modification allows us to prove stability and optimal error estimates under significantly weaker time regularity assumptions on data (cf. Remark 4.12 and 4.18). Finally, in Section 4.3.4 we give a new simple proof of discrete Korn's inequality, not requiring particular geometrical assumption on the mesh. The interest of these results goes beyond the specific application considered here.

The material is organized as follows. In Section 4.2 we present the assumptions on the stress-strain law and the variational formulation of the nonlinear poroelasticity problem. In Section 4.3 we define the space and time meshes and the discrete spaces for the displacement and the pressure fields. In Section 4.4 we define the discrete counterparts of the elasticity, Darcy, and hydromechanical coupling operators and formulate the discrete problem. In Section 4.5 we prove the well-posedness of the scheme by deriving an a priori estimate on the discrete solution that holds also when the specific storage coefficient vanishes. The convergence analysis of the method is carried out in Section 4.6.

4.2 Continuous setting

In this section we introduce the notation for function spaces, formulate the assumptions on the stress-strain law, and derive a weak formulation of problem (4.1).

4.2.1 Notation for function spaces

Let $X \subset \overline{\Omega}$. Spaces of functions, vector fields, and tensor fields defined over X are respectively denoted by italic capital, boldface Roman capital, and special Roman capital letters. The subscript "s" appended to a special Roman capital letter denotes a space of symmetric tensor fields. Thus, for example, $L^2(X)$, $\mathbf{L}^2(X)$, and $\mathbb{L}^2_{\mathrm{s}}(X)$ denote the spaces of square integrable functions, vector fields, and symmetric tensor fields over X, respectively. For any measured set X and any $m \in \mathbb{Z}$, we denote by $H^m(X)$ the usual Sobolev space of functions that have weak partial derivatives of order up to m in $L^2(X)$, with the convention that $H^0(X) \coloneqq L^2(X)$, while $C^m(X)$ and $C^\infty_{\mathrm{c}}(X)$ denote, respectively, the usual spaces of m-times continuously differentiable functions and infinitely continuously differentiable functions with compact support on X. We denote by $(\cdot,\cdot)_X$ and $(\cdot,\cdot)_{m,X}$ the usual scalar products in $L^2(X)$ and $H^m(X)$ respectively, and by $\|\cdot\|_X$ and $\|\cdot\|_{m,X}$ the induced norms.

For a vector space V with scalar product $(\cdot,\cdot)_V$, the space $C^m(V) \coloneqq C^m([0,t_{\mathrm{F}}];V)$ is spanned by V-valued functions that are m-times continuously differentiable in the time interval $[0,t_{\mathrm{F}}]$. The space $C^m(V)$ is a Banach space when equipped with the norm

$$\|\varphi\|_{C^m(V)} \coloneqq \max_{0\leq i\leq m} \max_{t\in[0,t_{\mathrm{F}}]} \|\mathrm{d}_t^i\varphi(t)\|_V.$$

Similarly, the Hilbert space $H^m(V) := H^m((0,t_\mathrm{F});V)$ is spanned by V-valued functions of the time interval, and the norm $\|\cdot\|_{H^m(V)}$ is induced by the scalar product

$$(\varphi,\psi)_{H^m(V)} = \sum_{j=0}^{m} \int_0^{t_\mathrm{F}} (\mathrm{d}_t^j\varphi(t), \mathrm{d}_t^j\psi(t))_V \mathrm{d}t \qquad \forall \varphi,\psi \in H^m(V).$$

4.2.2 Stress-strain law

The following assumptions on the stress-strain relation are required to obtain a well-posed weak formulation of the nonlinear poroelasticity problem.

Assumption 4.1 (Stress-strain relation)**.** We assume that the stress function $\boldsymbol{\sigma} : \Omega \times \mathbb{R}_\mathrm{s}^{d\times d} \to \mathbb{R}_\mathrm{s}^{d\times d}$ is a Carathéodory function, i.e., $\boldsymbol{\sigma}(\boldsymbol{x},\cdot)$ is continuous on $\mathbb{R}_\mathrm{s}^{d\times d}$ for almost every $\boldsymbol{x}\in\Omega$ and $\boldsymbol{\sigma}(\cdot,\boldsymbol{\tau})$ is measurable on Ω for all $\boldsymbol{\tau}\in\mathbb{R}_\mathrm{s}^{d\times d}$. Moreover, there exist real numbers $C_\mathrm{gr}, C_\mathrm{cv} \in (0,+\infty)$ such that, for a.e. $\boldsymbol{x}\in\Omega$ and all $\boldsymbol{\tau},\boldsymbol{\eta}\in\mathbb{R}_\mathrm{s}^{d\times d}$, the following conditions hold:

$$|\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau})|_{d\times d} \le C_\mathrm{gr}|\boldsymbol{\tau}|_{d\times d}, \qquad \text{(growth)} \tag{4.2a}$$

$$\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau}) : \boldsymbol{\tau} \ge C_\mathrm{cv}^2|\boldsymbol{\tau}|_{d\times d}^2, \qquad \text{(coercivity)} \tag{4.2b}$$

$$(\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\eta})) : (\boldsymbol{\tau}-\boldsymbol{\eta}) > 0 \text{ if } \boldsymbol{\eta}\neq\boldsymbol{\tau}. \qquad \text{(monotonicity)} \tag{4.2c}$$

Above, we have introduced the Frobenius product such that, for all $\boldsymbol{\tau},\boldsymbol{\eta}\in\mathbb{R}^{d\times d}$, $\boldsymbol{\tau}:\boldsymbol{\eta} := \sum_{1\le i,j\le d}\tau_{ij}\eta_{ij}$ with corresponding matrix norm such that, for all $\boldsymbol{\tau}\in\mathbb{R}^{d\times d}$, $|\boldsymbol{\tau}|_{d\times d} := (\boldsymbol{\tau}:\boldsymbol{\tau})^{\frac{1}{2}}$.

Three meaningful examples for the stress-strain relation $\boldsymbol{\sigma} : \Omega\times\mathbb{R}_\mathrm{s}^{d\times d} \to \mathbb{R}_\mathrm{s}^{d\times d}$ in (4.1a) are:

- The (possibly heterogeneous) *linear elasticity model* given by the usual Hooke's law

$$\boldsymbol{\sigma}(\cdot,\boldsymbol{\tau}) = \lambda(\cdot)\,\mathrm{tr}(\boldsymbol{\tau})\boldsymbol{I}_d + 2\mu(\cdot)\boldsymbol{\tau}, \tag{4.3}$$

  where $\mu : \Omega \to [\mu_*,\mu^*]$, with $0<\mu_*\le\mu^*<+\infty$, and $\lambda:\Omega\to\mathbb{R}_+$ are the Lamé parameters.

- The *nonlinear Hencky–Mises model* of [105, 157] corresponding to the mechanical behavior law

$$\boldsymbol{\sigma}(\cdot,\boldsymbol{\tau}) = \tilde{\lambda}(\cdot,\mathrm{dev}(\boldsymbol{\tau}))\,\mathrm{tr}(\boldsymbol{\tau})\boldsymbol{I}_d + 2\tilde{\mu}(\cdot,\mathrm{dev}(\boldsymbol{\tau}))\boldsymbol{\tau}, \tag{4.4}$$

  with nonlinear Lamé scalar functions $\tilde{\mu}:\Omega\times\mathbb{R}_+\to[\mu_*,\mu^*]$ and $\tilde{\lambda}:\Omega\times\mathbb{R}_+\to\mathbb{R}_+$ depending on the deviatoric part of the strain $\mathrm{dev}(\boldsymbol{\tau}) := \mathrm{tr}(\boldsymbol{\tau}^2) - \frac{1}{d}\mathrm{tr}(\boldsymbol{\tau})^2$.

- The *isotropic reversible hyperelastic damage model* [50], for which the stress-strain relation reads

$$\boldsymbol{\sigma}(\cdot,\boldsymbol{\tau}) = (1-D(\cdot,\boldsymbol{\tau}))\mathfrak{C}(\cdot)\boldsymbol{\tau}. \tag{4.5}$$

  where $D:\Omega\times\mathbb{R}_\mathrm{s}^{d\times d}\to[0,1]$ is the scalar damage function and $\mathfrak{C}:\Omega\to\mathbb{R}^{d^4}$ is a fourth-order symmetric and uniformly elliptic tensor field, namely, for some strictly positive constants $\underline{C}$ and $\overline{C}$, it holds

$$\underline{C}|\boldsymbol{\tau}|_{d\times d}^2 \le \mathfrak{C}(\boldsymbol{x})\boldsymbol{\tau}:\boldsymbol{\tau} \le \overline{C}|\boldsymbol{\tau}|_{d\times d}^2 \quad \forall\boldsymbol{\tau}\in\mathbb{R}^{d\times d},\ \forall\boldsymbol{x}\in\Omega.$$

Being linear, the Cauchy stress tensor in (4.3) clearly satisfies the previous assumptions. Moreover, under some mild requirements (cf. Appendix B and [87]) on the nonlinear Lamé scalar functions $\tilde{\mu}$ and $\tilde{\lambda}$ in (4.4) and on the damage function D in (4.5), it can be proven that also the Hencky–Mises model and the isotropic reversible damage model satisfy Assumption 4.1.

4.2.3 Weak formulation

At each time $t \in [0, t_F]$, the natural functional spaces for the displacement $\boldsymbol{u}(t)$ and pore pressure $p(t)$ taking into account the boundary condition (4.1c) and the zero average constraint (4.1f) are, respectively,

$$\boldsymbol{U} \coloneqq \mathbf{H}^1_0(\Omega) \qquad \text{and} \qquad P \coloneqq \begin{cases} H^1(\Omega) & \text{if } C_0 > 0, \\ H^1(\Omega) \cap L^2_0(\Omega) & \text{if } C_0 = 0, \end{cases}$$

with $\mathbf{H}^1_0(\Omega) \coloneqq \left\{ \boldsymbol{v} \in \mathbf{H}^1(\Omega) \ : \ \boldsymbol{v}_{|\partial\Omega} = \mathbf{0} \right\}$ and $L^2_0(\Omega) \coloneqq \left\{ q \in L^2(\Omega) \ : \ \int_\Omega q = 0 \right\}$. We consider the following weak formulation of problem (4.1): For a loading term $\boldsymbol{f} \in L^2(\mathbf{L}^2(\Omega))$, a fluid source $g \in L^2(L^2(\Omega))$, and an initial datum $\phi^0 \in L^2(\Omega)$ that verify (4.1f), find $\boldsymbol{u} \in L^2(\boldsymbol{U})$ and $p \in L^2(P)$ such that, for all $\boldsymbol{v} \in \boldsymbol{U}$, all $q \in P$, and all $\varphi \in C^\infty_c((0, t_F))$

$$\int_0^{t_F} a(\boldsymbol{u}(t), \boldsymbol{v})\, \varphi(t) \mathrm{d}t + \int_0^{t_F} b(\boldsymbol{v}, p(t))\, \varphi(t) \mathrm{d}t = \int_0^{t_F} (\boldsymbol{f}(t), \boldsymbol{v})_\Omega\, \varphi(t) \mathrm{d}t, \tag{4.6a}$$

$$\int_0^{t_F} \left[b(\boldsymbol{u}(t), q) - C_0 (p(t), q)_\Omega \right] \mathrm{d}_t \varphi(t) \mathrm{d}t + \int_0^{t_F} c(p(t), q)\, \varphi(t) \mathrm{d}t = \int_0^{t_F} (g(t), q)_\Omega\, \varphi(t) \mathrm{d}t, \tag{4.6b}$$

$$(C_0 p(0) + \boldsymbol{\nabla}\cdot \boldsymbol{u}(0), q)_\Omega = (\phi^0, q)_\Omega, \tag{4.6c}$$

where we have defined the nonlinear function $a : \boldsymbol{U} \times \boldsymbol{U} \to \mathbb{R}$ and the bilinear forms $b : \boldsymbol{U} \times P \to \mathbb{R}$ and $c : P \times P \to \mathbb{R}$ such that, for all $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{U}$ and all $q, r \in P$,

$$a(\boldsymbol{v}, \boldsymbol{w}) \coloneqq (\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_s \boldsymbol{v}), \boldsymbol{\nabla}_s \boldsymbol{w})_\Omega, \qquad b(\boldsymbol{v}, q) \coloneqq -(\boldsymbol{\nabla}\cdot \boldsymbol{v}, q)_\Omega, \qquad c(q, r) \coloneqq (\boldsymbol{\kappa}(\cdot) \boldsymbol{\nabla} r, \boldsymbol{\nabla} q)_\Omega.$$

The first term in (4.6a) is well defined thanks to the growth assumption (4.2a). Moreover, owing to (4.2b) together with Korn's first inequality and Poincaré's inequality, $a(\cdot, \cdot)$ and $c(\cdot, \cdot)$ are coercive on $\boldsymbol{U}$ and P, respectively. The strict monotonicity assumption (4.2c) guarantees the uniqueness of the weak solution.

Remark 4.2 (Regularity of the fluid content and of the pore pressure). Using an integration by parts in time in (4.6b), it is inferred that

$$\mathrm{d}_t \left[C_0 (p, q)_\Omega - b(\boldsymbol{u}, q) \right] + c(p, q) = (g, q)_\Omega \quad \forall q \in P \quad \text{ in } L^2((0, t_F)). \tag{4.7}$$

Therefore, defining the fluid content $\phi \coloneqq C_0 p + \boldsymbol{\nabla}\cdot \boldsymbol{u}$, we have that $t \mapsto (\phi(t), q)_\Omega \in H^1((0, t_F)) \subset C^0([0, t_F])$ for all $q \in P$, and, as a result, (4.6c) makes sense. Moreover, in the

case $C_0 > 0$, taking $q = 1$ in (4.7) and owing to the definition of the bilinear form c and the homogeneous Dirichlet condition (4.1c), we infer that

$$\mathrm{d}_t \left(C_0 \int_\Omega p(\cdot,t) \right) = \int_\Omega g(\cdot,t) \quad \text{in } L^2((0,t_\mathrm{F})).$$

Thus, $t \mapsto \int_\Omega p(\cdot,t) \in H^1((0,t_\mathrm{F}))$, namely the average of the pore pressure on the medium Ω is a continuous function in $[0,t_\mathrm{F}]$.

4.3 Discrete setting

In this section we define the space and time meshes, recall the definition and properties of L^2-orthogonal projectors on local and broken polynomial spaces, and introduce the discrete spaces for the displacement and the pressure.

4.3.1 Space mesh

We consider here polygonal or polyhedral meshes corresponding to couples $\mathcal{M}_h := (\mathcal{T}_h, \mathcal{F}_h)$, where $\mathcal{T}_h$ is a finite collection of polygonal elements T such that $h := \max_{T\in\mathcal{T}_h} h_T > 0$ with h_T denoting the diameter of T, while $\mathcal{F}_h$ is a finite collection of hyperplanar faces F. It is assumed henceforth that the mesh $\mathcal{M}_h$ matches the geometrical requirements detailed in [86, Definition 7.2]; see also [82, Section 2]. We additionally assume here that the mesh elements are star-shaped with respect to a ball of diameter uniformly comparable to the element diameter in order to use the results of [32, Appendix A]. To avoid dealing with jumps of the permeability coefficient inside elements, we additionally assume that $\mathcal{M}_h$ is compliant with the partition P_Ω on which $\boldsymbol{\kappa}$ is piecewise constant meaning that, for every $T \in \mathcal{T}_h$, there exists a unique subdomain $\omega \in P_\Omega$ such that $T \subset \omega$. For every mesh element $T \in \mathcal{T}_h$, we denote by $\mathcal{F}_T$ the subset of $\mathcal{F}_h$ containing the faces that lie on the boundary ∂T of T. For each face $F \in \mathcal{F}_T$, $\boldsymbol{n}_{TF}$ is the (constant) unit normal vector to F pointing out of T. Boundary faces lying on $\partial\Omega$ and internal faces contained in Ω are collected in the sets $\mathcal{F}_h^\mathrm{b}$ and $\mathcal{F}_h^\mathrm{i}$, respectively.

Our focus is on the so-called h-convergence analysis, so we consider a sequence of refined meshes that is regular in the sense of [82, Definition 3]. The corresponding strictly positive regularity parameter, uniformly bounded away from zero, is denoted by ϱ. We additionally assume that, for each mesh in the sequence, all the elements are star-shaped with respect to every point of a ball of radius uniformly comparable to the diameter of the element. The mesh regularity assumption implies, in particular, that the diameter h_T of a mesh element $T \in \mathcal{T}_h$ is uniformly comparable to the diameter h_F of each face $F \in \mathcal{F}_T$, and that the number of faces in $\mathcal{F}_T$ is bounded above by an integer N_∂ independent of h.

4.3.2 Time mesh

We subdivide $(0,t_\mathrm{F})$ into $N \in \mathbb{N}^*$ uniform subintervals, and introduce the timestep $\tau := t_\mathrm{F}/N$ and the discrete times $t^n := n\tau$ for all $0 \le n \le N$. We define the space of piecewise H^1

functions on $(0,t_\mathrm{F})$ by

$$H^1(\mathcal{T}_\tau) \coloneqq \left\{\varphi \in L^2((0,t_\mathrm{F})) \;:\; \varphi_{|(t^n,t^{n-1})} \in H^1((t^{n-1},t^n)) \text{ for all } 1 \le n \le N\right\}.$$

Since each $\psi \in H^1((t^{n-1},t^n))$ has an absolutely continuous representative in $[t^{n-1},t^n]$, we can identify $\varphi \in H^1(\mathcal{T}_\tau)$ with a left continuous function in $(0,t_\mathrm{F})$. Therefore, for any vector space V and any $\varphi \in H^1(\mathcal{T}_\tau;V)$, we set $\varphi^0 \coloneqq \varphi(0)$ and, for all $1 \le n \le N$,

$$\varphi^n \coloneqq \lim_{t\to(t^n)^-} \varphi(t) \in V.$$

If $\varphi \in C^0(V)$, this simply amounts to setting $\varphi^n \coloneqq \varphi(t^n)$. For all $n \ge 1$ and $\psi \in L^1(V)$, we define the time average of ψ in (t^{n-1},t^n) as

$$\overline{\psi}^n \coloneqq \tau^{-1}\int_{t^{n-1}}^{t^n} \psi(t)\mathrm{d}t \in V, \tag{4.8}$$

with the convention that $\overline{\psi}^0 = 0 \in V$. We also let, for all $(\varphi^i)_{0\le i\le N} \in V^{N+1}$ and all $1 \le n \le N$,

$$\delta_t\varphi^n \coloneqq \frac{\varphi^n - \varphi^{n-1}}{\tau} \in V$$

denote the backward approximation of the first derivative of φ at time t^n.

We note a preliminary result that will be used in the convergence analysis of Section 4.6. Let $\psi \in H^1(\mathcal{T}_\tau)$ and $1 \le n \le N$. Identifying $\psi_{|(t^{n-1},t^n)} \in H^1((t^{n-1},t^n))$ with its absolutely continuous representative in $(t^{n-1},t^n]$, one can use the fundamental theorem of calculus to infer that

$$\psi^n - \overline{\psi}^n = \psi(t^n) - \frac{1}{\tau}\int_{t^{n-1}}^{t^n}\left(\psi(t^n) - \int_s^{t^n} \mathrm{d}_t\psi(t)\mathrm{d}t\right)\mathrm{d}s = \frac{1}{\tau}\int_{t^{n-1}}^{t^n}\int_s^{t^n}\mathrm{d}_t\psi(t)\mathrm{d}t\,\mathrm{d}s \le \int_{t^{n-1}}^{t^n}|\mathrm{d}_t\psi(t)|\mathrm{d}t.$$

Thus, applying the previous result together with the Jensen inequality yields, for all $\varphi \in H^1(\mathcal{T}_\tau;L^2(\Omega))$,

$$\|\varphi^n - \overline{\varphi}^n\|_\Omega^2 \le \int_\Omega\left(\int_{t^{n-1}}^{t^n}|\mathrm{d}_t\varphi(\boldsymbol{x},t)|\mathrm{d}t\right)^2\mathrm{d}\boldsymbol{x} \le \tau\int_{t^{n-1}}^{t^n}\|\mathrm{d}_t\varphi(t)\|_\Omega^2\mathrm{d}t \le \tau\|\varphi\|^2_{H^1((t^{n-1},t^n);L^2(\Omega))}, \tag{4.9}$$

where $\varphi(\boldsymbol{x},t)$ is a shorthand notation for $(\varphi(t))(\boldsymbol{x})$. As a result of (4.9), we get

$$\sum_{n=1}^N \tau\|\varphi^n - \overline{\varphi}^n\|_\Omega^2 \le \tau^2\sum_{n=1}^N\|\varphi\|^2_{H^1((t^{n-1},t^n);L^2(\Omega))} =: \tau^2\|\varphi\|^2_{H^1(\mathcal{T}_\tau;L^2(\Omega))}.$$

4.3.3 L^2-orthogonal projectors on local and broken polynomial spaces

For $X \subset \overline{\Omega}$ and $k \in \mathbb{N}$, we denote by $P^k(X)$ the space spanned by the restriction to X of scalar-valued, d-variate polynomials of total degree k. The L^2-projector $\pi_X^k : L^1(X) \to P^k(X)$ is defined such that, for all $v \in L^1(X)$,

$$\int_X(\pi_X^k v - v)w = 0 \qquad \forall w \in P^k(X). \tag{4.10}$$

As a projector, π_X^k is linear and idempotent so that, for all $v \in P^k(X)$, $\pi_X^k v = v$. When dealing with the vector-valued polynomial space $\mathbf{P}^k(X)$ or with the tensor-valued polynomial space $\mathbb{P}^k(X)$, we use the boldface notation $\boldsymbol{\pi}_X^k$ for the corresponding L^2-orthogonal projectors acting component-wise. At the global level, we denote by $P^k(\mathcal{T}_h)$, $\mathbf{P}^k(\mathcal{T}_h)$, and $\mathbb{P}^k(\mathcal{T}_h)$, respectively, the spaces of fully discontinuous scalar-valued, vector-valued, and tensor-valued broken polynomial functions on $\mathcal{T}_h$ of total degree k, and by π_h^k and $\boldsymbol{\pi}_h^k$ the L^2-projectors on $P^k(\mathcal{T}_h)$ and $\mathbf{P}^k(\mathcal{T}_h)$. The following optimal approximation properties for the L^2-projector π_X^k follow from [69, Lemmas 3.4 and 3.6]: There exists a strictly positive real number C_{ap} independent of h such that, for all $T \in \mathcal{T}_h$, all $l \in \{1, \ldots, k+1\}$, and all $v \in H^l(T)$,

$$|v - \pi_T^k v|_{H^m(T)} + h_T^{\frac{1}{2}} |v - \pi_T^k v|_{H^m(\mathcal{F}_T)} \le C_{\mathrm{ap}} h_T^{l-m} |v|_{H^l(T)} \qquad \forall m \in \{0, \ldots, l\}, \tag{4.11}$$

where $|\cdot|_{H^m(\mathcal{F}_T)}$ is the broken Sobolev seminorm on $\mathcal{F}_T$.

4.3.4 Discrete spaces

In this section we define the discrete spaces upon which the HHO method corresponding to a polynomial degree $k \ge 1$ is built.

4.3.4.1 Displacement

The discrete unknowns for the displacement are collected in the space

$$\underline{\boldsymbol{U}}_h^k := \left\{ \underline{\boldsymbol{v}}_h = ((\boldsymbol{v}_T)_{T\in\mathcal{T}_h}, (\boldsymbol{v}_F)_{F\in\mathcal{F}_h}) \ : \ \boldsymbol{v}_T \in \mathbf{P}^k(T) \ \forall T \in \mathcal{T}_h \text{ and } \boldsymbol{v}_F \in \mathbf{P}^k(F) \ \forall F \in \mathcal{F}_h \right\}.$$

For any $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$, we denote by $\boldsymbol{v}_h \in \mathbf{P}^k(\mathcal{T}_h)$ the broken polynomial vector field obtained patching element-based unknowns, so that

$$(\boldsymbol{v}_h)_{|T} = \boldsymbol{v}_T \qquad \forall T \in \mathcal{T}_h.$$

The discrete unknowns corresponding to a function $\boldsymbol{v} \in \mathbf{H}^1(\Omega)$ are obtained by means of the interpolator $\underline{\boldsymbol{I}}_h^k : \mathbf{H}^1(\Omega) \to \underline{\boldsymbol{U}}_h^k$ such that

$$\underline{\boldsymbol{I}}_h^k \boldsymbol{v} := ((\boldsymbol{\pi}_T^k \boldsymbol{v}_{|T})_{T\in\mathcal{T}_h}, (\boldsymbol{\pi}_F^k \boldsymbol{v}_{|F})_{F\in\mathcal{F}_h}). \tag{4.12}$$

For all $T \in \mathcal{T}_h$, we denote by $\underline{\boldsymbol{U}}_T^k$ and $\underline{\boldsymbol{I}}_T^k$ the restrictions to T of $\underline{\boldsymbol{U}}_h^k$ and $\underline{\boldsymbol{I}}_h^k$, respectively and, for any $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$, we let $\underline{\boldsymbol{v}}_T := (\boldsymbol{v}_T, (\boldsymbol{v}_F)_{F\in\mathcal{F}_T})$ collect the local discrete unknowns attached to T. At each time step, the displacement is sought in the following subspace of $\underline{\boldsymbol{U}}_h^k$ that strongly accounts for the homogeneous Dirichlet condition (4.1c):

$$\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k := \left\{ \underline{\boldsymbol{v}}_h = ((\boldsymbol{v}_T)_{T\in\mathcal{T}_h}, (\boldsymbol{v}_F)_{F\in\mathcal{F}_h}) \in \underline{\boldsymbol{U}}_h^k \ : \ \boldsymbol{v}_F = \boldsymbol{0} \quad \forall F \in \mathcal{F}_h^{\mathrm{b}} \right\}.$$

We next prove a discrete version of Korn's first inequality on $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ that will play a key role in the analysis. To this purpose, we endow the space $\underline{\boldsymbol{U}}_h^k$ with the discrete strain seminorm $\|\cdot\|_{\boldsymbol{\varepsilon},h}$ defined, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$, such that

$$\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h} := \left[\sum_{T\in\mathcal{T}_h} \left(\|\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}_T\|_T^2 + \sum_{F\in\mathcal{F}_T} \frac{\|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F^2}{h_F} \right) \right]^{\frac{1}{2}}. \tag{4.13}$$

We will also need the following continuous trace inequality, whose proof follows the arguments of [76, Lemma 1.49] (where a slightly different notion of mesh faces is considered): There exists a strictly positive real number C_{tr}, independent of h but possibly depending on ϱ, such that, for all $T \in \mathcal{T}_h$, all $\boldsymbol{v}_T \in \mathbf{H}^1(T)$, and all $F \in \mathcal{F}_T$,

$$\|\boldsymbol{v}_T\|_F^2 \le C_{\mathrm{tr}}^2 \left(\|\boldsymbol{\nabla}\boldsymbol{v}_T\|_T + h_T^{-1}\|\boldsymbol{v}_T\|_T\right) \|\boldsymbol{v}_T\|_T. \tag{4.14}$$

Proposition 4.3 (Discrete Korn's first inequality)**.** *There is a real number $C_{\mathrm{K}} > 0$, only depending on Ω, d, and ϱ such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$,*

$$\|\boldsymbol{v}_h\|_\Omega \le C_{\mathrm{K}} \|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}. \tag{4.15}$$

Remark 4.4 (Strain norm). An immediate consequence of (4.15) is that the map $\|\cdot\|_{\boldsymbol{\varepsilon},h}$ is a norm in $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$.

Proof. Let $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$. Since the divergence operator $\boldsymbol{\nabla}\cdot : \mathbb{H}^1_{\mathrm{s}}(\Omega) \to \mathbf{L}^2(\Omega)$ is onto (c.f. [31, 56]), there exists $\boldsymbol{\tau}_{\boldsymbol{v}_h} \in \mathbb{H}^1_{\mathrm{s}}(\Omega)$ such that $\boldsymbol{\nabla}\cdot\boldsymbol{\tau}_{\boldsymbol{v}_h} = \boldsymbol{v}_h$ and $\|\boldsymbol{\tau}_{\boldsymbol{v}_h}\|_{1,\Omega} \le C_{\mathrm{sj}}\|\boldsymbol{v}_h\|_\Omega$, with $C_{\mathrm{sj}} > 0$ independent of h. It follows that

$$\|\boldsymbol{v}_h\|_\Omega^2 = (\boldsymbol{v}_h, \boldsymbol{\nabla}\cdot\boldsymbol{\tau}_{\boldsymbol{v}_h})_\Omega = \sum_{T\in\mathcal{T}_h}\left(-\int_T \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_T : \boldsymbol{\tau}_{\boldsymbol{v}_h} + \sum_{F\in\mathcal{F}_T}\int_F (\boldsymbol{v}_T - \boldsymbol{v}_F)\cdot\boldsymbol{\tau}_{\boldsymbol{v}_h}\boldsymbol{n}_{TF}\right),$$

where we have integrated by parts element by element and used the fact that $\boldsymbol{\tau}_{\boldsymbol{v}_h}$ has continuous normal components across interfaces and that boundary unknowns are set to zero in order to insert $\boldsymbol{v}_F$ into the boundary term. Applying a Cauchy–Schwarz inequality on the integrals over the element, a generalized Hölder inequality with exponents $(2, 2, +\infty)$ on the integrals over faces, using the fact that $\|\boldsymbol{n}_{TF}\|_{\mathbf{L}^\infty(F)} \le 1$, and invoking a discrete Cauchy–Schwarz inequality on the sum over $T \in \mathcal{T}_h$, we infer that

$$\begin{aligned}
\|\boldsymbol{v}_h\|_\Omega^2 &\le \sum_{T\in\mathcal{T}_h}\left(\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_T\|_T\|\boldsymbol{\tau}_{\boldsymbol{v}_h}\|_T + \sum_{F\in\mathcal{F}_T}\frac{\|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F}{h_F^{\frac12}} h_F^{\frac12}\|\boldsymbol{\tau}_{\boldsymbol{v}_h}\|_F\right)\\
&\le \left(\sum_{T\in\mathcal{T}_h}\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}_T\|_T^2\right)^{\frac12}\|\boldsymbol{\tau}_{\boldsymbol{v}_h}\|_\Omega + \left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\frac{\|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F^2}{h_F}\right)^{\frac12}\left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} h_F\|\boldsymbol{\tau}_{\boldsymbol{v}_h}\|_F^2\right)^{\frac12}\\
&\le \sqrt{2N_\partial}C_{\mathrm{tr}}\left(\sum_{T\in\mathcal{T}_h}\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}\|_T^2 + \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}\frac{\|\boldsymbol{v}_F - \boldsymbol{v}_T\|_F^2}{h_F}\right)^{\frac12}\|\boldsymbol{\tau}_{\boldsymbol{v}_h}\|_{1,\Omega}\\
&= \sqrt{2N_\partial}C_{\mathrm{tr}}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\|\boldsymbol{\tau}_{\boldsymbol{v}_h}\|_{1,\Omega},
\end{aligned}$$

where, to pass to the third line, we have estimated $\|\boldsymbol{\tau}_{\boldsymbol{v}_h}\|_F$ using the continuous trace inequality (4.14) together with the fact that $h_F \le h_T \le \mathrm{diam}(\Omega) = 1$ for any $T \in \mathcal{T}_h$ and $F \in \mathcal{F}_T$. Thus, invoking the boundedness of the divergence operator, we get

$$\|\boldsymbol{v}_h\|_\Omega^2 \le \sqrt{2N_\partial}C_{\mathrm{sj}}C_{\mathrm{tr}}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\|\boldsymbol{v}_h\|_\Omega,$$

which yields the conclusion with $C_{\mathrm{K}} = \sqrt{2N_\partial}C_{\mathrm{sj}}C_{\mathrm{tr}}$. □

4.3.4.2 Pore pressure

At each time step, the discrete pore pressure is sought in the space

$$P_h^k := \begin{cases} P^k(\mathcal{T}_h) & \text{if } C_0 > 0, \\ P_0^k(\mathcal{T}_h) := \left\{ q_h \in P^k(\mathcal{T}_h) \ : \ \int_\Omega q_h = 0 \right\} & \text{if } C_0 = 0. \end{cases}$$

For any internal face $F \in \mathcal{F}_h^{\mathrm{i}}$, we denote by $T_{F,1}, T_{F,2} \in \mathcal{T}_h$ the two mesh elements that share F as a face, that is to say $F \subset \partial T_{F,1} \cap \partial T_{F,2}$ and $T_{F,1} \neq T_{F,2}$ (the ordering of the elements is arbitrary but fixed), and we set

$$\kappa_{F,i} := \left(\boldsymbol{\kappa}_{|T_{F,i}} \boldsymbol{n}_{T_{F,i}F} \right) \cdot \boldsymbol{n}_{T_{F,i}F} \quad \text{for } i \in \{1,2\}, \qquad \kappa_F := \frac{2\kappa_{F,1}\kappa_{F,2}}{\kappa_{F,1} + \kappa_{F,2}}. \tag{4.16}$$

For all $q_h \in P_h^k$, we denote by q_T the restriction of q_h to an element $T \in \mathcal{T}_h$ and we define the discrete seminorm

$$\|q_h\|_{\boldsymbol{\kappa},h} := \left(\sum_{T \in \mathcal{T}_h} \|\boldsymbol{\kappa}^{\frac{1}{2}} \boldsymbol{\nabla} q_T\|_T^2 + \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} \frac{\kappa_F}{h_F} \|q_{T_{F,1}} - q_{T_{F,2}}\|_F^2 \right)^{\frac{1}{2}}. \tag{4.17}$$

The fact that in (4.17) boundary terms only appear on internal faces reflects the homogeneous Neumann boundary condition (4.1d).

Using the surjectivity of the divergence operator $\boldsymbol{\nabla}\cdot : \mathbf{H}_0^1(\Omega) \to L_0^2(\Omega)$ and proceeding as in the proof of the discrete Korn inequality (4.15), a discrete Poincaré-Wirtinger inequality in $P^k(\mathcal{T}_h)$ is readily inferred, namely one has the existence of $C_{\mathrm{P}} > 0$, only depending on Ω, d, and ϱ such that, for all $q_h \in P^k(\mathcal{T}_h)$,

$$\|q_h - \pi_\Omega^0 q_h\|_\Omega \le C_{\mathrm{P}} \underline{\kappa}^{-\frac{1}{2}} \|q_h\|_{\boldsymbol{\kappa},h}.$$

This result ensures, in particular, that the seminorm $\|\cdot\|_{\boldsymbol{\kappa},h}$ defined in (4.17) is a norm on $P_0^k(\mathcal{T}_h)$. For a proof of more general Sobolev inequalities on broken polynomial spaces, we refer the reader to [75] and [76, Section 5.1.2].

4.4 Discretization

In this section we define the discrete counterparts of the elasticity, hydro-mechanical coupling, and Darcy operators, and formulate the HHO–dG scheme for problem (4.6).

4.4.1 Nonlinear elasticity operator

The discretization of the nonlinear elasticity operator closely follows the one presented in Section 3.4. We define the local symmetric gradient reconstruction $\boldsymbol{G}_{\mathrm{s},T}^k : \underline{\boldsymbol{U}}_T^k \to \mathbb{P}_{\mathrm{s}}^k(T)$ such that, for a given $\underline{\boldsymbol{v}}_T = (\boldsymbol{v}_T, (\boldsymbol{v}_F)_{F \in \mathcal{F}_T}) \in \underline{\boldsymbol{U}}_T^k$, $\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T \in \mathbb{P}_{\mathrm{s}}^k(T)$ solves

$$\int_T \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T : \boldsymbol{\tau} = -\int_T \boldsymbol{v}_T \cdot (\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) + \sum_{F \in \mathcal{F}_T} \int_F \boldsymbol{v}_F \cdot (\boldsymbol{\tau}\boldsymbol{n}_{TF}) \qquad \forall \boldsymbol{\tau} \in \mathbb{P}_{\mathrm{s}}^k(T). \tag{4.18}$$

Existence and uniqueness of $\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T$ follow from the Riesz representation theorem in $\mathbb{P}_{\mathrm{s}}^k(T)$ for the $L^2(T)^{d\times d}$-inner product. This definition is motivated by the following property.

Proposition 4.5 (Commuting property for the local symmetric gradient reconstruction)**.** *For all $\boldsymbol{v} \in \boldsymbol{H}^1(T)$, it holds that*

$$\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{I}}_T^k \boldsymbol{v} = \boldsymbol{\pi}_T^k(\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}). \tag{4.19}$$

Remark 4.6 (Approximation properties of the local symmetric gradient reconstruction). The commuting property (4.19) combined with (4.11) shows that $\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{I}}_T^k \boldsymbol{v}$ optimally approximates $\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}$ in $\mathbb{P}_{\mathrm{s}}^k(T)$.

Proof. We start by noticing that it holds, for all $\boldsymbol{\alpha} \in \mathbb{R}_{\mathrm{s}}^{d\times d}$ and all $\boldsymbol{\beta} \in \mathbb{R}^{d\times d}$, denoting by $\boldsymbol{\beta}_{\mathrm{s}} \coloneqq \frac{1}{2}(\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}})$ the symmetric part of $\boldsymbol{\beta}$,

$$\boldsymbol{\alpha} : \boldsymbol{\beta} = \boldsymbol{\alpha} : \boldsymbol{\beta}_{\mathrm{s}}. \tag{4.20}$$

For all $\boldsymbol{\tau} \in \mathbb{P}^k(T)$ we can then write

$$\begin{aligned}
\int_T \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{I}}_T^k \boldsymbol{v} : \boldsymbol{\tau} &= \int_T \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{I}}_T^k \boldsymbol{v} : \boldsymbol{\tau}_{\mathrm{s}} \\
&= -\int_T \boldsymbol{\pi}_T^k \boldsymbol{v} \cdot (\boldsymbol{\nabla}\cdot\boldsymbol{\tau}_{\mathrm{s}}) + \sum_{F\in\mathcal{F}_T} \int_F \boldsymbol{\pi}_F^k \boldsymbol{v} \cdot (\boldsymbol{\tau}_{\mathrm{s}} \boldsymbol{n}_{TF}) \\
&= -\int_T \boldsymbol{v} \cdot (\boldsymbol{\nabla}\cdot\boldsymbol{\tau}_{\mathrm{s}}) + \sum_{F\in\mathcal{F}_T} \int_F \boldsymbol{v} \cdot (\boldsymbol{\tau}_{\mathrm{s}} \boldsymbol{n}_{TF}) \\
&= \int_T \boldsymbol{\nabla}\boldsymbol{v} : \boldsymbol{\tau}_{\mathrm{s}} = \int_T \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v} : \boldsymbol{\tau}_{\mathrm{s}} = \int_T \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v} : \boldsymbol{\tau} = \int_T \boldsymbol{\pi}_T^k(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}) : \boldsymbol{\tau},
\end{aligned}$$

where we have used (4.20) with $\boldsymbol{\alpha} = \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{I}}_T^k \boldsymbol{v}$ and $\boldsymbol{\beta} = \boldsymbol{\tau}$ in the first line, the definition (4.18) of the local symmetric gradient with $\underline{\boldsymbol{v}}_T = \underline{\boldsymbol{I}}_T^k \boldsymbol{v}$ in the second line, and definition (4.10) after observing that $\boldsymbol{\nabla}\cdot\boldsymbol{\tau}_{\mathrm{s}} \in \mathbf{P}^{k-1}(T) \subset \mathbf{P}^k(T)$ and $\boldsymbol{\tau}_{\mathrm{s}} \boldsymbol{n}_{TF} \in \mathbf{P}^k(F)$ for all $F \in \mathcal{F}_T$ to remove the L^2-orthogonal projectors in the third line. In the fourth line, we have used an integration by parts, then invoked (4.20) first with $\boldsymbol{\alpha} = \boldsymbol{\tau}_{\mathrm{s}}$ and $\boldsymbol{\beta} = \boldsymbol{\nabla}\boldsymbol{v}$, then with $\boldsymbol{\alpha} = \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}$ and $\boldsymbol{\beta} = \boldsymbol{\tau}$, and we have used the definition (4.10) of $\boldsymbol{\pi}_T^k$ to conclude. □

From $\boldsymbol{G}_{\mathrm{s},T}^k$, we define the local displacement reconstruction operator $\boldsymbol{r}_T^{k+1} : \underline{\boldsymbol{U}}_T^k \to \mathbf{P}^{k+1}(T)$ such that, for all $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}_T^k$,

$$\begin{gathered}
\int_T (\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{r}_T^{k+1} \underline{\boldsymbol{v}}_T - \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T) : \boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{w} = 0 \qquad \forall \boldsymbol{w} \in \mathbf{P}^{k+1}(T), \\
\int_T \boldsymbol{r}_T^{k+1} \underline{\boldsymbol{v}}_T = \int_T \boldsymbol{v}_T, \qquad \int_T \boldsymbol{\nabla}\times(\boldsymbol{r}_T^{k+1} \underline{\boldsymbol{v}}_T) = \sum_{F\in\mathcal{F}_T} \int_F \boldsymbol{n}_{TF} \times \boldsymbol{v}_F,
\end{gathered}$$

where $\boldsymbol{\nabla}\times$ denotes the curl operator and $\boldsymbol{v} \times \boldsymbol{w}$ the vector product of two vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$. Optimal approximation properties for $\boldsymbol{r}_T^{k+1} \underline{\boldsymbol{I}}_T^k$ have been recently proved in [32, Appendix A]

generalizing the ones of [74, Lemma 2]. The optimal approximation properties of $\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{I}}_T^k$ are required to infer (4.46) below.

The discretization of the nonlinear elasticity operator is realized by the function $a_h : \underline{\boldsymbol{U}}_h^k \times \underline{\boldsymbol{U}}_h^k \to \mathbb{R}$ such that, for all $\underline{\boldsymbol{w}}_h, \underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$,

$$a_h(\underline{\boldsymbol{w}}_h, \underline{\boldsymbol{v}}_h) \coloneqq \sum_{T\in\mathcal{T}_h}\left(\int_T \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{u}}_T) : \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T + \sum_{F\in\mathcal{F}_T}\frac{\gamma}{h_F}\int_F \boldsymbol{\Delta}_{TF}^k\underline{\boldsymbol{u}}_T\cdot\boldsymbol{\Delta}_{TF}^k\underline{\boldsymbol{v}}_T\right), \tag{4.21}$$

where $\gamma > 0$ denotes a user-dependent parameter and we penalize in a least-square sense the face-based residual $\boldsymbol{\Delta}_{TF}^k : \underline{\boldsymbol{U}}_T^k \to \mathbf{P}^k(F)$ such that, for all $T \in \mathcal{T}_h$, all $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}_T^k$, and all $F \in \mathcal{F}_T$,

$$\boldsymbol{\Delta}_{TF}^k\underline{\boldsymbol{v}}_T \coloneqq \boldsymbol{\pi}_F^k(\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T - \boldsymbol{v}_F) - \boldsymbol{\pi}_T^k(\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T - \boldsymbol{v}_T).$$

This definition ensures that $\boldsymbol{\Delta}_{TF}^k$ vanishes whenever its argument is of the form $\underline{\boldsymbol{I}}_T^k\boldsymbol{w}$ with $\boldsymbol{w} \in \mathbf{P}^{k+1}(T)$, a crucial property to obtain high-order error estimates (cf. Theorems 2.12, 3.16). For further use, we note the following seminorm equivalence, which can be proved using the arguments of [74, Lemma 4]: For all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$,

$$C_{\mathrm{eq}}^{-2}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}^2 \le \sum_{T\in\mathcal{T}_h}\left(\|\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T\|_T^2 + \sum_{F\in\mathcal{F}_T}h_F^{-1}\|\boldsymbol{\Delta}_{TF}^k\underline{\boldsymbol{v}}_T\|_F^2\right) \le C_{\mathrm{eq}}^2\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}^2, \tag{4.22}$$

where $C_{\mathrm{eq}} > 0$ is independent of h, and the discrete strain seminorm $\|\cdot\|_{\boldsymbol{\varepsilon},h}$ is defined by (4.13). By (4.2b), this implies the coercivity of a_h.

Remark 4.7 (Choice of the stabilization parameter). The constants $C_{\mathrm{gr}}, C_{\mathrm{cv}}$ appearing in (4.2) satisfy $C_{\mathrm{cv}}^2 \le C_{\mathrm{gr}}$. Indeed, owing to (4.2b), the Cauchy–Schwarz inequality, and (4.2a), it holds for all $\boldsymbol{\tau} \in \mathbb{R}_{\mathrm{s}}^{d\times d}$,

$$C_{\mathrm{cv}}^2|\boldsymbol{\tau}|_{d\times d}^2 \le \boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau}) : \boldsymbol{\tau} \le |\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau})|_{d\times d}|\boldsymbol{\tau}|_{d\times d} \le C_{\mathrm{gr}}|\boldsymbol{\tau}|_{d\times d}^2. \tag{4.23}$$

Thus, we choose the stabilization parameter γ in (4.21) such that

$$\gamma \in [C_{\mathrm{cv}}^2, C_{\mathrm{gr}}]. \tag{4.24}$$

For the linear elasticity model (4.3), we have $C_{\mathrm{gr}} = 2\mu + d\lambda$ and $C_{\mathrm{cv}} = \sqrt{2\mu}$, so that a natural choice for the stabilization parameter is $\gamma = 2\mu$.

4.4.2 Hydro-mechanical coupling

The hydro-mechanical coupling is realized by means of the bilinear form b_h on $\underline{\boldsymbol{U}}_h^k \times P^k(\mathcal{T}_h)$ such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$ and all $q_h \in P^k(\mathcal{T}_h)$,

$$b_h(\underline{\boldsymbol{v}}_h, q_h) \coloneqq \sum_{T\in\mathcal{T}_h}\left(\int_T \boldsymbol{v}_T\cdot\boldsymbol{\nabla}q_T - \sum_{F\in\mathcal{F}_T}\int_F \boldsymbol{v}_F\cdot q_T\boldsymbol{n}_{TF}\right), \tag{4.25}$$

where we remind the reader that $q_T := q_{h|T}$. It can be checked using Cauchy–Schwarz inequalities together with the definition (4.13) of the discrete strain seminorm and discrete trace inequalities that there exists $C_{\rm bd} > 0$ independent of h such that

$$b_h(\underline{\boldsymbol{v}}_h, q_h) \le C_{\rm bd} \|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h} \|q_h\|_\Omega.$$

Additionally, using the strongly enforced boundary condition in $\underline{\boldsymbol{U}}_{h,\rm D}^k$, it can be proved that

$$b_h(\underline{\boldsymbol{v}}_h, 1) = 0, \qquad \forall \underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\rm D}^k. \tag{4.26}$$

Finally, we note the following lemma stating that the hybrid interpolator $\underline{\boldsymbol{I}}_h^k : \mathbf{H}^1(\Omega) \to \underline{\boldsymbol{U}}_h^k$ is a Fortin operator.

Lemma 4.8 (Fortin operator)**.** *For all $\boldsymbol{v} \in \boldsymbol{H}^1(\Omega)$ and all $q_h \in P^k(\mathcal{T}_h)$, the interpolator $\underline{\boldsymbol{I}}_h^k$ satisfies*

$$\|\underline{\boldsymbol{I}}_h^k \boldsymbol{v}\|_{\boldsymbol{\varepsilon},h} \le C_{\rm st} |\boldsymbol{v}|_{1,\Omega}, \tag{4.27a}$$

$$b_h(\underline{\boldsymbol{I}}_h^k \boldsymbol{v}, q_h) = b(\boldsymbol{v}, q_h), \tag{4.27b}$$

where the strictly positive real number $C_{\rm st}$ is independent of h.

Proof. (i) *Proof of* (4.27a). Recalling the definitions (4.13) of the discrete strain seminorm and (4.12) of the global interpolator, we can write

$$\begin{aligned}
\|\underline{\boldsymbol{I}}_h^k \boldsymbol{v}\|_{\boldsymbol{\varepsilon},h}^2 &= \sum_{T\in\mathcal{T}_h} \left(\|\boldsymbol{\nabla}_{\rm s} \boldsymbol{\pi}_T^k \boldsymbol{v}\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1} \|\boldsymbol{\pi}_F^k \boldsymbol{v} - \boldsymbol{\pi}_T^k \boldsymbol{v}\|_F^2 \right) \\
&\le \sum_{T\in\mathcal{T}_h} \left(2\|\boldsymbol{\nabla}_{\rm s}(\boldsymbol{\pi}_T^k \boldsymbol{v} - \boldsymbol{v})\|_T^2 + 2\|\boldsymbol{\nabla}_{\rm s} \boldsymbol{v}\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1} \|\boldsymbol{v} - \boldsymbol{\pi}_T^k \boldsymbol{v}\|_F^2 \right) \le C_{\rm st} |\boldsymbol{v}|_{1,\Omega},
\end{aligned}$$

with $C_{\rm st} > 0$ independent of h. To pass to the second line, we have used a triangle inequality after inserting $\pm\boldsymbol{\nabla}_{\rm s} \boldsymbol{v}$ into the first term, and we have used the linearity, idempotency, and boundedness of $\boldsymbol{\pi}_F^k$ to write $\|\boldsymbol{\pi}_F^k \boldsymbol{v} - \boldsymbol{\pi}_T^k \boldsymbol{v}\|_F = \|\boldsymbol{\pi}_F^k (\boldsymbol{v} - \boldsymbol{\pi}_T^k \boldsymbol{v})\|_F \le \|\boldsymbol{v} - \boldsymbol{\pi}_T^k \boldsymbol{v}\|_F$. To conclude, we have used (4.11) respectively with $l = m = 1$ and with $l = 1$ and $m = 0$ to bound the first and third terms inside the summation.

(ii) *Proof of* (4.27b). Recalling the definitions (4.25) of $b_h(\cdot,\cdot)$ and (4.12) of the global interpolator, we can write for all $q_h \in P^k(\mathcal{T}_h)$,

$$\begin{aligned}
b_h(\underline{\boldsymbol{I}}_h^k \boldsymbol{v}, q_h) &= \sum_{T\in\mathcal{T}_h} \left(\int_T \boldsymbol{\pi}_T^k \boldsymbol{v} \cdot \boldsymbol{\nabla} q_T - \int_F \boldsymbol{\pi}_F^k \boldsymbol{v}_{|F} \cdot q_T \boldsymbol{n}_{TF} \right) \\
&= \sum_{T\in\mathcal{T}_h} \left(\int_T \boldsymbol{v} \cdot \boldsymbol{\nabla} q_T - \int_F \boldsymbol{v} \cdot q_T \boldsymbol{n}_{TF} \right) = b(\boldsymbol{v}, q_h),
\end{aligned}$$

where we have used definition (4.10) after observing that $\boldsymbol{\nabla} q_T \in \mathbf{P}^{k-1}(T) \subset \mathbf{P}^k(T)$ and $q_{T|F} \boldsymbol{n}_{TF} \in \mathbf{P}^k(F)$ to remove the L^2-orthogonal projectors in the second line, and integration by parts over $T \in \mathcal{T}_h$ to conclude. □

As a result of the previous Lemma, one has the following inf-sup condition, cf. Proposition A.1 for the proof.

Proposition 4.9. *There is a strictly positive real number β independent of h such that, for all $q_h \in P_0^k(\mathcal{T}_h)$,*

$$\|q_h\|_\Omega \le \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\underline{\boldsymbol{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, q_h)}{\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}}. \tag{4.28}$$

4.4.3 Darcy operator

The discretization of the Darcy operator is based on the Symmetric Weighted Interior Penalty method of [77], cf. also [76, Section 4.5]. For all $F \in \mathcal{F}_h^{\mathrm{i}}$ and all $q_h \in P^k(\mathcal{T}_h)$, we define the jump and weighted average operators such that

$$[q_h]_F \coloneqq q_{T_{F,1}} - q_{T_{F,2}}, \qquad \{q_h\}_F \coloneqq \frac{\sqrt{\kappa_{F,2}}}{\sqrt{\kappa_{F,1}} + \sqrt{\kappa_{F,2}}} q_{T_{F,1}} + \frac{\sqrt{\kappa_{F,1}}}{\sqrt{\kappa_{F,1}} + \sqrt{\kappa_{F,2}}} q_{T_{F,2}},$$

with $T_{F,1}, T_{F,2} \in \mathcal{T}_h$, $T_{F,1} \neq T_{F,2}$, such that $F \subset \partial T_{F,1} \cap \partial T_{F,2}$ and $\kappa_{F,1}, \kappa_{F,2}$ defined in (4.16). The bilinear form c_h on $P^k(\mathcal{T}_h) \times P^k(\mathcal{T}_h)$ is defined such that, for all $q_h, r_h \in P^k(\mathcal{T}_h)$,

$$\begin{aligned} c_h(r_h, q_h) \coloneqq & \int_\Omega \boldsymbol{\kappa} \boldsymbol{\nabla}_h r_h \cdot \boldsymbol{\nabla}_h q_h + \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} \frac{\varsigma \kappa_F}{h_F} \int_F [r_h]_F [q_h]_F \\ & - \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} \int_F \left([r_h]_F \{\boldsymbol{\kappa} \boldsymbol{\nabla}_h q_h\}_F + [q_h]_F \{\boldsymbol{\kappa} \boldsymbol{\nabla}_h r_h\}_F \right) \cdot \boldsymbol{n}_{T_1 F}, \end{aligned}$$

where we have introduced the broken gradient operator $\boldsymbol{\nabla}_h$ on $\mathcal{T}_h$ and we have denoted by $\varsigma > \underline{\varsigma} > 0$ a user-defined penalty parameter chosen large enough to ensure the coercivity of c_h (the proof is similar to [76, Lemma 4.51]):

$$c_h(q_h, q_h) \ge (\varsigma - \underline{\varsigma})(1 + \varsigma)^{-1} \|q_h\|_{\boldsymbol{\kappa},h}^2, \qquad \forall q_h \in P_h^k.$$

Since, under this condition, c_h is a symmetric positive definite bilinear form on the broken polynomial space P_h^k, we can define an associated norm by setting $\|\cdot\|_{c,h} \coloneqq c_h(\cdot,\cdot)^{\frac{1}{2}}$.

The following consistency result can be proved adapting the arguments of [76, Chapter 4] to homogeneous Neumann boundary conditions and will be instrumental for the analysis. We define the functional spaces $P_* \coloneqq \left\{ r \in H^1(\Omega) \cap H^2(P_\Omega) \ : \ \boldsymbol{\kappa} \boldsymbol{\nabla} r \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega \right\}$ and set $P_{*h}^k \coloneqq P_* + P_h^k$. Extending the bilinear form c_h to $P_{*h}^k \times P_{*h}^k$, it is inferred that, for all $r \in P_*$,

$$-(\boldsymbol{\nabla} \cdot (\boldsymbol{\kappa} \boldsymbol{\nabla} r), q)_\Omega = c_h(r, q) \qquad \forall q \in P_{*h}. \tag{4.29}$$

4.4.4 Discrete problem

For all $1 \le n \le N$, the discrete solution $(\underline{\boldsymbol{u}}_h^n, p_h^n) \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times P_h^k$ at time t^n is such that, for all $(\underline{\boldsymbol{v}}_h, q_h) \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times P^k(\mathcal{T}_h)$,

$$a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{v}}_h) + b_h(\underline{\boldsymbol{v}}_h, p_h^n) = (\overline{\boldsymbol{f}}^n, \boldsymbol{v}_h)_\Omega, \tag{4.30a}$$

$$C_0(\delta_t p_h^n, q_h)_\Omega - b_h(\delta_t \underline{\boldsymbol{u}}_h^n, q_h) + c_h(p_h^n, q_h) = (\overline{g}^n, q_h)_\Omega, \tag{4.30b}$$

with $\overline{\boldsymbol{f}}^n \in \mathbf{L}^2(\Omega)$ and $\overline{g}^n \in L^2(\Omega)$ defined in Section 4.3.2. In order to start the time-stepping scheme we need to initialize the discrete fluid content. This is done by setting ϕ_h^0 equal to the L^2-orthogonal projection of ϕ^0 on $P^k(\mathcal{T}_h)$ according to (4.6c). This implies in particular that

$$C_0(p_h^0, q_h)_\Omega - b_h(\underline{\boldsymbol{u}}_h^0, q_h) \coloneqq (\phi^0, q_h)_\Omega \qquad \forall q_h \in P^k(\mathcal{T}_h). \tag{4.30c}$$

Remark 4.10 (Advance in time scheme). The modified backward Euler scheme obtained by taking time averages instead of pointwise evaluation of the right-hand sides in (4.30) can be interpreted as a low-order discontinuous Galerkin time-stepping method, cf. [179, 183].

Notice that other time discretizations could be used, but we have decided to focus on the backward Euler scheme to keep the proofs as simple as possible. From the practical point of view, at each time step n, the discrete nonlinear system (4.30) can be solved by the Newton method using as initial guess the solution at step $(n-1)$. The size of the linear system to be solved at each Newton iteration can be reduced by statically condensing a large part of the unknowns as described in Section 2.5.

4.5 Stability and well-posedness

In this section we study the stability of problem (4.30) and prove its well-posedness. We start with an a priori estimate on the discrete solutions not requiring conditions on the time step τ and robust with respect to vanishing storage coefficients and small permeability fields.

Proposition 4.11 (A priori estimate)**.** *Denote by $(\underline{\boldsymbol{u}}_h^n, p_h^n)_{1\le n\le N}$ the solution to (4.30). Under Assumption 4.1 on the stress-strain relation and the regularity on the data $\boldsymbol{f}$, g, and ϕ^0 assumed in Section 4.2.3, it holds*

$$\sum_{n=1}^N \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + \sum_{n=1}^N \tau \left(\|p_h^n - \pi_\Omega^0 p_h^n\|_\Omega^2 + C_0 \|p_h^n\|_\Omega^2 \right) + \|s_h^N\|_{c,h}^2 \le$$
$$C \left(\|\boldsymbol{f}\|_{L^2(\boldsymbol{L}^2(\Omega))}^2 + t_{\mathrm{F}}^2 \|g\|_{L^2(L^2(\Omega))}^2 + t_{\mathrm{F}} \|\phi^0\|_\Omega^2 + \frac{t_{\mathrm{F}}^2}{C_0} \|\pi_\Omega^0 g\|_{L^2(L^2(\Omega))}^2 + \frac{t_{\mathrm{F}}}{C_0} \|\pi_\Omega^0 \phi^0\|_\Omega^2 . \right). \tag{4.31}$$

where $C > 0$ denotes a real number independent of h, τ, the physical parameters C_0 and $\boldsymbol{\kappa}$, and the final time t_{F}. In (4.31), we have defined $s_h^N \coloneqq \sum_{n=1}^N \tau p_h^n$ and we have adopted the convention that $C_0^{-1} \|\pi_\Omega^0 g\|_{L^2(L^2(\Omega))}^2 = 0$ and $C_0^{-1} \|\pi_\Omega^0 \phi^0\|_\Omega^2 = 0$ if $C_0 = 0$.

Remark 4.12. In order to prove the a priori bound (4.31), no additional time regularity assumption on the loading term $\boldsymbol{f}$ and the mass source g are needed, whereas the stability estimate of Lemma 2.7, valid for linear stress-strain relation, requires $\boldsymbol{f} \in C^1(\mathbf{L}^2(\Omega))$ and $g \in C^0(L^2(\Omega))$. On the other hand, Proposition 4.11 gives an estimate of the discrete displacement and pressure in the L^2-norm in time, while Lemma 2.7 ensures a control in the L^∞-norm in time. However, under additional requirements on the stress-strain law (for instance Assumption 4.15) and H^1-regularity in time of $\boldsymbol{f}$, a stronger version of (4.31) can be inferred, establishing in particular an estimate in the L^∞-norm in time.

Proof. (i) *Estimate of* $\|p_h^n - \pi_\Omega^0 p_h^n\|_\Omega$. The growth property of the stress-strain function (4.2a) together with the Cauchy–Schwarz inequality, assumption (4.24), and the second inequality in (4.22) yield, for all $1 \le n \le N$ and all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$,

$$\begin{aligned} a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{v}}_h) &= \sum_{T\in\mathcal{T}_h}\left(\int_T \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{u}}_T^n) : \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T + \sum_{F\in\mathcal{F}_T} \frac{\gamma}{h_F}\int_F \boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{u}}_T^n \cdot \boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{v}}_T\right) \\ &\le C_{\mathrm{gr}} \sum_{T\in\mathcal{T}_h}\left(\|\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{u}}_T^n\|_T \|\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T\|_T + \sum_{F\in\mathcal{F}_T}\frac{1}{h_F}\|\boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{u}}_T^n\|_F \|\boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{v}}_T\|_F\right) \\ &\le C_{\mathrm{gr}} C_{\mathrm{eq}}^2 \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h} \|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}. \end{aligned} \tag{4.32}$$

Using the inf-sup condition (4.28), (4.26), and the mechanical equilibrium equation (4.30a), we get, for any $1 \le n \le N$,

$$\|p_h^n - \pi_\Omega^0 p_h^n\|_\Omega \le \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\underline{\boldsymbol{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, p_h^n - \pi_\Omega^0 p_h^n)}{\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}} = \beta \sup_{\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \setminus \{\underline{\boldsymbol{0}}\}} \frac{(\overline{\boldsymbol{f}}^n, \boldsymbol{v}_h)_\Omega - a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{v}}_h)}{\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}}.$$

Therefore, owing to the discrete Korn inequality (4.15) and to (4.32), we infer from the previous bound that

$$\|p_h^n - \pi_\Omega^0 p_h^n\|_\Omega \le \beta\left(C_{\mathrm{K}}\|\overline{\boldsymbol{f}}^n\|_\Omega + C_{\mathrm{gr}} C_{\mathrm{eq}}^2 \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}\right). \tag{4.33}$$

(ii) *Energy balance.* For all $1 \le n \le N$, summing (4.30b) at times $1 \le i \le n$ and taking $q_h = \tau^2 p_h^n$ as a test function yields

$$\tau C_0(p_h^n, p_h^n)_\Omega - \tau b_h(\underline{\boldsymbol{u}}_h^n, p_h^n) + \sum_{i=1}^n \tau^2 c_h(p_h^i, p_h^n) = \sum_{i=1}^n \tau^2(\overline{g}^i, p_h^n)_\Omega + \tau(\phi^0, p_h^n)_\Omega. \tag{4.34}$$

Moreover, using the linearity of c_h and the formula $2x(x-y) = x^2 + (x-y)^2 - y^2$, the third term in the left-hand side of (4.34) can be rewritten as

$$\sum_{i=1}^n \tau^2 c_h(p_h^i, p_h^n) = \tau c_h\left(\sum_{i=1}^n \tau p_h^i, p_h^n\right) = \tau c_h(s_h^n, \delta_t s_h^n) = \frac{1}{2}\left(\|s_h^n\|_{c,h}^2 + \|\delta_t s_h^n\|_{c,h}^2 - \|s_h^{n-1}\|_{c,h}^2\right),$$

where we have set $s_h^0 := 0$, $s_h^n := \sum_{i=1}^n \tau p_h^i$ for any $1 \le n \le N$, and observed that $p_h^n = \delta_t s_h^n$. Therefore, summing (4.34) and (4.30a) at discrete time n with $\underline{\boldsymbol{v}}_h = \tau \underline{\boldsymbol{u}}_h^n$, leads to

$$\tau a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{u}}_h^n) + \tau C_0 \|p_h^n\|_\Omega^2 + \frac{1}{2}\left(\|s_h^n\|_{c,h}^2 - \|s_h^{n-1}\|_{c,h}^2\right) \le \tau(\overline{\boldsymbol{f}}^n, \boldsymbol{u}_h^n)_\Omega + \sum_{i=1}^n \tau^2(\overline{g}^i, p_h^n)_\Omega + \tau(\phi^0, p_h^n)_\Omega.$$

Summing the previous relation for $1 \le n \le N$, telescoping out the appropriate summands, and using the coercivity property (4.2b), assumption (4.24), and the first inequality in (4.22), we get

$$\frac{C_{\mathrm{cv}}^2}{C_{\mathrm{eq}}^2}\sum_{n=1}^N \tau\|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + C_0\sum_{n=1}^N \tau\|p_h^n\|_\Omega^2 + \frac{1}{2}\|s_h^N\|_{c,h}^2 \le \sum_{n=1}^N \tau(\overline{\boldsymbol{f}}^n, \boldsymbol{u}_h^n)_\Omega + \sum_{n=1}^N \tau(G^n + \phi^0, p_h^n)_\Omega, \tag{4.35}$$

with the notation $G^n \coloneqq \sum_{i=1}^n \tau \overline{g}^i = \int_0^{t^n} g(t)\mathrm{d}t$. We denote by $\mathcal{R}$ the right-hand side of (4.35) and proceed to find a suitable upper bound.

(iii) *Upper bound for* $\mathcal{R}$. For the first term in the right-hand side of (4.35), using the Cauchy–Schwarz, discrete Korn (4.15), and Young inequalities, we obtain

$$\begin{aligned}\sum_{n=1}^N \tau(\overline{\boldsymbol{f}}^n, \boldsymbol{u}_h^n)_\Omega &\le C_{\mathrm{K}} \left(\sum_{n=1}^N \tau \|\overline{\boldsymbol{f}}^n\|_\Omega^2\right)^{\frac{1}{2}} \left(\sum_{n=1}^N \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2\right)^{\frac{1}{2}} \\ &\le \frac{C_{\mathrm{K}}^2 C_{\mathrm{eq}}^2}{C_{\mathrm{cv}}^2} \|\boldsymbol{f}\|_{L^2(\mathbf{L}^2(\Omega))}^2 + \frac{C_{\mathrm{cv}}^2}{4C_{\mathrm{eq}}^2} \sum_{n=1}^N \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2,\end{aligned} \tag{4.36}$$

where we have used the Jensen inequality to infer that

$$\sum_{n=1}^N \tau \|\overline{\boldsymbol{f}}^n\|_\Omega^2 = \sum_{n=1}^N \frac{1}{\tau} \int_\Omega \left(\int_{t^{n-1}}^{t^n} \boldsymbol{f}(\boldsymbol{x},t)\mathrm{d}t\right)^2 \mathrm{d}\boldsymbol{x} \le \sum_{n=1}^N \int_{t^{n-1}}^{t^n} \|\boldsymbol{f}(t)\|_\Omega^2 \mathrm{d}t = \|\boldsymbol{f}\|_{L^2(\mathbf{L}^2(\Omega))}^2.$$

We estimate the second term in the right-hand side of (4.35) by splitting it into two contributions as follows:

$$\sum_{n=1}^N \tau(G^n + \phi^0, p_h^n)_\Omega = \sum_{n=1}^N \tau(G^n + \phi^0, p_h^n - \pi_\Omega^0 p_h^n)_\Omega + \sum_{n=1}^N \tau(\pi_\Omega^0(G^n + \phi^0), p_h^n)_\Omega \coloneqq \mathfrak{T}_1 + \mathfrak{T}_2,$$

where we have used its definition (4.10) to move π_Ω^0 from p_h^n to $(G^n + \phi^0)$ in the second term. Owing to the Cauchy–Schwarz, triangle, Jensen, and Young inequalities, and using (4.33), we have

$$\begin{aligned}|\mathfrak{T}_1| &\le \left(\sum_{n=1}^N \tau \|G^n + \phi^0\|_\Omega^2\right)^{\frac{1}{2}} \left(\sum_{n=1}^N \tau \|p_h^n - \pi_\Omega^0 p_h^n\|_\Omega^2\right)^{\frac{1}{2}} \\ &\le \sqrt{2}\beta \left(\sum_{n=1}^N \tau \int_\Omega \left(\int_0^{t^n} g(\boldsymbol{x},t)\mathrm{d}t\right)^2 \mathrm{d}\boldsymbol{x} + t_{\mathrm{F}} \|\phi^0\|_\Omega^2\right)^{\frac{1}{2}} \left(\sum_{n=1}^N \tau \left(C_{\mathrm{K}} \|\overline{\boldsymbol{f}}^n\|_\Omega + C_{\mathrm{gr}} C_{\mathrm{eq}}^2 \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}\right)^2\right)^{\frac{1}{2}} \\ &\le 2\beta \left(t_{\mathrm{F}} \|g\|_{L^2(L^2(\Omega))} + t_{\mathrm{F}}^{\frac{1}{2}} \|\phi^0\|_\Omega\right) \left[C_{\mathrm{K}} \|\boldsymbol{f}\|_{L^2(\mathbf{L}^2(\Omega))} + C_{\mathrm{gr}} C_{\mathrm{eq}}^2 \left(\sum_{n=1}^N \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2\right)^{\frac{1}{2}}\right] \\ &\le t_{\mathrm{F}} \beta^2 \left(1 + \frac{C_{\mathrm{gr}}^2 C_{\mathrm{eq}}^6}{C_{\mathrm{cv}}^2}\right) \left(t_{\mathrm{F}}^{\frac{1}{2}} \|g\|_{L^2(L^2(\Omega))} + \|\phi^0\|_\Omega\right)^2 + C_{\mathrm{K}}^2 \|\boldsymbol{f}\|_{L^2(\mathbf{L}^2(\Omega))}^2 + \frac{C_{\mathrm{cv}}^2}{4C_{\mathrm{eq}}^2} \sum_{n=1}^N \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2.\end{aligned} \tag{4.37}$$

Owing to the compatibility condition (4.1f) and the linearity of the L^2-projector, $\mathfrak{T}_2 = 0$ if $C_0 = 0$. Otherwise, using again the Cauchy–Schwarz, triangle, Jensen, and Young

inequalities, leads to

$$
\begin{aligned}
|\mathfrak{T}_2| &\le \left(\sum_{n=1}^{N} \tau \|\pi_\Omega^0 G^n + \pi_\Omega^0 \phi^0\|_\Omega^2\right)^{\frac{1}{2}} \left(\sum_{n=1}^{N} \tau \|p_h^n\|_\Omega^2\right)^{\frac{1}{2}} \\
&\le \sqrt{2}\left(t_{\mathrm{F}}^2 \|\pi_\Omega^0 g\|_{L^2(L^2(\Omega))}^2 + t_{\mathrm{F}} \|\pi_\Omega^0 \phi^0\|_\Omega^2\right)^{\frac{1}{2}} \left(\sum_{n=1}^{N} \tau \|p_h^n\|_\Omega^2\right)^{\frac{1}{2}} \\
&\le \frac{2t_{\mathrm{F}}^2}{3C_0} \|\pi_\Omega^0 g\|_{L^2(L^2(\Omega))}^2 + \frac{2t_{\mathrm{F}}}{3C_0} \|\pi_\Omega^0 \phi^0\|_\Omega^2 + \frac{3C_0}{4} \sum_{n=1}^{N} \tau \|p_h^n\|_\Omega^2.
\end{aligned}
\tag{4.38}
$$

Finally, from (4.36), (4.37), (4.38), it follows that

$$
\begin{aligned}
\mathcal{R} \le{}& \frac{C_{\mathrm{cv}}^2}{2C_{\mathrm{eq}}^2} \sum_{n=1}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + \frac{3C_0}{4} \sum_{n=1}^{N} \tau \|p_h^n\|_\Omega^2 + \frac{2t_{\mathrm{F}}^2}{3C_0} \|\pi_\Omega^0 g\|_{L^2(L^2(\Omega))}^2 + \frac{2t_{\mathrm{F}}}{3C_0} \|\pi_\Omega^0 \phi^0\|_\Omega^2 \\
&+ \frac{C_{\mathrm{K}}^2}{C_{\mathrm{cv}}^2} \left(C_{\mathrm{eq}}^2 + C_{\mathrm{cv}}^2\right) \|\boldsymbol{f}\|_{L^2(\mathbf{L}^2(\Omega))}^2 + \frac{t_{\mathrm{F}} \beta^2}{C_{\mathrm{cv}}^2} \left(4C_{\mathrm{gr}}^2 C_{\mathrm{eq}}^6 + C_{\mathrm{cv}}^2\right) \left(t_{\mathrm{F}}^{\frac{1}{2}} \|g\|_{L^2(L^2(\Omega))} + \|\phi^0\|_\Omega\right)^2.
\end{aligned}
\tag{4.39}
$$

(iv) *Conclusion.* Passing the first two terms in the right-hand side of (4.39) to the left-hand side of (4.35) and multiplying both sides by a factor 4, we obtain

$$
\frac{2C_{\mathrm{cv}}^2}{C_{\mathrm{eq}}^2} \sum_{n=1}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + C_0 \sum_{n=1}^{N} \tau \|p_h^n\|_\Omega^2 + 2\|s_h^N\|_{c,h}^2 \le 4C,
\tag{4.40}
$$

where we have denoted by C the last four summands in the right-hand side of (4.39). In order to conclude we apply again (4.33) to obtain a bound of the L^2-norm of the discrete pressure independent of the storage coefficient C_0. Indeed, owing to (4.33) and $C_{\mathrm{cv}}^2 \le C_{\mathrm{gr}}$, it is inferred that

$$
\frac{C_{\mathrm{cv}}^2}{2C_{\mathrm{gr}}^2 C_{\mathrm{eq}}^6} \sum_{n=1}^{N} \tau \|p_h^n - \pi_\Omega^0 p_h^n\|_\Omega^2 \le \frac{C_{\mathrm{cv}}^2}{C_{\mathrm{eq}}^2} \sum_{n=1}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + \frac{\beta^2 C_{\mathrm{K}}^2}{C_{\mathrm{cv}}^2 C_{\mathrm{eq}}^6} \|\boldsymbol{f}\|_{L^2(\mathbf{L}^2(\Omega))}^2.
$$

Summing the previous relation to (4.40) yields

$$
\begin{aligned}
&\frac{C_{\mathrm{cv}}^2}{C_{\mathrm{eq}}^2} \sum_{n=1}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + \frac{C_{\mathrm{cv}}^2}{2C_{\mathrm{gr}}^2 C_{\mathrm{eq}}^6} \sum_{n=1}^{N} \tau \|p_h^n - \pi_\Omega^0 p_h^n\|_\Omega^2 + C_0 \sum_{n=1}^{N} \tau \|p_h^n\|_\Omega^2 + 2\|s_h^N\|_{c,h}^2 \le \\
&\qquad \frac{8t_{\mathrm{F}}^2}{3C_0} \|\pi_\Omega^0 g\|_{L^2(L^2(\Omega))}^2 + \frac{8t_{\mathrm{F}}}{3C_0} \|\pi_\Omega^0 \phi^0\|_\Omega^2 + C_{\mathrm{K}}^2 \left(\frac{4C_{\mathrm{eq}}^2}{C_{\mathrm{cv}}^2} + \frac{\beta^2}{C_{\mathrm{eq}}^6 C_{\mathrm{cv}}^2} + 4\right) \|\boldsymbol{f}\|_{L^2(\mathbf{L}^2(\Omega))}^2 \\
&\qquad\qquad + 4t_{\mathrm{F}} \beta^2 \left(\frac{4C_{\mathrm{gr}}^2 C_{\mathrm{eq}}^6}{C_{\mathrm{cv}}^2} + 1\right) \left(t_{\mathrm{F}}^{\frac{1}{2}} \|g\|_{L^2(L^2(\Omega))} + \|\phi^0\|_\Omega\right)^2.
\end{aligned}
$$

Thus, multiplying both sides of the previous relation by $\max\{C_{\mathrm{eq}}^2 C_{\mathrm{cv}}^{-2}, 2C_{\mathrm{gr}}^2 C_{\mathrm{eq}}^6 C_{\mathrm{cv}}^{-2}, 1\}$ gives (4.31). □

Remark 4.13 (A priori bound for $C_0 = 0$). When $C_0 = 0$, the a priori bound (4.31) reads

$$\sum_{n=1}^{N} \tau \|\underline{\boldsymbol{u}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + \sum_{n=1}^{N} \tau \|p_h^n\|_\Omega^2 + \|s_h^N\|_{c,h}^2 \le C \left(\|\boldsymbol{f}\|_{L^2(\mathbf{L}^2(\Omega))}^2 + t_{\mathrm{F}}^2 \|g\|_{L^2(L^2(\Omega))}^2 + t_{\mathrm{F}} \|\phi^0\|_\Omega^2 \right).$$

The conventions $C_0^{-1}\|\pi_\Omega^0 g\|_{L^2(L^2(\Omega))}^2 = 0$ and $C_0^{-1}\|\pi_\Omega^0 \phi^0\|_\Omega^2 = 0$ if $C_0 = 0$ are justified since the term $\mathfrak{T}_2$ in point (3) of the previous proof vanishes in this case thanks to the compatibility condition (4.1f).

We next proceed to discuss the existence and uniqueness of the discrete solutions. The proof of the following theorem hinges on the arguments of [64, Theorem 3.3].

Theorem 4.14 (Existence and uniqueness)**.** *Let Assumption 4.1 hold and let $(\mathcal{M}_h)_{h\in\mathcal{H}}$ be a regular mesh sequence. Then, for all $h \in \mathcal{H}$ and all $N \in \mathbb{N}^*$, there exists a unique solution $(\underline{\boldsymbol{u}}_h^n, p_h^n)_{1\le n\le N} \in (\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times P_h^k)^N$ to (4.30).*

Proof. We define the linear stress-strain function $\boldsymbol{\sigma}^{\mathrm{lin}} : \mathbb{R}_{\mathrm{s}}^{d\times d} \to \mathbb{R}_{\mathrm{s}}^{d\times d}$ such that, for all $\boldsymbol{\tau} \in \mathbb{R}_{\mathrm{s}}^{d\times d}$,

$$\boldsymbol{\sigma}^{\mathrm{lin}}(\boldsymbol{\tau}) = \frac{C_{\mathrm{cv}}^2}{2}\boldsymbol{\tau} + \frac{C_{\mathrm{cv}}^2}{2d}\operatorname{tr}(\boldsymbol{\tau})\boldsymbol{I}_d,$$

where C_{cv} is the coercivity constant of $\boldsymbol{\sigma}$ (see (4.2b)), and we denote by a_h^{lin} the bilinear form obtained by replacing $\boldsymbol{\sigma}$ with $\boldsymbol{\sigma}^{\mathrm{lin}}$ in (4.21). We consider the following auxiliary linear problem:c4: For all $1 \le n \le N$, find $(\underline{\boldsymbol{y}}_h^n, p_h^n) \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times P_h^k$ such that

$$\begin{aligned} a_h^{\mathrm{lin}}(\underline{\boldsymbol{y}}_h^n, \underline{\boldsymbol{v}}_h) + b_h(\underline{\boldsymbol{v}}_h, p_h^n) &= (\overline{\boldsymbol{f}}^n, \boldsymbol{v}_h)_\Omega \qquad \forall \underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k, \\ C_0(\delta_t p_h^n, q_h)_\Omega - b_h(\delta_t \underline{\boldsymbol{y}}_h^n, q_h) + c_h(p_h^n, q_h) &= (\overline{g}^n, q_h)_\Omega \qquad \forall q_h \in P_h^k, \end{aligned} \tag{4.41}$$

with initial condition as in (4.30c). Since the previous system is linear and square and its solution satisfies the a priori estimate of Proposition 4.11, it is readily inferred that problem (4.41) admits a unique solution.

Now we observe that, thanks to the norm equivalence (4.22), $a_h^{\mathrm{lin}}(\cdot,\cdot)$ is a scalar product on $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, and we define the mapping $\Phi : \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \to \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$,

$$a_h^{\mathrm{lin}}(\Phi(\underline{\boldsymbol{v}}_h), \underline{\boldsymbol{w}}_h) = a_h(\underline{\boldsymbol{v}}_h, \underline{\boldsymbol{w}}_h), \qquad \forall \underline{\boldsymbol{w}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k.$$

We want to show that Φ is an isomorphism. Let $\underline{\boldsymbol{v}}_h, \underline{\boldsymbol{z}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ be such that $\Phi(\underline{\boldsymbol{v}}_h) = \Phi(\underline{\boldsymbol{z}}_h)$. If $\underline{\boldsymbol{v}}_h \ne \underline{\boldsymbol{z}}_h$, owing to the norm equivalence (4.22) and the fact that $\|\cdot\|_{\boldsymbol{\varepsilon},h}$ is a norm on $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, there is at least one $T \in \mathcal{T}_h$ such that $\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T \ne \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{z}}_T$ or $\boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{v}}_T \ne \boldsymbol{\Delta}_{TF}^k \underline{\boldsymbol{z}}_T$ for some $F \in \mathcal{F}_T$. In both cases, owing to the definition of a_h and the strict monotonicity assumption (4.2c), it holds

$$0 < a_h(\underline{\boldsymbol{v}}_h, \underline{\boldsymbol{v}}_h - \underline{\boldsymbol{z}}_h) - a_h(\underline{\boldsymbol{z}}_h, \underline{\boldsymbol{v}}_h - \underline{\boldsymbol{z}}_h) = a_h^{\mathrm{lin}}(\Phi(\underline{\boldsymbol{v}}_h) - \Phi(\underline{\boldsymbol{z}}_h), \underline{\boldsymbol{v}}_h - \underline{\boldsymbol{z}}_h) = 0.$$

Thus, we infer by contradiction that $\underline{\boldsymbol{v}}_h = \underline{\boldsymbol{z}}_h$ and, as a result, Φ is injective. In order to prove that Φ is also onto, we recall the following result: If $(E, (\cdot,\cdot)_E)$ is a Euclidean space

and $\Psi : E \to E$ is a continuous map such that $\frac{(\Psi(x),x)_E}{\|x\|_E} \to +\infty$ as $\|x\|_E \to +\infty$, then Ψ is surjective. Since $(\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k, a_h^{\mathrm{lin}}(\cdot,\cdot))$ is a Euclidean space and the coercivity (4.2b) of $\boldsymbol{\sigma}$ together with the definition of $\boldsymbol{\sigma}^{\mathrm{lin}}$ yields $a_h^{\mathrm{lin}}(\Phi(\underline{\boldsymbol{v}}_h), \underline{\boldsymbol{v}}_h) \geq a_h^{\mathrm{lin}}(\underline{\boldsymbol{v}}_h, \underline{\boldsymbol{v}}_h)$ for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, we deduce that Φ is an isomorphism. Let, for all $1 \leq n \leq N$, $(\underline{\boldsymbol{y}}_h^n, p_h^n) \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times P_h^k$ be the unique solution to problem (4.41). By the surjectivity and injectivity of Φ, for all $1 \leq n \leq N$, there exists a unique $\underline{\boldsymbol{u}}_h^n \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ such that $\Phi(\underline{\boldsymbol{u}}_h^n) = \underline{\boldsymbol{y}}_h^n$. By definition of Φ and $(\underline{\boldsymbol{y}}_h^n)_{1\leq n\leq N}$, $(\underline{\boldsymbol{u}}_h^n, p_h^n)_{1\leq n\leq N}$ is therefore the unique solution of the discrete problem (4.30). □

4.6 Convergence analysis

In this section we study the convergence of problem (4.30) and prove optimal error estimates under the following additional assumptions on the stress-strain function $\boldsymbol{\sigma}$.

Assumption 4.15 (Stress-strain relation II)**.** There exist real numbers $C_{\mathrm{lp}}, C_{\mathrm{mn}} \in (0, +\infty)$ such that, for a.e. $\boldsymbol{x} \in \Omega$, and all $\boldsymbol{\tau}, \boldsymbol{\eta} \in \mathbb{R}_{\mathrm{s}}^{d\times d}$,

$$|\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\eta})|_{d\times d} \leq C_{\mathrm{lp}}|\boldsymbol{\tau}-\boldsymbol{\eta}|_{d\times d}, \qquad \text{(Lipschitz continuity)} \tag{4.42a}$$

$$(\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\eta})) : (\boldsymbol{\tau}-\boldsymbol{\eta}) \geq C_{\mathrm{mn}}^2|\boldsymbol{\tau}-\boldsymbol{\eta}|_{d\times d}^2. \qquad \text{(strong monotonicity)} \tag{4.42b}$$

Remark 4.16 (Lipschitz continuity and strong monotonocity). It is readily seen, by taking $\boldsymbol{\eta} = \mathbf{0}$ in (4.42), that Lipschitz continuity and strong monotonicity imply respectively the growth and coercivity properties of Assumption 4.1. Therefore, recalling (4.23), it is inferred that the constants appearing in (4.2a), (4.2b), (4.42a), and (4.42b) satisfy

$$C_{\mathrm{mn}}^2 \leq C_{\mathrm{cv}}^2 \leq C_{\mathrm{gr}} \leq C_{\mathrm{lp}}. \tag{4.43}$$

It was proved in [13, Lemma 4.1] that the stress-strain relation for the Hencky–Mises model is strongly monotone and Lipschitz-continuous. Also the isotropic damage model satisfies Assumption 4.15 if the damage function in (4.5) is, for instance, such that

$$D(\boldsymbol{x}, |\boldsymbol{\tau}|) = 1 - (1 + |\mathfrak{C}(\boldsymbol{x})\boldsymbol{\tau}|_{d\times d})^{-\frac{1}{2}} \quad \forall \boldsymbol{x} \in \Omega.$$

In order to prove a convergence rate of $(k+1)$ in space for both the displacement and pressure errors, we assume from this point on that the permeability tensor field $\boldsymbol{\kappa}$ is constant on Ω, and that the following elliptic regularity holds (which is the case, e.g., when Ω is convex [115, 149]): There is a real number $C_{\mathrm{el}} > 0$ only depending on Ω such that, for all $\psi \in L_0^2(\Omega)$, the unique function $\zeta \in P$ solution of the homogeneous Neumann problem

$$-\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}\zeta) = \psi \quad \text{in } \Omega, \qquad \boldsymbol{\kappa}\boldsymbol{\nabla}\zeta \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega,$$

is such that

$$\|\zeta\|_{H^2(\Omega)} \leq C_{\mathrm{el}}\underline{\kappa}^{-\frac{1}{2}}\|\psi\|_\Omega. \tag{4.44}$$

Let $(\underline{\boldsymbol{u}}_h^n, p_h^n)_{1\leq n\leq N}$ be the solution to (4.30). We consider, for all $1 \leq n \leq N$, the discrete error components defined as

$$\underline{\boldsymbol{e}}_h^n := \underline{\boldsymbol{u}}_h^n - \underline{\boldsymbol{I}}_h^k\overline{\boldsymbol{u}}^n, \qquad \epsilon_h^n := p_h^n - \widehat{p}_h^n, \tag{4.45}$$

where the global elliptic projection $\widehat{p}_h^n \in P_h^k$ is defined as the solution to

$$c_h(\widehat{p}_h^n, q_h) = c_h(\overline{p}^n, q_h) \quad \forall q_h \in P_h^k \quad \text{and} \quad \int_\Omega \widehat{p}_h^n = \int_\Omega \overline{p}^n.$$

Before proving the convergence of the scheme, we recall two preliminary approximation results for the projector $\underline{\boldsymbol{I}}_h^k$ and the projection $\widehat{p}_h^n$ that have been proved in [33, Theorem 16] and [30, Lemma 11], respectively. There is a strictly positive constant C_{pj} depending only on Ω, k, and the mesh regularity parameter, such that,

- Assuming (4.42) and $\boldsymbol{u} \in L^2(\boldsymbol{U} \cap \mathbf{H}^{k+2}(\mathcal{T}_h))$ with $\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) \in L^2(\mathbb{H}_{\mathrm{s}}^{k+1}(\mathcal{T}_h))$, for a.e. $t \in (0, t_{\mathrm{F}})$ and all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, it holds

$$\left|a_h(\underline{\boldsymbol{I}}_h^k \boldsymbol{u}(\cdot,t), \underline{\boldsymbol{v}}_h) + (\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}(\cdot,t)), \boldsymbol{v}_h)\right| \le C_{\mathrm{pj}} h^{k+1} \left(|\boldsymbol{u}(\cdot,t)|_{\mathbf{H}^{k+2}(\mathcal{T}_h)} + |\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}(\cdot,t))|_{\mathbb{H}^{k+1}(\mathcal{T}_h)}\right) \|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}. \quad (4.46)$$

- Assuming the elliptic regularity (4.44) and $\overline{p}^n \in P \cap H^{k+1}(\mathcal{T}_h)$, for all $1 \le n \le N$, it holds

$$h\|\widehat{p}_h^n - \overline{p}^n\|_{c,h} + \underline{\kappa}^{\frac{1}{2}} \|\widehat{p}_h^n - \overline{p}^n\|_\Omega \le C_{\mathrm{pj}} h^{k+1} \overline{\kappa}^{\frac{1}{2}} |\overline{p}^n|_{H^{k+1}(\mathcal{T}_h)}. \quad (4.47)$$

Now we have all the ingredients to estimate the discrete errors defined in (4.45).

Theorem 4.17 (Error estimate)**.** *Let $(\boldsymbol{u}, p)$ denote the unique solution to (4.6), for which we assume*

$$\boldsymbol{u} \in H^1(\mathcal{T}_\tau; \boldsymbol{U}) \cap L^2(\boldsymbol{H}^{k+2}(\mathcal{T}_h)), \qquad \boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) \in L^2(\mathbb{H}_{\mathrm{s}}^{k+1}(\mathcal{T}_h)),$$
$$p \in L^2(P \cap H^{k+1}(\mathcal{T}_h)), \qquad \phi \in H^1(\mathcal{T}_\tau; L^2(\Omega)),$$

with $\phi = C_0 p + \boldsymbol{\nabla}\cdot\boldsymbol{u}$. If $C_0 > 0$, we further assume $\pi_\Omega^0 p \in H^1(\mathcal{T}_\tau; P^0(\Omega)) =: H^1(\mathcal{T}_\tau)$. Then, under Assumption 4.15 and the elliptic regularity (4.44), it holds

$$\sum_{n=1}^N \tau \|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + \sum_{n=1}^N \tau \left(\|\epsilon_h^n - \pi_\Omega^0 \epsilon_h^n\|_\Omega^2 + C_0 \|\epsilon_h^n\|_\Omega^2\right) + \|z_h^N\|_{c,h}^2 \le C \left(h^{2k+2} C_1 + \tau^2 C_2\right), \quad (4.48)$$

where C is a strictly positive constant independent of h, τ, C_0, $\boldsymbol{\kappa}$, and t_{F}, and, for the sake of brevity, we have defined $z_h^N := \sum_{n=1}^N \tau \epsilon_h^n$ and introduced the bounded quantities

$$C_1 := |\boldsymbol{u}|_{L^2(\boldsymbol{H}^{k+2}(\mathcal{T}_h))}^2 + |\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})|_{L^2(\mathbb{H}^{k+1}(\mathcal{T}_h))}^2 + (1 + C_0)\frac{\overline{\kappa}}{\underline{\kappa}} |p|_{L^2(H^{k+1}(\mathcal{T}_h))}^2,$$
$$C_2 := \|\boldsymbol{u}\|_{H^1(\mathcal{T}_\tau; \boldsymbol{H}^1(\Omega))}^2 + \|\phi\|_{H^1(\mathcal{T}_\tau; L^2(\Omega))}^2 + C_0 \|\pi_\Omega^0 p\|_{H^1(\mathcal{T}_\tau)}^2.$$

Remark 4.18 (Time regularity). In order to prove the previous error estimate, we only require the displacement $\boldsymbol{u}$ and the fluid content ϕ solving problem (4.6) to be piecewise H^1-regular in $(0, t_{\mathrm{F}})$, whereas 2.12 is established under the much stronger regularity $\boldsymbol{u} \in C^2(\mathbf{H}^1(\Omega))$ and, if $C_0 > 0$, $p \in C^2(L^2(\Omega))$. Moreover, the assumptions $\phi \in H^1(\mathcal{T}_\tau; L^2(\Omega))$ and, if $C_0 > 0$, $\pi_\Omega^0 p \in H^1(\mathcal{T}_\tau)$ are consistent with the time regularity results observed in Remark 4.2.

Proof. (i) *Estimate of* $\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}^2$. First we observe that, owing to (4.1a) and the definition of b_h given in (4.25), for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ and all $1 \le n \le N$, we have

$$(\overline{\boldsymbol{f}}^n, \boldsymbol{v}_h)_\Omega = -(\boldsymbol{\nabla}\cdot\overline{\boldsymbol{\sigma}}^n(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}), \boldsymbol{v}_h)_\Omega + (\boldsymbol{\nabla}\overline{p}^n, \boldsymbol{v}_h)_\Omega = a_h(\underline{\boldsymbol{I}}_h^k\overline{\boldsymbol{u}}^n, \underline{\boldsymbol{v}}_h) + b_h(\underline{\boldsymbol{v}}_h, \widehat{p}_h^n) - \mathcal{R}^n(\underline{\boldsymbol{v}}_h), \tag{4.49}$$

where the residual linear form $\mathcal{R}^n : \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \to \mathbb{R}$ is defined such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$,

$$\mathcal{R}^n(\underline{\boldsymbol{v}}_h) \coloneqq a_h(\underline{\boldsymbol{I}}_h^k\overline{\boldsymbol{u}}^n, \underline{\boldsymbol{v}}_h) + (\boldsymbol{\nabla}\cdot\overline{\boldsymbol{\sigma}}^n(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}), \boldsymbol{v}_h)_\Omega, + b_h(\underline{\boldsymbol{v}}_h, \widehat{p}_h^n) - (\boldsymbol{\nabla}\overline{p}^n, \boldsymbol{v}_h)_\Omega. \tag{4.50}$$

Using the norm equivalence (4.22), the strong monotonicity (4.42b) of $\boldsymbol{\sigma}$ along with assumption (4.24) on the stabilization parameter, the discrete mechanical equilibrium (4.30a), and (4.49), yields

$$\begin{aligned}
\frac{C_{\mathrm{mn}}^2}{C_{\mathrm{eq}}^2}\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 &= C_{\mathrm{mn}}^2 \sum_{T\in\mathcal{T}_h}\left(\|\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{e}}_T^n\|_T^2 + \sum_{F\in\mathcal{F}_T}\frac{1}{h_F}\|\boldsymbol{\Delta}_{TF}^k\underline{\boldsymbol{e}}_T^n\|_F^2\right)\\
&\le a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{e}}_h^n) - a_h(\underline{\boldsymbol{I}}_h^k\overline{\boldsymbol{u}}^n, \underline{\boldsymbol{e}}_h^n)\\
&= (\overline{\boldsymbol{f}}^n, \boldsymbol{e}_h^n)_\Omega - b_h(\underline{\boldsymbol{e}}_h^n, p_h^n) - a_h(\underline{\boldsymbol{I}}_h^k\overline{\boldsymbol{u}}^n, \underline{\boldsymbol{e}}_h^n) = -b_h(\underline{\boldsymbol{e}}_h^n, \epsilon_h^n) - \mathcal{R}^n(\underline{\boldsymbol{e}}_h^n).
\end{aligned}$$

Thus, owing to the previous relation and defining the dual norm

$$\|\mathcal{R}^n\|_{\boldsymbol{\varepsilon},h,*} \coloneqq \sup_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k\setminus\{\underline{\boldsymbol{0}}\}}\frac{\mathcal{R}^n(\underline{\boldsymbol{v}}_h^n)}{\|\underline{\boldsymbol{v}}_h^n\|_{\boldsymbol{\varepsilon},h}},$$

we have that

$$\frac{C_{\mathrm{mn}}^2}{C_{\mathrm{eq}}^2}\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + b_h(\underline{\boldsymbol{e}}_h^n, \epsilon_h^n) \le \|\mathcal{R}^n\|_{\boldsymbol{\varepsilon},h,*}\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h} \le \frac{C_{\mathrm{eq}}^2}{2C_{\mathrm{mn}}^2}\|\mathcal{R}^n\|_{\boldsymbol{\varepsilon},h,*}^2 + \frac{C_{\mathrm{mn}}^2}{2C_{\mathrm{eq}}^2}\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}^2,$$

where the conclusion follows from Young's inequality. Hence, rearranging, we arrive at

$$\frac{C_{\mathrm{mn}}^2}{2C_{\mathrm{eq}}^2}\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + b_h(\underline{\boldsymbol{e}}_h^n, \epsilon_h^n) \le \frac{C_{\mathrm{eq}}^2}{2C_{\mathrm{mn}}^2}\|\mathcal{R}^n\|_{\boldsymbol{\varepsilon},h,*}^2 \tag{4.51}$$

(ii) *Estimate of* $C_0\|\epsilon_h^n\|_\Omega^2$. Using (4.1b), the fact that $(\boldsymbol{\nabla}\cdot\boldsymbol{u}(t), 1)_\Omega = 0$ to insert $\pi_\Omega^0 q_h$, and the consistency property (4.29), we infer that, for all $q_h \in P_h^k$ and all $1 \le i \le N$,

$$\begin{aligned}
(\overline{g}^i, q_h)_\Omega &= (C_0\mathrm{d}_t\overline{p}^i, q_h)_\Omega + (\boldsymbol{\nabla}\cdot(\mathrm{d}_t\overline{\boldsymbol{u}}^i), q_h - \pi_\Omega^0 q_h)_\Omega - (\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}\overline{p}^i), q_h)_\Omega\\
&= \tau^{-1}\int_{t^{i-1}}^{t^i}\mathrm{d}_t\left[C_0(p(t), q_h)_\Omega + (\boldsymbol{\nabla}\cdot\boldsymbol{u}(t), q_h - \pi_\Omega^0 q_h)_\Omega\right]\mathrm{d}t + c_h(\overline{p}^i, q_h)\\
&= \delta_t\left[C_0(p^i, q_h)_\Omega + (\boldsymbol{\nabla}\cdot\boldsymbol{u}^i, q_h - \pi_\Omega^0 q_h)_\Omega\right] + c_h(\overline{p}^i, q_h).
\end{aligned} \tag{4.52}$$

Therefore, using the discrete mass conservation equation (4.30b), (4.26), the Fortin property (4.27b) , the definition of the elliptic projection $\widehat{p}_h^n$, and (4.52), we obtain

$$\begin{aligned}
&C_0(\delta_t\epsilon_h^i, q_h)_\Omega - b_h(\delta_t\underline{\boldsymbol{e}}_h^i, q_h) + c_h(\epsilon_h^i, q_h)\\
&\qquad = (\overline{g}^i, q_h)_\Omega - C_0(\delta_t\widehat{p}_h^i, q_h)_\Omega + b_h(\delta_t(\underline{\boldsymbol{I}}_h^k\overline{\boldsymbol{u}}^i), q_h - \pi_\Omega^0 q_h) - c_h(\widehat{p}_h^i, q_h)\\
&\qquad = (\overline{g}^i, q_h)_\Omega - \delta_t\left[C_0(\widehat{p}_h^i, q_h)_\Omega - (\boldsymbol{\nabla}\cdot\overline{\boldsymbol{u}}^i, q_h - \pi_\Omega^0 q_h)_\Omega\right] - c_h(\overline{p}^i, q_h)\\
&\qquad = \delta_t\left[C_0(p^i - \widehat{p}_h^i, q_h)_\Omega + (\boldsymbol{\nabla}\cdot\boldsymbol{u}^i - \boldsymbol{\nabla}\cdot\overline{\boldsymbol{u}}^i, q_h - \pi_\Omega^0 q_h)_\Omega\right]\\
&\qquad = \delta_t\left[C_0(\overline{p}^i - \widehat{p}_h^i, q_h)_\Omega + C_0(\pi_\Omega^0(p^i - \overline{p}^i), q_h)_\Omega + (\phi^i - \overline{\phi}^i, q_h - \pi_\Omega^0 q_h)_\Omega\right],
\end{aligned} \tag{4.53}$$

where, in order to pass to the last line, we have inserted $\pm\overline{p}^i$ into the first term inside brackets in the third line, we have defined, according to (4.1e), $\phi^i := C_0 p^i + \boldsymbol{\nabla}\cdot\boldsymbol{u}^i$ for all $0 \le i \le N$, and we have used the definition of the global L^2-projector π_Ω^0. Moreover, setting $\widehat{p}_h^0 := 0$, it follows from the initial condition (4.30c), the boundary condition (4.1c), and (4.26) that

$$C_0(\epsilon_h^0, q_h)_\Omega - b_h(\underline{\boldsymbol{e}}_h^0, q_h) = (\phi^0, q_h)_\Omega = (\phi^0, q_h - \pi_\Omega^0 q_h)_\Omega + C_0(\pi_\Omega^0 p^0, q_h)_\Omega. \tag{4.54}$$

For all $1 \le n \le N$, summing (4.53) for $1 \le i \le n$ with the choice $q_h = \tau\epsilon_h^n$, using (4.54), and proceeding as in the second step of the proof of Proposition 4.11, leads to

$$\begin{aligned}
&C_0\|\epsilon_h^n\|_\Omega^2 - b_h(\underline{\boldsymbol{e}}_h^n, \epsilon_h^n) + \frac{1}{2\tau}\left(\|z_h^n\|_{c,h}^2 - \|z_h^{n-1}\|_{c,h}^2\right)\\
&\qquad \le C_0(\overline{p}^n - \widehat{p}_h^n, \epsilon_h^n)_\Omega + C_0(\pi_\Omega^0(p^n - \overline{p}^n), \epsilon_h^n)_\Omega + (\phi^n - \overline{\phi}^n, \epsilon_h^n - \pi_\Omega^0\epsilon_h^n)_\Omega,
\end{aligned} \tag{4.55}$$

where $z_h^n := \sum_{i=1}^n \tau\epsilon_h^i$ if $n \ge 1$ and $z_h^0 := 0$. We bound the first term in the right-hand side of (4.55) applying the Cauchy–Schwarz and Young inequalities followed by the approximation result (4.47), yielding

$$\begin{aligned}
C_0(\overline{p}^n - \widehat{p}_h^n, \epsilon_h^n)_\Omega &\le C_{\mathrm{pj}}^2 h^{2(k+1)}\frac{\overline{\kappa}}{\underline{\kappa}}C_0|\overline{p}^n|_{H^{k+1}(\mathcal{T}_h)}^2 + \frac{C_0}{4}\|\epsilon_h^n\|_\Omega^2\\
&\le C_{\mathrm{pj}}^2\frac{h^{2(k+1)}}{\tau}\left(\frac{\overline{\kappa}}{\underline{\kappa}}\right)C_0|p|_{L^2((t^{n-1},t^n);H^{k+1}(\mathcal{T}_h))}^2 + \frac{C_0}{4}\|\epsilon_h^n\|_\Omega^2,
\end{aligned} \tag{4.56}$$

where, in order to pass to the second line, we have used the Cauchy–Schwarz inequality and adopted the notation $|\cdot|_{L^2((t^{n-1},t^n);H^m(\mathcal{T}_h))} := \||\cdot|_{H^m(\mathcal{T}_h)}\|_{L^2((t^{n-1},t^n))}$, for any $m \in \mathbb{N}$. We estimate the second and third terms using the Cauchy–Schwarz and the Young inequalities together with the time approximation result (4.9) as follows:

$$\begin{aligned}
C_0(\pi_\Omega^0(p^n - \overline{p}^n), \epsilon_h^n)_\Omega &\le C_0\tau\|\pi_\Omega^0 p\|_{H^1((t^{n-1},t^n))}^2 + \frac{C_0}{4}\|\epsilon_h^n\|_\Omega^2,\\
(\phi^n - \overline{\phi}^n, \epsilon_h^n - \pi_\Omega^0\epsilon_h^n)_\Omega &\le \eta\tau\|\phi\|_{H^1((t^{n-1},t^n);L^2(\Omega))}^2 + \frac{1}{4\eta}\|\epsilon_h^n - \pi_\Omega^0\epsilon_h^n\|_\Omega^2,
\end{aligned} \tag{4.57}$$

with η denoting a positive real number that will be fixed later on in the proof. The relation obtained by plugging (4.56) and (4.57) into (4.55) reads

$$\frac{C_0}{2}\|\epsilon_h^n\|_\Omega^2 - b_h(\underline{\boldsymbol{e}}_h^n, \epsilon_h^n) + \frac{1}{2\tau}\left(\|z_h^n\|_{c,h}^2 - \|z_h^{n-1}\|_{c,h}^2\right) - \frac{1}{4\eta}\|\epsilon_h^n - \pi_\Omega^0\epsilon_h^n\|_\Omega^2 \le$$
$$C_{\mathrm{pj}}^2\frac{h^{2(k+1)}}{\tau}\left(\frac{\overline{\kappa}}{\underline{\kappa}}\right)C_0|p|^2_{L^2((t^{n-1},t^n);H^{k+1}(\mathcal{T}_h))} + C_0\tau\|\pi_\Omega^0 p\|^2_{H^1((t^{n-1},t^n))} + \eta\tau\|\phi\|^2_{H^1((t^{n-1},t^n);L^2(\Omega))}. \tag{4.58}$$

(iii) *Estimate of* $\|\epsilon_h^n - \pi_\Omega^0\epsilon_h^n\|_\Omega^2$. We proceed as in the first step of the proof of Proposition 4.11. Using the inf-sup condition (4.28), (4.26) followed by the definition (4.45) of the pressure error, the linearity of b_h, the mechanical equilibrium equation (4.30a), and (4.49), we get, for all $1 \le n \le N$,

$$\begin{aligned}
\|\epsilon_h^n - \pi_\Omega^0\epsilon_h^n\|_\Omega &\le \beta \sup_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k\setminus\{\underline{\boldsymbol{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, \epsilon_h^n - \pi_\Omega^0\epsilon_h^n)}{\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}} \\
&= \beta \sup_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k\setminus\{\underline{\boldsymbol{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, p_h^n - \widehat{p}_h^n)}{\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}} \\
&= \beta \sup_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k\setminus\{\underline{\boldsymbol{0}}\}} \frac{(\overline{\boldsymbol{f}}^n, \boldsymbol{v}_h)_\Omega - a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{v}}_h) - b_h(\underline{\boldsymbol{v}}_h, \widehat{p}_h^n)}{\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}} \\
&= \beta \sup_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k\setminus\{\underline{\boldsymbol{0}}\}} \frac{a_h(\underline{\boldsymbol{I}}_h^k\overline{\boldsymbol{u}}^n, \underline{\boldsymbol{v}}_h) - a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{v}}_h) - \mathcal{R}^n(\underline{\boldsymbol{v}}_h)}{\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}}.
\end{aligned} \tag{4.59}$$

Moreover, the Lipschitz continuity of the stress-strain function (4.42a), the Cauchy–Schwarz inequality, assumption (4.24) on the stabilization parameter γ together with (4.43), and the second inequality in (4.22), lead to

$$\begin{aligned}
&a_h(\underline{\boldsymbol{I}}_h^k\overline{\boldsymbol{u}}^n, \underline{\boldsymbol{v}}_h) - a_h(\underline{\boldsymbol{u}}_h^n, \underline{\boldsymbol{v}}_h) \\
&\quad= \sum_{T\in\mathcal{T}_h}\left(\int_T(\boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{I}}_T^k\overline{\boldsymbol{u}}^n) - \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{u}}_T^n)) : \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T + \sum_{F\in\mathcal{F}_T}\frac{\gamma}{h_F}\int_F\boldsymbol{\Delta}_{TF}^k(\underline{\boldsymbol{I}}_T^k\overline{\boldsymbol{u}}^n - \underline{\boldsymbol{u}}_T^n)\cdot\boldsymbol{\Delta}_{TF}^k\underline{\boldsymbol{v}}_T\right) \\
&\quad\le C_{\mathrm{lp}}C_{\mathrm{eq}}^2\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}.
\end{aligned} \tag{4.60}$$

Therefore, plugging the previous bound into the last line of (4.59), yields

$$\|\epsilon_h^n - \pi_\Omega^0\epsilon_h^n\|_\Omega \le \beta C_{\mathrm{lp}}C_{\mathrm{eq}}^2\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h} + \beta\|\mathcal{R}^n\|_{\boldsymbol{\varepsilon},h,*}.$$

Squaring and rearranging the previous relation and recalling that, owing to (4.43), $C_{\mathrm{mn}}^2 \le C_{\mathrm{lp}}$, it is inferred that

$$\frac{\|\epsilon_h^n - \pi_\Omega^0\epsilon_h^n\|_\Omega^2}{2\beta^2C_{\mathrm{mn}}^{-2}C_{\mathrm{lp}}^2C_{\mathrm{eq}}^6} \le \frac{C_{\mathrm{mn}}^2}{C_{\mathrm{eq}}^2}\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}^2 + \frac{\|\mathcal{R}^n\|_{\boldsymbol{\varepsilon},h,*}^2}{C_{\mathrm{mn}}^2C_{\mathrm{eq}}^6}. \tag{4.61}$$

(iv) *Estimate of the dual norm of the residual.* We split the residual linear form $\mathcal{R}^n$ defined in (4.50) into three contributions $\mathcal{R}^n := \mathcal{R}_1^n + \mathcal{R}_2^n + \mathcal{R}_3^n$, defined, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ and all $1 \le n \le N$, such that

$$\mathcal{R}_1^n(\underline{\boldsymbol{v}}_h) := a_h(\underline{\boldsymbol{I}}_h^k \overline{\boldsymbol{u}}^n, \underline{\boldsymbol{v}}_h) - \frac{1}{\tau}\int_{t^{n-1}}^{t^n} a_h(\underline{\boldsymbol{I}}_h^k \boldsymbol{u}(t), \underline{\boldsymbol{v}}_h)\,\mathrm{d}t, \tag{4.62a}$$

$$\mathcal{R}_2^n(\underline{\boldsymbol{v}}_h) := (\boldsymbol{\nabla}\cdot\overline{\boldsymbol{\sigma}}^n(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}), \boldsymbol{v}_h)_\Omega + \frac{1}{\tau}\int_{t^{n-1}}^{t^n} a_h(\underline{\boldsymbol{I}}_h^k \boldsymbol{u}(t), \underline{\boldsymbol{v}}_h)\,\mathrm{d}t, \tag{4.62b}$$

$$\mathcal{R}_3^n(\underline{\boldsymbol{v}}_h) := b_h(\underline{\boldsymbol{v}}_h, \widehat{p}_h^n) - (\boldsymbol{\nabla}\overline{p}^n, \boldsymbol{v}_h)_\Omega. \tag{4.62c}$$

The first contribution can be bounded proceeding as in (4.60), then using the stability property (4.27a) of the interpolator $\underline{\boldsymbol{I}}_h^k$, the Cauchy–Schwarz inequality, and a Poincaré–Wirtinger inequality on the time interval (t^{n-1}, t^n). By doing so, we get

$$\begin{aligned}
\mathcal{R}_1^n(\underline{\boldsymbol{v}}_h) &= \frac{1}{\tau}\int_{t^{n-1}}^{t^n} a_h(\underline{\boldsymbol{I}}_h^k \overline{\boldsymbol{u}}^n, \underline{\boldsymbol{v}}_h) - a_h(\underline{\boldsymbol{I}}_h^k \boldsymbol{u}(t), \underline{\boldsymbol{v}}_h)\,\mathrm{d}t \\
&\le \frac{1}{\tau}\int_{t^{n-1}}^{t^n} \left(C_{\mathrm{lp}} C_{\mathrm{eq}}^2 \|\underline{\boldsymbol{I}}_h^k(\overline{\boldsymbol{u}}^n - \boldsymbol{u}(t))\|_{\boldsymbol{\varepsilon},h} \|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\right)\mathrm{d}t \\
&\le \frac{C_{\mathrm{lp}} C_{\mathrm{eq}}^2 C_{\mathrm{st}}}{\tau}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\int_{t^{n-1}}^{t^n} \|\overline{\boldsymbol{u}}^n - \boldsymbol{u}(t)\|_{1,\Omega}\mathrm{d}t \\
&\le \frac{C_{\mathrm{lp}} C_{\mathrm{eq}}^2 C_{\mathrm{st}}}{\sqrt{\tau}}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\|\overline{\boldsymbol{u}}^n - \boldsymbol{u}\|_{L^2((t^{n-1},t^n);\mathbf{H}^1(\Omega))} \\
&\le C_{\mathrm{lp}} C_{\mathrm{eq}}^2 C_{\mathrm{st}} C_{\mathrm{ap}}\sqrt{\tau}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\|\boldsymbol{u}\|_{H^1((t^{n-1},t^n);\mathbf{H}^1(\Omega))}.
\end{aligned} \tag{4.63}$$

We estimate the residual linear form $\mathcal{R}_2^n$ defined in (4.62b) using the consistency property (4.46) and the Cauchy–Schwarz inequality, obtaining

$$\begin{aligned}
\mathcal{R}_2^n(\underline{\boldsymbol{v}}_h) &= \frac{1}{\tau}\int_{t^{n-1}}^{t^n} \left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}(t)), \underline{\boldsymbol{v}}_h\right)_\Omega + a_h(\underline{\boldsymbol{I}}_h^k \boldsymbol{u}(t), \underline{\boldsymbol{v}}_h)\,\mathrm{d}t \\
&\le C_{\mathrm{pj}}\frac{h^{k+1}}{\tau}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\int_{t^{n-1}}^{t^n} \left(|\boldsymbol{u}(t)|_{\mathbf{H}^{k+2}(\mathcal{T}_h)} + |\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}(t))|_{\mathbb{H}^{k+1}(\mathcal{T}_h)}\right)\mathrm{d}t \\
&\le C_{\mathrm{pj}}\frac{h^{k+1}}{\sqrt{\tau}}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\left(|\boldsymbol{u}|_{L^2((t^{n-1},t^n);\mathbf{H}^{k+2}(\mathcal{T}_h))} + |\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})|_{L^2((t^{n-1},t^n);\mathbb{H}^{k+1}(\mathcal{T}_h))}\right).
\end{aligned} \tag{4.64}$$

Finally, the third term in (4.62) can be bounded integrating by parts element-wise and using the Cauchy–Schwarz inequality, the trace inequality (4.14), the consistency property (4.47),

and again the Cauchy–Schwarz inequality, namely

$$\begin{aligned}\mathcal{R}_3^n(\underline{\boldsymbol{v}}_h) &\le \sum_{T\in\mathcal{T}_h}\int_T \boldsymbol{\nabla}\cdot\boldsymbol{v}_T(\overline{p}^n-\widehat{p}_h^n)+\sum_{F\in\mathcal{F}_T}\int_F(\boldsymbol{v}_F-\boldsymbol{v}_T)\cdot(\overline{p}^n-\widehat{p}_h^n)\boldsymbol{n}_{TF}\\ &\le C_{\mathrm{pj}}h^{k+1}\left(\frac{\overline{\kappa}}{\underline{\kappa}}\right)^{\frac12}|\overline{p}^n|_{H^{k+1}(\mathcal{T}_h)}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\\ &\le C_{\mathrm{pj}}\frac{h^{k+1}}{\sqrt{\tau}}\|\underline{\boldsymbol{v}}_h\|_{\boldsymbol{\varepsilon},h}\left(\frac{\overline{\kappa}}{\underline{\kappa}}\right)^{\frac12}|p|_{L^2((t^{n-1},t^n);H^{k+1}(\mathcal{T}_h))}.\end{aligned}\tag{4.65}$$

Therefore, combining (4.63), (4.64), and (4.65), it is inferred that

$$\|\mathcal{R}^n\|_{\boldsymbol{\varepsilon},h,*}^2=\|\mathcal{R}_1^n+\mathcal{R}_2^n+\mathcal{R}_3^n\|_{\boldsymbol{\varepsilon},h,*}^2\le 4C_{\mathrm{lp}}^2C_{\mathrm{eq}}^4C_{\mathrm{st}}^2C_{\mathrm{ap}}^2\tau\|\boldsymbol{u}\|_{H^1((t^{n-1},t^n);\mathbf{H}^1(\Omega))}^2+4C_{\mathrm{pj}}^2\tau^{-1}h^{2(k+1)}\widetilde{C}_1^n,\tag{4.66}$$

with

$$\widetilde{C}_1^n:=|\boldsymbol{u}|_{L^2((t^{n-1},t^n);\mathbf{H}^{k+2}(\mathcal{T}_h))}^2+|\boldsymbol{\sigma}(\cdot,\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})|_{L^2((t^{n-1},t^n);\mathbb{H}^{k+1}(\mathcal{T}_h))}^2+\left(\frac{\overline{\kappa}}{\underline{\kappa}}\right)|p|_{L^2((t^{n-1},t^n);H^{k+1}(\mathcal{T}_h))}^2.$$

(v) *Conclusion.* Adding (4.51) to (4.58) with $\eta=4\beta^2C_{\mathrm{mn}}^{-2}C_{\mathrm{lp}}^2C_{\mathrm{eq}}^6$, using (4.61) and (4.66), summing the resulting equation over $1\le n\le N$, and multiplying both sides by 2τ, we obtain

$$\sum_{n=1}^N\tau\left(\frac{C_{\mathrm{mn}}^2}{2C_{\mathrm{eq}}^2}\|\underline{\boldsymbol{e}}_h^n\|_{\boldsymbol{\varepsilon},h}^2+C_0\|\epsilon_h^n\|_\Omega^2+\frac{\|\epsilon_h^n-\pi_\Omega^0\epsilon_h^n\|_\Omega^2}{8\beta^2C_{\mathrm{mn}}^{-2}C_{\mathrm{lp}}^2C_{\mathrm{eq}}^6}\right)+\|z_h^N\|_{c,h}^2\le\widetilde{C}\left(h^{2(k+1)}C_1+\tau^2C_2\right),\tag{4.67}$$

with $\widetilde{C}:=\max\left\{1,\,C_{\mathrm{pj}}^2,\,2C_{\mathrm{pj}}^2C_{\mathrm{mn}}^{-2}(2C_{\mathrm{eq}}^2+C_{\mathrm{eq}}^{-6}),\,2C_{\mathrm{lp}}^2C_{\mathrm{st}}^2C_{\mathrm{ap}}^2(2C_{\mathrm{eq}}^6+C_{\mathrm{eq}}^{-2}),\,4\beta^2C_{\mathrm{lp}}^2C_{\mathrm{eq}}^6C_{\mathrm{mn}}^{-2}\right\}$ and

$$\begin{aligned}C_1&:=\sum_{n=1}^N\left(\widetilde{C}_1^n+C_0\frac{\overline{\kappa}}{\underline{\kappa}}|p|_{L^2((t^{n-1},t^n);H^{k+1}(\mathcal{T}_h))}^2\right),\\ &=|\boldsymbol{u}|_{L^2(\mathbf{H}^{k+2}(\mathcal{T}_h))}^2+|\boldsymbol{\sigma}(\cdot,\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})|_{L^2(\mathbb{H}^{k+1}(\mathcal{T}_h))}^2+(1+C_0)\frac{\overline{\kappa}}{\underline{\kappa}}|p|_{L^2(H^{k+1}(\mathcal{T}_h))}^2\\ C_2&:=\sum_{n=1}^N\left(\|\boldsymbol{u}\|_{H^1((t^{n-1},t^n);\mathbf{H}^1(\Omega))}^2+\|\phi\|_{H^1((t^{n-1},t^n);L^2(\Omega))}^2+C_0\|\pi_\Omega^0p\|_{H^1((t^{n-1},t^n))}^2\right)\\ &=\|\boldsymbol{u}\|_{H^1(\mathcal{T}_\tau;\mathbf{H}^1(\Omega))}^2+\|\phi\|_{H^1(\mathcal{T}_\tau;L^2(\Omega))}^2+C_0\|\pi_\Omega^0p\|_{H^1(\mathcal{T}_\tau)}^2.\end{aligned}$$

Finally, multiplying both sides of (4.67) by $2(C_{\mathrm{pj}}^{-2}+1)\widetilde{C}$ yields (4.48) with $C=2(C_{\mathrm{pj}}^{-2}+1)\widetilde{C}^2$. □

Chapter 5

Poroelasticity with uncertain coefficients

This chapter contains some preliminary results on the numerical solution of the Biot problem with random poroelastic coefficients in the context of uncertainty quantification. It collects part of the ongoing work carried out during the nine month internship at the BRGM that took place from January to September 2018.

Contents

5.1 Introduction

The aim of this chapter is to study poroelasticity problems where the model coefficients are uncertain. The interest of this type of hydro-mechanical coupled models is particularly manifest in geosciences applications [123, 126, 150, 151], where subsurface fluid flows induce a deformation of the rock matrix. We rely here on the linear Biot model [27, 190] that describes the Darcean fluid flow in saturated porous media under the assumptions of small deformations and small variations of the porosity and the fluid density. This model depends on physical parameters that are poorly known justifying a stochastic description, due to space heterogeneities, measurement inaccuracies, and sometimes the ill-posedness of inverse problems inherent in parameter estimation techniques. Although there is an extensive literature on the poroelasticity model (we mention, in particular, the comprehensive textbooks [61, 68, 198]) and its numerical approximation, to our knowledge few works have addressed the role of uncertainty in Biot's formulation of hydromechanical coupling. In [52, 122] the authors include uncertainty in one-dimensional consolidation analysis by incorporating heterogeneity in the consolidation coefficient. The stochastic consolidation model has been extended in [159] to nonlinear uncertain soil parameters. Variability in the initial pore pressure and in the heterogeneous hydraulic mobility has been considered in [62] and [100], respectively. We also mention [65], where a stochastic Galerkin approach is proposed to solve the poroelasticity equations with randomness in all material parameters and tested on a one-dimensional problem.

Uncertainty Quantification (UQ) methods have been developed in the last decades to take into account the effect of random input quantities of a model on the quantity of interest. They enable to obtain information on the model output that is richer than in a deterministic context, since they provide the statistical moments (mean and variance), the probability distribution, and the sensitivity analysis. Among the techniques designed for UQ in numerical models, the stochastic spectral methods have received a considerable attention. The principle of these methods is to decompose a random variable on suitable approximation bases. In particular, Polynomial Chaos (PC) expansion represents the solution as a finite sum of orthogonal polynomials. PC were initially introduced by Wiener [201] and applied by Ghanem and Spanos [112] in solid mechanics and by Le Maître and Knio [137] in fluid mechanics. They were subsequently used to treat a large variety of problems, including elliptic models (*e.g.* [10, 148]), flow and transport in porous media (*e.g.* [110, 111]), thermal problems (*e.g.* [119, 141]), and hyperbolic systems (see, *e.g.* [192]). In this chapter, we focus on the non-intrusive spectral projection method, that only requires to use a deterministic solver as a black box to construct the spectral expansion of the solution. Specifically, (i) a sparse grid allows to efficiently sample the parametric space, (ii) the numerical code is run for each sample (offline stage), and (iii) the model outputs at each sample are collected and assembled in order to construct the spectral expansion of the solution.

The numerical solution of the poroelasticity coupled system requires a discretization method able to (i) treat a complex geometry with polyhedral meshes and nonconforming interfaces, (ii) handle possible heterogeneities of the poromechanical parameters, and (iii) prevent localized pressure oscillations arising in the case of low-permeable and low-

compressible porous media. We choose to discretize the poroelasticity problem using the Hybrid High-Order–discontinuous Galerkin coupled method developed in Chapter 2, which satisfies these requirements. The contribution of this work is threefold. First, a probabilistic framework is introduced to study the Biot model with uncertain coefficients. Special attention is paid to providing a physically admissible set of poroelastic parameters. Second, the well-posedness of the stochastic Biot model is proven at the continuous level. Third, a non-intrusive polynomial chaos approach is implemented in order to investigate the effect of the random poroelastic coefficients on the displacement and the pressure fields for two test cases.

The material is organized as follows. In Section 5.2 we present the poroelasticity model and identify the relations among the poromechanical coefficients allowing to express the constrained specific storage coefficient as a function of the other parameters and deterministic primary variables. In Section 5.3 we introduce the probabilistic framework and investigate how the uncertainty of the model parameters propagates on the storage coefficient. In Section 5.4 we present the linear poroelasticity problem with random coefficients in strong and weak form. We then give the assumptions on the random model parameter yielding the well-posedness of the variational problem. In Section 5.5 we outline the features of the polynomial chaos expansion and of the pseudo-spectral projection method. This procedure allows to reduce the stochastic problem to a finite set of parametric deterministic problems discretized by the HHO–dG scheme. Finally, we present numerical results in Section 5.6 to illustrate the performance of the method and to perform the sensitivity analysis.

5.2 The Biot model

In this section we introduce the model problem and present some preliminary relations between the coefficients characterizing the porous medium. The linear poroelasticity model, usually referred as Biot model, consists of two coupled governing equations; one describing the mass balance of the fluid and the other expressing the momentum equilibrium of the porous medium.

5.2.1 Fluid mass balance

The fluid mass conservation law in a *fully saturated* porous medium reads

$$\mathrm{d}_t(\rho_\mathrm{f}\varphi) + \boldsymbol{\nabla}\cdot(\rho_\mathrm{f}\varphi\boldsymbol{v}) = \rho_\mathrm{f} g, \tag{5.1}$$

where d_t denotes the time derivative, ρ_f [kg/m^3] is the fluid density, φ [−] the soil porosity, $\boldsymbol{v}$ [ms^{-1}] the velocity field, and g [s^{-1}] the fluid source. The state equation for *slightly compressible* fluids (cf. [61, Section 3.1] and [206]) yields

$$\mathrm{d}_t\rho_\mathrm{f} = \frac{\rho_\mathrm{f}}{K_\mathrm{f}}\mathrm{d}_t p, \tag{5.2}$$

with p [Pa] the pore pressure and $K_\mathrm{f} > 0$ [Pa] the fluid bulk modulus. Moreover, under the assumptions of an *isotropic and isothermal* conditions, *infinitesimal strains*, and *small*

relative variations of porosity, it is inferred, following [61, Section 4.1], that the change of the porosity is caused by a fluid and a mechanical effect as follows:

$$\mathrm{d}_t\varphi = \frac{1}{M}\mathrm{d}_t p + \alpha \mathrm{d}_t(\boldsymbol{\nabla}\cdot\boldsymbol{u}), \tag{5.3}$$

where $\boldsymbol{u}$ [m] denotes the displacement field. We rely on the definitions of the Biot-Willis coefficient $\alpha \in (0, 1]$ [−] and the Biot tangent modulus $M > 0$ [Pa^{-1}] given in [61, 68]:

$$\alpha := 1 - \frac{K_\mathrm{d}}{K_\mathrm{m}}, \qquad M := \frac{K_\mathrm{m}}{\alpha - \varphi}, \tag{5.4}$$

with $K_\mathrm{d}, K_\mathrm{m} > 0$ [Pa] denoting the bulk moduli of the drained medium and the solid matrix, respectively. The coefficient α quantifies the amount of fluid that can be forced into the medium by a variation of the pore volume for a constant fluid pressure, while M measures the amount of fluid that can be forced into the medium by pressure increments due to the compressibility of the structure. The case of a solid matrix with incompressible grains ($K_\mathrm{m} \to +\infty$) corresponds to the limit value $1/M = 0$. From (5.2) and (5.3), it follows that the variation of fluid content in the medium is given by

$$\mathrm{d}_t(\rho_\mathrm{f}\varphi) = \varphi \mathrm{d}_t\rho_\mathrm{f} + \rho_\mathrm{f}\mathrm{d}_t\varphi = \rho_\mathrm{f}\left(c_0 \mathrm{d}_t p + \alpha \mathrm{d}_t(\boldsymbol{\nabla}\cdot\boldsymbol{u})\right).$$

where, according to (5.2), (5.3), and (5.4) the constrained specific storage coefficient is defined by

$$c_0 := \frac{\alpha - \varphi}{K_\mathrm{m}} + \frac{\varphi}{K_\mathrm{f}}. \tag{5.5}$$

Therefore, plugging the previous relation into (5.1) and assuming that the fluid density is *uniform* in the medium, we obtain

$$c_0 \mathrm{d}_t p + \alpha \mathrm{d}_t(\boldsymbol{\nabla}\cdot\boldsymbol{u}) + \boldsymbol{\nabla}\cdot(\varphi \boldsymbol{v}) = g, \tag{5.6}$$

The fluid velocity is related to the pore pressure through the well-known Darcy's law (see for instance [68, Section 4.1.2]). Consistent with the small perturbations hypothesis adopted here, Darcy's law can be considered in its simplest form

$$\varphi \boldsymbol{v} = -\frac{\mathbb{K}}{\mu_\mathrm{f}}\boldsymbol{\nabla} p,$$

where $\mathbb{K}$ [m^2] is the tensor-valued intrinsic permeability and μ_f [Pa · s] is the fluid viscosity. Thus, defining the hydraulic mobility as $\boldsymbol{\kappa} := \mathbb{K}/\mu_\mathrm{f}$, the mass conservation equation (5.6) becomes

$$c_0 \mathrm{d}_t p + \alpha \mathrm{d}_t(\boldsymbol{\nabla}\cdot\boldsymbol{u}) - \boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla} p) = g. \tag{5.7}$$

For the sake of simplicity, we also assume in what follows that the mobility field is isotropic and strictly positive, namely $\boldsymbol{\kappa} = \kappa \boldsymbol{I}_d$, with $\kappa > 0$.

5.2.2 Momentum balance

The momentum conservation equation under the *quasi-static* assumption, namely when the inertia effects in the elastic structure are negligible, reads

$$-\boldsymbol{\nabla}\cdot\widetilde{\boldsymbol{\sigma}} = \boldsymbol{f}, \tag{5.8}$$

where $\widetilde{\boldsymbol{\sigma}}$ [Pa] denotes the total stress tensor and $\boldsymbol{f}$ [Pa/m] is the loading force (*e.g.* gravity). Owing to the Terzaghi's decomposition [190], the stress tensor is modeled as the sum of an effective term (mechanical effect) and a pressure term (fluid effect), yielding

$$\widetilde{\boldsymbol{\sigma}} = \boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) - \alpha p\boldsymbol{I}_d, \tag{5.9}$$

where $\boldsymbol{I}_d$ is the identity matrix. The symmetric part of the gradient of the displacement field $\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}$ measures the strain accordingly to the small deformation hypothesis. In the context of *linear isotropic* poroelasticity, the soil mechanical behavior is described through the Cauchy strain-stress relation defined, for all $\boldsymbol{\epsilon} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$, by

$$\boldsymbol{\sigma}(\boldsymbol{\epsilon}) = 2\mu\boldsymbol{\epsilon} + \lambda\,\mathrm{tr}(\boldsymbol{\epsilon})\boldsymbol{I}_d = 2\mu\,\mathbf{dev}(\boldsymbol{\epsilon}) + K\,\mathrm{tr}(\boldsymbol{\epsilon})\boldsymbol{I}_d, \tag{5.10}$$

where $\mu > 0$ [Pa] and $\lambda > 0$ [Pa] are the Lamé's coefficients, $K = {}^{2\mu}/_{d} + \lambda$ [Pa] is the bulk modulus, and

$$\mathrm{tr}(\boldsymbol{\epsilon}) \coloneqq \sum_{i=1}^{d} \epsilon_{ii}, \quad \mathbf{dev}(\boldsymbol{\epsilon}) \coloneqq \boldsymbol{\epsilon} - \frac{\mathrm{tr}(\boldsymbol{\epsilon})\boldsymbol{I}_d}{d}$$

are the trace and deviator operator, respectively. As noted in [22, 28], physical and experimental investigations suggest that the mechanical behavior of porous solids may be nonlinear. More general stress-strain relations could be used in place of the linear law (5.10), as we have done in Chapter 3 and 4. Plugging (5.10) and (5.9) into (5.8), leads to

$$-\boldsymbol{\nabla}\cdot\left(2\mu\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u} + (\lambda\boldsymbol{\nabla}\cdot\boldsymbol{u} - \alpha p)\boldsymbol{I}_d\right) = \boldsymbol{f}\,. \tag{5.11}$$

Finally, we mention the Gassmann equation (cf. [103, 150]), that relates the bulk moduli K and K_{d} to the Biot–Willis and storage coefficients α, c_0:

$$\alpha^2 = c_0(K - K_{\mathrm{d}}). \tag{5.12}$$

5.2.3 Constrained specific storage coefficient

Investigating the relations between the poroelastic coefficients, we propose to express the specific storage coefficient c_0 as a function of other physical parameters. Using (5.4) to express the drained bulk modulus K_{d} as a function of α and K_{m}, then plugging into (5.12), leads to

$$K_{\mathrm{m}} = \frac{Kc_0 - \alpha^2}{c_0(1-\alpha)}. \tag{5.13}$$

Plugging the previous equation into (5.5) and rearranging, yields the following second-order equation in c_0:

$$K c_0^2 - \left(\alpha + \alpha\varphi + \frac{\varphi K}{K_\mathrm{f}} - \varphi\right) c_0 + \frac{\alpha^2 \varphi}{K_\mathrm{f}} = 0. \tag{5.14}$$

The parameters φ, K_f, K, and α are then used to evaluate c_0 by solving the previous equation. Some conditions need to be prescribed in order to avoid non-physical solutions. Owing to the definition of the Biot–Willis coefficient as a volume change ratio, it holds that $0 \leq \varphi \leq \alpha \leq 1$. As observed in [206, Section 3], a stricter lower bound can be imposed, namely

$$\frac{3\varphi}{2+\varphi} \leq \alpha. \tag{5.15}$$

Lemma 5.1 (Existence)**.** *Let $K_\mathrm{f}, K > 0$, and φ, α satisfy condition (5.15). Then, there exists $c_0 \in \mathbb{R}^+$ solution of (5.14).*

Proof. We prove existence by assessing the positivity of the discriminant $\mathcal{D}$ associated to (5.14). Computing $\mathcal{D}$ and rearranging, leads to

$$\begin{aligned}
\mathcal{D} &= \left(\alpha + \alpha\varphi + \frac{\varphi K}{K_\mathrm{f}} - \varphi\right)^2 - \frac{4\alpha^2 \varphi K}{K_\mathrm{f}} \\
&= \varphi^2 \left({}^{K}/_{K_\mathrm{f}} - 1\right)^2 + 2\alpha\varphi \left(1 + \varphi - 2\alpha\right) \left({}^{K}/_{K_\mathrm{f}} - 1\right) + \alpha^2 (1-\varphi)^2 \\
&= \left[\varphi \left({}^{K}/_{K_\mathrm{f}} - 1\right) + \alpha(1 + \varphi - 2\alpha)\right]^2 + \alpha^2(1-\varphi)^2 - \alpha^2(1+\varphi-2\alpha)^2 \\
&= \left[\varphi \left({}^{K}/_{K_\mathrm{f}} - 1\right) + \alpha(1 + \varphi - 2\alpha)\right]^2 + 4\alpha^2(1-\alpha)(\alpha-\varphi).
\end{aligned}$$

Since $0 \leq \varphi < \alpha \leq 1$ owing to (5.15), the second term in the previous sum is positive and, as a result, $\mathcal{D} \geq 0$. □

Remark 5.2. The previous lemma yields the existence of two real solutions $c_0^- \leq c_0^+$. We consider c_0^+ as the unique solution to (5.14) because, for admissible values of $(K, K_\mathrm{f}, \varphi, \alpha)$, we might have $c_0^- < {}^{\varphi}/_{K_\mathrm{f}}$ violating (5.5). For instance, this is the case if $\alpha = {}^{2\varphi}/_{(1+\varphi)}$ for any $\phi \leq {}^{3}/_{4}$ and $K, K_\mathrm{f} > 0$.

5.3 Probabilistic framework

In many applications, only limited information about the poroelastic coefficients in (5.7) and (5.11) is available. In the context of geomechanics, even when it is possible to carry out a large number of measurements, the actual knowledge of the soil properties typically suffers from inaccuracies. Indeed, due to the presence of different layers, the physical properties can have strong variations that are difficult to estimate. Accounting for uncertainties is therefore a fundamental issue that we propose to treat in a probabilistic framework. In this section we present the probabilistic setting and we perform some preliminary investigations on the possible choices to parametrize the model coefficients. In this chapter we adopt the notations of [10, 12, 24, 53]. Thus, Ω denotes the probability space for the parametrization of uncertainty, while the physical domain is denoted by D.

5.3.1 Notations and basic results

Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$ a measurable set. Spaces of functions, vector fields, and tensor fields defined over X are respectively denoted by italic capital, boldface Roman capital, and special Roman capital letters. Thus, for example, $L^2(X)$, $\mathbf{L}^2(X)$, $\mathbb{L}^2(X)$ denote the spaces of square integrable functions, vector fields, and tensor fields over X respectively. We introduce an abstract probability space $(\Theta, \mathcal{B}, \mathcal{P})$, where Θ is the set of possible outcomes, $\mathcal{B}$ a σ-algebra of events, and $\mathcal{P} : \mathcal{B} \to [0, 1]$ a probability measure. For any random variable $h : \Theta \to \mathbb{R}$ defined on the abstract probability space, the expectation of h is

$$E(h) \coloneqq \int_\Theta h(\theta) \mathrm{d}\mathcal{P}(\theta).$$

We assume hereafter that all random quantities are second-order ones, namely they belong to

$$L^2_{\mathcal{P}}(\Theta) \coloneqq \left\{ h : \Theta \to \mathbb{R} \ : \ E(h^2) < +\infty \right\}.$$

We further introduce a parametrization of the random input data using a random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N) : \Theta \to \Omega \coloneqq [0, 1]^N$ such that, for all $i \in \{1, \ldots, N\}$, $\xi_i \in L^2_{\mathcal{P}}(\Theta)$ and has zero mean. For convenience, the random variables ξ_i are assumed independent, so that the probability law $\mathcal{P}_{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$ factorizes. We denote by $\mathcal{B}_{\boldsymbol{\xi}}$ the Borel σ-algebra on Ω and by $(\Omega, \mathcal{B}_{\boldsymbol{\xi}}, \mathcal{P}_{\boldsymbol{\xi}})$ the image probability space. If, furthermore, each random variable ξ_i possesses the density function $\rho_i : [0, 1] \to [0, +\infty)$, we may define the space of second-order random variables on $(\Omega, \mathcal{B}_{\boldsymbol{\xi}}, \mathcal{P}_{\boldsymbol{\xi}})$ as the weighted Lebesgue space $L^2_\rho(\Omega)$, where the weight function

$$\rho(\boldsymbol{\xi}) \coloneqq \prod_{i=1}^{n} \rho_i(\xi_i)$$

is the joint density function of $\boldsymbol{\xi}$. The expectation operator on the image space is denoted using brackets and is related to the expectation on $(\Theta, \mathcal{B}, \mathcal{P})$ through the identity

$$\langle h \rangle \coloneqq \int_\Omega h(\boldsymbol{\omega}) \rho(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} = \int_\Theta h(\boldsymbol{\xi}(\theta)) \mathrm{d}\mathcal{P}(\theta) = E(h \circ \boldsymbol{\xi}).$$

The variance operator of $h \in (\Omega, \mathcal{B}_{\boldsymbol{\xi}}, \mathcal{P}_{\boldsymbol{\xi}})$ is then defined by

$$\mathbb{V}\mathrm{ar}(h) \coloneqq \left\langle (h - \langle h \rangle)^2 \right\rangle.$$

Let $D \subset \mathbb{R}^d$, $d \in \{2, 3\}$, denote the spatial domain occupied by the porous medium. For a given vector space $V(D)$ of real-valued functions on D, equipped with the norm $\| \cdot \|_{V(D)}$, we also define the Bochner space of second-order random fields by

$$L^2_\rho(\Omega, V(D)) \coloneqq \left\{ v : D \times \Omega \to \mathbb{R} \ : \ \|v\|_{L^2_\rho(\Omega, V(D))} \coloneqq \left\langle \|v\|^2_{V(D)} \right\rangle^{\frac{1}{2}} < \infty \right\}.$$

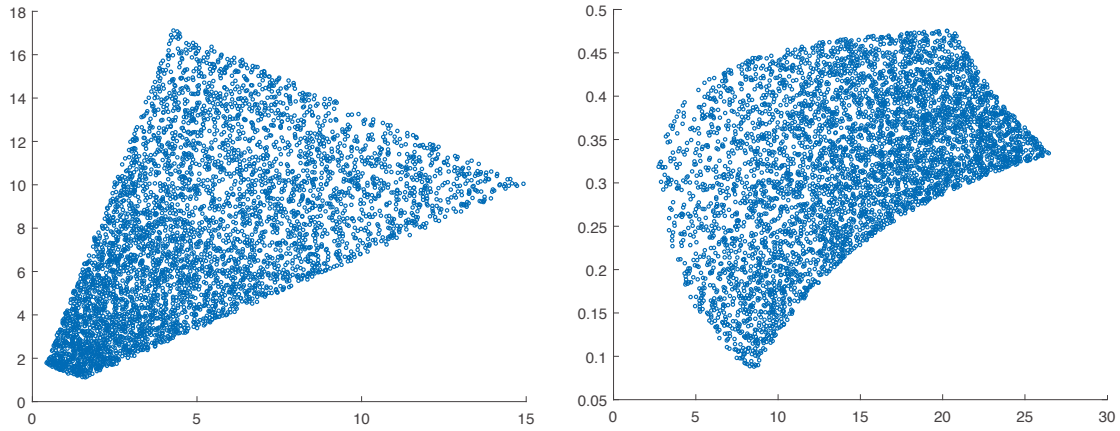


Figure 5.1: Left: Scatter plot of 5000 realizations (μ, λ) for $K \sim \mathcal{U}([1, 20])$ GPa and $\nu \sim \mathcal{U}([0.1, 0.4])$. Right: Scatter plot of 5000 realizations (K, ν) for $\mu \sim \mathcal{U}([1, 10])$ GPa and $\lambda \sim \mathcal{U}([2, 20])$ GPa.

5.3.2 Uncertain poroelastic coefficients

In order to account for the uncertainty of the poroelastic material we model the Lamé's constants, the Biot–Willis coefficient, and the hydraulic mobility as spatially varying random fields depending on the finite-dimensional noise $\boldsymbol{\xi}$. We represent the random fields $(\mu, \lambda, \alpha, \kappa) : (\boldsymbol{x}, \boldsymbol{\xi}) \in D \times \Omega \to \mathbb{R}^4$, as affine combinations of the parameters $(\xi_1, \dots, \xi_N) \in \Omega$:

$$\begin{aligned} \mu(\boldsymbol{x}, \boldsymbol{\xi}) &\coloneqq \mu_0(\boldsymbol{x}) + \sum_{i=1}^{N} \mu_i(\boldsymbol{x})\xi_i, & \lambda(\boldsymbol{x}, \boldsymbol{\xi}) &\coloneqq \lambda_0(\boldsymbol{x}) + \sum_{i=1}^{N} \lambda_i(\boldsymbol{x})\xi_i, \\ \alpha(\boldsymbol{x}, \boldsymbol{\xi}) &\coloneqq \alpha_0(\boldsymbol{x}) + \sum_{i=1}^{N} \alpha_i(\boldsymbol{x})\xi_i, & \kappa(\boldsymbol{x}, \boldsymbol{\xi}) &\coloneqq \kappa_0(\boldsymbol{x}) + \sum_{i=1}^{N} \kappa_i(\boldsymbol{x})\xi_i. \end{aligned} \tag{5.16a}$$

One well-established approach yielding such functional dependence on the random parametrization is the truncated Karhunen–Loéve expansion [142, 180]. According to the results presented in Section 5.2.3, the constrained specific storage coefficient c_0 can be expressed in terms of the elastic moduli λ, μ and the coupling Biot coefficient α (see also [60, 114] for theoretical and empirical investigations on the storage coefficient). Therefore, we let

$$c_0(\boldsymbol{x}, \boldsymbol{\xi}) \coloneqq c_0(\mu(\boldsymbol{x}, \boldsymbol{\xi}), \lambda(\boldsymbol{x}, \boldsymbol{\xi}), \alpha(\boldsymbol{x}, \boldsymbol{\xi})). \tag{5.16b}$$

In practice, the choice of the elastic parameters describing the mechanical properties of the medium is mostly one of convenience. For instance, one might be interested in the Young modulus E (*e.g.* in [128]), or the bulk modulus K, and the Poisson ratio ν (*e.g.* in [60, 173]) instead of μ and λ. For the sake of simplicity, here we choose the Lamé's coefficients as primary variables to avoid dealing with nonlinear parametrizations of the model coefficient inducing a warping of the (μ, λ) admissible set as depicted in Figure 5.1. However, using the truncated Karhunen–Loéve or Polynomial Chaos expansions, it is possible to obtain a parametrization of the model parameter in the affine combination form (5.16a) even if the

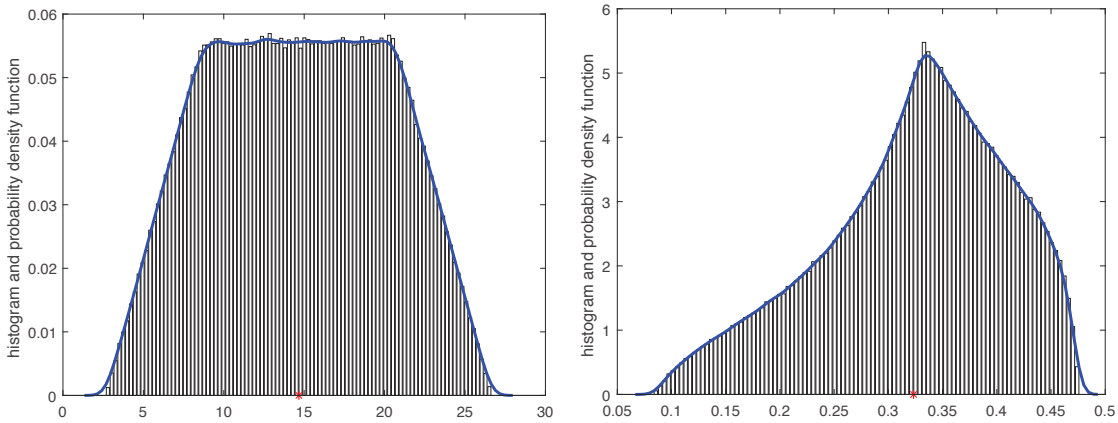


Figure 5.2: Distribution of the bulk modulus K (left) and the Poisson ration ν (right) corresponding to 10^6 realizations for $\mu \sim \mathcal{U}([1, 10])$GPa and $\lambda \sim \mathcal{U}([2, 20])$GPa.

distributions of other elastic moduli are given as entry data. In order to assess the effect of the perturbation of the Lamé's coefficients (μ, λ) as in (5.16a) on the couple (K, ν), one can use the relations

$$K = \frac{2}{d}\mu + \lambda \quad \text{and} \quad \nu = \frac{\lambda}{2(\mu + \lambda)}.$$

We notice that taking (μ, λ) as uniformly distributed primary mechanical variables, necessarily yields non-uniform distribution of (K, ν) as illustrated in Figure 5.2.

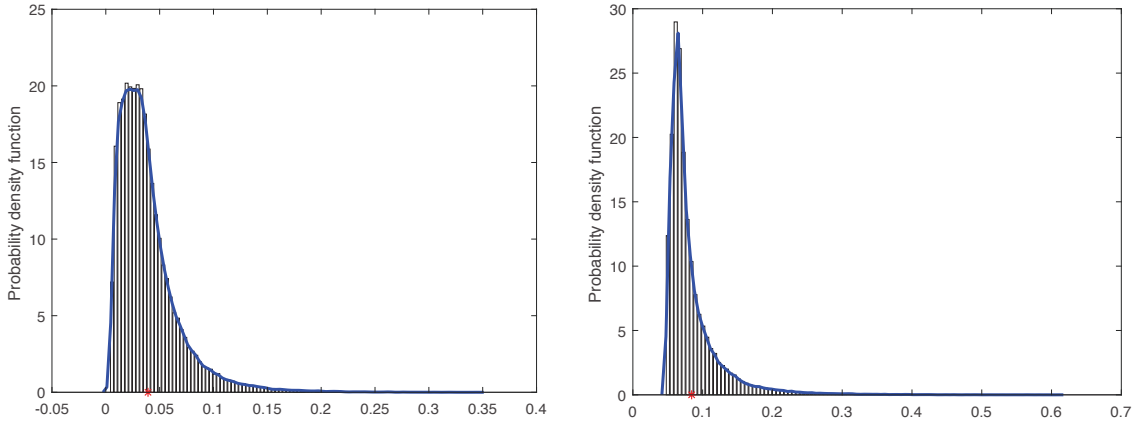
In what follows, we investigate the effect of the parametrization (5.16b) of c_0 on its probability distribution. We consider the poromechanical coefficients illustrated in Table 5.1. They are inspired from [173, Table 1], [206, Section 3], and [61, Table 4.1], and corresponds to two different layers of an ideal sandstone soil. Since in geophysical applications the fluid bulk modulus and the porosity are usually measurable with sufficient precision, we assume that K_f and φ are deterministic quantities, whereas λ, μ, and α are random variables with uniform distribution, whose ranges are reported in Table 5.1. As a particular case of (5.16a),

	Water filled sandstone	Water filled deep sand
K_f	2.2 GPa	2.2 GPa
φ	10%	30%
μ	$\mathcal{U}([1, 10])$ GPa	$\mathcal{U}([5, 80])$ KPa
λ	$\mathcal{U}([2, 20])$ GPa	$\mathcal{U}([12, 150])$ KPa
α	$\mathcal{U}([3\varphi(2+\varphi)^{-1}, 1])$	$\mathcal{U}([3\varphi(2+\varphi)^{-1}, 1])$

Table 5.1: Poromechanical properties

we consider homogeneous random Lamé's and Biot–Willis coefficients defined such that

$$\begin{aligned}
\mu(\boldsymbol{\xi}) &= \mu_{\min} + \xi_1(\mu_{\max} - \mu_{\min}),\\
\lambda(\boldsymbol{\xi}) &= \lambda_{\min} + \xi_2(\lambda_{\max} - \lambda_{\min}),\\
\alpha(\boldsymbol{\xi}) &= \frac{3\varphi}{2+\varphi} + \xi_3\frac{2(1-\varphi)}{2+\varphi},
\end{aligned}$$



(a) Sandstone soil, c_0

(b) Deep sand soil, c_0

Figure 5.3: Distribution of 10^5 realizations of c_0 obtained by solving (5.14) with uniformly distributed random input coefficients λ, μ, and α.

where $\boldsymbol{\xi} \in \Omega = [0,1]^3$ is a vector of three i.i.d. uniform variables. For each realization of $\boldsymbol{\xi}$, we solve (5.14) in order to compute c_0. We are interested in assessing how the uncertainty on this input parameters propagates on c_0. In Figure 5.3 we plot the histograms and probability density functions of c_0 corresponding to 10^5 realizations of $\boldsymbol{\xi}$ considering the poromechanical coefficients of Table 5.1. In particular, as expected, we observe that the computed probability density function (obtained with kernel density estimation) of c_0 is far from being uniform.

5.4 The Biot problem with random coefficients

In this section we present the linear poroelasticity problem, discuss the assumptions on the random parameters, derive a weak formulation, and prove its well-posedness.

5.4.1 Strong and weak formulation

The Biot problem consists in finding a vector-valued displacement field $\boldsymbol{u}$ and a scalar-valued pressure field p solutions of (5.7) and (5.11). We now consider the stochastic version obtained by taking random poroelastic coefficients as in (5.16). We assume that $D \subset \mathbb{R}^d$ is a bounded connected polyhedral domain with boundary ∂D and outward normal $\boldsymbol{n}$. In order to close the problem, we enforce boundary conditions corresponding to a medium that is clamped on $\Gamma_D \subset \partial D$, traction-free on $\Gamma_N := \partial D \setminus \Gamma_D$, permeable with free drainage on $\Gamma_d \subset \partial D$, and impermeable on $\Gamma_n := \partial D \setminus \Gamma_d$, as well as an initial condition prescribing the initial fluid

content ϕ_0. For a given finite time $t_F > 0$, the resulting problem is given by

$$
\begin{aligned}
-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\xi},t) + \boldsymbol{\nabla}(\alpha(\boldsymbol{x},\boldsymbol{\xi})p(\boldsymbol{x},\boldsymbol{\xi},t)) &= \boldsymbol{f}(\boldsymbol{x},t), && (\boldsymbol{x},\boldsymbol{\xi},t) \in D\times\Omega\times(0,t_F], && \text{(5.17a)}\\
\mathrm{d}_t\phi(\boldsymbol{x},\boldsymbol{\xi},t) - \boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}(\boldsymbol{x},\boldsymbol{\xi})\boldsymbol{\nabla}p(\boldsymbol{x},\boldsymbol{\xi},t)) &= g(\boldsymbol{x},t), && (\boldsymbol{x},\boldsymbol{\xi},t) \in D\times\Omega\times(0,t_F], && \text{(5.17b)}\\
\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\xi},t) &= \boldsymbol{0}, && (\boldsymbol{x},\boldsymbol{\xi},t) \in \Gamma_D\times\Omega\times(0,t_F], && \text{(5.17c)}\\
\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\xi},t)\boldsymbol{n} + \alpha(\boldsymbol{x},\boldsymbol{\xi})p(\boldsymbol{x},\boldsymbol{\xi},t))\boldsymbol{n}, &= \boldsymbol{0} && (\boldsymbol{x},\boldsymbol{\xi},t) \in \Gamma_N\times\Omega\times(0,t_F], && \text{(5.17d)}\\
p(\boldsymbol{x},\boldsymbol{\xi},t) &= 0, && (\boldsymbol{x},\boldsymbol{\xi},t) \in \Gamma_d\times\Omega\times(0,t_F], && \text{(5.17e)}\\
\boldsymbol{\kappa}(\boldsymbol{x},\boldsymbol{\xi})\boldsymbol{\nabla}p(\boldsymbol{x},\boldsymbol{\xi},t)\cdot\boldsymbol{n} &= 0, && (\boldsymbol{x},\boldsymbol{\xi},t) \in \Gamma_n\times\Omega\times(0,t_F], && \text{(5.17f)}\\
\phi(\boldsymbol{x},\boldsymbol{\xi},0) &= \phi_0(\boldsymbol{x}), && (\boldsymbol{x},\boldsymbol{\xi}) \in D\times\Omega, && \text{(5.17g)}
\end{aligned}
$$

where the stress tensor in (5.17a) and (5.17d), and the fluid content in (5.17b) and (5.17g), are defined, for all $(\boldsymbol{x},\boldsymbol{\xi},t) \in D\times\Omega\times(0,t_F]$, such that

$$
\begin{aligned}
\boldsymbol{\sigma}(\boldsymbol{x},\boldsymbol{\xi},t) &= 2\mu(\boldsymbol{x},\boldsymbol{\xi})\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\xi},t) + \lambda(\boldsymbol{x},\boldsymbol{\xi})(\boldsymbol{\nabla}\cdot\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\xi},t))\boldsymbol{I}_d, && \text{(5.17h)}\\
\phi(\boldsymbol{x},\boldsymbol{\xi},t) &= c_0(\boldsymbol{x},\boldsymbol{\xi})p(\boldsymbol{x},\boldsymbol{\xi},t) + \alpha(\boldsymbol{x},\boldsymbol{\xi})\boldsymbol{\nabla}\cdot\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\xi},t). && \text{(5.17i)}
\end{aligned}
$$

If $\Gamma_N = \Gamma_d = \emptyset$ and $c_0 = 0$, owing to (5.17b) and the homogeneous Neumann condition (5.17f), we need the following compatibility conditions on g and ϕ_0 and zero-average constraint on p:

$$
\int_D \phi_0(\cdot) = 0, \qquad \int_D g(\cdot,t) = 0, \quad \text{and} \quad \int_D p(\cdot,\cdot,t) = 0 \qquad \forall t \in (0,t_F). \tag{5.17j}
$$

For the sake of simplicity, we exclude the case of Γ_D with zero $(d-1)$-dimensional Hausdorff measure. Indeed, this situation simply requires an additional compatibility condition on $\boldsymbol{f}$ and the prescription of the rigid-body motions of the medium. We remark that other types of boundary conditions, such as inhomogeneous ones, can also be considered up to minor modifications.

Before giving the variational formulation of problem (5.17), we introduce some notations. Let $n \in \mathbb{N}$ and $X \subset \mathbb{R}^n$ measurable set. For any $m \in \mathbb{N}$, we denote by $H^m(X)$ the usual Sobolev space of functions that have weak partial derivatives of order up to m in $L^2(X)$, with the convention that $H^0(X) := L^2(X)$, while $C^m(X)$ and $C_{\mathrm{c}}^\infty(X)$ denote, respectively, the usual spaces of m-times continuously differentiable functions and infinitely continuously differentiable functions with compact support on X. Since the Biot problem is unsteady, we also need to introduce, for a vector space V with scalar product $(\cdot,\cdot)_V$, the Bochner space $L^2((0,t_{\mathrm{F}});V)$ spanned by square integrable V-valued functions of the time interval $(0,t_{\mathrm{F}})$. Similarly, the Hilbert space $H^1(V) := H^1((0,t_{\mathrm{F}});V)$ is spanned by V-valued functions of the time interval having first-order weak derivative, and the norm $\|\cdot\|_{H^1(V)}$ is induced by the scalar product

$$
(\phi,\psi)_{H^1(V)} = \int_0^{t_{\mathrm{F}}} (\phi(t),\psi(t))_V + (\mathrm{d}_t\phi(t),\mathrm{d}_t\psi(t))_V \mathrm{d}t \qquad \forall \phi,\psi \in H^1(V).
$$

At each time $t \in (0,t_{\mathrm{F}}]$, the natural functional spaces for the random displacement field $\boldsymbol{u}(t) : D\times\Omega \to \mathbb{R}^d$ and pressure field $p(t) : D\times\Omega \to \mathbb{R}$ taking into account the boundary

conditions (5.17c)–(5.17f) are, respectively,

$$\boldsymbol{U} \coloneqq L^2_\rho(\Omega, \mathbf{H}^1_{0,\Gamma_D}(D)),$$
$$P \coloneqq \begin{cases} L^2_\rho(\Omega, H^1(D) \cap \mathbb{R}(D)^\perp) & \text{if } \Gamma_d = \Gamma_N = \emptyset \text{ and } c_0 = 0, \\ L^2_\rho(\Omega, H^1_{0,\Gamma_d}(D)) & \text{otherwise,} \end{cases}$$

with $\mathbf{H}^1_{0,\Gamma_D}(D) \coloneqq \left\{ \boldsymbol{v} \in \mathbf{H}^1(D) \ : \ \boldsymbol{v}_{|\Gamma_D} = \mathbf{0} \right\}$, $H^1_{0,\Gamma_d}(D) \coloneqq \left\{ q \in H^1(D) \ : \ q_{|\Gamma_d} = 0 \right\}$, and

$$\mathbb{R}(D)^\perp \coloneqq \left\{ q \in L^2(D) \ : \ \int_D q = 0 \right\}.$$

We consider the following weak formulation of problem (5.17): For a loading term $\boldsymbol{f} \in L^2((0,t_\mathrm{F}), \mathbf{L}^2(D))$, a fluid source $g \in L^2((0,t_\mathrm{F}), L^2(D))$, and an initial datum $\phi_0 \in L^2(D)$ that verify (5.17j), find $\boldsymbol{u} \in L^2((0,t_\mathrm{F}), \boldsymbol{U})$ and $p \in L^2((0,t_\mathrm{F}), P)$ such that, for all $\boldsymbol{v} \in \boldsymbol{U}$, all $q \in P$, and all $\psi \in C^\infty_\mathrm{c}((0,t_\mathrm{F}))$

$$\int_0^{t_\mathrm{F}} a(\boldsymbol{u}(t), \boldsymbol{v})\, \psi(t)\mathrm{d}t + \int_0^{t_\mathrm{F}} b(\boldsymbol{v}, p(t))\, \psi(t)\mathrm{d}t = \int_0^{t_\mathrm{F}} \langle (\boldsymbol{f}(t), \boldsymbol{v})_D \rangle \psi(t)\mathrm{d}t, \tag{5.18a}$$

$$\int_0^{t_\mathrm{F}} \left[b(\boldsymbol{u}(t), q) - c(p(t), q) \right] \mathrm{d}_t\psi(t)\mathrm{d}t + \int_0^{t_\mathrm{F}} d(p(t), q)\, \psi(t)\mathrm{d}t = \int_0^{t_\mathrm{F}} \langle (g(t), q)_D \rangle \psi(t)\mathrm{d}t, \tag{5.18b}$$

$$c(p(0), q) - b(\boldsymbol{u}(0), q) = \langle (\phi_0, q)_D \rangle, \tag{5.18c}$$

where $(\cdot,\cdot)_D$ denotes the usual inner product in $L^2(D)$ and the bilinear forms $a : \boldsymbol{U} \times \boldsymbol{U} \to \mathbb{R}$, $b : \boldsymbol{U} \times P \to \mathbb{R}$, and $c : P \times P \to \mathbb{R}$ are defined such that, for all $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{U}$ and all $q, r \in P$,

$$a(\boldsymbol{v}, \boldsymbol{w}) \coloneqq \int_\Omega \int_D \left(2\mu(\boldsymbol{x}, \boldsymbol{\xi}) \boldsymbol{\nabla}_\mathrm{s} \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\xi}) : \boldsymbol{\nabla}_\mathrm{s} \boldsymbol{w}(\boldsymbol{x}, \boldsymbol{\xi}) + \lambda(\boldsymbol{x}, \boldsymbol{\xi}) \boldsymbol{\nabla}\cdot \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\xi}) \boldsymbol{\nabla}\cdot \boldsymbol{w}(\boldsymbol{x}, \boldsymbol{\xi}) \right) \rho(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\xi},$$
$$b(\boldsymbol{v}, q) \coloneqq -\int_\Omega \int_D \alpha(\boldsymbol{x}, \boldsymbol{\xi}) \boldsymbol{\nabla}\cdot \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\xi}) q(\boldsymbol{x}, \boldsymbol{\xi}) \rho(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\xi},$$
$$c(q, r) \coloneqq \int_\Omega \int_D c_0(\boldsymbol{x}, \boldsymbol{\xi}) q(\boldsymbol{x}, \boldsymbol{\xi}) r(\boldsymbol{x}, \boldsymbol{\xi}) \rho(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\xi},$$
$$d(q, r) \coloneqq \int_\Omega \int_D \kappa(\boldsymbol{x}, \boldsymbol{\xi}) \boldsymbol{\nabla} r(\boldsymbol{x}, \boldsymbol{\xi}) \cdot \boldsymbol{\nabla} q(\boldsymbol{x}, \boldsymbol{\xi}) \rho(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\xi}.$$

Above, we have introduced the Frobenius product such that, for all $\boldsymbol{\tau}, \boldsymbol{\eta} \in \mathbb{R}^{d\times d}$, $\boldsymbol{\tau} : \boldsymbol{\eta} \coloneqq \sum_{1\le i,j\le d} \tau_{ij}\eta_{ij}$ with corresponding norm such that, for all $\boldsymbol{\tau} \in \mathbb{R}^{d\times d}$, $|\boldsymbol{\tau}|_{d\times d} \coloneqq (\boldsymbol{\tau} : \boldsymbol{\tau})^{1/2}$.

5.4.2 Well-posedness

The aim of this section is to infer a stability estimate on the displacement and pressure $(\boldsymbol{u}, p) \in L^2((0,t_\mathrm{F}), \boldsymbol{U}) \times L^2((0,t_\mathrm{F}), P)$ solving (5.18), yielding, in particular, the well-posedness of the weak problem. The existence and uniqueness of a solution to the deterministic Biot

problem has been studied in [181, 205]. The results therein establish the existence. for almost every (a.e.) $\boldsymbol{\xi} \in \Omega$, of a solution $(\boldsymbol{u}(\boldsymbol{\xi}), p(\boldsymbol{\xi}))$ to (5.17). Thus, proving that the mapping $\boldsymbol{\xi} \to (\boldsymbol{u}(\boldsymbol{\xi}), p(\boldsymbol{\xi}))$ is measurable and bounded in $L^2_\rho(\Omega)$ gives the existence of a solution to (5.18).

We assume some additional conditions on the input random fields defined in (5.16) that will be needed in the proofs. First, we recall that, for all $\boldsymbol{x} \in D$ and all $\boldsymbol{\xi} \in \Omega$ the coupling coefficient $\alpha(\boldsymbol{x}, \boldsymbol{\xi}) \in (0, 1]$ satisfies the lower bound (5.15). We assume that the reference porosity of the medium is strictly positive (otherwise the Biot problem would reduce to a decoupled one), so that we have $0 < \underline{\alpha} := \frac{3\varphi}{2+\varphi} < \alpha(\boldsymbol{x}, \boldsymbol{\xi})$. We also assume that the Lamé's coefficients satisfies the following:

Assumption 5.3 (Elastic moduli)**.** The shear modulus $\mu \in L^\infty(D \times \Omega)$ is uniformly bounded from above and away from zero, *i.e.* there exist positive constants $\underline{\mu}$ and $\overline{\mu}$ such that

$$0 < \underline{\mu} \leq \mu(\boldsymbol{x}, \boldsymbol{\xi}) \leq \overline{\mu} < \infty, \qquad \text{a.e. in } D \times \Omega. \tag{5.19}$$

The dilatation modulus $\lambda \in L^\infty(D \times \Omega)$ is uniformly bounded, *i.e.* there exist $\overline{\lambda} > 0$ such that

$$0 < \lambda(\boldsymbol{x}, \boldsymbol{\xi}) \leq \overline{\lambda} < \infty, \qquad \text{a.e. in } D \times \Omega. \tag{5.20}$$

The next result establishes the coercivity and the inf-sup condition respectively of the bilinear form a and b appearing in (5.18a).

Lemma 5.4 (Coercivity and inf-sup)**.** *Under Assumption 5.3, the following bounds hold*

$$a(\boldsymbol{v}, \boldsymbol{v}) \geq 2\underline{\mu}\, C_{\mathrm{K}}^{-1} \|\boldsymbol{v}\|_{\boldsymbol{U}}^2, \qquad \forall \boldsymbol{v} \in \boldsymbol{U}, \tag{5.21}$$

$$\sup_{\boldsymbol{0} \neq \boldsymbol{v} \in \boldsymbol{U}} \frac{b(\boldsymbol{v}, q)}{\|\boldsymbol{v}\|_{\boldsymbol{U}}} \geq \underline{\alpha}\, C_{\mathrm{is}} \|q\|_{L^2_\rho(\Omega, L^2(D))}, \qquad \forall q \in L^2_\rho(\Omega, L^2(D)), \tag{5.22}$$

where $C_{\mathrm{K}} > 0$ denotes the constant in Korn's first inequality, and $C_{\mathrm{is}} > 0$ the inf-sup constant. In the case $\Gamma_D = \partial D$, (5.22) holds for all $q \in L^2_\rho(\Omega, L^2(D) \cap \mathbb{R}(D)^\perp)$.

Proof. The first bound directly follows from Assumption 5.3. Indeed, since $|\Gamma_D|_{d-1} > 0$, we can apply Korn's first inequality, yielding

$$2\underline{\mu} \|\boldsymbol{\nabla} \boldsymbol{v}\|^2_{L^2_\rho(\Omega, \mathbb{L}^2(D))} \leq 2\underline{\mu} C_{\mathrm{K}} \left\langle \|\boldsymbol{\nabla}_{\mathrm{s}} \boldsymbol{v}\|^2_{\mathbb{L}^2(D)} \right\rangle \leq C_{\mathrm{K}} a(\boldsymbol{v}, \boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \boldsymbol{U}.$$

To obtain (5.22), we use [24, Lemma 7.2] establishing the existence, for any $q \in L^2_\rho(\Omega, L^2(D))$ (or $q \in L^2_\rho(\Omega, L^2(D) \cap \mathbb{R}(D)^\perp)$ in the case $\Gamma_d = \partial D$), of $\boldsymbol{v}_q \in \boldsymbol{U}$ such that $\boldsymbol{\nabla} \cdot \boldsymbol{v}_q = q$ and $C_{\mathrm{is}} \|\boldsymbol{v}_q\|_{\boldsymbol{U}} \leq \|q\|_{L^2_\rho(\Omega, L^2(D))}$, with $C_{\mathrm{is}} > 0$ depending on D. Thus, we conclude

$$\sup_{\boldsymbol{0} \neq \boldsymbol{v} \in \boldsymbol{U}} \frac{b(\boldsymbol{v}, q)}{\|\boldsymbol{v}\|_{\boldsymbol{U}}} \geq \frac{-b(\boldsymbol{v}_q, q)}{\|\boldsymbol{v}_q\|_{\boldsymbol{U}}} \geq \frac{\underline{\alpha} \|q\|^2_{L^2_\rho(\Omega, L^2(D))}}{\|\boldsymbol{v}_q\|_{\boldsymbol{U}}} \geq \underline{\alpha}\, C_{\mathrm{is}} \|q\|_{L^2_\rho(\Omega, L^2(D))}.$$

□

In order to prove the stability estimate of Proposition 5.5 below, we infer from (5.19) and (5.20) a lower bound of the specific storage coefficient c_0. We rewrite (5.13), derived from the definition of α in (5.4) and the Gassmann equation (5.12), in the form

$$\frac{(1-\alpha)(K_{\mathrm{m}}-K)}{\alpha} = K - \frac{\alpha}{c_0}.$$

Since the left-hand side of the previous relation is always non-negative (according to [61, Chapter 4] we have $K_{\mathrm{d}} \le K \le K_{\mathrm{m}}$), we infer that ${}^{\alpha}/_{c_0} \le K$ and, as a result,

$$c_0^{-1}(\boldsymbol{x},\boldsymbol{\xi}) \le \frac{K(\boldsymbol{x},\boldsymbol{\xi})}{\alpha(\boldsymbol{x},\boldsymbol{\xi})} = \frac{2\mu(\boldsymbol{x},\boldsymbol{\xi}) + d\lambda(\boldsymbol{x},\boldsymbol{\xi})}{d\alpha(\boldsymbol{x},\boldsymbol{\xi})} \le \frac{2\overline{\mu} + d\overline{\lambda}}{d\underline{\alpha}} \qquad \text{a.e. in } D\times\Omega. \tag{5.23}$$

Proposition 5.5 (A priori estimate)**.** *Let* $(\boldsymbol{u},p) \in L^2((0,t_{\mathrm{F}}),\boldsymbol{U}\times P)$ *solve* (5.18)*. Under Assumption 5.3, it holds*

$$\int_0^{t_{\mathrm{F}}} \left[a(\boldsymbol{u}(t),\boldsymbol{u}(t)) + c(p(t),p(t))\right]\mathrm{d}t \le \int_0^{t_{\mathrm{F}}} \left[\frac{C_{\mathrm{K}}}{2\underline{\mu}}\|\boldsymbol{f}\|^2_{\boldsymbol{L}^2(D)} + \frac{2\overline{\mu}+d\overline{\lambda}}{d\underline{\alpha}}\|G\|^2_{L^2(D)}\right]\mathrm{d}t, \tag{5.24}$$

with $G : (0,t_{\mathrm{F}}) \to L^2(D)$ *defined by*

$$G(t) := \int_0^t g(s)\mathrm{d}s + \phi_0. \tag{5.25}$$

Proof. From (5.18a) we infer that $a(\boldsymbol{u}(t),\boldsymbol{v}) + b(\boldsymbol{v},p(t)) = (\boldsymbol{f}(t),\langle\boldsymbol{v}\rangle)_D$, for a.e. $t \in (0,t_{\mathrm{F}})$ and all $\boldsymbol{v} \in \boldsymbol{U}$. Setting $\boldsymbol{v}(\boldsymbol{x},\boldsymbol{\xi}) = \boldsymbol{u}(\boldsymbol{x},\boldsymbol{\xi},t)$ and integrating the previous relation on $(0,t_{\mathrm{F}})$, yields

$$\int_0^{t_{\mathrm{F}}} a(\boldsymbol{u}(t),\boldsymbol{u}(t))\mathrm{d}t + \int_0^{t_{\mathrm{F}}} b(\boldsymbol{u}(t),p(t))\mathrm{d}t = \int_0^{t_{\mathrm{F}}} (\boldsymbol{f}(t),\langle\boldsymbol{u}(t)\rangle)_D\,\mathrm{d}t. \tag{5.26}$$

Adapting the argument of Remark 4.2 we ensure that $b(\boldsymbol{u}(t),q) - c(p(t),q) \in H^1((0,t_{\mathrm{F}}))$ for all $q \in P$. Thus, we can integrate by parts (5.18b) and obtain

$$\mathrm{d}_t\left[c(p(s),q) - b(\boldsymbol{u}(s),q)\right] + d(p(s),q) = (g(s),\langle q\rangle)_D, \qquad \text{for a.e. } s \in (0,t_{\mathrm{F}}).$$

Letting $t \in (0,t_{\mathrm{F}})$, integrating the previous identity on $(0,t)$, and then taking $q(\boldsymbol{x},\boldsymbol{\xi}) = p(\boldsymbol{x},\boldsymbol{\xi},t)$, leads to

$$c(p(t),p(t)) - b(\boldsymbol{u}(t),p(t)) + \int_0^t d(p(s),p(t))\mathrm{d}s = \int_0^t (g(s),\langle p(t)\rangle)_D\,\mathrm{d}s + (\phi_0,\langle p(t)\rangle)_D\,.$$

We define $z(t) = \int_0^t p(s)\mathrm{d}s$ and observe that $\mathrm{d}_t z(t) = p(t)$ and $z(0) = 0$. Using the linearity of d and the formula $\mathrm{d}_t z^2(t) = 2z(t)\mathrm{d}_t z(t)$ to rewrite the third term in the left hand side of the previous relation, and recalling (5.25) we get

$$c(p(t),p(t)) - b(\boldsymbol{u}(t),p(t)) + \frac{1}{2}\mathrm{d}_t\left[d(z(t),z(t))\right] = (G(t),\langle p(t)\rangle)_D\,\mathrm{d}s.$$

Then, integrating on $(0, t_F)$ and summing the resulting identity to (5.26), gives

$$\int_0^{t_F} a(\boldsymbol{u}(t), \boldsymbol{u}(t))\mathrm{d}t + \int_0^{t_F} c(p(t), p(t))\mathrm{d}t + \frac{1}{2} d(z(t_F), z(t_F)) = \int_0^{t_F} (\boldsymbol{f}(t), \langle \boldsymbol{u}(t)\rangle)_D \,\mathrm{d}t + \int_0^{t_F} (G(t), \langle p(t)\rangle)_D \,\mathrm{d}t. \quad (5.27)$$

To establish the result, we now bound the right-hand side of the previous relation. First, we observe that, owing to the Jensen inequality, for all $\boldsymbol{v} \in \boldsymbol{U}$ it holds

$$\|\langle \boldsymbol{v}\rangle\|^2_{\mathbf{L}^2(D)} = \int_D \left(\int_\Omega \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\xi})\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}\right)^2 \mathrm{d}\boldsymbol{x} \le \int_\Omega \int_D \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\xi})^2 \rho(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\xi} \le \|\boldsymbol{v}\|^2_{\boldsymbol{U}}.$$

Using the previous relation and the Cauchy–Schwarz and Young inequalities followed by (5.21), it is inferred that

$$\begin{aligned}\int_0^{t_F} (\boldsymbol{f}(t), \langle \boldsymbol{u}(t)\rangle)_D \,\mathrm{d}t &\le \int_0^{t_F} (2\underline{\mu} C_K^{-1})^{-1/2} \|\boldsymbol{f}\|_{\mathbf{L}^2(D)} (2\underline{\mu} C_K^{-1})^{1/2} \|\boldsymbol{u}\|_U \mathrm{d}t \\ &\le \frac{C_K}{4\underline{\mu}} \int_0^{t_F} \|\boldsymbol{f}\|^2_{\mathbf{L}^2(D)} \mathrm{d}t + \frac{1}{2} \int_0^{t_F} a(\boldsymbol{u}(t), \boldsymbol{u}(t))\mathrm{d}t.\end{aligned}$$

Proceeding similarly for the second term in the right-hand side of (5.27) and recalling the lower bound (5.23), one has

$$\begin{aligned}\int_0^{t_F} (G(t), \langle p(t)\rangle)_D \,\mathrm{d}t &\le \int_0^{t_F} \|c_0^{-1/2} G\|_{L^2_\rho(\Omega, L^2(D))} \|c_0^{1/2} p\|_{L^2_\rho(\Omega, L^2(D))} \mathrm{d}t \\ &\le \frac{2\overline{\mu} + d\overline{\lambda}}{2d\underline{\alpha}} \int_0^{t_F} \|G\|^2_{L^2(D)} \mathrm{d}t + \frac{1}{2} \int_0^{t_F} c(p(t), p(t))\mathrm{d}t.\end{aligned}$$

Plugging the previous two estimates in (5.27), multiplying by a factor 2, and using the non-negativity of κ to infer $d(z(t_F), z(t_F)) \ge 0$, yields the conclusion. □

Some remarks are in order.

Remark 5.6 (Inf-sup conditon). We observe that it is possible to derive the previous a priori estimate without (5.23). Indeed the inf-sup condition (5.22) together with (5.18a) allows to bound the second term in the right-hand side of (5.27) using the following: for a.e. $t \in (0, t_F)$,

$$\begin{aligned}\|p(t)\|_{L^2_\rho(\Omega, L^2(D))} &\le (\underline{\alpha} C_{is})^{-1} \sup_{\mathbf{0} \ne \boldsymbol{v} \in \boldsymbol{U}} \frac{b(\boldsymbol{v}, p(t))}{\|\boldsymbol{v}\|_U} = (\underline{\alpha} C_{is})^{-1} \sup_{\mathbf{0} \ne \boldsymbol{v} \in \boldsymbol{U}} \frac{(\boldsymbol{f}(t), \langle \boldsymbol{v}\rangle)_D - a(\boldsymbol{u}(t), \boldsymbol{v})}{\|\boldsymbol{v}\|_U} \\ &\le \frac{1}{\underline{\alpha} C_{is}} \|\boldsymbol{f}(t)\|_{\mathbf{L}^2(D)} + \frac{(2\overline{\mu} + d\overline{\lambda})^{1/2}}{d^{1/2} \underline{\alpha} C_{is}} a(\boldsymbol{u}(t), \boldsymbol{u}(t))^{1/2}.\end{aligned}$$

The resulting stability estimate would have, compared to (5.24), an additional dependence on $\underline{\alpha} C_{is}$.

Remark 5.7 (Quasi-incompressible media). In order to prove the stability estimate (5.24) no additional assumption on the mobility $\kappa : D \times \Omega \to \mathbb{R}^+$ is required. Thus, Proposition 5.5 can handle the case of locally poorly permeable media (*i.e.* $\kappa \to 0$). However, assuming (5.20) does not allow to consider quasi-incompressible materials for which $\lambda \to +\infty$. To obtain a robust estimate in the case of λ unbounded, we can proceed as in [130, Theorem 1]. Assuming $\boldsymbol{f} \in H^1((0,t_F), \mathbf{L}^2(D))$ and κ uniformly bounded away from zero, the Darcy term gives a $L^2((0,t_F), P)$ estimate of the pressure and, as a consequence, (5.20) is not needed. We remark that, in a medium featuring very low permeability, an incompressible fluid cannot flow unless the material is compressible. Therefore, the two limit cases $\kappa \to 0$ and $\lambda \to +\infty$ cannot occur simultaneously.

Remark 5.8 (Lamé's coefficients). The well-posedness of problem (5.18) holds under weaker assumption on μ and λ. More precisely, one can assume instead of (5.19) and (5.20), that ρ-a.e. in Ω there holds

$$0 < \underline{\mu}(\boldsymbol{\xi}) \le \mu(\boldsymbol{x},\boldsymbol{\xi}) \le \overline{\mu}(\boldsymbol{\xi}) < \infty, \quad 0 \le \lambda(\boldsymbol{x},\boldsymbol{\xi}) \le \overline{\lambda}(\boldsymbol{\xi}) < \infty \quad \text{a.e. in } D \times \Omega,$$

where $\underline{\mu}(\boldsymbol{\xi})$, $\overline{\mu}(\boldsymbol{\xi})$, and $\overline{\lambda}(\boldsymbol{\xi})$ are second-order random variables. An assumption of this type is convenient when λ is an unbounded (*e.g.* Gaussian or lognormal) random variable at $\boldsymbol{x} \in D$. See [11, Lemma 1.2] and [53] for a discussion in the context of elliptic PDEs with random data.

5.5 Discrete setting

In this section we introduce polynomial chaos expansions and outline the construction of the Pseudo-Spectral Projection (PSP) algorithm. The PSP method is one of the so-called non-intrusive techniques, which consist in performing an ensemble of deterministic model simulations to estimate the expansion coefficients. In the numerical tests of Section 5.6, each simulation is performed by solving the discrete problem of Section 2.2.5.

5.5.1 Polynomial chaos expansion

Let us denote $\{\phi_{\boldsymbol{k}}(\boldsymbol{\xi}) : \boldsymbol{k} \in \mathbb{N}^N\}$ an Hilbertian basis of $L^2_\rho(\Omega)$, where $\phi_{\boldsymbol{k}}$ is a multivariate polynomial in $\boldsymbol{\xi}$ and $\boldsymbol{k} = (k_1, \dots, k_N)$ is a multi-index indicating the polynomial degree in the ξ_i's. The total degree of $\phi_{\boldsymbol{k}}$ is denoted $|\boldsymbol{k}| := \sum_{i=1}^N k_i$. The basis functions are commonly chosen to be orthogonal with respect to the inner product in $L^2_\rho(\Omega)$ characterized by the probability density function $\rho : \Omega \to \mathbb{R}^+$,

$$\int_\Omega \phi_{\boldsymbol{k}}(\boldsymbol{\xi})\phi_{\boldsymbol{l}}(\boldsymbol{\xi})\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \delta_{\boldsymbol{k},\boldsymbol{l}}\langle\phi^2_{\boldsymbol{k}}\rangle. \tag{5.28}$$

For instance, Legendre and Hermite polynomials are used for uniform and Gaussian densities, respectively (cf. [203]). If $X(\boldsymbol{\xi}) \in L^2_\rho(\Omega)$ is a second-order random variable, then it admits the so-called Polynomial Chaos (PC) expansion [112, 136],

$$X(\boldsymbol{\xi}) = \sum_{\boldsymbol{k}\in\mathbb{N}^N} X_{\boldsymbol{k}}\phi_{\boldsymbol{k}}(\boldsymbol{\xi}), \quad X_{\boldsymbol{k}} := \langle X, \phi_{\boldsymbol{k}}\rangle = \int_\Omega X(\boldsymbol{\xi})\phi_{\boldsymbol{k}}(\boldsymbol{\xi})\rho(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}, \tag{5.29}$$

where the deterministic coefficients of the series $\{X_{\boldsymbol{k}} : \boldsymbol{k} \in \mathbb{N}^N\}$, named the *spectral modes*, are defined as the projection of $X(\boldsymbol{\xi})$ onto the basis functions. The PC approximation $X_{\mathcal{K}}(\boldsymbol{\xi})$ of $X(\boldsymbol{\xi})$ is obtained by truncating the expansion above to a finite series,

$$X_{\mathcal{K}}(\boldsymbol{\xi}) \coloneqq \sum_{\boldsymbol{k}\in\mathcal{K}} X_{\boldsymbol{k}}\phi_{\boldsymbol{k}}(\boldsymbol{\xi}), \tag{5.30}$$

where $\mathcal{K} \subset \mathbb{N}^N$ is the set of multi-indices related to the expansion. As a result, the expectation and variance operators applied to PC approximations simplify as follows:

$$\langle X_{\mathcal{K}} \rangle = X_{\mathbf{0}}, \tag{5.31}$$

$$\mathbb{V}\mathrm{ar}(X_{\mathcal{K}}) = \sum_{\boldsymbol{k}\in\mathcal{K}\setminus\mathbf{0}} X_{\boldsymbol{k}}^2 \langle \phi_{\boldsymbol{k}}^2 \rangle. \tag{5.32}$$

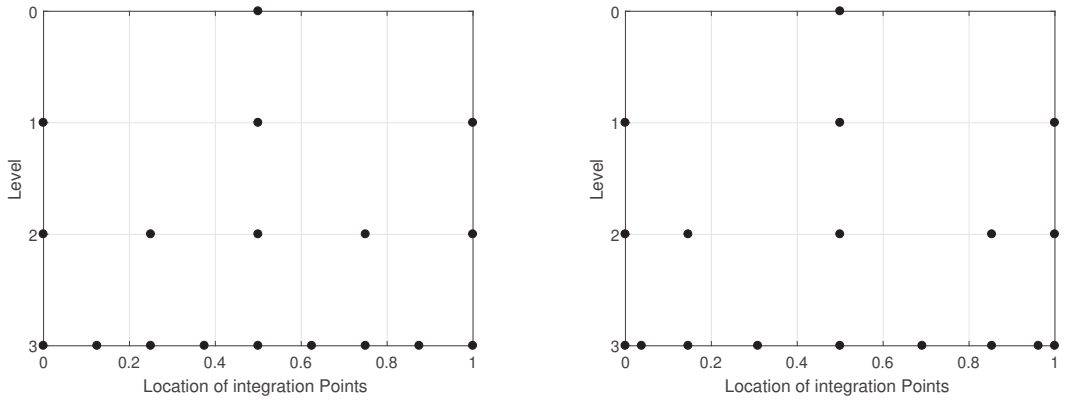
The truncation strategy, *i.e.* the selection of the set $\mathcal{K}$, depends on the method used to compute the spectral modes and the error $e_{\mathcal{K}}(\boldsymbol{\xi}) := X(\boldsymbol{\xi}) - X_{\mathcal{K}}(\boldsymbol{\xi})$ decreases when the PC degree increases. The convergence rate of the PC expansion of X with respect to the PC degree depends on the regularity of X. In particular, an exponential convergence is expected for analytical variables; more details about the convergence conditions can be found in [46, 98]. Several methods can be implemented to compute the spectral modes of the PC expansion [125, 136, 137] such as pseudo-spectral projection, least squares regression, and compressed sensing. These procedures mainly differ in the specific discrete norm used for minimizing the truncation error $e_{\mathcal{K}}(\boldsymbol{\xi})$. They are non-intrusively implemented by evaluating $X(\boldsymbol{\xi})$ at a finite number of values of $\boldsymbol{\xi}$ (the design of experiments) and therefore require only the availability of a deterministic solver. We focus here on the non-intrusive spectral projection approach [58, 59, 171] designed to avoid internal aliasing (the violation of the orthogonality condition (5.28) at the discrete level). The spectral projection method uses a quadrature rule in order to compute the spectral modes $X_{\boldsymbol{k}}$ by approximating (5.29),

$$X_{\boldsymbol{k}} = \langle X, \phi_{\boldsymbol{k}} \rangle \simeq \sum_{q=1}^{N_p} w_q X(\boldsymbol{\xi}^{(q)}) \phi_{\boldsymbol{k}}(\boldsymbol{\xi}^{(q)}), \tag{5.33}$$

where $\{\boldsymbol{\xi}^{(q)}\}$ and $\{w_q\}$ are the N_p integration points and associated weights. The previous equation shows that the main computational burden of the non-intrusive spectral projection technique is to compute N_p model evaluations for $X(\boldsymbol{\xi}^{(j)})$. In traditional tensor-product quadrature rules, N_p scales exponentially with the number of parameters N (the so-called "Curse of Dimensionality") and is intractable for all but very low dimensional problem. This issue has motivated the use of sparse quadrature formulas, which significantly reduce the computational cost.

5.5.2 Sparse Pseudo-Spectral Projection

In the numerical tests of Section 5.6, we rely on the sparse Pseudo-Spectral Projection (PSP) method that is based on the partial tensorization of nested one-dimensional quadrature rules



(a) Uniform points (b) Clenshaw–Curtis points

Figure 5.4: Examples of 1-D nested sequences including 1, 3, 5, and 9 quadrature points.

via Smolyak's formula [184]. Consider $Q_0, Q_1, Q_2, \ldots$ a nested sequence of 1-D quadrature formulas having increasing polynomial exactness $p_0, p_1, p_2, \ldots$, that is

$$\int_0^1 F(\xi)\rho(\xi)d\xi = Q_l F = \sum_{q=1}^{N_p(l)} F\left(\xi^{(q,l)}\right) w_{1D}^{(q,l)}, \quad \forall F \in P^{p_l}([0,1]),$$

where $N_p(l)$ is the number of points in the formula Q_l and $P^{p_l}([0,1])$ denotes the space spanned by polynomials of total degree p_l. We call l the level of the formula Q_l. Figure 5.4 shows 1-D nested sequences over the interval $[0,1]$ for level $1 \leq l \leq 3$ associated with uniform and Clenshaw–Curtis points.

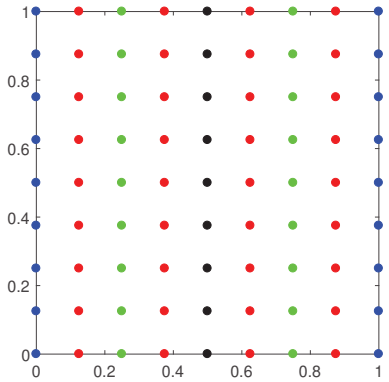
Prescribing the multi-index $\boldsymbol{l} = (l_1, \ldots, l_N) \in \mathbb{N}^N$, the full tensor-product quadrature of an integrable function F, can be written as follows:

$$\boldsymbol{Q}_{\boldsymbol{l}}^{\mathrm{FT}} F \coloneqq \left(Q_{l_1} \otimes \cdots \otimes Q_{l_N}\right) F = \sum_{q_1=1}^{N_p(l_1)} \cdots \sum_{q_d=1}^{N_p(l_N)} F\left(\xi_1^{(q_1,l_1)}, \ldots, \xi_N^{(q_N,l_N)}\right) w_{1D}^{(q_1,l_1)} \cdots w_{1D}^{(q_N,l_N)}. \tag{5.34}$$

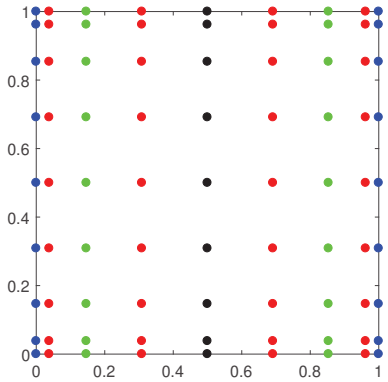
The previous quadrature rule can be used in (5.33) to compute the spectral modes $X_{\boldsymbol{k}}$, using $F = X\phi_{\boldsymbol{k}}$, for all $\boldsymbol{k} \in \mathcal{K}$. The complexity of $\boldsymbol{Q}_{\boldsymbol{l}}^{\mathrm{FT}}$ for such nested sequences is $N_p(\boldsymbol{l}) = \prod_{i=1}^N N_p(l_i)$, and so it increases exponentially with N. Sparse grids mitigate this complexity by first introducing the 1-D difference formulas between two successive levels, $\Delta_l^Q = Q_l - Q_{l-1}$, $\Delta_0^Q = Q_0$, such that (5.34) can be reformulated as

$$\boldsymbol{Q}_{\boldsymbol{l}}^{\mathrm{FT}} F = \left(\left(\sum_{i_1=0}^{l_1} \Delta_{i_1}^Q\right) \otimes \cdots \otimes \left(\sum_{i_N=0}^{l_N} \Delta_{i_N}^Q\right)\right) F = \sum_{i_1=0}^{l_1} \cdots \sum_{i_N=0}^{l_N} \left(\Delta_{i_1}^Q \otimes \cdots \otimes \Delta_{i_N}^Q\right) F = \sum_{\boldsymbol{i} \in \mathcal{L}_{\boldsymbol{l}}^{\mathrm{FT}}} \boldsymbol{\Delta}_{\boldsymbol{i}}^Q F,$$

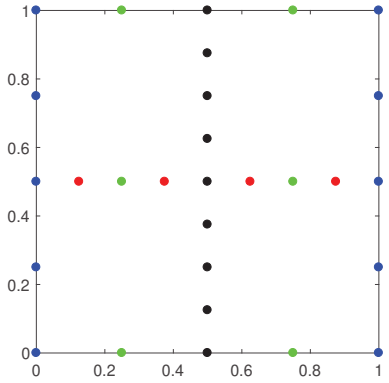
where $\mathcal{L}_{\boldsymbol{l}}^{\mathrm{FT}} = \{\boldsymbol{i} \in \mathbb{N}^N, i_j \leq l_j \text{ for } j = 1, \ldots, d\}$ is the multi-index set of full tensorizations. The sparse quadrature rule $\boldsymbol{Q}_{\mathcal{L}}$ is finally constructed by considering the summation over a



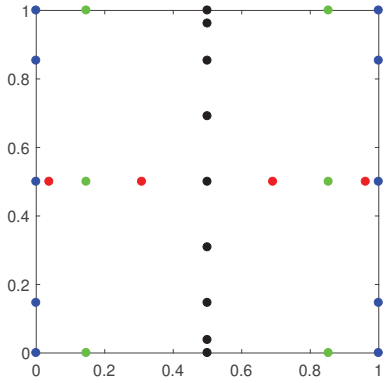
(a) Uniform points, full tensorization



(b) Clenshaw–Curtis points, full tensorization



(c) Uniform points, partial tensorization



(d) Clenshaw–Curtis points, partial tensorization

Figure 5.5: Examples of tensor-product quadrature grids for $N = 2$ and $\boldsymbol{l} = (3, 3)$. In the full tensorization (top) the colored points correspond to the colored contributions in $\boldsymbol{Q}_{\boldsymbol{l}}^{\mathrm{FT}} = Q_0 \otimes Q_3 + \Delta_1^Q \otimes Q_3 + \Delta_2^Q \otimes Q_3 + \Delta_3^Q \otimes Q_3$. In Smolyak's tensorization (bottom) the colored points correspond to the colored contributions in $\boldsymbol{Q}_{\mathcal{L}} = Q_0 \otimes Q_3 + \Delta_1^Q \otimes Q_2 + \Delta_2^Q \otimes Q_1 + \Delta_3^Q \otimes Q_0$.

subset $\mathcal{L}$ of tensorized quadrature differences:

$$\boldsymbol{Q}_{\mathcal{L}} F = \sum_{\boldsymbol{i} \in \mathcal{L}} \boldsymbol{\Delta}_{\boldsymbol{i}}^Q F, \quad \mathcal{L} \subset \mathcal{L}_{\boldsymbol{l}}^{\mathrm{FT}}.$$

The set of tensorizations $\mathcal{L}$ must be admissible in the sense that the following condition holds [108, 109]

$$\forall\, \boldsymbol{i} = (i_1, \ldots, i_N) \in \mathcal{L} : i_{1 \le j \le N} \ge 1 \Rightarrow \boldsymbol{i} - \boldsymbol{e}_j \in \mathcal{L},$$

where $\{\boldsymbol{e}_j : j = 1, \ldots, N\}$ is the set containing the canonical unit vectors of $\mathbb{N}^N$. This admissibility condition is necessary to preserve the telescopic property of the sum of quadrature differences. In the following, we denote by $\mathcal{G}(\mathcal{L})$ the set of nodes in the sparse grid,

$$\mathcal{G}(\mathcal{L}) \coloneqq \bigcup_{\boldsymbol{i} \in \mathcal{L}} \left\{ \left(\xi^{(q_1, i_1)}, \ldots, \xi^{(q_N, i_N)} \right), 1 \le q_j \le N_p(i_j), 1 \le j \le N \right\}, \quad N_p(\mathcal{L}) = |\mathcal{G}(\mathcal{L})|,$$

where $|\cdot|$ is the cardinality of a set. In Figure 5.5 we plot the grids corresponding to full and sparse tensor-product quadrature rules in the case of $N = 2$ and $\boldsymbol{l} = (3, 3)$.

A quadrature rule $\boldsymbol{Q}_{\mathcal{L}}$ is said to be *sufficiently exact* with respect to a polynomial multi-index set $\mathcal{K}$ when for any couple $(\boldsymbol{k}, \boldsymbol{k}') \in \mathcal{K} \times \mathcal{K}$, the basis orthonormality conditions (5.28) are recovered using the discrete inner product defined, for all $X, Y \in L^2_\rho(\Omega)$, by

$$\langle X, Y \rangle_{\boldsymbol{Q}_{\mathcal{L}}} \coloneqq \boldsymbol{Q}_{\mathcal{L}}(XY).$$

In the case of full tensorizations, the largest set $\mathcal{K}^*(\boldsymbol{l})$ for which internal aliasing errors vanish can be easily determined from the degrees of polynomial exactness of the 1-D sequence, namely

$$\mathcal{K}^*(\boldsymbol{l}) = \left\{ \boldsymbol{k} \in \mathbb{N}^N : k_i \leq \frac{p_{l_i}}{2},\ i = 1, \dots, N \right\}.$$

An immediate consequence of using the sparse quadrature formula with $\mathcal{L} \subset \mathcal{L}_{\boldsymbol{l}}^{\mathrm{FT}}$, is that $\boldsymbol{Q}_{\mathcal{L}}$ is not sufficiently exact with respect to $\mathcal{K}^*(\boldsymbol{l})$, as the sparse formula is not able to exactly integrate the product of high-order monomials. We observe that the use of sparse quadrature, while reducing the complexity ($N_p(\mathcal{L}) \ll \prod_{j=1}^N N_p(l_j)$), yields also a significant reduction of the (not uniquely-defined) largest set $\mathcal{K}^*(\mathcal{L})$ for which the quadrature is sufficiently exact.

The PSP method allows to consider polynomial spaces larger than $\mathcal{K}^*(\mathcal{L})$, for the same sparse grid $\mathcal{G}(\mathcal{L})$ and without introducing internal aliasing. Let $\{\Pi_l\}_{l \geq 1}$ denote the sequence of 1-D projection operators associated to the sequence $\{Q_l\} l \geq 1$ of 1-D nested quadrature, where

$$\Pi_l : X(\xi) \mapsto \Pi_l X(\xi) \coloneqq \sum_{k=0}^{p_l/2} X_k \phi_k(\xi) \in P^{\frac{p_l}{2}}([0,1]), \quad X_k = \sum_{q=1}^{N_p(l)} X(\xi^{(q,l)}) \phi_k(\xi^{(q,l)}) w_{1D}^{(q,l)}.$$

Note that $\Pi_{l \geq 1}$ is free of internal aliasing, owing to the dependence on l of the projection spaces, which ensures that the quadrature is sufficiently exact. Consider for $l \geq 1$ the difference of successive 1-D projection operators $\Delta_{l>1}^{\Pi} = \Pi_l - \Pi_{l-1}$, $\Delta_1^{\Pi} = \Pi_1$. It is possible to write the full tensor-product projection corresponding to the quadrature rule given by (5.34) in terms of a sum of tensorized difference operators by observing that

$$\begin{aligned} \boldsymbol{\Pi}_{\boldsymbol{l}}^{\mathrm{FT}} X &\coloneqq \sum_{\boldsymbol{k} \in \mathcal{K}^*(\boldsymbol{l})} \boldsymbol{Q}_{\boldsymbol{l}}^{\mathrm{FT}}(X \phi_{\boldsymbol{k}})(\boldsymbol{\xi}) = \left(\Pi_{l_1} \otimes \cdots \otimes \Pi_{l_N}\right) X \\ &= \sum_{i_1=1}^{l_1} \cdots \sum_{i_N=1}^{l_N} \left(\Delta_{i_1}^{\Pi} \otimes \cdots \otimes \Delta_{i_N}^{\Pi}\right) X = \sum_{\boldsymbol{i} \in \mathcal{L}_{\boldsymbol{l}}^{\mathrm{FT}}} \left(\Delta_{i_1}^{\Pi} \otimes \cdots \otimes \Delta_{i_N}^{\Pi}\right) X. \end{aligned}$$

The sparse PSP operator $\boldsymbol{\Pi}_{\mathcal{L}}$ is finally obtained by considering a summation over an admissible subset $\mathcal{L} \subset \mathcal{L}_{\boldsymbol{l}}^{\mathrm{FT}}$ of tensorized difference projection operators. This results in

$$\boldsymbol{\Pi}_{\mathcal{L}} X = \sum_{\boldsymbol{i} \in \mathcal{L}} \left(\Delta_{i_1}^{\Pi} \otimes \cdots \otimes \Delta_{i_N}^{\Pi}\right) X. \tag{5.35}$$

The key point is that the sparse PSP operator in (5.35) involves a telescopic sum of differences

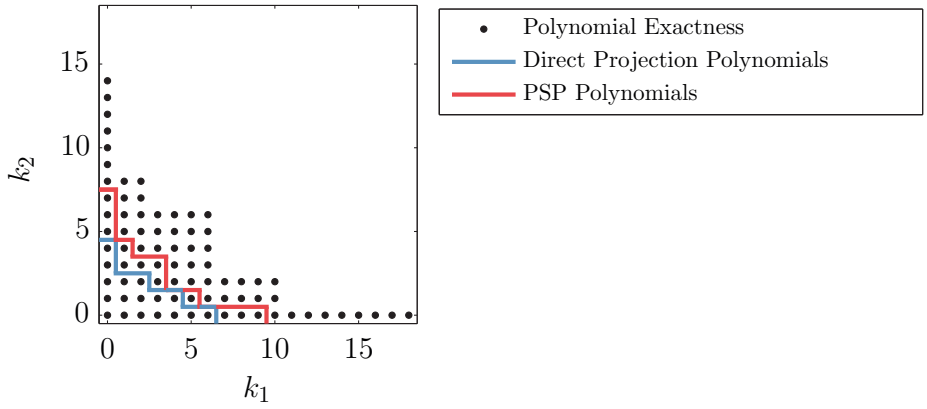


Figure 5.6: Comparison of polynomials multi-index sets $\mathcal{K}(\mathcal{L})$ (below the red line) and $\mathcal{K}^*(\mathcal{L})$ (below the blue line) using a quadrature rule $\boldsymbol{Q}_{\mathcal{L}}$ in $d = 2$ dimensions. Points are multi-indices (k_1, k_2) such $F \in P^{k_1}([0,1]) \times P^{k_1}([0,1])$ is exactly integrated by $\boldsymbol{Q}_{\mathcal{L}}$.

in the projection onto subspaces of increasing dimensions, where each individual tensorized quadrature projection is evaluated without internal aliasing. Moreover, $\boldsymbol{\Pi}_{\mathcal{L}}$ belongs to the span of $\{\phi_{\boldsymbol{k}} : \boldsymbol{k} \in \mathcal{K}(\mathcal{L})\}$, with $\mathcal{K}(\mathcal{L}) = \bigcup_{\boldsymbol{i} \in \mathcal{L}} \mathcal{K}^*(\boldsymbol{i})$. We have $\mathcal{K}(\mathcal{L}) \supseteq \mathcal{K}^*(\mathcal{L})$ and the projection space of the sparse PSP is usually significantly larger than that of the direct sparse projection. Nevertheless, the two methods have the same complexity relying both on the same sparse grid $\mathcal{G}(\mathcal{L})$. Note, however, that the global sparse quadrature rule $\boldsymbol{Q}_{\mathcal{L}}$ is generally not sufficiently accurate with respect to $\mathcal{K}(\mathcal{L})$. The inclusion of the sets $\mathcal{K}(\mathcal{L})$ and $\mathcal{K}^*(\mathcal{L})$ is illustrated in Figure 5.6.

5.6 Test cases

We assess the performance of the PSP method detailed in the previous section on two model problems. The first one is an injection test designed to investigate the effect of a fluid source on the deformation of the domain, while the second one is meant to evaluate the mechanical effect of a traction term on the flux. Our numerical experiments feature a simplified model with homogeneous random coefficients, namely the dependence of the spacial variable $\boldsymbol{x}$ in (5.16) is dropped. More realistic models, requiring (possibly strongly correlated) spatially varying input coefficients with non-standard distributions will be the object of future works. After presenting the description of the test cases, we numerically investigate the convergence of the probability error with respect to the level of the sparse grid. Finally, we perform a sensitivity analysis in order to rank the contribution of the input parameters on the variance of the solutions.

Our input uncertainty is parametrized by a vector $\boldsymbol{\xi}$ of dimension $N = 4$. In particular, we assume that μ, λ, and κ have a log-uniform distribution, while α is uniformly distributed. As proposed in Section 5.3.2, the specific storage coefficient c_0 is evaluated from the other uncertain input parameters. The advantage of this approach is twofold: (i) it ensures that

the set of poroelastic coefficients belong to the physical admissible set, and (ii) it allows to reduce the uncertainty dimension and then the size of the sampling (*i.e.* , number of sparse grid points). Even if the experiments presented here are mainly academic model problems, we try to consider realistic sets of parameters and have a particular regard for the underlying physical phenomena. The ranges of variation of the uncertain input data are inspired from an ideal water filled sand soil with a $\varphi = 20\%$ reference porosity. In both tests we set

$$\begin{aligned} \mu(\boldsymbol{\xi}) &= 10^{2\xi_1} \text{ KPa}, \\ \lambda(\boldsymbol{\xi}) &= 2 \cdot 10^{2\xi_2} \text{ KPa}, \\ \alpha(\boldsymbol{\xi}) &= \frac{3\varphi}{2+\varphi} + \xi_3 \frac{2(1-\varphi)}{2+\varphi}, \\ \kappa(\boldsymbol{\xi}) &= 10^{2(\xi_4-1)} \text{ m}^2\text{KPa}^{-1}\text{s}^{-1}, \end{aligned} \tag{5.36}$$

with uniformly distributed $(\xi_1, \ldots, \xi_4) \in [0,1]^4$. This choice yields

$$\begin{aligned} &\langle\mu\rangle \sim 21.5 \text{ KPa}, \quad \langle\lambda\rangle \sim 43 \text{ KPa}, \quad \langle\alpha\rangle \sim 0.64, \quad \text{and} \quad \langle\kappa\rangle \sim 0.22 \frac{\text{m}^2}{\text{KPa s}}, \\ &\text{c}_\text{v}(\mu) \sim 1.16, \qquad \text{c}_\text{v}(\lambda) \sim 1.16, \quad \text{c}_\text{v}(\alpha) \sim 0.33, \quad \text{and} \quad \text{c}_\text{v}(\kappa) \sim 1.16, \end{aligned} \tag{5.37}$$

where $\text{c}_\text{v}(\cdot) := \frac{\mathbb{V}\text{ar}(\cdot)^{1/2}}{\langle\cdot\rangle}$ denotes the coefficient of variation operator. In particular we observe that the previous average values have the same magnitude as the coefficients considered in [102, Table 2]. Our choice leads to a specific storage coefficient c_0 ranging in $(10^{-3}, 10^{-1})$.

5.6.1 Point injection

The first experiment is inspired by the Barry and Mercer test considered in Section 2.6.2 (see also [14, 166]). On the boundary of the unit square domain $D = [0,1]^2$ we impose the following time-independent homogeneous boundary conditions:

$$\boldsymbol{u} \cdot \boldsymbol{\tau} = 0, \qquad \boldsymbol{n}^\text{T} \boldsymbol{\nabla u}\, \boldsymbol{n} = 0, \qquad p = 0,$$

where $\boldsymbol{\tau}$ denotes the tangent vector on ∂D. The loading term $\boldsymbol{f}$ in (5.17a) and the initial condition ϕ_0 in (5.17g) are both chosen to be zero. The evolution of the displacement and pressure fields solving (5.17) is only driven by a stationary fluid source located at $\boldsymbol{x}_0 = (0.25, 0.25)$:

$$g = 10\, \delta(\boldsymbol{x} - \boldsymbol{x}_0),$$

with δ denoting the Dirac delta function. The following results have been obtained using the HHO–dG coupled method with polynomial degree $k = 3$ on a Cartesian mesh containing 1,024 elements. Using the static condensation procedure detailed in Section 2.5, the spacial discretization consists of 27,648 unknowns. We are interested in the stationary solution, namely the pressure and displacement fields manifesting after the initial transient phase. Therefore, we only focus on the configuration at the final time $t_\text{F} = 1\text{s}$ obtained after 10 time steps of the implicit Euler scheme.

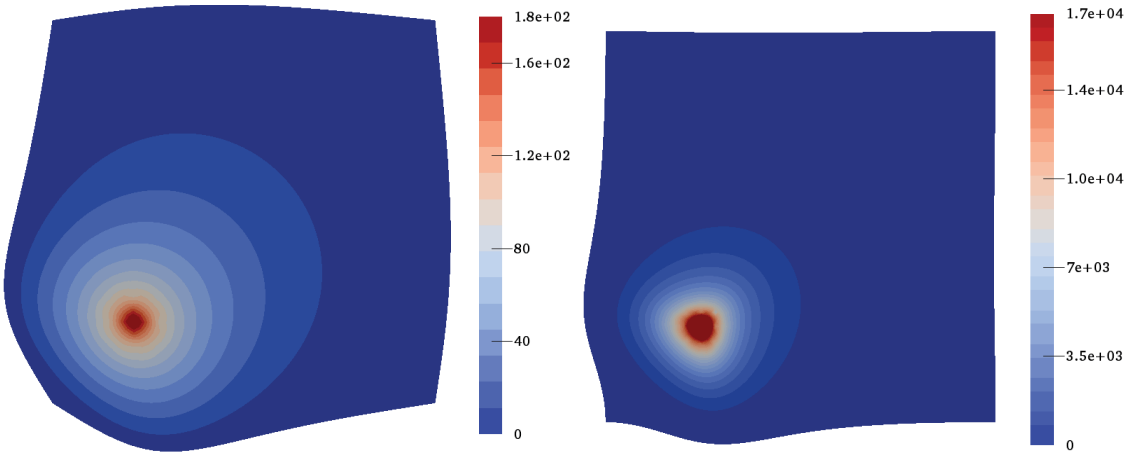


Figure 5.7: Pressure plotted on the deformed domain at $t = t_F = 1$ (data are in KPa). Mean (Left) and Variance (Right) of the PSP method with level $l = 5$ and $N_p(l) = 2561$.

Figure 5.7 represents the two first PC statistical moments given by (5.31) and (5.32) of the displacement and pressure solutions approximated using a level $l = 5$ sparse grid. We plot the pressure field on the deformed configuration obtained by applying the displacement to the reference unit square domain. As expected, we observe a dilatation of the domain caused by the injection. The computed solutions are symmetric with respect to the diagonal according to the source location and the homogeneous boundary conditions. The pressure mean field exhibits maximum values around the source, while no perturbation is observed near the boundary. The pressure variance field has the same behavior as the mean, with higher values close to the injection point. Moreover, we notice that the magnitude of the standard deviation (*i.e.* the square root of the variance) is comparable to the mean field range.

5.6.2 Poroelastic footing

As second example, we consider the 2D footing problem of [102, 155]. The simulation domain is a unit square porous soil that is assumed to be free to drain and fixed at the base and at the vertical edges. A uniform load is applied on a portion of the upper boundary $\Gamma_N = \{\boldsymbol{x} = (x_1, x_2) \in \partial D : x_2 = 1\}$ simulating a footing step compressing the medium. Thus, the boundary conditions data are given as follows:

$$\begin{aligned}
\boldsymbol{\sigma}\boldsymbol{n} &= (0, -5) \text{ KPa} \quad && \text{on } \Gamma_{N,1} := \{\boldsymbol{x} = (x_1, x_2) \in \partial D \ : \ 0.3 \le x_1 \le 0.7, \ x_2 = 1\}, \\
\boldsymbol{\sigma}\boldsymbol{n} &= \boldsymbol{0}, && \text{on } \Gamma_N \setminus \Gamma_{N,1}, \\
\boldsymbol{u} &= \boldsymbol{0}, && \text{on } \partial D \setminus \Gamma_N, \\
p &= 0, && \text{on } \partial D.
\end{aligned}$$

Also in this case we take $\boldsymbol{f} = \boldsymbol{0}$ and $\phi_0 = 0$, so that the displacement and pressure solutions are determined only by the prescription of the stress boundary condition. Thus, the situation is complementary to the one of the previous test, where the deformation of the medium was caused by a fluid term. We solve the Biot problem (5.17) with uncertain coefficients defined in (5.36). For the space discretization of the footing problem, we consider a triangular mesh

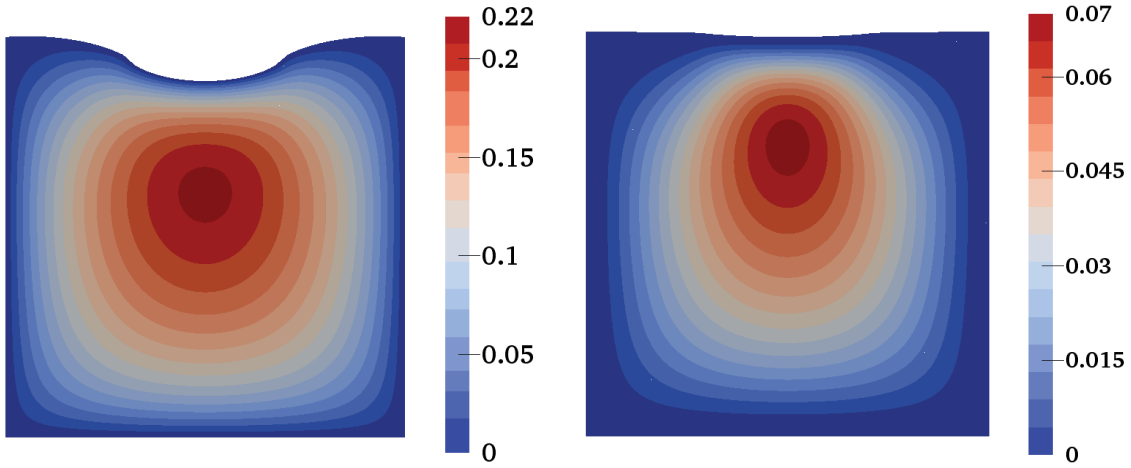


Figure 5.8: Pressure plotted on the deformed domain at $t = 0.2$. Mean (left) and Variance (right) computed using the PSP method with level $l = 5$ and $N_p(l) = 2561$.

with 3,584 elements and we use the HHO–dG method with $k = 2$. For each point of the sparse grid and each time iteration, the linear system to be solved has dimension 54,720. In this test case, we are particularly interested in the pressure profile at early times. In fact, owing to the free drainage boundary condition, after a few time steps the fluid is completely squeezed out of the soil, yielding an equilibrium configuration with no pressure and the displacement balancing the Neumann condition. We consider the approximated solution at time $t = 0.2$s obtained after two time steps of the implicit Euler method. We mention, in passing, that the unphysical pressure oscillations observed in [155] do not occur with our choice of the discretization (this is in agreement with the results of Section 2.6.2). We plot in Figure 5.8 the mean and variance fields corresponding to the PC approximation of the solution. Concerning the mean, we observe a significant deformation where the load is applied and a maximum pressure value near the center of the domain. The variance fields exhibit a similar configuration.

5.6.3 Convergence

Let $V(D)$ be a normed vector space of functions over D and $X \in L^2_\rho(\Omega, V(D))$ a random field. We evaluate the predictive capacity of the PC expansion $X_{\mathcal{K}}$ defined in (5.29) using the mean-squared error (MSE) computed from a validation set of $N^* = 500$ Latin Hypercube Samples (LHS) and defined as

$$\mathrm{MSE}(X - X_{\mathcal{K}})(\boldsymbol{x}) := \frac{1}{N^*} \sum_{i=1}^{N^*} \left(X(\boldsymbol{x}, \boldsymbol{\xi}^i) - X_{\mathcal{K}}(\boldsymbol{x}, \boldsymbol{\xi}^i) \right)^2 .$$

We plot in figures 5.10 and 5.11 the MSE fields of the PC solutions of the injection and footing test, respectively. Owing to the symmetry of the point injection test with respect to the diagonal, in 5.10 we only plot the first component of $\mathrm{MSE}(\boldsymbol{u} - \boldsymbol{u}_{\mathcal{K}})(\boldsymbol{x})$. Similarly, for the footing problem we consider the second component. The results are obtained using sparse

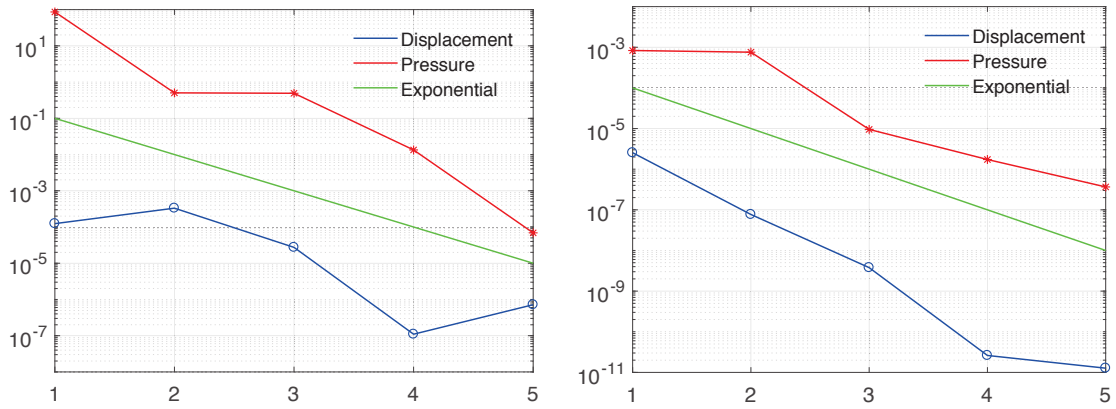


Figure 5.9: Errors $\|\mathrm{MSE}(\boldsymbol{u} - \boldsymbol{u}_{\mathcal{K}})\|_{\mathbf{L}^2(D)}$ and $\|\mathrm{MSE}(p - p_{\mathcal{K}})\|_{L^2(D)}$ vs. level l of the Sparse Grid. Point injection test (left) and footing test (right) with model coefficients $\mu, \lambda, \alpha, \kappa$ distributed according to (5.36).

grid of level $l = 1, 3, 5$. We observe a rapid decrease of the MSE field for both experiments and both the displacement and pressure. We especially note the strong decay of the maximal value with respect to the increasing sparse grid level. In Figure 5.9 we plot the $L^2(D)$-norm of the mean-squared error fields $\|\mathrm{MSE}(\boldsymbol{u} - \boldsymbol{u}_{\mathcal{K}})\|_{\mathbf{L}^2(D)}$ and $\|\mathrm{MSE}(p - p_{\mathcal{K}})\|_{L^2(D)}$. The linear trends observed in the semi-log scale of the sequential PC approximations suggest that the method achieves roughly exponential convergence with respect to the level of the sparse grid, as predicted by theory.

5.6.4 Sensitivity analysis

We conclude the numerical investigation by assessing the different contributions of the input parameters on the variance of the solutions. For each of the experiments presented above, we compute the first order conditional variances

$$\mathbb{V}\mathrm{ar}(\boldsymbol{u}_{\mathcal{K}}(\boldsymbol{x}, \boldsymbol{\xi})|\xi_i)_{1\leq i\leq 4}, \quad \text{and} \quad \mathbb{V}\mathrm{ar}(p_{\mathcal{K}}(\boldsymbol{x}, \boldsymbol{\xi})|\xi_i)_{1\leq i\leq 4},$$

of the PC expansions obtained with the finer sparse grid ($l = 5$). Since each ξ_i in (5.36) parametrizes only one of the poromechanical model coefficients, the previous quantities allows to evaluate and differentiate the effect of the perturbations of $\mu, \lambda, \alpha, \kappa$ on the solutions.

Concerning the injection test, we are mainly interested in assessing the effect of the fluid source on the deformation. Therefore, we perform the sensitivity analysis considering the displacement as the quantity of interest. On the other hand, in the footing experiment we focus on the pressure. In Figure 5.12 we compare the total variance $\mathbb{V}\mathrm{ar}(u_{\mathcal{K},1})$ defined in (5.32) with the sum of the first-order conditional variances $\sum_{i=1}^{4} \mathbb{V}\mathrm{ar}(u_1(\boldsymbol{x}, \boldsymbol{\xi})|\xi_i)$ of the PC approximated displacement solving the injection test. In Figure 5.14 we perform the same comparison for the footing problem considering the pressure variance $\mathbb{V}\mathrm{ar}(p_{\mathcal{K},1})$ and its first-order conditional contributions. The difference between the two fields indicates that the contribution due to second-order conditional variances is not negligible. This is probably due to the fact that computing c_0 using (5.14) yields a dependence $c_0 = c_0(\xi_1, \xi_2, \xi_3)$. We observe in Figures 5.13 and 5.15 that, in both experiments, the parameter with the main

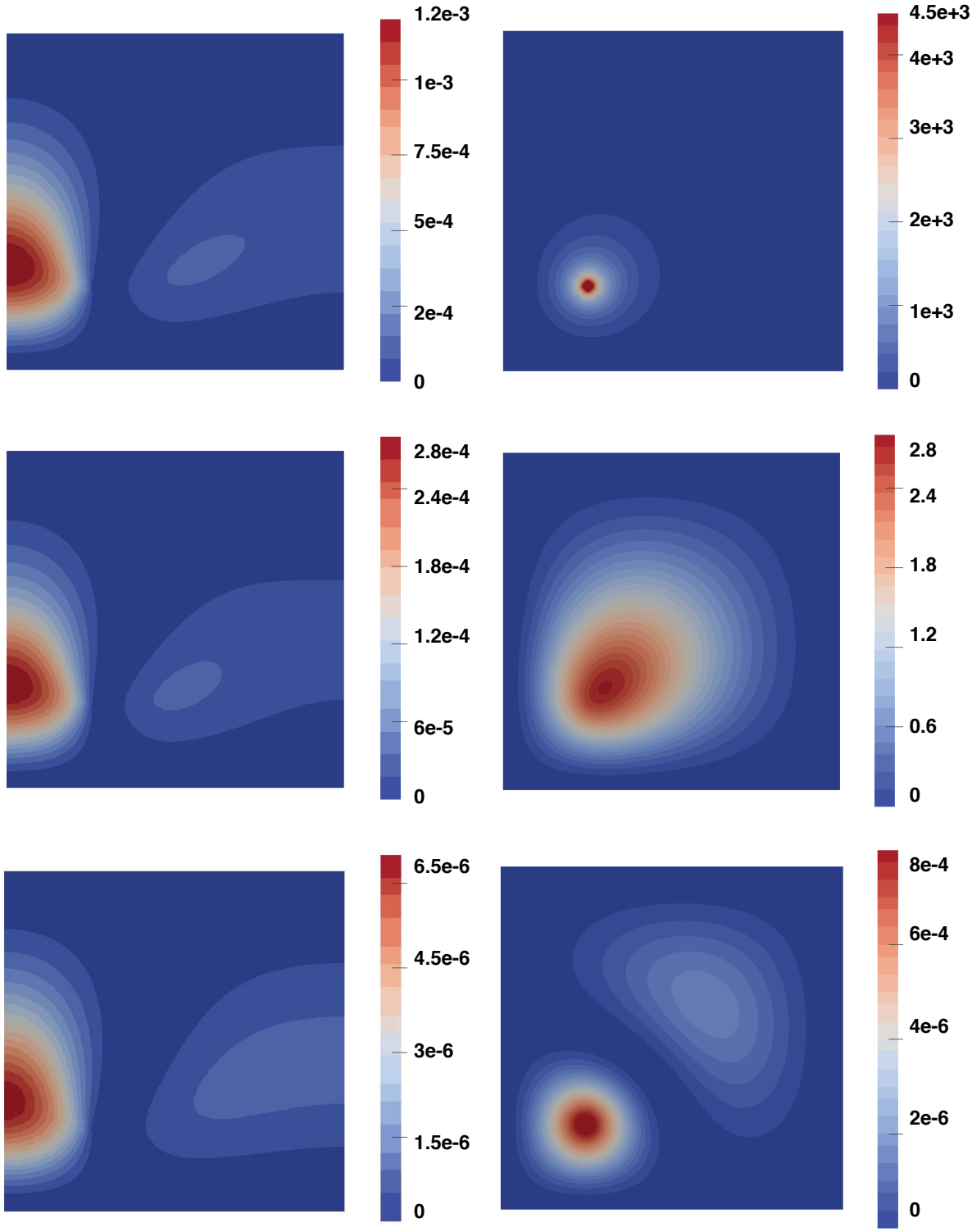


Figure 5.10: Displacement MSE($u_1 - u_{\mathcal{K},1}$) field (left) and pressure MSE($p - p_{\mathcal{K}}$) field (right) of the injection test case. Results are computed using the PSP method with $l = 1$ (top), $l = 3$ (middle), and $l = 5$ (bottom).

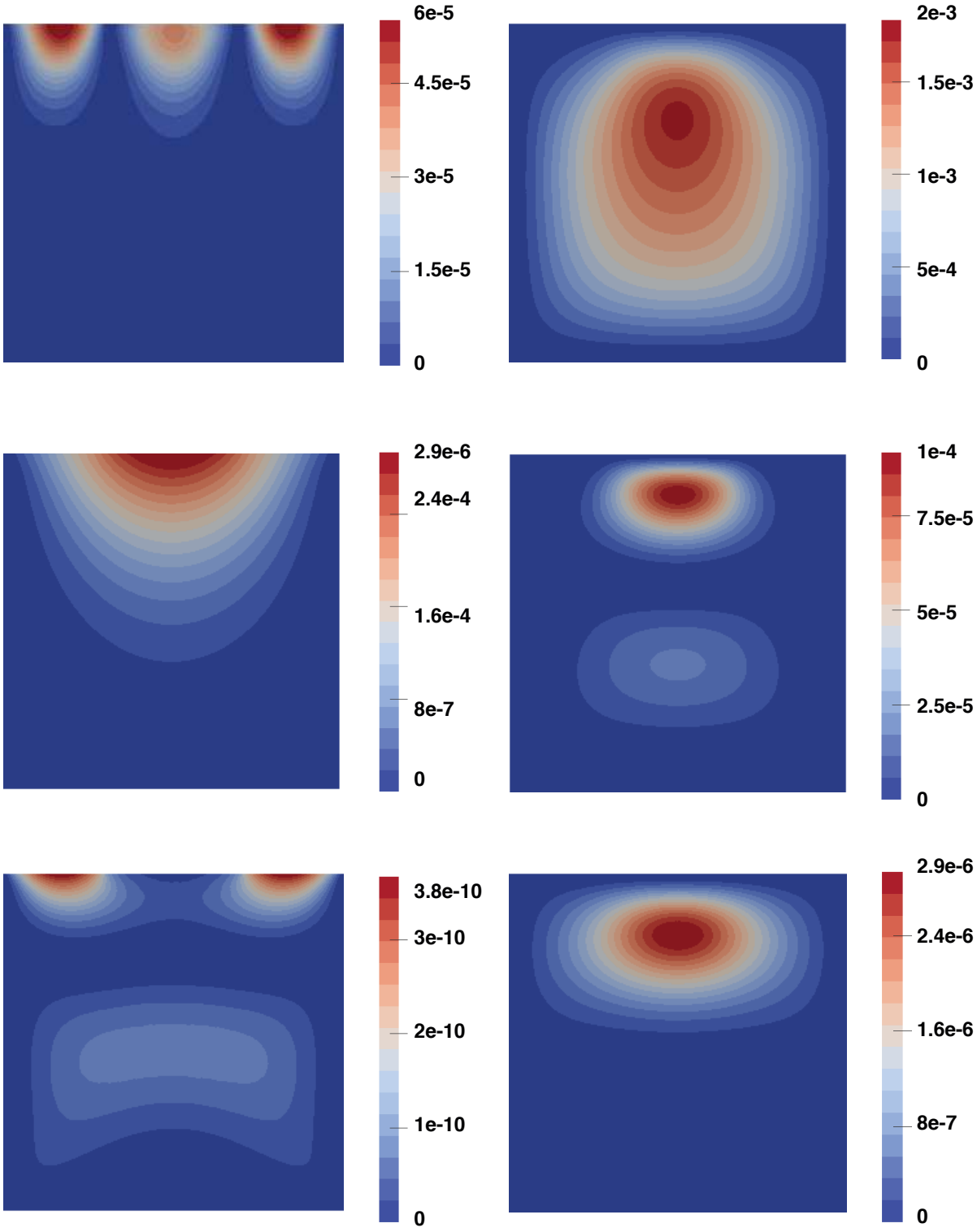


Figure 5.11: Displacement MSE($u_2 - u_{\mathcal{K},2}$) field (left) and pressure MSE($p - p_{\mathcal{K}}$) field (right) corresponding to the footing test case. Results computed using the PSP method with $l = 1$ (top), $l = 3$ (middle), and $l = 5$ (bottom).

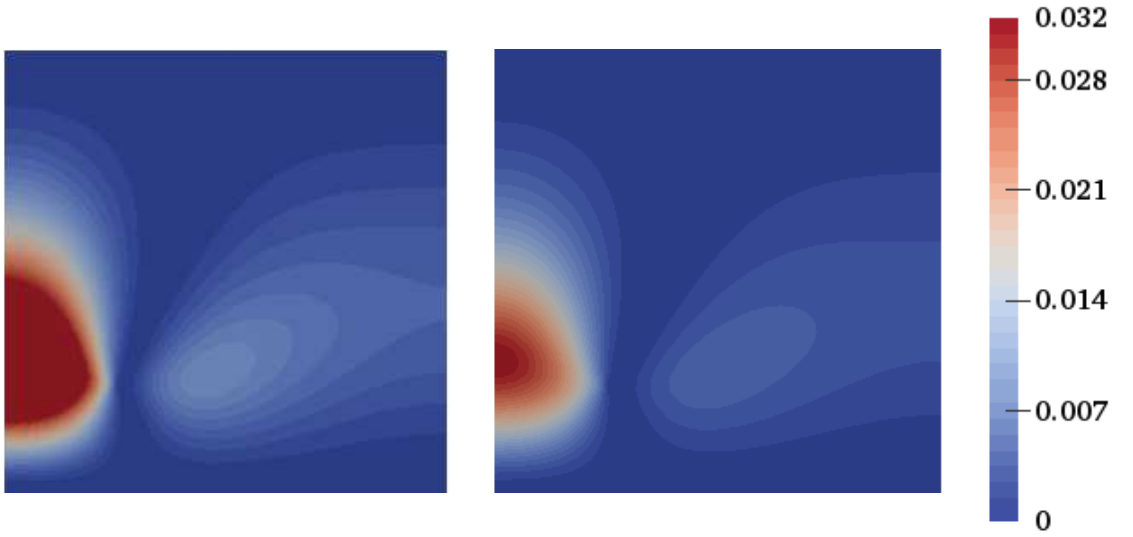


Figure 5.12: Point injection test: Comparison between the total variance $\mathbb{V}\mathrm{ar}(u_1(\boldsymbol{x},\boldsymbol{\xi}))$ and the sum of the first-order conditional variances $\sum_{i=1}^{4}\mathbb{V}\mathrm{ar}(u_1(\boldsymbol{x},\boldsymbol{\xi})|\xi_i)$ estimated using the PSP method with $l = 5$.

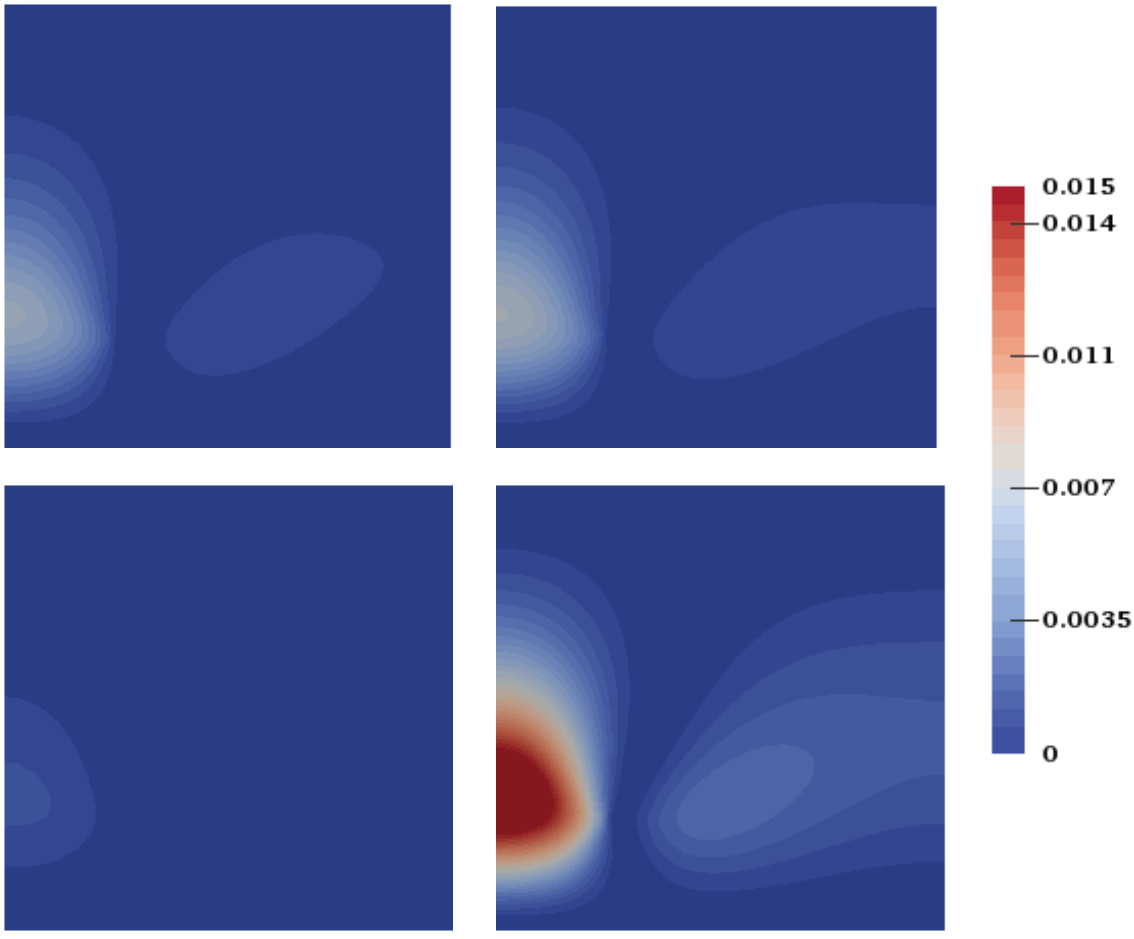


Figure 5.13: Point injection test: Conditional variances $\mathbb{V}\mathrm{ar}(u_1(\boldsymbol{x},\boldsymbol{\xi})|\xi_1)$ (top left) $\mathbb{V}\mathrm{ar}(u_1(\boldsymbol{x},\boldsymbol{\xi})|\xi_2)$ (top right) $\mathbb{V}\mathrm{ar}(u_1(\boldsymbol{x},\boldsymbol{\xi})|\xi_3)$ (bottom left) $\mathbb{V}\mathrm{ar}(u_1(\boldsymbol{x},\boldsymbol{\xi})|\xi_4)$ (bottom right) estimated using the PSP method with $l = 5$.

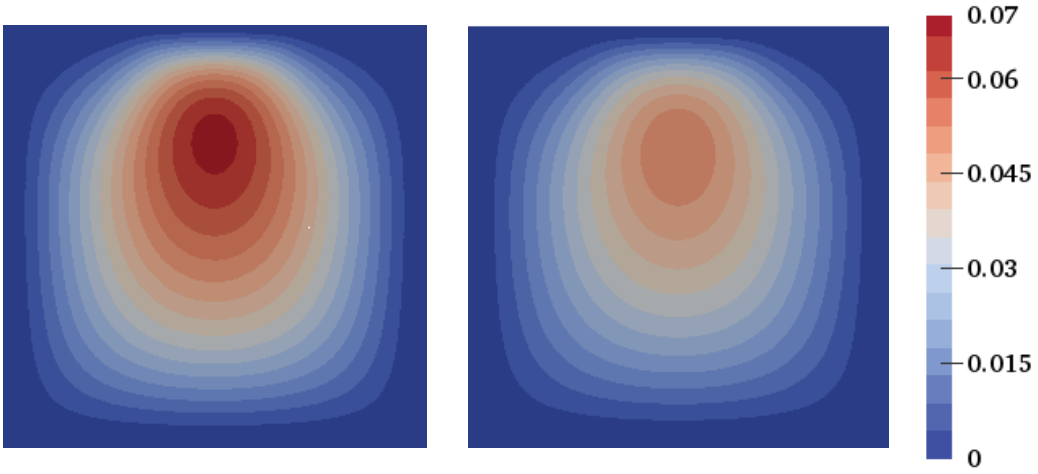


Figure 5.14: Poroelastic footing test: Comparison between the total variance $\mathbb{V}\text{ar}(p(\boldsymbol{x}, \boldsymbol{\xi}))$ and the sum of the first-order conditional variances $\sum_{i=1}^{4} \mathbb{V}\text{ar}(p(\boldsymbol{x}, \boldsymbol{\xi})|\xi_i)$ estimated using the PSP method with $l = 5$.

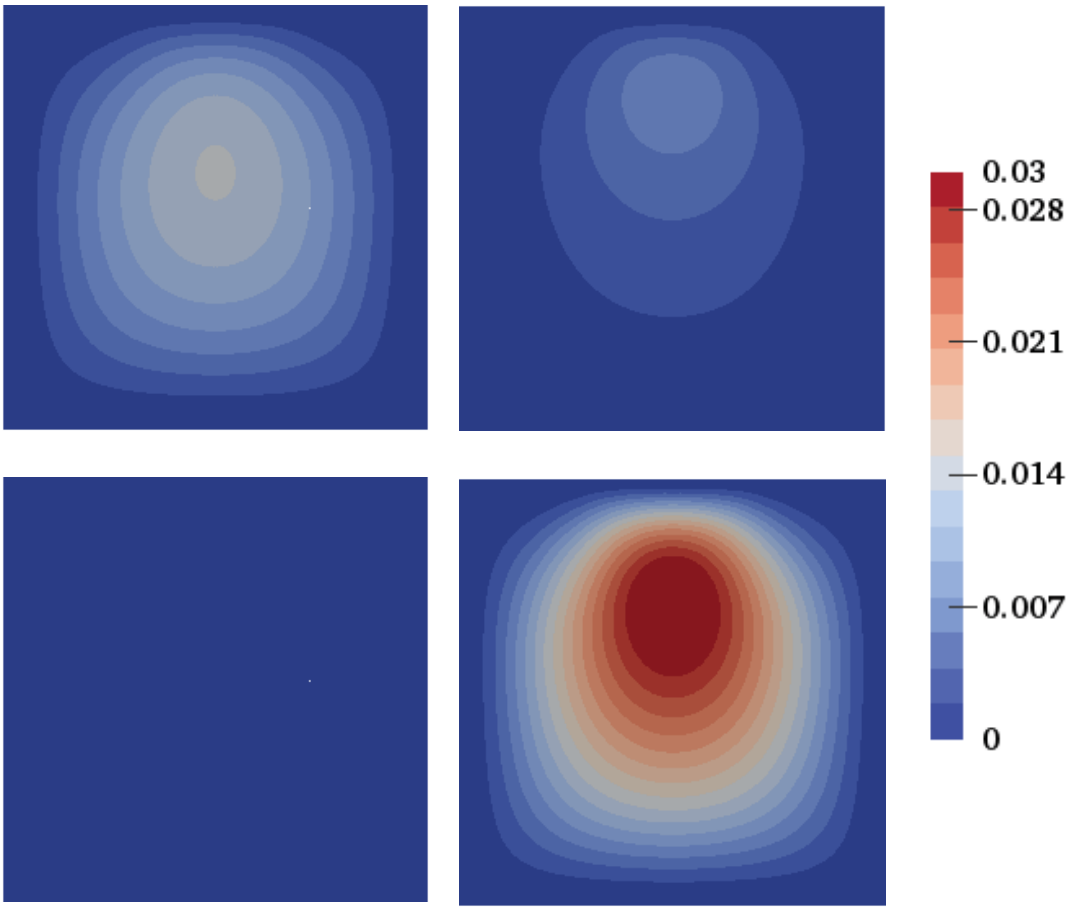


Figure 5.15: Poroelastic footing test: Conditional variances $\mathbb{V}\text{ar}(p(\boldsymbol{x}, \boldsymbol{\xi})|\xi_1)$ (top left) $\mathbb{V}\text{ar}(p(\boldsymbol{x}, \boldsymbol{\xi})|\xi_2)$ (top right) $\mathbb{V}\text{ar}(p(\boldsymbol{x}, \boldsymbol{\xi})|\xi_3)$ (bottom left) $\mathbb{V}\text{ar}(p(\boldsymbol{x}, \boldsymbol{\xi})|\xi_4)$ (bottom right) estimated using the PSP method with $l = 5$.

influence on the solutions is the hydraulic mobility κ (parametrized by ξ_4), while the effect of the Biot–Willis coefficient α (parametrized by ξ_3) on the variance is very weak. In particular, we notice in Figure 5.15 that α has no influence on the uncertainty of the pressure solving the footing problem. Thus, in this case α could be fixed at a deterministic value. One possible reason may be that the coefficient of variation $\mathrm{c_v}(\alpha)$ computed in (5.37) is smaller than the ones relative to the other model coefficients. Finally, we remark that the conditional variances corresponding to the Lamé's coefficients μ, λ are almost the same in the case of the injection problem (cf. Figure 5.13), but are considerably different in the footing test (cf. Figure 5.15).

Appendix A

A nonconforming high-order method for nonlinear poroelasticity

This appendix has been published in the following conference book (see [34])

Finite Volumes for Complex Applications VIII,
Hyperbolic, Elliptic and Parabolic Problems, 2017, Pages 537–545.

Contents

A.1 Introduction

We consider the nonlinear poroelasticity model obtained by generalizing the linear Biot's consolidation model of [27, 190] to nonlinear stress-strain constitutive laws. Our original motivation comes from applications in geosciences, where the support of polyhedral meshes and nonconforming interfaces is crucial.

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, denote a bounded connected polyhedral domain with Lipschitz boundary $\partial\Omega$ and outward normal $\boldsymbol{n}$. For a given finite time $t_F > 0$, volumetric load $\boldsymbol{f}$, fluid source g, the considered nonlinear poroelasticity problem consists in finding a vector-valued displacement field $\boldsymbol{u}$ and a scalar-valued pore pressure field p solution of

$$-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\cdot, \boldsymbol{\nabla}_s\boldsymbol{u}) + \boldsymbol{\nabla}p = \boldsymbol{f} \qquad \text{in } \Omega \times (0, t_F), \tag{A.1a}$$

$$c_0 \mathrm{d}_t p + \mathrm{d}_t \boldsymbol{\nabla}\cdot\boldsymbol{u} - \boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}(\cdot)\boldsymbol{\nabla}p) = g \qquad \text{in } \Omega \times (0, t_F), \tag{A.1b}$$

where ∇_s denotes the symmetric gradient, $c_0 \geq 0$ is the constrained specific storage coefficient, and $\boldsymbol{\kappa} : \Omega \to (0, \overline{\kappa}]$ is the scalar-valued permeability field. Equations (A.1a) and (A.1b) express, respectively, the momentum equilibrium and the fluid mass balance. For the sake of simplicity, we assume that $\boldsymbol{\kappa}$ is piecewise constant on a partition P_Ω of Ω into bounded disjoint polyhedra and we consider the following homogeneous boundary conditions:

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \partial\Omega \times (0, t_F), \tag{A.1c}$$

$$(\boldsymbol{\kappa}(\cdot)\nabla p) \cdot \boldsymbol{n} = 0 \qquad \text{on } \partial\Omega \times (0, t_F). \tag{A.1d}$$

The treatment of more general permeability fields and boundary conditions is possible up to minor modifications. Initial conditions are set prescribing $\boldsymbol{u}(\cdot, 0) = \boldsymbol{u}^0$ and, if $c_0 > 0$, $p(\cdot, 0) = p^0$. In the incompressible case $c_0 = 0$, we also need the following compatibility condition on g and zero-average constraint on p:

$$\int_\Omega g(\cdot, t) = 0 \quad \text{and} \quad \int_\Omega p(\cdot, t) = 0 \qquad \forall t \in (0, t_F).$$

We assume that the symmetric stress tensor $\boldsymbol{\sigma} : \Omega \times \mathbb{R}^{d\times d}_{\text{sym}} \to \mathbb{R}^{d\times d}_{\text{sym}}$ is a Caratheodory function such that there exist real numbers $\overline{\sigma}, \underline{\sigma} \in (0, +\infty)$ and, for a.e. $\boldsymbol{x} \in \Omega$, and all $\boldsymbol{\tau}, \boldsymbol{\eta} \in \mathbb{R}^{d\times d}_{\text{sym}}$, the following conditions hold:

$$\|\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{0})\|_{d\times d} \leq \overline{\sigma}\|\boldsymbol{\tau}\|_{d\times d}, \quad \text{(growth)} \tag{A.2a}$$

$$\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{\tau}) : \boldsymbol{\tau} \geq \underline{\sigma}\|\boldsymbol{\tau}\|^2_{d\times d}, \quad \text{(coercivity)} \tag{A.2b}$$

$$(\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{\tau}) - \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{\eta})) : (\boldsymbol{\tau} - \boldsymbol{\eta}) \geq 0, \quad \text{(monotonicity)} \tag{A.2c}$$

where $\boldsymbol{\tau} : \boldsymbol{\eta} \coloneqq \sum_{i,j=1}^d \boldsymbol{\tau}_{i,j}\boldsymbol{\eta}_{i,j}$ and $\|\boldsymbol{\tau}\|^2_{d\times d} = \boldsymbol{\tau} : \boldsymbol{\tau}$.

A.2 Mesh and notation

Denote by $\mathcal{H} \subset \mathbb{R}^+_*$ a countable set having 0 as unique accumulation point. We consider refined mesh sequences $(\mathcal{T}_h)_{h\in\mathcal{H}}$ where each $\mathcal{T}_h$ is a finite collection of disjoint open polyhedral elements T with boundary ∂T such that $\overline{\Omega} = \bigcup_{T\in\mathcal{T}_h} \overline{T}$ and $h = \max_{T\in\mathcal{T}_h} h_T$ with h_T diameter of T. We assume that mesh regularity holds in the sense of [76, Definition 1.38] and that, for all $h \in \mathcal{H}$, $\mathcal{T}_h$ is compatible with the partition P_Ω, so that jumps of the permeability coefficient do not occur inside mesh elements. Mesh faces are hyperplanar subsets of $\overline{\Omega}$ with positive $(d-1)$-dimensional Hausdorff measure and disjoint interiors. Interfaces are collected in the set $\mathcal{F}_h^{\text{i}}$, boundary faces in $\mathcal{F}_h^{\text{b}}$, and we assume that $\mathcal{F}_h \coloneqq \mathcal{F}_h^{\text{i}} \cup \mathcal{F}_h^{\text{b}}$ is such that $\bigcup_{T\in\mathcal{T}_h} \partial T = \bigcup_{F\in\mathcal{F}_h} F$. For all $T \in \mathcal{T}_h$, $\mathcal{F}_T \coloneqq \{F \in \mathcal{F}_h \ : \ F \subset \partial T\}$ denotes the set of faces contained in ∂T and, for all $F \in \mathcal{F}_T$, $\boldsymbol{n}_{TF}$ is the unit normal to F pointing out of T.

For $X \subset \overline{\Omega}$, we denote by $\|\cdot\|_X$ the norm in $L^2(X;\mathbb{R})$, $L^2(X;\mathbb{R}^d)$, and $L^2(X;\mathbb{R}^{d\times d})$. For $l \geq 0$, the space $\mathbb{P}^l_d(X;\mathbb{R})$ is spanned by the restriction to X of polynomials of total degree l. On regular mesh sequences, we have the following optimal approximation property for the L^2-projector $\pi^l_X : L^1(X;\mathbb{R}) \to \mathbb{P}^l_d(X;\mathbb{R})$: There exists $C_{\text{ap}} > 0$ such that, for all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$, all $s \in \{1, \dots, l+1\}$, and all $v \in H^s(T;\mathbb{R})$,

$$|v - \pi^l_T v|_{H^m(T;\mathbb{R})} \leq C_{\text{ap}} h_T^{s-m} |v|_{H^s(T;\mathbb{R})} \qquad \forall m \in \{0, \dots, s\}. \tag{A.3}$$

In what follows, for an integer $l \geq 0$, we denote by $\mathbb{P}_d^l(\mathcal{T}_h;\mathbb{R})$, $\mathbb{P}_d^l(\mathcal{T}_h;\mathbb{R}^d)$, and $\mathbb{P}_d^l(\mathcal{T}_h;\mathbb{R}^{d\times d})$, respectively, the space of scalar-, vector-, and tensor-valued broken polynomials of total degree l on $\mathcal{T}_h$. The space of broken vector-valued polynomials of total degree l on the mesh skeleton is denoted by $\mathbb{P}_d^l(\mathcal{F}_h;\mathbb{R}^d)$.

A.3 Discretization

In this section we define the discrete counterparts of the elasticity and Darcy operators and of the hydro-mechanical coupling terms.

A.3.1 Nonlinear elasticity operator

The discretization of the nonlinear elasticity operator is based on the Hybrid High-Order method of [74]. Let a polynomial degree $k \geq 1$ be fixed. The degrees of freedom (DOFs) for the displacement are collected in the space $\underline{\boldsymbol{U}}_h^k \coloneqq \mathbb{P}_d^k(\mathcal{T}_h;\mathbb{R}^d) \times \mathbb{P}_d^k(\mathcal{F}_h;\mathbb{R}^d)$. To account for the Dirichlet condition (A.1c) we define the subspace

$$\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \coloneqq \left\{ \underline{\boldsymbol{v}}_h \coloneqq ((\boldsymbol{v}_T)_{T\in\mathcal{T}_h}, (\boldsymbol{v}_F)_{F\in\mathcal{F}_h}) \in \underline{\boldsymbol{U}}_h^k \ : \ \boldsymbol{v}_F = \boldsymbol{0} \quad \forall F \in \mathcal{F}_h^{\mathrm{b}} \right\},$$

equipped with the discrete strain norm $\|\cdot\|_{\epsilon,h}$ defined in (2.13). The DOFs corresponding to a function $\boldsymbol{v} \in H_0^1(\Omega;\mathbb{R}^d)$ are obtained by means of the reduction map $\underline{I}_h^k : H_0^1(\Omega;\mathbb{R}^d) \to \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ such that $\underline{I}_h^k \boldsymbol{v} \coloneqq ((\boldsymbol{\pi}_T^k \boldsymbol{v})_{T\in\mathcal{T}_h}, (\boldsymbol{\pi}_F^k \boldsymbol{v})_{F\in\mathcal{F}_h})$. Using the H^1-stability of the L^2-projector and the trace inequality [76, Lemma 1.49], we infer the existence of $C_{\mathrm{st}} > 0$ independent of h such that, for all $\boldsymbol{v} \in H_0^1(\Omega;\mathbb{R}^d)$,

$$\|\underline{I}_h^k \boldsymbol{v}\|_{\epsilon,h} \leq C_{\mathrm{st}} \|\boldsymbol{v}\|_{H^1(\Omega;\mathbb{R}^d)}. \tag{A.4}$$

For all $T \in \mathcal{T}_h$, we denote by $\underline{\boldsymbol{U}}_T^k$ and $\underline{I}_T^k$ the restrictions to T of $\underline{\boldsymbol{U}}_h^k$ and $\underline{I}_h^k$, and we define the local symmetric gradient reconstruction $\boldsymbol{G}_{\mathrm{s},T}^k : \underline{\boldsymbol{U}}_T^k \to \mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$ as the unique solution of the pure traction problem: For a given $\underline{\boldsymbol{v}}_T = (\boldsymbol{v}_T, (\boldsymbol{v}_F)_{F\in\mathcal{F}_T}) \in \underline{\boldsymbol{U}}_T^k$, find $\boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T \in \mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$ such that, for all $\boldsymbol{\tau} \in \mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$,

$$\int_T \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T : \boldsymbol{\tau} = -\int_T \boldsymbol{v}_T \cdot (\boldsymbol{\nabla}\cdot\boldsymbol{\tau}) + \sum_{F\in\mathcal{F}_T} \int_F \boldsymbol{v}_F \cdot (\boldsymbol{\tau}\boldsymbol{n}_{TF}). \tag{A.5}$$

The definition of $\boldsymbol{G}_{\mathrm{s},T}^k$ is justified by the commuting property (4.19) that, combined with (A.3), shows that $\boldsymbol{G}_{\mathrm{s},T}^k \underline{I}_T^k$ has optimal approximation properties in $\mathbb{P}_d^k(T;\mathbb{R}_{\mathrm{sym}}^{d\times d})$. From $\boldsymbol{G}_{\mathrm{s},T}^k$ we define the local displacement reconstruction operator $\boldsymbol{r}_T^{k+1} : \underline{\boldsymbol{U}}_T^k \to \mathbb{P}_d^{k+1}(T;\mathbb{R}^d)$ as in (2.5).

The discretization of the nonlinear elasticity operator is realized by the function $a_h : \underline{\boldsymbol{U}}_h^k \times \underline{\boldsymbol{U}}_h^k \to \mathbb{R}$ defined such that, for all $\underline{\boldsymbol{w}}_h, \underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_h^k$,

$$a_h(\underline{\boldsymbol{w}}_h, \underline{\boldsymbol{v}}_h) \coloneqq \sum_{T\in\mathcal{T}_h} \left(\int_T \boldsymbol{\sigma}(\cdot, \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{u}}_T) : \boldsymbol{G}_{\mathrm{s},T}^k \underline{\boldsymbol{v}}_T + \sum_{F\in\mathcal{F}_T} \frac{\gamma}{h_F} \int_F \Delta_{TF}^k \underline{\boldsymbol{u}}_T \cdot \Delta_{TF}^k \underline{\boldsymbol{v}}_T \right), \tag{A.6}$$

where we penalize in a least-square sense the face-based residual $\underline{\mathit{\Delta}}_{TF}^k : \underline{\boldsymbol{U}}_T^k \to \mathbb{P}_d^k(F;\mathbb{R}^d)$ such that, for all $T \in \mathcal{T}_h$, all $\underline{\boldsymbol{v}}_T \in \underline{\boldsymbol{U}}_T^k$, and all $F \in \mathcal{F}_T$,

$$\underline{\mathit{\Delta}}_{TF}^k \underline{\boldsymbol{v}}_T \coloneqq \boldsymbol{\pi}_F^k(\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T - \boldsymbol{v}_F) - \boldsymbol{\pi}_T^k(\boldsymbol{r}_T^{k+1}\underline{\boldsymbol{v}}_T - \boldsymbol{v}_T).$$

This definition ensures that $\underline{\mathit{\Delta}}_{TF}^k$ vanishes whenever its argument is of the form $\underline{\boldsymbol{I}}_T^k \boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{P}_d^{k+1}(T;\mathbb{R}^d)$, a crucial property to obtain high-order error estimates (cf. Theorems 2.12, 3.16, and 4.17). A possible choice for the scaling parameter $\gamma > 0$ in (A.6) is $\gamma = \underline{\sigma}$. For all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$, it holds (the proof follows from [74, Lemma 4]):

$$C_{\mathrm{eq}}^{-1}\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2 \le \sum_{T\in\mathcal{T}_h}\left(\|\boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T\|_T^2 + \sum_{F\in\mathcal{F}_T}\frac{\gamma}{h_F}\|\underline{\mathit{\Delta}}_{TF}^k\underline{\boldsymbol{v}}_T\|_F^2\right) \le C_{\mathrm{eq}}\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}^2,$$

where $C_{\mathrm{eq}} > 0$ is independent of h. By (A.2b), this implies the coercivity of a_h.

A.3.2 Darcy operator

The discretization of the Darcy operator is based on the Symmetric Weighted Interior Penalty method of [77], cf. also [76, Sec. 4.5]. At each time step, the discrete pore pressure is sought in the broken polynomial space P_h^k defined in (2.17). For all $F \in \mathcal{F}_h^{\mathrm{i}}$ and all $q_h \in \mathbb{P}_d^k(\mathcal{T}_h;\mathbb{R})$, we define the jump and average operators such that, denoting by q_T and $\boldsymbol{\kappa}_T$ the restrictions of q_h and $\boldsymbol{\kappa}$ to an element $T \in \mathcal{T}_h$,

$$[q_h]_F \coloneqq q_{T_{F,1}} - q_{T_{F,2}}, \qquad \{q_h\}_F \coloneqq \frac{\boldsymbol{\kappa}_{T_{F,2}}}{\boldsymbol{\kappa}_{T_{F,1}}+\boldsymbol{\kappa}_{T_{F,2}}}q_{T_{F,1}} + \frac{\boldsymbol{\kappa}_{T_{F,1}}}{\boldsymbol{\kappa}_{T_{F,1}}+\boldsymbol{\kappa}_{T_{F,2}}}q_{T_{F,2}},$$

where $T_{F,1}, T_{F,2} \in \mathcal{T}_h$ are such that $F \subset T_{F,1} \cap T_{F,2}$. The bilinear form c_h on $P_h^k \times P_h^k$ is defined such that, for all $q_h, r_h \in P_h^k$,

$$\begin{aligned} c_h(r_h,q_h) \coloneqq &\int_\Omega \boldsymbol{\kappa}\boldsymbol{\nabla}_h r_h \cdot \boldsymbol{\nabla}_h q_h + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}}\frac{2\varsigma\boldsymbol{\kappa}_{T_{F,1}}\boldsymbol{\kappa}_{T_{F,2}}}{h_F(\boldsymbol{\kappa}_{T_{F,1}}+\boldsymbol{\kappa}_{T_{F,2}})}\int_F [r_h]_F[q_h]_F \\ &- \sum_{F\in\mathcal{F}_h^{\mathrm{i}}}\int_F \left([r_h]_F\{\boldsymbol{\kappa}\boldsymbol{\nabla}_h q_h\}_F + [q_h]_F\{\boldsymbol{\kappa}\boldsymbol{\nabla}_h r_h\}_F\right)\cdot \boldsymbol{n}_{T_{F,1}F}, \end{aligned} \tag{A.7}$$

where $\boldsymbol{\nabla}_h$ denotes the broken gradient and $\varsigma > 0$ is a user-defined penalty parameter chosen large enough to ensure the coercivity of c_h (cf. [76, Lemma 4.51]). In the numerical tests of Sec. A.5, we took $\varsigma = (N_\partial + 0.1)k^2$, with N_∂ equal to the maximum number of faces between the elements in $\mathcal{T}_h$. The fact that the boundary terms only appear on internal faces in (A.7) reflects the Neumann boundary condition (A.1d).

A.3.3 Hydro-mechanical coupling

The hydro-mechanical coupling is realized by means of the bilinear form b_h on $\underline{\boldsymbol{U}}_{h,\mathrm{D}}^k \times \mathbb{P}_d^k(\mathcal{T}_h;\mathbb{R})$ such that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}_{h,\mathrm{D}}^k$ and all $q_h \in \mathbb{P}_d^k(\mathcal{T}_h;\mathbb{R})$,

$$b_h(\underline{\boldsymbol{v}}_h, q_h) \coloneqq -\sum_{T\in\mathcal{T}_h}\int_T \boldsymbol{G}_{\mathrm{s},T}^k\underline{\boldsymbol{v}}_T : q_{h|T}\boldsymbol{I}_d,$$

where $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}_{\rm sym}$ is the identity matrix. A simple verification shows that there exists $C_{\rm bd} > 0$ independent of h such that $b_h(\underline{\boldsymbol{v}}_h, q_h) \le C_{\rm bd}\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}\|q_h\|_\Omega$. Additionally, using definition (A.5) of $\boldsymbol{G}^k_{{\rm s},T}$, it can be proved that, for all $\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}^k_{h,{\rm D}}$, $b_h(\underline{\boldsymbol{v}}_h, 1) = 0$. The following inf-sup condition expresses the stability of the coupling:

Proposition A.1. *There is a real β independent of h such that, for all $q_h \in \mathbb{P}^k_{d0}(\mathcal{T}_h;\mathbb{R})$,*

$$\|q_h\|_\Omega \le \beta \sup_{\underline{\boldsymbol{v}}_h\in\underline{\boldsymbol{U}}^k_{h,{\rm D}}\setminus\{\underline{\boldsymbol{0}}\}} \frac{b_h(\underline{\boldsymbol{v}}_h, q_h)}{\|\underline{\boldsymbol{v}}_h\|_{\epsilon,h}}. \tag{A.8}$$

Proof. Let $q_h \in \mathbb{P}^k_{d0}(\mathcal{T}_h;\mathbb{R})$. There is $\boldsymbol{v}_{q_h} \in H^1_0(\Omega;\mathbb{R}^d)$ such that $\boldsymbol{\nabla}\cdot\boldsymbol{v}_{q_h} = q_h$ and $\|\boldsymbol{v}_{q_h}\|_{H^1(\Omega;\mathbb{R}^d)} \le C_{\rm sj}\|q_h\|_\Omega$, with $C_{\rm sj} > 0$ independent of h. Owing to (A.4) we get

$$\|\underline{I}^k_h\boldsymbol{v}_{q_h}\|_{\epsilon,h} \le C_{\rm st}\|\boldsymbol{v}_{q_h}\|_{H^1(\Omega;\mathbb{R}^d)} \le C_{\rm st}C_{\rm sj}\|q_h\|_\Omega.$$

Therefore, using the commuting property (4.19), denoting by S the supremum in (A.8), and using the previous inequality, it is inferred that

$$\|q_h\|^2_\Omega = \sum_{T\in\mathcal{T}_h}\int_T(\boldsymbol{\nabla}_{\rm s}\boldsymbol{v}_{q_h} : q_h\boldsymbol{I}_d)_{|T} = -b_h(\underline{I}^k_h\boldsymbol{v}_{q_h}, q_h) \le \mathsf{S}\|\underline{I}^k_h\boldsymbol{v}_{q_h}\|_{\epsilon,h} \le C_{\rm st}C_{\rm sj}\mathsf{S}\|q_h\|_\Omega. \qquad \square$$

If a HHO discretization were used also for the Darcy operator, only cell DOFs would be controlled by the inf-sup condition.

A.4 Formulation of the method

For the time discretization, we consider a uniform mesh of the time interval $(0, t_{\rm F})$ of step $\tau := t_{\rm F}/N$ with $N \in \mathbb{N}^*$, and introduce the discrete times $t^n := n\tau$ for all $0 \le n \le N$. For any $\varphi \in C^l([0,t_{\rm F}];V)$, we set $\varphi^n := \varphi(t^n) \in V$ and let, for all $1 \le n \le N$, $\delta_t\varphi^n := {}^{(\varphi^n-\varphi^{n-1})}/_\tau \in V$. For all $1 \le n \le N$, the discrete solution $(\underline{\boldsymbol{u}}^n_h, p^n_h) \in \underline{\boldsymbol{U}}^k_{h,{\rm D}} \times P^k_h$ at time t^n is such that, for all $(\underline{\boldsymbol{v}}_h, q_h) \in \underline{\boldsymbol{U}}^k_{h,{\rm D}} \times \mathbb{P}^k_d(\mathcal{T}_h;\mathbb{R})$,

$$\begin{aligned} a_h(\underline{\boldsymbol{u}}^n_h, \underline{\boldsymbol{v}}_h) + b_h(\underline{\boldsymbol{v}}_h, p^n_h) &= \sum_{T\in\mathcal{T}_h}\int_T \boldsymbol{f}^n\cdot\boldsymbol{v}_T,\\ c_0\int_\Omega(\delta_t p^n_h)q_h - b_h(\delta_t\underline{\boldsymbol{u}}^n_h, q_h) + c_h(p^n_h, q_h) &= \int_\Omega g^n q_h. \end{aligned} \tag{A.9}$$

If $c_0 = 0$, we set the initial discrete displacement as $\underline{\boldsymbol{u}}^0_h = \underline{I}^k_h\boldsymbol{u}^0$. If $c_0 > 0$, the usual way to enforce the initial condition is to compute a displacement from the given initial pressure p^0. We let $p^0_h := \pi^k_h p^0$ and set $\underline{\boldsymbol{u}}^0_h \in \underline{\boldsymbol{U}}^k_{h,{\rm D}}$ as the solution of

$$a_h(\underline{\boldsymbol{u}}^0_h, \underline{\boldsymbol{v}}_h) = \sum_{T\in\mathcal{T}_h}\int_T \boldsymbol{f}^0\cdot\boldsymbol{v}_T - b_h(\underline{\boldsymbol{v}}_h, p^0_h) \qquad \forall\underline{\boldsymbol{v}}_h \in \underline{\boldsymbol{U}}^k_{h,{\rm D}}.$$

At each time step n the discrete nonlinear equations (A.9) are solved by the Newton's method using as initial guess the solution at step $n-1$. At each Newton's iteration the Jacobian matrix is computed analytically and in the linearized system the displacement element unknowns can be statically condensed (cf. Section 2.5).

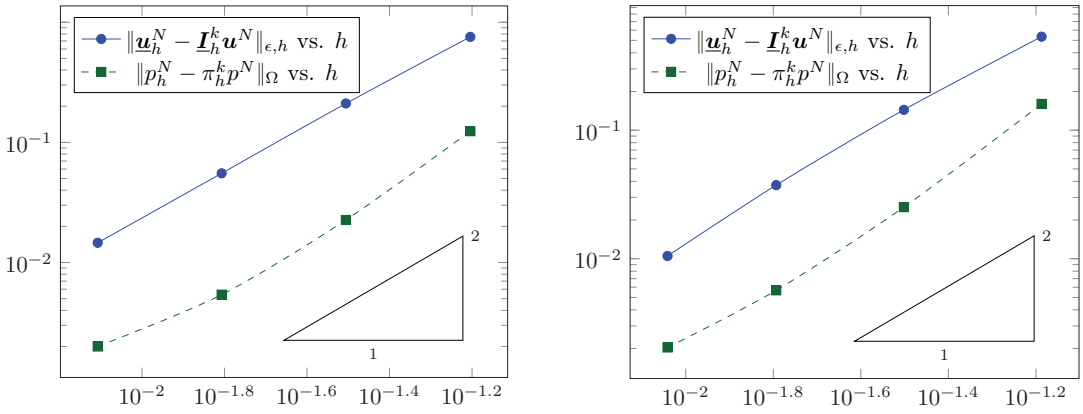


Figure A.1: Convergence tests on a Cartesian mesh family (left) and on a Voronoi mesh family (right).

A.5 Numerical results

We consider a regular exact solution in order to assess the convergence of the method for polynomial degree $k = 1$. Specifically, we solve problem (A.1) in the square domain $\Omega = (0,1)^2$ with $t_F = 1$ and physical parameters $c_0 = 0$ and $\boldsymbol{\kappa} = 1$. As nonlinear constitutive law we take the Hencky–Mises relation given by

$$\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}) = (2\exp^{-\operatorname{dev}(\boldsymbol{\nabla}_s \boldsymbol{u})} - 1)\operatorname{tr}(\boldsymbol{\nabla}_s \boldsymbol{u})\boldsymbol{I}_d + (4 - 2\exp^{-\operatorname{dev}(\boldsymbol{\nabla}_s \boldsymbol{u})})\boldsymbol{\nabla}_s \boldsymbol{u},$$

where $\operatorname{tr}(\boldsymbol{\tau}) \coloneqq \boldsymbol{\tau} : \boldsymbol{I}_d$ and $\operatorname{dev}(\boldsymbol{\tau}) = \operatorname{tr}(\boldsymbol{\tau}^2) - \frac{1}{d}\operatorname{tr}(\boldsymbol{\tau})^2$ are the trace and deviatoric operators. It can be checked that the previous stress-strain relation satisfies (A.2). The exact displacement $\boldsymbol{u}$ and exact pressure p are given by

$$\begin{aligned}\boldsymbol{u}(\boldsymbol{x},t) &= t^2(\sin(\pi x_1)\sin(\pi x_2),\ \sin(\pi x_1)\sin(\pi x_2)),\\ p(\boldsymbol{x},t) &= -\pi^{-1}t(\sin(\pi x_1)\cos(\pi x_2) + \cos(\pi x_1)\sin(\pi x_2)).\end{aligned}$$

The volumetric load $\boldsymbol{f}$, the source term g, and the boundary conditions are inferred from the exact solutions. The time step τ on the coarsest mesh is 0.2 and it decreases with the mesh size h according to $\tau_1/\tau_2 = 2h_1/h_2$. In Fig. A.1 we display the convergence results obtained on two mesh families. The method exhibits second order convergence with respect to the mesh size h for both the energy norm of the displacement and the L^2-norm of the pressure at final time t_F.

Appendix B

A posteriori error estimation for nonlinear elasticity

This appendix has been published in the following peer-reviewed journal (see [36])

Computational Methods in Applied Mathematics,
Published online: 20 June 2018, DOI: 10.1515/cmam-2018-0012.

Contents

B.1 Introduction

In this appendix we develop equilibrated $H(\mathrm{div})$-conforming stress tensor reconstructions for a class of (linear and) nonlinear elasticity problems in the small deformation regime. Based on these reconstructions, we can derive an a posteriori error estimate distinguishing the discretization and linearization errors, as proposed in [95] for nonlinear diffusion problems. Thanks to this distinction we can, at each iteration, compare these two error contributions

and stop the linearization algorithm once its contribution is negligible compared to the discretization error.

Let $\Omega \in \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded, simply connected polyhedron, which is occupied by a body subjected to a volumetric force field $\boldsymbol{f} \in [L^2(\Omega)]^d$. For the sake of simplicity, we assume that the body is fixed on its boundary $\partial\Omega$. The nonlinear elasticity problem consists in finding a vector-valued displacement field $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ solving

$$-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{\nabla}_s\boldsymbol{u}) = \boldsymbol{f} \qquad \text{in } \Omega, \tag{B.1a}$$

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \partial\Omega, \tag{B.1b}$$

where $\boldsymbol{\nabla}_s\boldsymbol{u} = \frac{1}{2}((\boldsymbol{\nabla}\boldsymbol{u})^T + \boldsymbol{\nabla}\boldsymbol{u})$ denotes the symmetric gradient and expresses the strain tensor associated to $\boldsymbol{u}$. The stress-strain law $\boldsymbol{\sigma} : \Omega \times \mathbb{R}^{d\times d}_{\text{sym}} \to \mathbb{R}^{d\times d}_{\text{sym}}$ is assumed to satisfy regularity requirements inspired by [33, 69]. Problem (B.1) describes the mechanical behavior of soft materials [191] and metal alloys [169]. In these applications, the solution is often approximated using H^1-conforming finite elements. For nonlinear mechanical behavior laws, the resulting discrete nonlinear equation can then be solved using an iterative linearization algorithm yielding at each iteration a linear algebraic system to be solved, until the residual of the nonlinear equation lies under a predefined threshold.

The a posteriori error estimate is based on equilibrated stress reconstructions. It is well known that, in contrast to the analytical solution, the discrete stress tensor resulting from the conforming finite element method does not have continuous normal components across mesh interfaces, and that its divergence is not locally in equilibrium with the source term $\boldsymbol{f}$ on mesh elements. We consider here the stress reconstruction proposed in [175] for linear elasticity to restore these two properties. This reconstruction uses the Arnold–Falk–Winther mixed finite element spaces [7], leading to weakly symmetric tensors. In Section B.3 we apply this reconstruction to the nonlinear case by constructing two stress tensors: one playing the role of the discrete stress and one expressing the linearization error. They are obtained by summing up the solutions of minimization problems on cell patches around each mesh vertex, so that they are $H(\mathrm{div})$-conforming and the sum of the two reconstructions verifies locally the mechanical equilibrium (B.1a). The patch-wise equilibration technique was introduced in [38, 67] for the Poisson problem using the Raviart–Thomas finite element spaces. In [83] it is extended to linear elasticity without any symmetry constraint by using linewise Raviart–Thomas reconstructions. Elementwise reconstructions from local Neumann problems requiring some pre-computations to determine the fluxes to obtain an equilibrated stress tensor can be found in [4, 51, 134, 161], whereas in [158] the direct prescription of the degrees of freedom in the Arnold–Winther finite element space is considered.

Based on the equilibrated stress reconstructions, we develop the a posteriori error estimate in Section B.4 and prove that it is efficient, meaning that, up to a generic constant, it is also a local lower bound for the error. The idea goes back to [170] and was advanced amongst others by [2, 132, 133, 172] for the upper bound.Local lower error bounds are derived in [38, 67, 94, 97, 144]. Using equilibrated fluxes for a posteriori error estimation offers several advantages. The first one is, as mentioned above, the possible distinction of error components by expressing them in terms of fluxes. Secondly, the error upper bound is obtained with computable constants. In our case these constants depend only on the parameters of the

stress-strain relation. Thirdly, since the estimate is based on the discrete stress (and not the displacement), it does not depend on the mechanical behavior law (except for the constant in the upper bound). In addition, equilibrated error estimates were proven to be polynomial-degree robust for several linear problems in 2D, as the Poisson problem in [37, 97], linear elasticity in [83] and the related Stokes problem in [49] and recently in 3D in [96].

The material is organized as follows. In Section B.2 we introduce the weak and the discrete formulations of problem (B.1), along with some useful notation. In Section B.3 we present the equilibrated stress tensor reconstructions, first assuming that we solve the nonlinear discrete problem exactly, and then distinguish its discrete and its linearization error part at each iteration of a linearization solver. In Section B.4 we derive the a posteriori error estimate distinguishing the different error components. We then propose an algorithm equilibrating the error sources using adaptive stopping criteria for the linearization and adaptive remeshing. We finally show the efficiency of the error estimate. In Section B.5 we evaluate the performance of the estimates. We perform a linear test with analytical solution in order to compare the error estimator and the analytical error and to show that effectivity indices are close to one. We then apply the proposed adaptive algorithm to application-oriented tests with nonlinear stress-strain relations.

B.2 Setting

In this section we write the weak and the considered discrete formulation of problem (B.1). In the following, we will adopt the notation that bold greek letters are tensor-valued functions and bold latin letters are vector-valued functions. For $X \subset \overline{\Omega}$, we respectively denote by $(\cdot,\cdot)_X$ and $\|\cdot\|_X$ the standard inner product and norm in $L^2(X)$, with the convention that the subscript is omitted whenever $X = \Omega$. The same notation is used in the vector- and tensor-valued cases. $[H^1(\Omega)]^d$ and $\mathbb{H}(\mathrm{div},\Omega)$ stand for the Sobolev spaces composed of vector-valued $[L^2(\Omega)]^d$ functions with weak gradient in $[L^2(\Omega)]^{d\times d}$, and tensor-valued $[L^2(\Omega)]^{d\times d}$ functions with weak divergence in $[L^2(\Omega)]^d$, respectively. In what follows, we consider stress-strain functions $\boldsymbol{\sigma}$ satisfying the following assumption.

Assumption B.1 (Stress-strain relation)**.** The symmetric stress tensor $\boldsymbol{\sigma} : \mathbb{R}^{d\times d}_{\mathrm{sym}} \to \mathbb{R}^{d\times d}_{\mathrm{sym}}$ is continuous on $\mathbb{R}^{d\times d}_{\mathrm{sym}}$, satisfies the growth and strong monotonicity assumptions (3.2c) and (3.40b), and $\boldsymbol{\sigma}(\mathbf{0}) = \mathbf{0}$.

Multiplying (B.1a) by a test function $\boldsymbol{v} \in [H^1_0(\Omega)]^d$ and integrating by parts yields

$$(\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}), \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}). \tag{B.2}$$

Owing to (3.2c), for all $\boldsymbol{v}, \boldsymbol{w} \in [H^1(\Omega)]^d$, the form $a(\boldsymbol{v},\boldsymbol{w}) \coloneqq (\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}), \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w})$ is well defined and, from equation (B.2), we can derive the following weak formulation of (B.1):

$$\text{Given } \boldsymbol{f} \in [L^2(\Omega)]^d, \text{ find } \boldsymbol{u} \in [H^1_0(\Omega)]^d \text{ s.t., } \forall \boldsymbol{v} \in [H^1_0(\Omega)]^d,\ a(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{f},\boldsymbol{v}). \tag{B.3}$$

From (B.3) it is clear that the analytical stress tensor $\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u})$ lies in the space $\mathbb{H}_{\mathrm{s}}(\mathrm{div},\Omega) \coloneqq \{\boldsymbol{\tau} \in [L^2(\Omega)]^{d\times d} \mid \boldsymbol{\nabla}\cdot\boldsymbol{\tau} \in [L^2(\Omega)]^d \text{ and } \boldsymbol{\tau} \text{ is symmetric}\}$.

The discretization of (B.3) is based on a conforming triangulation $\mathcal{T}_h$ of Ω, i.e. a set of closed triangles or tetrahedra with union $\overline{\Omega}$ and such that, for any distinct $T_1, T_2 \in \mathcal{T}_h$, the set

$T_1 \cap T_2$ is either a common edge, a vertex, the empty set or, if $d = 3$, a common face. The set of vertices of the mesh is denoted by $\mathcal{V}_h$; it is decomposed into interior vertices $\mathcal{V}_h^{\text{int}}$ and boundary vertices $\mathcal{V}_h^{\text{ext}}$. For all $a \in \mathcal{V}_h$, $\mathcal{T}_a$ is the patch of elements sharing the vertex a, ω_a the corresponding open subdomain in Ω and $\mathcal{V}_a$ the set of vertices in ω_a. For all $T \in \mathcal{T}_h$, $\mathcal{V}_T$ denotes the set of vertices of T, h_T its diameter and $\boldsymbol{n}_T$ its unit outward normal vector.

For all $p \in \mathbb{N}$ and all $T \in \mathcal{T}_h$, we denote by $\mathbb{P}^p(T)$ the space of d-variate polynomials in T of total degree at most p and by $\mathbb{P}^p(\mathcal{T}_h) = \{\varphi \in L^2(\Omega) \mid \varphi_{|T} \in \mathbb{P}^p(T)\ \forall T \in \mathcal{T}_h\}$ the corresponding broken space over $\mathcal{T}_h$. In what follows we will focus on conforming discretizations of problem (B.2) of polynomial degree $p \geq 2$ to avoid numerical locking, cf [196]. The discrete formulation reads: find $\boldsymbol{u}_h \in [H_0^1(\Omega)]^d \cap [\mathbb{P}^p(\mathcal{T}_h)]^d$ such that

$$\forall \boldsymbol{v}_h \in [H_0^1(\Omega)]^d \cap [\mathbb{P}^p(\mathcal{T}_h)]^d, \quad a(\boldsymbol{u}_h, \boldsymbol{v}_h) = (\boldsymbol{f}, \boldsymbol{v}_h). \tag{B.4}$$

This problem is usually solved using some iterative linearization algorithm defining at each iteration $k \geq 1$ a linear approximation $\boldsymbol{\sigma}_{\text{lin}}^{k-1}$ of $\boldsymbol{\sigma}$. Then the linearized formulation reads: find $\boldsymbol{u}_h^k \in [H_0^1(\Omega)]^d \cap [\mathbb{P}^p(\mathcal{T}_h)]^d$ such that

$$\forall \boldsymbol{v}_h \in [H_0^1(\Omega)]^d \cap [\mathbb{P}^p(\mathcal{T}_h)]^d, \quad (\boldsymbol{\sigma}_{\text{lin}}^{k-1}(\boldsymbol{\nabla}_{\text{s}} \boldsymbol{u}_h^k), \boldsymbol{\nabla}_{\text{s}} \boldsymbol{v}_h) = (\boldsymbol{f}, \boldsymbol{v}_h). \tag{B.5}$$

For the Newton algorithm the linearized stress $\boldsymbol{\sigma}_{\text{lin}}^{k-1}$ is defined, for all $\boldsymbol{\eta} \in \mathbb{R}^{d\times d}_{\text{sym}}$, such that

$$\boldsymbol{\sigma}_{\text{lin}}^{k-1}(\boldsymbol{\eta}) \coloneqq \frac{\partial \boldsymbol{\sigma}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}}|_{\boldsymbol{\tau}=\boldsymbol{\nabla}_{\text{s}} \boldsymbol{u}_h^{k-1}} (\boldsymbol{\eta} - \boldsymbol{\nabla}_{\text{s}} \boldsymbol{u}_h^{k-1}) + \boldsymbol{\sigma}(\boldsymbol{\nabla}_{\text{s}} \boldsymbol{u}_h^{k-1}).$$

B.3 Equilibrated stress reconstruction

In general, the discrete stress tensor $\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\text{s}} \boldsymbol{u}_h)$ resulting from (B.4) does not lie in $\mathbb{H}(\text{div}, \Omega)$ and thus cannot verify the equilibrium equation (B.1a). In this section we will reconstruct from $\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\text{s}} \boldsymbol{u}_h)$ a discrete stress tensor $\boldsymbol{\sigma}_h$ satisfying these properties. Based on this reconstruction, we then devise two equilibrated stress tensors representing the discrete stress and the linearization error respectively, which will be useful for the distinction of error components in the a posteriori error estimate of Section B.4.

B.3.1 Patchwise construction in the Arnold–Falk–Winther spaces

Let us for now suppose that $\boldsymbol{u}_h$ solves (B.4) exactly, before considering iterative linearization methods such as (B.5) in Section B.3.2. For the stress reconstruction we will use mixed finite element formulations on patches around mesh vertices in the spirit of [174, 175]. The goal is to obtain a stress tensor $\boldsymbol{\sigma}_h$ in a suitable (i.e. $H(\text{div})$-conforming) finite element space by summing up these local solutions. The local problems are posed such that this global stress tensor is close to the discrete stress tensor $\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\text{s}} \boldsymbol{u}_h)$ obtained from (B.4), and that it satisfies the mechanical equilibrium on each element. In [174] the stress tensor is reconstructed in the Arnold–Winther finite element space [8], directly providing symmetric tensors, but requiring high computational effort. As done in [175], we weaken the symmetry constraint and impose

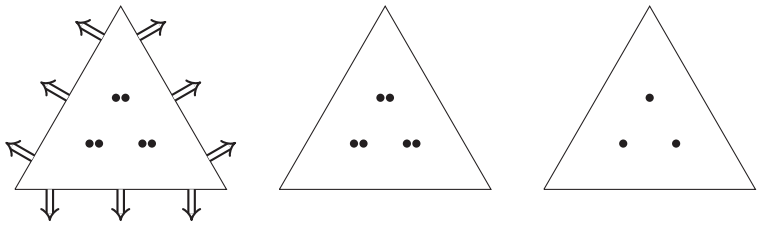


Figure B.1: Element diagrams for $(\boldsymbol{\Sigma}_T, \boldsymbol{V}_T, \boldsymbol{\Lambda}_T)$ in the case $d = q = 2$

it weakly, as proposed in [7]: for each element $T \in \mathcal{T}_h$, the local Arnold–Falk–Winther mixed finite element spaces of degree $q \geq 1$ hinge on the Brezzi–Douglas–Marini mixed finite element spaces [41] for each line of the stress tensor and are defined by

$$\boldsymbol{\Sigma}_T \coloneqq [\mathbb{P}^q(T)]^{d\times d}, \quad \boldsymbol{V}_T \coloneqq [\mathbb{P}^{q-1}(T)]^d, \quad \boldsymbol{\Lambda}_T \coloneqq \{\boldsymbol{\mu} \in [\mathbb{P}^{q-1}(T)]^{d\times d} \mid \boldsymbol{\mu} = -\boldsymbol{\mu}^T\}. \tag{B.6}$$

For $q = 2$, the degrees of freedom are displayed in Figure B.1. On a patch ω_a the global space $\boldsymbol{\Sigma}_h(\omega_a)$ is the subspace of $\mathbb{H}(\mathrm{div}, \omega_a)$ composed of functions belonging piecewise to $\boldsymbol{\Sigma}_T$. The spaces $\boldsymbol{V}_h(\omega_a)$ and $\boldsymbol{\Lambda}_h(\omega_a)$ consist of functions lying piecewise in $\boldsymbol{V}_T$ and $\boldsymbol{\Lambda}_T$ respectively, with no continuity conditions between two elements.

Let now $q \coloneqq p$. On each patch we consider subspaces where a zero normal component is enforced on the stress tensor on the boundary of the patch, so that the sum of the local solutions will have continuous normal component across any mesh interface. Since the boundary condition (B.1b) prescribes the displacement and not the normal stress, we distinguish the case whether a is an interior vertex or a boundary vertex. If $a \in \mathcal{V}_h^{\mathrm{int}}$ we set

$$\begin{aligned}
\boldsymbol{\Sigma}_h^a &\coloneqq \{\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_h(\omega_a) \mid \boldsymbol{\tau}_h \boldsymbol{n}_{\omega_a} = \boldsymbol{0} \text{ on } \partial\omega_a\},\\
\boldsymbol{V}_h^a &\coloneqq \{\boldsymbol{v}_h \in \boldsymbol{V}_h(\omega_a) \mid (\boldsymbol{v}_h, \boldsymbol{z})_{\omega_a} = 0 \ \forall \boldsymbol{z} \in \boldsymbol{RM}^d\},\\
\boldsymbol{\Lambda}_h^a &\coloneqq \boldsymbol{\Lambda}_h(\omega_a),
\end{aligned}$$

where $\boldsymbol{RM}^2 \coloneqq \{\boldsymbol{b} + c(x_2, -x_1)^T \mid \boldsymbol{b} \in \mathbb{R}^2, c \in \mathbb{R}\}$ and $\boldsymbol{RM}^3 \coloneqq \{\boldsymbol{b} + \boldsymbol{a} \times \boldsymbol{x} \mid \boldsymbol{b} \in \mathbb{R}^3, \boldsymbol{a} \in \mathbb{R}^3\}$ are the spaces of rigid-body motions respectively for $d = 2$ and $d = 3$. If $a \in \mathcal{V}_h^{\mathrm{ext}}$ we set

$$\begin{aligned}
\boldsymbol{\Sigma}_h^a &\coloneqq \{\boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_h(\omega_a) \mid \boldsymbol{\tau}_h \boldsymbol{n}_{\omega_a} = \boldsymbol{0} \text{ on } \partial\omega_a \backslash \partial\Omega\},\\
\boldsymbol{V}_h^a &\coloneqq \boldsymbol{V}_h(\omega_a),\\
\boldsymbol{\Lambda}_h^a &\coloneqq \boldsymbol{\Lambda}_h(\omega_a).
\end{aligned}$$

For each vertex $a \in \mathcal{V}_h$ we define its hat function $\psi_a \in \mathbb{P}^1(\mathcal{T}_h)$ as the piecewise linear function taking value one at the vertex a and zero on all other mesh vertices.

Construction B.2 (Stress tensor reconstruction)**.** Let $\boldsymbol{u}_h$ solve (B.4). For each $a \in \mathcal{V}_h$ find $(\boldsymbol{\sigma}_h^a, \boldsymbol{r}_h^a, \boldsymbol{\lambda}_h^a) \in \boldsymbol{\Sigma}_h^a \times \boldsymbol{V}_h^a \times \boldsymbol{\Lambda}_h^a$ such that for all $(\boldsymbol{\tau}_h, \boldsymbol{v}_h, \boldsymbol{\mu}_h) \in \boldsymbol{\Sigma}_h^a \times \boldsymbol{V}_h^a \times \boldsymbol{\Lambda}_h^a$,

$$(\boldsymbol{\sigma}_h^a, \boldsymbol{\tau}_h)_{\omega_a} + (\boldsymbol{r}_h^a, \boldsymbol{\nabla}\cdot\boldsymbol{\tau}_h)_{\omega_a} + (\boldsymbol{\lambda}_h^a, \boldsymbol{\tau}_h)_{\omega_a} = (\psi_a \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h), \boldsymbol{\tau}_h)_{\omega_a}, \tag{B.7a}$$

$$(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h^a, \boldsymbol{v}_h)_{\omega_a} = (-\psi_a \boldsymbol{f} + \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h)\boldsymbol{\nabla}\psi_a, \boldsymbol{v}_h)_{\omega_a}, \tag{B.7b}$$

$$(\boldsymbol{\sigma}_h^a, \boldsymbol{\mu}_h)_{\omega_a} = 0. \tag{B.7c}$$

Then, extending $\boldsymbol{\sigma}_h^a$ by zero outside ω_a, set $\boldsymbol{\sigma}_h \coloneqq \sum_{a\in\mathcal{V}_h} \boldsymbol{\sigma}_h^a$.

For interior vertices, the source term in (B.7b) has to verify the Neumann compatibility condition

$$(-\psi_a \boldsymbol{f} + \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h)\boldsymbol{\nabla}\psi_a, \boldsymbol{z})_{\omega_a} = 0 \quad \forall \boldsymbol{z} \in \boldsymbol{RM}^d. \tag{B.8}$$

Taking $\psi_a \boldsymbol{z}$ as a test function in (B.4), we see that (B.8) holds and we obtain the following result.

Lemma B.3 (Properties of $\boldsymbol{\sigma}_h$)**.** *Let $\boldsymbol{\sigma}_h$ be prescribed by Construction B.2. Then $\boldsymbol{\sigma}_h \in \mathbb{H}(\mathrm{div}, \Omega)$, and for all $T \in \mathcal{T}_h$, the following holds:*

$$(\boldsymbol{f} + \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h, \boldsymbol{v})_T = 0 \quad \forall \boldsymbol{v} \in \boldsymbol{V}_T \ \ \forall T \in \mathcal{T}_h. \tag{B.9}$$

Proof. All the fields $\boldsymbol{\sigma}_h^a$ are in $\mathbb{H}(\mathrm{div}, \omega_a)$ and satisfy appropriate zero normal conditions so that their zero-extension to Ω is in $\mathbb{H}(\mathrm{div}, \Omega)$. Hence, $\boldsymbol{\sigma}_h \in \mathbb{H}(\mathrm{div}, \Omega)$. Let us prove (B.9). Since (B.8) holds for all $a \in \mathcal{V}_h^{\mathrm{int}}$, we infer that (B.7b) is actually true for all $\boldsymbol{v}_h \in \boldsymbol{V}_h(\omega_a)$. The same holds if $a \in \mathcal{V}_h^{\mathrm{ext}}$ by definition of $\boldsymbol{V}_h^a$. Since $\boldsymbol{V}_h(\omega_a)$ is composed of piecewise polynomials that can be chosen independently in each cell $T \in \mathcal{T}_a$, and using $\boldsymbol{\sigma}_h|_T = \sum_{a\in\mathcal{V}_T} \boldsymbol{\sigma}_h^a|_T$ and the partition of unity $\sum_{a\in\mathcal{V}_T} \psi_a = 1$, we infer that $(\boldsymbol{f} + \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h, \boldsymbol{v})_T = 0$ for all $\boldsymbol{v} \in \boldsymbol{V}_T$ and all $T \in \mathcal{T}_h$. □

Remark B.4 (Choice of q). In contrast to flux reconstructions in the Raviart–Thomas space (cf. [38] or [83]), the choice $q = p-1$ in (B.6) does not lead to optimal convergence rates of the resulting error estimate in Section B.4. This is due to the fact that the Brezzi–Douglas–Marini space is smaller than the Raviart–Thomas space of the same degree.

B.3.2 Discretization and linearization error stress reconstructions

Let now, for $k \geq 1$, $\boldsymbol{u}_h^k$ solve (B.5). We will construct two different equilibrated $H(\mathrm{div})$-conforming stress tensors. The first one, $\boldsymbol{\sigma}_{h,\mathrm{disc}}^k$, represents as above the discrete stress tensor $\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$, for which we will have to modify Construction B.2, because the Neumann compatibility condition (B.8) is not satisfied anymore. The second stress tensor $\boldsymbol{\sigma}_{h,\mathrm{lin}}^k$ will be a measure for the linearization error and approximate $\boldsymbol{\sigma}_{\mathrm{lin}}^{k-1}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k) - \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$. The matrix resulting from the left side of (B.7) will stay unchanged and we will only modify the source terms.

We denote by $\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$ the L^2-orthogonal projection of $\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$ onto $[\mathbb{P}^{p-1}(\mathcal{T}_h)]^{d\times d}$ such that $(\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k) - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k), \boldsymbol{\tau}_h) = 0$ for any $\boldsymbol{\tau}_h \in [\mathbb{P}^{p-1}(\mathcal{T}_h)]^{d\times d}$.

Construction B.5 (Discrete stress reconstruction)**.** For each $a \in \mathcal{V}_h$ solve (B.7) with $\boldsymbol{u}_h^k$ instead of $\boldsymbol{u}_h$, $\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$ instead of $\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$ and the source term in (B.7b) replaced by

$$-\psi_a \boldsymbol{f} + \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)\boldsymbol{\nabla}\psi_a - \boldsymbol{y}_{\mathrm{disc}}^k,$$

where $\boldsymbol{y}_{\mathrm{disc}}^k \in \boldsymbol{RM}^d$ is the unique solution of

$$(\boldsymbol{y}_{\mathrm{disc}}^k, \boldsymbol{z})_{\omega_a} = -(\boldsymbol{f}, \psi_a \boldsymbol{z})_{\omega_a} + (\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k), \boldsymbol{\nabla}_s(\psi_a \boldsymbol{z}))_{\omega_a} \quad \forall \boldsymbol{z} \in \boldsymbol{RM}^d. \tag{B.10}$$

The so obtained problem reads: find $(\boldsymbol{\sigma}_h^a, \boldsymbol{r}_h^a, \boldsymbol{\lambda}_h^a) \in \boldsymbol{\Sigma}_h^a \times \boldsymbol{V}_h^a \times \boldsymbol{\Lambda}_h^a$ such that for all $(\boldsymbol{\tau}_h, \boldsymbol{v}_h, \boldsymbol{\mu}_h) \in \boldsymbol{\Sigma}_h^a \times \boldsymbol{V}_h^a \times \boldsymbol{\Lambda}_h^a$,

$$\begin{aligned}(\boldsymbol{\sigma}_h^a, \boldsymbol{\tau}_h)_{\omega_a} + (\boldsymbol{r}_h^a, \boldsymbol{\nabla}\cdot\boldsymbol{\tau}_h)_{\omega_a} + (\boldsymbol{\lambda}_h^a, \boldsymbol{\tau}_h)_{\omega_a} &= (\psi_a \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k), \boldsymbol{\tau}_h)_{\omega_a},\\ (\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h^a, \boldsymbol{v}_h)_{\omega_a} &= (-\psi_a \boldsymbol{f} + \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)\boldsymbol{\nabla}\psi_a - \boldsymbol{y}_{\mathrm{disc}}^k, \boldsymbol{v}_h)_{\omega_a},\\ (\boldsymbol{\sigma}_h^a, \boldsymbol{\mu}_h)_{\omega_a} &= 0.\end{aligned}$$

Then set $\boldsymbol{\sigma}_{h,\mathrm{disc}}^k \coloneqq \sum_{a\in\mathcal{V}_h} \boldsymbol{\sigma}_h^a$.

Construction B.6 (Linearization error stress reconstruction)**.** For each $a \in \mathcal{V}_h$ solve (B.7) with $\boldsymbol{u}_h^k$ instead of $\boldsymbol{u}_h$, the source term in (B.7a) replaced by

$$\psi_a(\boldsymbol{\sigma}_{\mathrm{lin}}^{k-1}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k) - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)),$$

and the source term in (B.7b) replaced by

$$(\boldsymbol{\sigma}_{\mathrm{lin}}^{k-1}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k) - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k))\boldsymbol{\nabla}\psi_a + \boldsymbol{y}_{\mathrm{disc}}^k,$$

where $\boldsymbol{y}_{\mathrm{disc}}^k \in \boldsymbol{RM}^d$ is defined by (B.10). The corresponding local problem is to find $(\boldsymbol{\sigma}_h^a, \boldsymbol{r}_h^a, \boldsymbol{\lambda}_h^a) \in \boldsymbol{\Sigma}_h^a \times \boldsymbol{V}_h^a \times \boldsymbol{\Lambda}_h^a$ such that for all $(\boldsymbol{\tau}_h, \boldsymbol{v}_h, \boldsymbol{\mu}_h) \in \boldsymbol{\Sigma}_h^a \times \boldsymbol{V}_h^a \times \boldsymbol{\Lambda}_h^a$,

$$\begin{aligned}(\boldsymbol{\sigma}_h^a, \boldsymbol{\tau}_h)_{\omega_a} + (\boldsymbol{r}_h^a, \boldsymbol{\nabla}\cdot\boldsymbol{\tau}_h)_{\omega_a} + (\boldsymbol{\lambda}_h^a, \boldsymbol{\tau}_h)_{\omega_a} &= (\psi_a(\boldsymbol{\sigma}_{\mathrm{lin}}^{k-1}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k) - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)), \boldsymbol{\tau}_h)_{\omega_a},\\ (\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h^a, \boldsymbol{v}_h)_{\omega_a} &= ((\boldsymbol{\sigma}_{\mathrm{lin}}^{k-1}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k) - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k))\boldsymbol{\nabla}\psi_a + \boldsymbol{y}_{\mathrm{disc}}^k, \boldsymbol{v}_h)_{\omega_a},\\ (\boldsymbol{\sigma}_h^a, \boldsymbol{\mu}_h)_{\omega_a} &= 0.\end{aligned}$$

Then set $\boldsymbol{\sigma}_{h,\mathrm{lin}}^k \coloneqq \sum_{a\in\mathcal{V}_h} \boldsymbol{\sigma}_h^a$.

Notice that the role of $\boldsymbol{y}_{\mathrm{disc}}^k$ is to guarantee that, for interior vertices, the source terms in Constructions B.5 and B.6 satisfy the Neumann compatibility conditions

$$\begin{aligned}(-\psi_a \boldsymbol{f} + \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)\boldsymbol{\nabla}\psi_a - \boldsymbol{y}_{\mathrm{disc}}^k, \boldsymbol{z})_{\omega_a} &= 0 \quad \forall \boldsymbol{z} \in \boldsymbol{RM}^d,\\ ((\boldsymbol{\sigma}_{\mathrm{lin}}^{k-1}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k) - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k))\boldsymbol{\nabla}\psi_a + \boldsymbol{y}_{\mathrm{disc}}^k, \boldsymbol{z})_{\omega_a} &= 0 \quad \forall \boldsymbol{z} \in \boldsymbol{RM}^d.\end{aligned}$$

Lemma B.7 (Properties of the discretization and linearization error stress reconstructions)**.** *Let $\boldsymbol{\sigma}_{h,\mathrm{disc}}^k$ and $\boldsymbol{\sigma}_{h,\mathrm{lin}}^k$ be prescribed by Constructions B.5 and B.6. Then it holds*

1. $\boldsymbol{\sigma}_{h,\mathrm{disc}}^k, \boldsymbol{\sigma}_{h,\mathrm{lin}}^k \in \mathbb{H}(\mathrm{div}, \Omega)$,
2. $(\boldsymbol{f} + \boldsymbol{\nabla}\cdot(\boldsymbol{\sigma}_{h,\mathrm{disc}}^k + \boldsymbol{\sigma}_{h,\mathrm{lin}}^k), \boldsymbol{v})_T = 0 \quad \forall \boldsymbol{v} \in \boldsymbol{V}_T \ \ \forall T \in \mathcal{T}_h$,
3. *As the Newton solver converges,* $\boldsymbol{\sigma}_{h,\mathrm{lin}}^k \to \boldsymbol{0}$.

Proof. The proof is similar to the proof of Lemma B.3. The first property is again satisfied due to the definition of $\boldsymbol{\Sigma}_h^a$. In order to show that the second property holds, we add the two equations (B.7b) obtained for each of the constructions. The right hand side of this sum then reads $(-\psi_a \boldsymbol{f} + \boldsymbol{\sigma}_{\mathrm{lin}}^{k-1}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)\boldsymbol{\nabla}\psi_a, \boldsymbol{v}_h)_{\omega_a}$. Once again we can, for any $\boldsymbol{z} \in \boldsymbol{RM}^d$, take $\psi_a \boldsymbol{z}$ as a test function in (B.5) to show that this term is zero if $\boldsymbol{v}_h \in \boldsymbol{RM}^d$, and so the equation holds for all $\boldsymbol{v}_h \in \boldsymbol{V}_h(\omega_a)$. Then we proceed as in the proof of Lemma B.3. □

B.4 A posteriori error estimate and adaptive algorithm

In this section we first derive an upper bound on the error between the analytical solution of (B.3) and the solution $\boldsymbol{u}_h$ of (B.4), in which we then identify and distinguish the discretization and linearization error components at each Newton iteration for the solution $\boldsymbol{u}_h^k$ of (B.5). Based on this distinction, we present an adaptive algorithm stopping the Newton iterations once the linearization error estimate is dominated by the estimate of the discretization error. Finally, we show the efficiency of the error estimate.

B.4.1 Guaranteed upper bound

We measure the error in the energy norm $\|\boldsymbol{v}\|_{\mathrm{en}}^2 := a(\boldsymbol{v},\boldsymbol{v}) = (\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}),\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v})$, for which we obtain the properties

$$\sigma_*^2 C_K^{-2}\|\boldsymbol{\nabla}\boldsymbol{v}\|^2 \le \|\boldsymbol{v}\|_{\mathrm{en}}^2 \le \overline{\sigma}\|\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}\|^2, \tag{B.13}$$

by applying (3.40b) and the Korn inequality for the left inequality, and the Cauchy–Schwarz inequality and (3.2c) for the right one. In our case it holds $C_K = \sqrt{2}$, owing to (B.1b).

Theorem B.8 (Basic a posteriori error estimate)**.** *Let $\boldsymbol{u}$ be the solution of* (B.3) *and $\boldsymbol{u}_h$ the solution of* (B.4)*. Let $\boldsymbol{\sigma}_h$ be the stress tensor defined in Construction B.2. Then,*

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\mathrm{en}} \le \sqrt{2}\overline{\sigma}\sigma_*^{-3}\left(\sum_{T\in\mathcal{T}_h}\left(\frac{h_T}{\pi}\|\boldsymbol{f}+\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h\|_T + \|\boldsymbol{\sigma}_h-\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h)\|_T\right)^2\right)^{1/2}. \tag{B.14}$$

Remark B.9 (Constants $\overline{\sigma}$ and σ_*). For the estimate to be computable, the constants $\overline{\sigma}$ and σ_* have to be specified. For the linear elasticity model (3.3) we set $\overline{\sigma} := 2\mu + d\lambda$ and $\sigma_* := \sqrt{2\mu}$, whereas for the Hencky–Mises model (3.4) we set $\overline{\sigma} := 2\tilde{\mu}(0) + d\tilde{\lambda}(0)$ and $\sigma_* := \sqrt{2\tilde{\mu}(0)}$. Following [175], we obtain a sharper bound in the case of linear elasticity, with $\mu^{-1/2}$ instead of $\sqrt{2}\overline{\sigma}\sigma_*^{-3}$ in (B.14).

Proof of Theorem B.8. We start by bounding the energy norm of the error by the dual norm of the residual of the weak formulation (B.3). Using (B.13), (3.40b), the linearity of a in its second argument, and (B.3) we obtain

$$\begin{aligned}
\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\mathrm{en}}^2 &\le \overline{\sigma}\|\boldsymbol{\nabla}_{\mathrm{s}}(\boldsymbol{u}-\boldsymbol{u}_h)\|^2 \le \overline{\sigma}\sigma_*^{-2}|a(\boldsymbol{u},\boldsymbol{u}-\boldsymbol{u}_h)-a(\boldsymbol{u}_h,\boldsymbol{u}-\boldsymbol{u}_h)| \\
&= \overline{\sigma}\sigma_*^{-2}\|\boldsymbol{\nabla}(\boldsymbol{u}-\boldsymbol{u}_h)\|\left|a\left(\boldsymbol{u},\frac{\boldsymbol{u}-\boldsymbol{u}_h}{\|\boldsymbol{\nabla}(\boldsymbol{u}-\boldsymbol{u}_h)\|}\right)-a\left(\boldsymbol{u}_h,\frac{\boldsymbol{u}-\boldsymbol{u}_h}{\|\boldsymbol{\nabla}(\boldsymbol{u}-\boldsymbol{u}_h)\|}\right)\right| \\
&\le \overline{\sigma}\sigma_*^{-3}C_K\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\mathrm{en}} \sup_{\boldsymbol{v}\in[H_0^1(\Omega)]^d,\|\boldsymbol{\nabla}\boldsymbol{v}\|=1}\{a(\boldsymbol{u},\boldsymbol{v})-a(\boldsymbol{u}_h,\boldsymbol{v})\} \\
&= \overline{\sigma}\sigma_*^{-3}C_K\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\mathrm{en}} \sup_{\boldsymbol{v}\in[H_0^1(\Omega)]^d,\|\boldsymbol{\nabla}\boldsymbol{v}\|=1}\{(\boldsymbol{f},\boldsymbol{v})-(\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h),\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v})\}.
\end{aligned}$$

and thus

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\mathrm{en}} \le \overline{\sigma}\sigma_*^{-3}C_K \sup_{\boldsymbol{v}\in[H_0^1(\Omega)]^d,\|\boldsymbol{\nabla}\boldsymbol{v}\|=1}\{(\boldsymbol{f},\boldsymbol{v})-(\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h),\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v})\}. \tag{B.15}$$

Note that, due to the symmetry of $\boldsymbol{\sigma}$ we can replace $\boldsymbol{\nabla}_s \boldsymbol{v}$ by $\boldsymbol{\nabla}\boldsymbol{v}$ in the second term inside the supremum. Now fix $\boldsymbol{v} \in [H_0^1(\Omega)]^d$, such that $\|\boldsymbol{\nabla}\boldsymbol{v}\| = 1$. Since $\boldsymbol{\sigma}_h \in \mathbb{H}(\mathrm{div}, \Omega)$, we can insert $(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h, \boldsymbol{v}) + (\boldsymbol{\sigma}_h, \boldsymbol{\nabla}\boldsymbol{v}) = 0$ into the term inside the supremum and obtain

$$(\boldsymbol{f}, \boldsymbol{v}) - (\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h), \boldsymbol{\nabla}\boldsymbol{v}) = (\boldsymbol{f} + \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h, \boldsymbol{v}) + (\boldsymbol{\sigma}_h - \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h), \boldsymbol{\nabla}\boldsymbol{v}). \tag{B.16}$$

For the first term of the right hand side of (B.16) we obtain, using (B.9) on each $T \in \mathcal{T}_h$ to insert $\boldsymbol{\Pi}_T^0 \boldsymbol{v}$, which denotes the $\boldsymbol{L}^2$-projection of $\boldsymbol{v}$ onto $[\mathbb{P}^0(T)]^d$, the Cauchy–Schwarz inequality and the Poincaré inequality on simplices,

$$|(\boldsymbol{f} + \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h, \boldsymbol{v})| \leq \Big| \sum_{T\in\mathcal{T}_h} (\boldsymbol{f} + \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h, \boldsymbol{v} - \boldsymbol{\Pi}_T^0 \boldsymbol{v})_T \Big| \leq \sum_{T\in\mathcal{T}_h} \frac{h_T}{\pi} \|\boldsymbol{f} + \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h\|_T \|\boldsymbol{\nabla}\boldsymbol{v}\|_T, \tag{B.17}$$

whereas the Cauchy–Schwarz inequality applied to the second term directly yields

$$|(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h), \boldsymbol{\nabla}\boldsymbol{v})| \leq \sum_{T\in\mathcal{T}_h} \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h)\|_T \|\boldsymbol{\nabla}\boldsymbol{v}\|_T.$$

Inserting in (B.15) and applying again the Cauchy–Schwarz inequality yields the result. □

The goal of the following result is to elaborate the error estimate (B.14) so as to distinguish different error components using the equilibrated stress tensors of Constructions B.5 and B.6. This distinction is essential for the development of Algorithm B.12, where the mesh and the stopping criteria for the iterative solver are chosen adaptively. Notice that in Theorem B.8 we don't necessarily need $\boldsymbol{\sigma}_h$ to be the stress tensor obtained in Construction B.2. We only need it to satisfy two properties: First, equation (B.16) requires $\boldsymbol{\sigma}_h$ to lie in $\mathbb{H}(\mathrm{div}, \Omega)$. Second, in order to be able to apply the Poincaré inequality in (B.17), $\boldsymbol{\sigma}_h$ has to satisfy the local equilibrium relation

$$(\boldsymbol{f} - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h, \boldsymbol{v})_T = 0 \quad \forall \boldsymbol{v} \in [\mathbb{P}^0(T)]^d \ \forall T \in \mathcal{T}_h.$$

Thus, the theorem also holds for $\boldsymbol{\sigma}_h := \boldsymbol{\sigma}^k_{h,\mathrm{disc}} + \boldsymbol{\sigma}^k_{h,\mathrm{lin}}$, where $\boldsymbol{\sigma}^k_{h,\mathrm{disc}}$ and $\boldsymbol{\sigma}^k_{h,\mathrm{lin}}$ are defined in Constructions B.5.

Theorem B.10 (A posteriori error estimate distinguishing different error sources)**.** *Let $\boldsymbol{u}$ be the solution of* (B.3)*, $\boldsymbol{u}_h^k$ the solution of* (B.5)*, and $\boldsymbol{\sigma}_h := \boldsymbol{\sigma}^k_{h,\mathrm{disc}} + \boldsymbol{\sigma}^k_{h,\mathrm{lin}}$. Then,*

$$\|\boldsymbol{u} - \boldsymbol{u}_h^k\|_{\mathrm{en}} \leq \sqrt{2\overline{\sigma}}\,\sigma_*^{-3} (\eta^k_{\mathrm{disc}} + \eta^k_{\mathrm{lin}} + \eta^k_{\mathrm{quad}} + \eta^k_{\mathrm{osc}}),$$

where the local discretization, linearization, quadrature and oscillation error estimators on each $T \in \mathcal{T}_h$ are defined as

$$\begin{aligned}
\eta^k_{\mathrm{disc},T} &:= \|\boldsymbol{\sigma}^k_{h,\mathrm{disc}} - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)\|_T, \\
\eta^k_{\mathrm{lin},T} &:= \|\boldsymbol{\sigma}^k_{h,\mathrm{lin}}\|_T, \\
\eta^k_{\mathrm{quad},T} &:= \|\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k) - \boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)\|_T, \\
\eta^k_{\mathrm{osc},T} &:= h_T \pi^{-1} \|\boldsymbol{f} - \boldsymbol{\Pi}_T^{p-1} \boldsymbol{f}\|_T,
\end{aligned}$$

with $\boldsymbol{\Pi}_T^{p-1}$ denoting the $\boldsymbol{L}^2$-projection onto $[\mathbb{P}^{p-1}(T)]^d$, and for each error source the global estimator is given by $\eta^k_{\cdot} := 2\Big(\sum_{T\in\mathcal{T}_h} (\eta^k_{\cdot,T})^2 \Big)^{1/2}$.

Remark B.11 (Quadrature error). In practice, the projection $\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$ of $\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$ onto $[\mathbb{P}^{p-1}(\mathcal{T}_h)]^{d\times d}$ for a general nonlinear stress-strain relation cannot be computed exactly. The quadrature error estimator $\eta_{\mathrm{quad},T}^k$ measures the quality of this projection and is computed by using a considerably higher number of Gauss points.

B.4.2 Adaptive algorithm

Based on the error estimate of Theorem B.10, we propose an adaptive algorithm where the mesh size is locally adapted, and a dynamic stopping criterion is used for the linearization iterations. The idea is to compare the estimators for the different error sources with each other in order to concentrate the computational effort on reducing the dominant one. For this purpose, let $\gamma_{\mathrm{lin}}, \gamma_{\mathrm{quad}} \in (0,1)$ be user-given weights representing the relative size of the quadrature and linearization errors with respect to the total error.

Algorithm B.12 (Mesh adaptation algorithm)**.**

1. Choose an initial function $\boldsymbol{u}_h^0 \in [H_0^1(\Omega)]^d \cap [\mathbb{P}^p(\mathcal{T}_h)]^d$ and set $k := 1$
2. Set the initial quadrature precision $\nu := 2p$ (exactness for polynomials up to degree ν)
3. Linearization iterations
   (a) Calculate $\boldsymbol{\sigma}_{\mathrm{lin}}^{k-1}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$, $\boldsymbol{u}_h^k$, and $\boldsymbol{\sigma}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$
   (b) Calculate $\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_s \boldsymbol{u}_h^k)$, the stress reconstructions $\boldsymbol{\sigma}_{h,\mathrm{disc}}^k$ and $\boldsymbol{\sigma}_{h,\mathrm{lin}}^k$, and the error estimators η_{disc}^k, η_{lin}^k, η_{osc}^k and η_{quad}^k
   (c) Improve the quadrature rule (setting $\nu := \nu + 1$) and go back to step 3(b) until
   $$\eta_{\mathrm{quad}}^k \le \gamma_{\mathrm{quad}}(\eta_{\mathrm{disc}}^k + \eta_{\mathrm{lin}}^k + \eta_{\mathrm{osc}}^k) \tag{B.18a}$$
   (d) Set $k := k + 1$ and go back to step 3(a) until
   $$\eta_{\mathrm{lin}}^k \le \gamma_{\mathrm{lin}}(\eta_{\mathrm{disc}}^k + \eta_{\mathrm{osc}}^k) \tag{B.18b}$$
4. Refine or coarsen the mesh $\mathcal{T}_h$ such that the local discretization error estimators $\eta_{\mathrm{disc},T}^k$ are distributed evenly

Instead of using the global stopping criteria (B.18a) and (B.18b), which are evaluated over all mesh elements, we can also define the local criteria

$$\begin{aligned} \eta_{\mathrm{quad},T}^k &\le \gamma_{\mathrm{quad}}(\eta_{\mathrm{disc},T}^k + \eta_{\mathrm{lin},T}^k + \eta_{\mathrm{osc},T}^k) \quad \forall T \in \mathcal{T}_h, \\ \eta_{\mathrm{lin},T}^k &\le \gamma_{\mathrm{lin}}(\eta_{\mathrm{disc},T}^k + \eta_{\mathrm{osc},T}^k) \quad \forall T \in \mathcal{T}_h, \end{aligned} \tag{B.19}$$

where it is also possible to define local weights $\gamma_{\mathrm{lin},T}$ and $\gamma_{\mathrm{quad},T}$ for each element. These local stopping criteria are necessary to establish the local efficiency of the error estimators in the following section, whereas the global criteria are only sufficient to prove global efficiency.

B.4.3 Local efficiency

Let us start by introducing some additional notation used in this section. For a given element $T \in \mathcal{T}_h$, the set $\mathcal{T}_T$ collects the elements sharing at least a vertex with T. The set $\mathcal{F}_h$ contains all faces (if $d = 2$ we will, for simplicity, refer to the edges as faces) of the mesh and is decomposed into boundary faces $\mathcal{F}_h^{\mathrm{ext}}$ and interfaces $\mathcal{F}_h^{\mathrm{int}}$. For any $T \in \mathcal{T}_h$ the set $\mathcal{F}_T$ contains the faces of T, whereas $\mathcal{F}_{\mathcal{T}_T}$ collects all faces sharing at least a vertex with T and we denote $\mathcal{F}_{\mathcal{T}_T}^{\mathrm{int}} = \mathcal{F}_{\mathcal{T}_T} \cap \mathcal{F}_h^{\mathrm{int}}$. We also assume that $\mathcal{T}_h$ verifies the minimum angle condition, i.e., there exists $\alpha_{\min} > 0$ uniform with respect to all considered meshes such that the minimum angle α_T of each $T \in \mathcal{T}_h$ satisfies $\alpha_T \geq \alpha_{\min}$. In what follows we let $a \lesssim b$ stand for $a \leq Cb$ with a generic constant C, which is independent of the mesh size, the domain Ω and the stress-strain relation, but that can depend on $\alpha_{\min}$ and on the polynomial degree p.

To prove efficiency, we will use a posteriori error estimators of residual type. Following [194, 195] we define for $X \subseteq \Omega$ the functional $\mathcal{R}_X : [H^1(X)]^d \to [H^{-1}(X)]^d$ such that, for all $\boldsymbol{v} \in [H^1(X)]^d$, $\boldsymbol{w} \in [H_0^1(X)]^d$,

$$\langle \mathcal{R}_X(\boldsymbol{v}), \boldsymbol{w} \rangle_X \coloneqq (\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{v}), \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w})_X - (\boldsymbol{f}, \boldsymbol{w})_X.$$

Define, for each $T \in \mathcal{T}_h$,

$$(\eta_{\sharp,T}^k)^2 \coloneqq \sum_{T' \in \mathcal{T}_T} h_{T'}^2 \|\boldsymbol{\nabla}\cdot\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k) + \boldsymbol{\Pi}_T^p \boldsymbol{f}\|_{T'}^2 + \sum_{F \in \mathcal{F}_{\mathcal{T}_T}^{\mathrm{int}}} h_F \|[\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k)\boldsymbol{n}_F]_F\|_F^2,$$

$$(\eta_{\flat,T}^k)^2 \coloneqq \sum_{T' \in \mathcal{T}_T} h_{T'}^2 \|\boldsymbol{\nabla}\cdot(\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k) - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k))\|_{T'}^2 + \sum_{F \in \mathcal{F}_{\mathcal{T}_T}^{\mathrm{int}}} h_F \|[(\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k) - \overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k))\boldsymbol{n}_F]_F\|_F^2,$$

with $[\cdot]_F$ denoting the jump operators defined in (2.18). The quantity $\eta_{\flat,T}^k$ measures the quality of the approximation of $\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k)$ by $\overline{\boldsymbol{\sigma}}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k)$ and can be estimated explicitly. The following result is shown in [194, Section 3.3]. Denoting by $\eta_{\mathrm{osc},\mathcal{T}_T}^k \coloneqq \{2\sum_{T' \in \mathcal{T}_T} (\eta_{\mathrm{osc},T'}^k)^2\}^{1/2}$, it holds

$$\eta_{\sharp,T}^k \lesssim \|\mathcal{R}_{\mathcal{T}_T}(\boldsymbol{u}_h^k)\|_{[H^{-1}(\mathcal{T}_T)]^d} + \eta_{\flat,T}^k + \eta_{\mathrm{osc},\mathcal{T}_T}^k. \tag{B.20}$$

In order to bound the dual norm of the residual, we need to additionally assume that the stress-strain relation $\boldsymbol{\sigma}$ is Lipschitz continuous, cf. Assumption 3.14. We recall that the three stress-strain relations presented in Examples 3.2, 3.3, and 3.4 satisfy the previous Lipschitz continuity assumptions. Owing to the definition of the functional $\mathcal{R}_{\mathcal{T}_T}$ and to the fact that $-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}) = \boldsymbol{f} \in L^2(\mathcal{T}_T)^d$, using the Cauchy–Schwarz inequality and the Lipschitz continuity (3.40a) of $\boldsymbol{\sigma}$, it is inferred that

$$\begin{aligned}
\|\mathcal{R}_{\mathcal{T}_T}(\boldsymbol{u}_h^k)\|_{[H^{-1}(\mathcal{T}_T)]^d} &\coloneqq \sup_{\boldsymbol{w} \in [H_0^1(\mathcal{T}_T)]^d,\, \|\boldsymbol{w}\|_{[H_0^1(\mathcal{T}_T)]^d} \leq 1} (\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k), \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w})_{\mathcal{T}_T} - (\boldsymbol{f}, \boldsymbol{w})_{\mathcal{T}_T} \\
&= \sup_{\boldsymbol{w} \in [H_0^1(\mathcal{T}_T)]^d,\, \|\boldsymbol{w}\|_{[H_0^1(\mathcal{T}_T)]^d} \leq 1} (\boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}_h^k) - \boldsymbol{\sigma}(\boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{u}), \boldsymbol{\nabla}_{\mathrm{s}}\boldsymbol{w})_{\mathcal{T}_T} \\
&\leq \sigma^* \|\boldsymbol{\nabla}_{\mathrm{s}}(\boldsymbol{u} - \boldsymbol{u}_h^k)\|_{\mathcal{T}_T}.
\end{aligned}$$

Thus, by (B.20), the previous bound, and the strong monotonicity (3.40b) it holds

$$\eta^k_{\sharp,T} \lesssim \sigma^* \sigma_*^{-1} \|\boldsymbol{u} - \boldsymbol{u}^k_h\|_{\mathrm{en},\mathcal{T}_T} + \eta^k_{\flat,T} + \eta^k_{\mathrm{osc},\mathcal{T}_T}. \tag{B.21}$$

It is well known that there exist nonconforming finite element methods which are equivalent to mixed finite element methods using the Brezzi–Douglas–Marini spaces (see e.g. [5]). Following the ideas of [91, 95, 118] and references therein, we use these spaces to prove that $\eta^k_{\mathrm{disc},T} \lesssim \eta^k_{\sharp,T}$. As a result of this bound together with (B.21), we obtain the following Theorem (see [36, Section 4.4] for the detailed proof).

Theorem B.13 (Local efficiency)**.** *Let* $\boldsymbol{u} \in [H^1_0(\Omega)]^d$ *be the solution of* (B.3), $\boldsymbol{u}^k_h \in [H^1_0(\Omega)]^d \cap [\mathbb{P}^p(\mathcal{T}_h)]^d$ *be arbitrary and* $\boldsymbol{\sigma}^k_{h,\mathrm{disc}}$ *and* $\boldsymbol{\sigma}^k_{h,\mathrm{lin}}$ *defined by Constructions B.5 and B.6. Let the local stopping criteria* (B.19) *be verified. Then it holds for all* $T \in \mathcal{T}_h$*,*

$$\eta^k_{\mathrm{disc},T} + \eta^k_{\mathrm{lin},T} + \eta^k_{\mathrm{quad},T} + \eta^k_{\mathrm{osc},T} \lesssim \sigma^* \sigma_*^{-1} \|\boldsymbol{u} - \boldsymbol{u}^k_h\|_{\mathrm{en},\mathcal{T}_T} + \eta^k_{\flat,T} + \eta^k_{\mathrm{osc},\mathcal{T}_T}.$$

B.5 Numerical results

In this section we illustrate numerically our results on two test cases, both performed with the `Code_Aster`[1] software, which uses conforming finite elements of degree $p = 2$. Our intention is, first, to show the relevance of the discretization error estimators used as mesh refinement indicators, and second, to propose a stopping criterion for the Newton iterations based on the linearization error estimator. All the triangulations are conforming, since in the remeshing progress hanging nodes are removed by bisecting the neighboring element.

B.5.1 L-shaped domain

Following [4, 129, 158], we consider the L-shaped domain $\Omega = (-1, 1)^2 \setminus ([0, 1] \times [-1, 0])$, where for the linear elasticity case an analytical solution is given by

$$\boldsymbol{u}(r, \theta) = \frac{1}{2\mu} r^\alpha \begin{pmatrix} \cos(\alpha\theta) - \cos((\alpha - 2)\theta) \\ A \sin(\alpha\theta) + \sin((\alpha - 2)\theta) \end{pmatrix},$$

with the parameters

$$\mu = 1.0, \quad \lambda = 5.0, \quad \alpha = 0.6, \quad A = 31/9.$$

This solution is imposed as Dirichlet boundary condition on $\partial\Omega$, together with $\boldsymbol{f} = \boldsymbol{0}$ in Ω. We perform this test for two different stress-strain relations. First on the linear elasticity model (3.3), where we can compare the error estimate (B.14) to the analytical error $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathrm{en}}$. The second relation is the nonlinear Hencky–Mises model (3.4), for which we distinguish the discretization and linearization error components and use the adaptive algorithm from Section B.4.2.

We compute the analytical error and its estimate on two series of unstructured meshes. Starting with the same initial mesh, we use uniform mesh refinement for the first one and

[1] `http://web-code-aster.org`

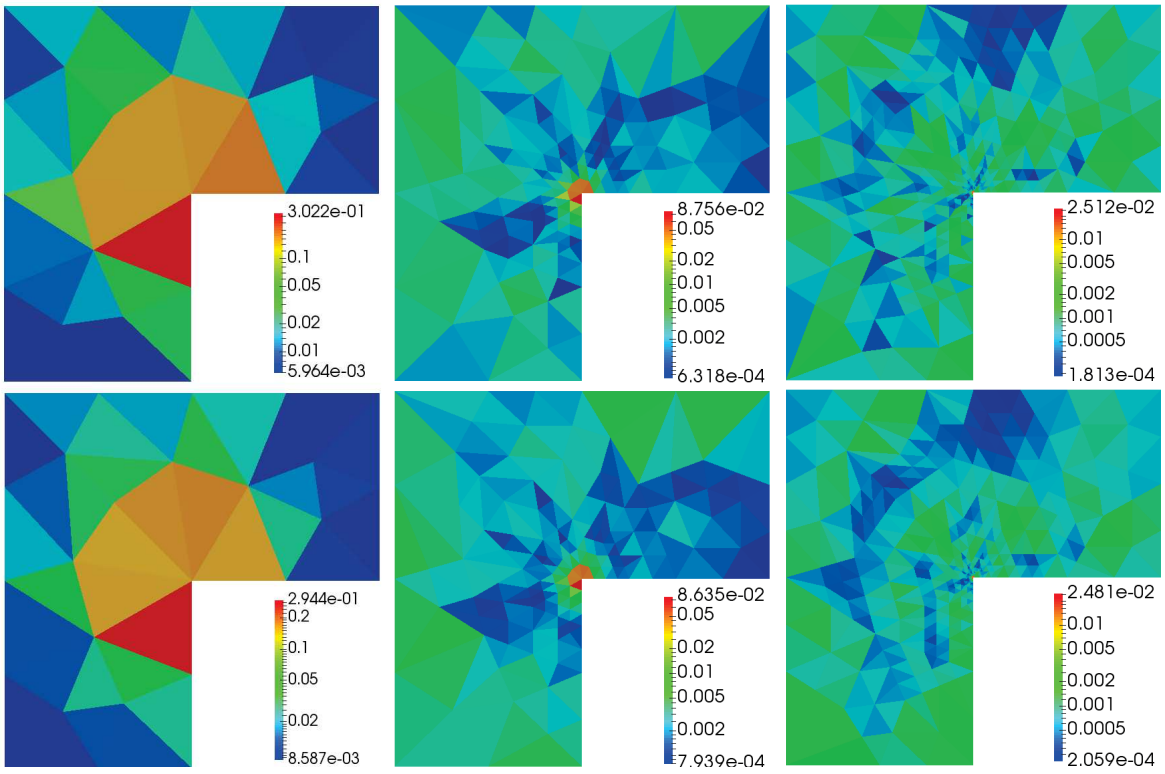
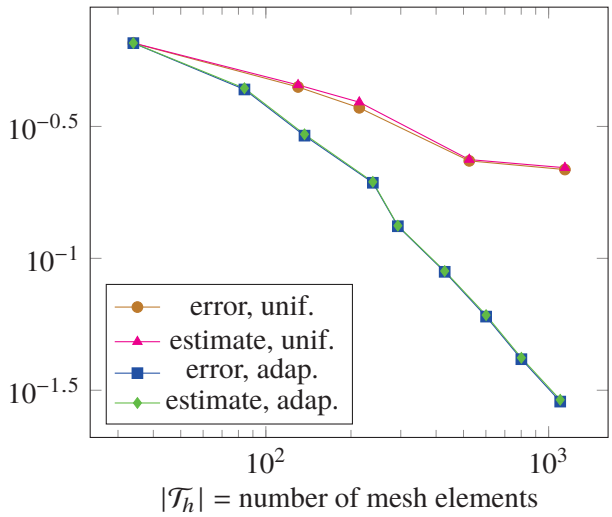
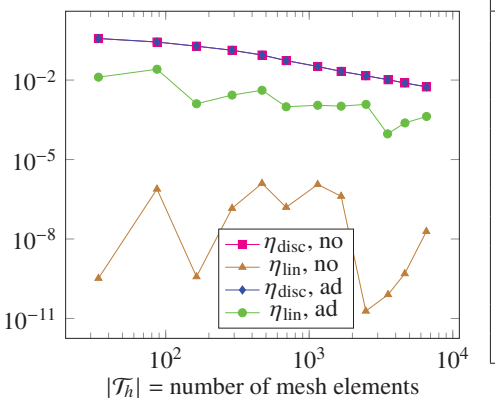


Figure B.2: L-shaped domain with linear elasticity model. Distribution of the error estimators (top) and the analytical error (bottom) for the initial mesh (left) and after three (middle) and six (right) adaptive mesh refinements.



$\lvert\mathcal{T}_h\rvert$	I_{eff}
34	1.00
84	1.01
130	*1.02*
137	1.01
214	*1.05*
239	1.01
293	1.00
429	1.01
524	*1.01*
601	1.01
801	1.01
1099	1.02
1142	*1.02*

Figure B.3: L-shaped domain with linear elasticity model. *Left:* Comparison of the error estimate (B.14) and $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\text{en}}$ on two series of meshes, obtained by uniform and adaptive remeshing. *Right:* Effectivity indices of the estimate for each mesh, with the meshes stemming from uniform refinement in italic.



	unif.	adap.
0	4	2
1	4	2
2	5	3
3	5	3
4	5	3
5	6	4
6	6	4
7	6	4
8	7	4
9	7	5
10	8	5
11	9	5

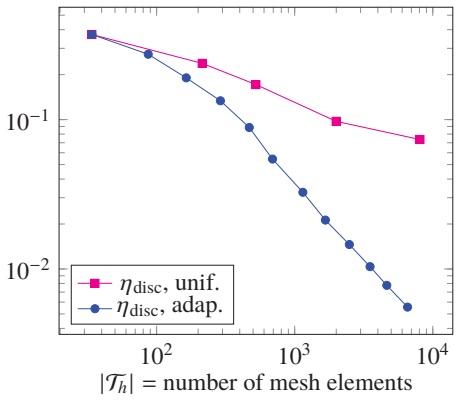


Figure B.4: L-shaped domain with Hencky–Mises model. *Left:* Comparison of the discretization and linearization error estimators on a series of meshes, without and with adaptive stopping criterion for the Newton algorithm. *Middle:* Number of Newton iterations without and with adaptive stopping criterion for each mesh. *Right:* Discretization error estimate for uniform and adaptive remeshing.

adaptive refinement based on the error estimate for the second series. Figure B.2 compares the distribution of the error and the estimators on the initial and two adaptively refined meshes. The error estimators reflect the distribution of the analytical error, which makes them a good indicator for adaptive remeshing. Figure B.3 shows the global estimates and errors for each mesh, as well as their effectivity index corresponding to the ratio of the estimate to the error. We obtain effectivity indices close to one, showing that the estimated error value lies close to the actual one, what we can also observe in the graphics on the left. As expected, the adaptively refined mesh series has a higher convergence rate, with corresponding error an order of magnitude lower for 10^3 elements.

For the Hencky–Mises model we choose the Lamé functions

$$\tilde{\mu}(\rho) := a + b(1+\rho^2)^{-1/2}, \quad \tilde{\lambda}(\rho) := \kappa - \frac{3}{2}\tilde{\mu}(\rho),$$

corresponding to the Carreau law for elastoplastic materials (see, e.g. [104, 143, 177]), and we set $a = 1/20$, $b = 1/2$, and $\kappa = 17/3$ so that the shear modulus reduces progressively to approximately 10% of its initial value. This model allows us to soften the singularity observed in the linear case and to validate our results on more homogeneous error distributions. We apply Algorithm B.12 with $\gamma_{\text{lin}} = 0.1$ and compare the obtained results to those without the adaptive stopping criterion for the Newton solver. In both cases, we use adaptive remeshing based on the spatial error estimators.

The results are shown in Figure B.4. In the left graphic we observe that the linearization error estimate in the adaptive case is much higher than in the one without adaptive stopping criterion. We see that this does not affect the discretization error estimator. The table shows the number of performed Newton iterations for both cases. The algorithm using the adaptive stopping criterion is more efficient. In the right graphic we compare the discretization error estimate on two series of meshes, one refined uniformly and the other one adaptively, based on the local discretization error estimators. Again the convergence rate is higher for the adaptively refined mesh series.

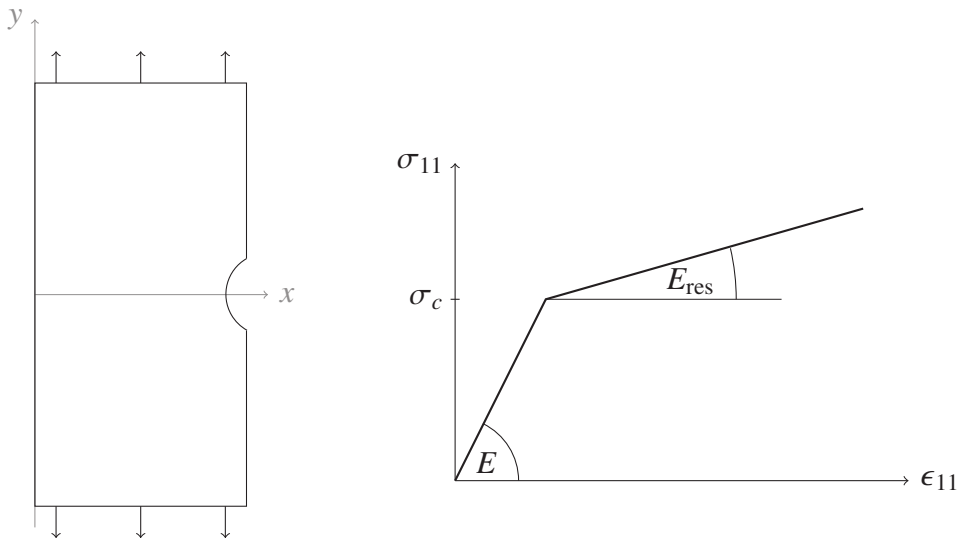


Figure B.5: *Left:* Notched specimen plate. *Right:* Uniaxial traction curve

B.5.2 Notched specimen plate

In our second test we use the two nonlinear models of Examples 3.3 and 3.4 on a more application-oriented test. The idea is to set a special sample geometry yielding to a model discrimination test, namely different physical results for different models. We simulate the uniform traction of a notched specimen under plain strain assumption (cf. Figure B.5). The notch is meant to favor strain localization phenomenon. We consider a domain $\Omega = (0, 10\text{m}) \times (-10\text{m}, 10\text{m}) \setminus \{\boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x}\text{m} - (0, 11\text{m})^T\| \leq 2\text{m}\}$, we take $\boldsymbol{f} = \boldsymbol{0}$, and we prescribe a displacement on the boundary leading to the following Dirichlet conditions:

$$u_x = 0\text{m} \text{ if } x = 0\text{m}, \quad u_y = -1.1 \cdot 10^{-3}\text{m} \text{ if } y = -10\text{m}, \quad u_y = 1.1 \cdot 10^{-3}\text{m} \text{ if } y = 10\text{m}.$$

In many applications, the information about the material properties are obtained in uniaxial experiments, yielding a relation between σ_{ii} and ϵ_{ii} for a space direction x_i. Since we only consider isotropic materials, we can choose $i = 1$. From this curve one can compute the nonlinear Lamé functions of (3.4) and the damage function in (3.5). Although the uniaxial relation is the same, the resulting stress-strain relations will be different. In our test, we use the $\sigma_{11} - \epsilon_{11}$ – relation indicated in the right of Figure B.5 with

$$\sigma_c = 3 \cdot 10^4\text{Pa}, \quad \text{E} = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu} = 3 \cdot 10^8\text{Pa}, \quad \text{E}_{\text{res}} = 3 \cdot 10^7\text{Pa},$$

corresponding to the Lamé parameters $\mu = \frac{3}{26} \cdot 10^9\text{Pa}$ and $\lambda = \frac{9}{52} \cdot 10^9\text{Pa}$. For both stress-strain relations we apply Algorithm B.12 with $\gamma_{\text{lin}} = 0.1$. We first compare the results to a computation on a very fine mesh to evaluate the remeshing based on the discretization error estimators. Secondly, we perform adaptive remeshing based on these estimators but without applying the adaptive stopping of the Newton iterations and compare the two series of meshes. As in Section B.5.1, we verify if the reduced number of iterations impacts the discretization error.

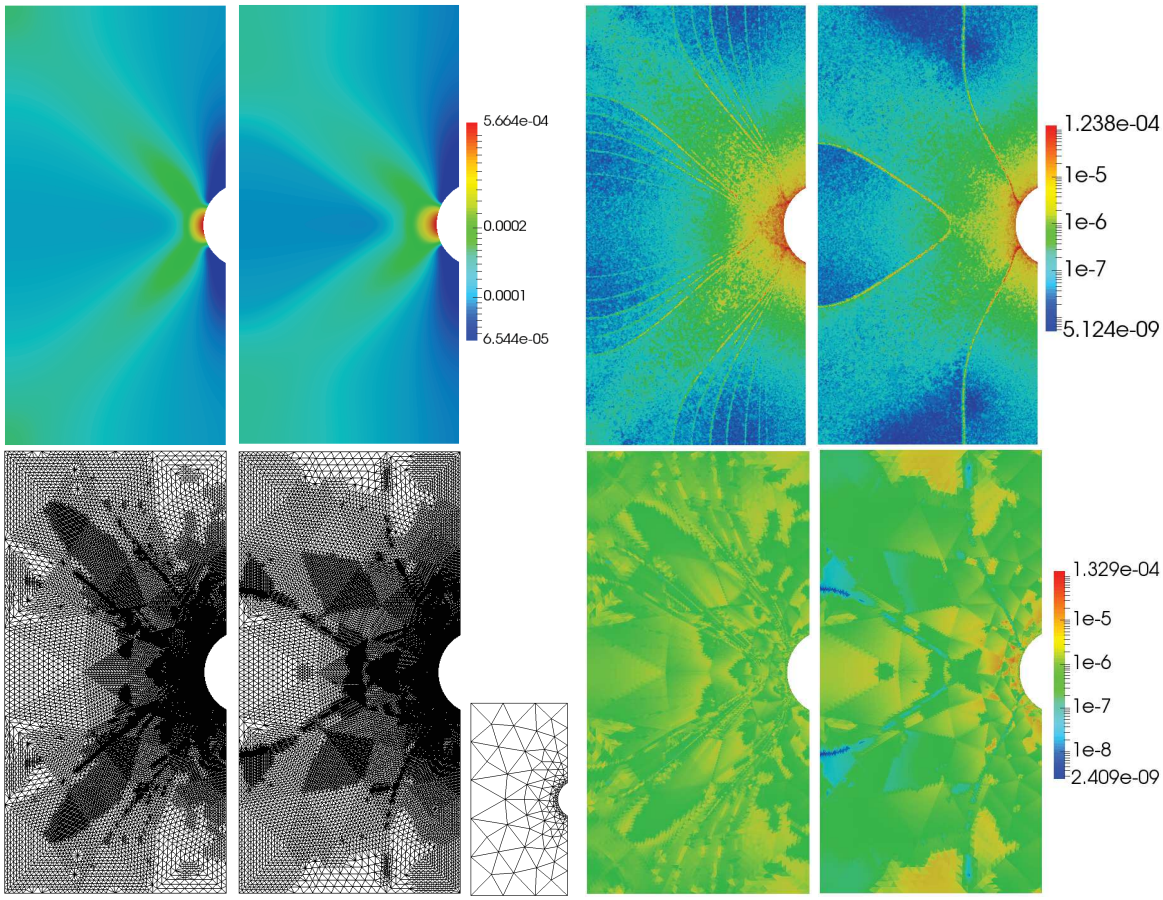
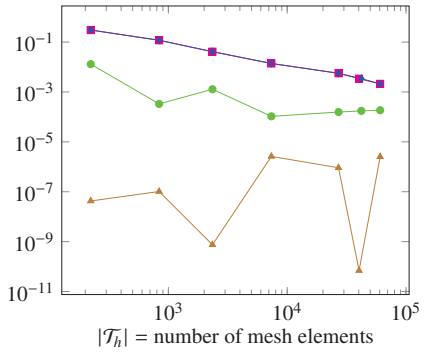
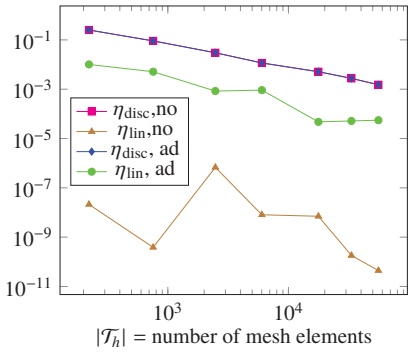


Figure B.6: Notched specimen plate, comparison between Hencky–Mises (left in each picture) and damage model (right). *Top left:* $\mathrm{tr}(\boldsymbol{\nabla}_s \boldsymbol{u}_h)$. *Top right:* η_{disc} on a fine mesh (no adaptive refinement). *Bottom left:* meshes after six adaptive refinements. *Bottom middle:* initial mesh. *Bottom right:* η_{disc} on the adaptively refined meshes.

Figure B.6 shows the result of the first part of the test. In each of the four images the left specimen corresponds to the Hencky–Mises and the right to the isotropic damage model. To illustrate the difference of the two models, the top left picture shows the trace of the strain tensor. This scalar value is a good indicator for both models, representing locally the relative volume increase which could correspond to either a damage or shear band localization zone. In the top right picture we see the distribution of the discretization error estimators in the reference computation (209,375 elements), whereas the distribution of the estimators on the sixth adaptively refined mesh is shown in the bottom right picture (60,618 elements for Hencky–Mises, 55,718 elements for the damage model). The corresponding meshes and the initial mesh for the adaptive algorithm are displayed in the bottom left of the figure. To ensure a good discretization of the notch after repeated mesh refinement, the initial mesh cannot be too corse in this curved area. We observe that the adaptively refined meshes match the distribution of the discretization error estimators on the uniform mesh, and that the estimators are more evenly distributed on these meshes.

The results of the second part of the test are illustrated in Figure B.7. As for the L-shape





	Hencky–Mises		Damage	
	nor.	adap.	nor.	adap.
0	5	3	6	4
1	5	4	6	4
2	6	4	6	5
3	6	5	7	5
4	6	5	7	6
5	7	5	8	6
6	6	5	8	6

Figure B.7: Notched specimen plate. Comparison of the global discretization and linearization error estimators without and with adaptive stopping criterion for the Hencky–Mises model (left) and the damage model (middle), and comparison of the number of performed Newton iterations (right).

test, we observe that the reduced number of Newton iterations does not affect the discretization error estimate, nor the overall error estimate which is dominated by the discretization error estimate if the Newton algorithm is stopped.

List of Figures

Bibliography

[1] M. Abbas, A. Ern, and N. Pignet. *Hybrid High-Order methods for finite deformations of hyperelastic materials*. Submitted. 2018.

[2] M. Ainsworth and J.T. Oden. "A posteriori error estimation in finite element analysis". In: *Pure and Applied Mathematics, Wiley-Interscience [John Wiley & Sons], New York* (2000).

[3] M. Ainsworth and J.T. Oden. "A posteriori error estimators for second order elliptic systems Part 2. An optimal order process for calculating self-equilibrating fluxes". In: *Computers & Mathematics with Applications* 26.9 (1993), pp. 75–87. DOI: 10.1016/0898-1221(93)90007-I.

[4] M. Ainsworth and Rankin R. "Guaranteed computable error bounds for conforming and nonconforming finite element analysis in planar elasticity". In: *Internat. J. Numer. Methods Engrg* 82 (2010), pp. 1114–1157.

[5] T. Arbogast and Z. Chen. "On the implementation of mixed methods as nonconforming methods for second-order elliptic problems". In: *Math. Comp.* 64.211 (1995), pp. 943–972.

[6] C.G. Armstrong, W.M. Lai, and V.C. Mow. "An analysis of the unconfined compression of articular cartilage". In: *J. Biomech. Eng.* 106.2 (1984), pp. 165–173.

[7] D. N. Arnold, R. S. Falk, and R. Winther. "Mixed finite element methods for linear elasticity with weakly imposed symmetry". In: *Math. Comp.* 76 (2007), pp. 1699–1723.

[8] D. N. Arnold and Winther R. "Mixed finite elements for elasticity". In: *Numer. Math.* 92 (2002), pp. 401–419.

[9] J.-L. Auriault and E. Sanchez-Palencia. "Étude du comportement macroscopique d'un milieu poreux saturé déformable". In: *J. Méc.* 16.4 (1977), pp. 575–603.

[10] I. Babuška and P. Chatzipantelidis. "On solving elliptic stochastic partial differential equations". In: *Comput. Methods Appl. Mech. Engrg.* 191.37 (2002), pp. 4093–4122. ISSN: 0045-7825. DOI: 10.1016/S0045-7825(02)00354-7.

[11] I. Babuška, F. Nobile, and R. Tempone. "A stochastic collocation method for elliptic partial differential equations with random input data". In: *SIAM J. Numer. Anal.* 45.3 (2007), pp. 1005–1034. DOI: 10.1137/050645142.

[12] I. Babuška, R. Tempone, and G.E. Zouraris. "Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations". In: *SIAM J. Numer. Anal.* 42.2 (2004), pp. 800–825. ISSN: 0036-1429. DOI: 10.1137/S0036142902418680.

[13] A. M. Barrientos, N. G. Gatica, and P. E. Stephan. "A mixed finite element method for nonlinear elasticity: two-fold saddle point approach and a-posteriori error estimate". In: *Numer. Math.* 91.2 (2002), pp. 197–222. DOI: 10.1007/s002110100337.

[14] S. Barry and G. Mercer. "Exact solution for two-dimensional time dependent flow and deformation within a poroelastic medium". In: *J. Appl. Mech.* 66.2 (1999), pp. 536–540.

[15] S.I. Barry and G.K. Aldis. "Comparison of models for flow induced deformation of soft biological tissue". In: *J. Biomech.* 23.7 (1990), pp. 647–654.

[16] F. Bassi, L. Botti, A. Colombo, D. A. Di Pietro, and P. Tesini. "On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations". In: *J. Comput. Phys.* 231.1 (2012), pp. 45–65.

[17] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. "A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows". In: *Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics* (1997), pp. 99–109.

[18] J. Bear. *Dynamics of Fluids In Porous Media*. Vol. 1. American Elsevier Publishing Company, New York, 1972.

[19] L. Beirão Da Veiga. "A mimetic discretization method for linear elasticity". In: *M2AN Math. Model. Numer. Anal.* 44.2 (2010), pp. 231–250.

[20] L. Beirão da Veiga, F. Brezzi, and L. D. Marini. "Virtual elements for linear elasticity problems". In: *SIAM J. Numer. Anal.* 2.51 (2013), pp. 794–812.

[21] L. Beirão da Veiga, C. Lovadina, and D. Mora. "A Virtual Element Method for elastic and inelastic problems on polytope meshes". In: *Comput. Methods Appl. Mech. Engrg.* 295 (2015), pp. 327–346.

[22] E. Bemer, M. Boutéca, O. Vincké, N. Hoteit, and O. Ozanam. "Poromechanics: From linear to nonlinear poroelasticity and poroviscoelasticity". In: *Oil & Gas Science and Technologies– Rev. IFP* 56.6 (2001), pp. 531–544. DOI: 10.2516/OGST:2001043.

[23] L. Berger, R. Bordas, D. Kay, and S. Tavener. "A stabilized finite element method for finite-strain three-field poroelasticity". In: *Comput. Mech.* 60 (2017), pp. 51–68. DOI: 10.1007/s00466-017-1381-8.

[24] A. Bespalov, C.E. Powell, and D. Silvester. "A priori error analysis of stochastic Galerkin mixed approximations of elliptic PDEs with random data". In: *SIAM J. Numer. Anal.* 50.4 (2012), pp. 2039–2063. DOI: 10.1137/110854898.

[25] C. Bi and Y. Lin. "Discontinuous Galerkin method for monotone nonlinear elliptic problems". In: *Int. J. Numer. Anal. Model* 9 (2012), pp. 999–1024.

[26] S.O.R. Biabanaki, A.R. Khoei, and P. Wriggers. "Polygonal finite element methods for contact-impact problems on non-conformal meshes". In: *Comput. Meth. Appl. Mech. Engrg.* 269 (2014), pp. 198–221. DOI: 10.1016/j.cma.2013.10.025.

[27] M. A. Biot. "General theory of threedimensional consolidation". In: *J. Appl. Phys.* 12.2 (1941), pp. 155–164.

[28] M. A. Biot. "Nonlinear and semilinear rheology of porous solids". In: *J. Geoph. Res.* 78.23 (1973), pp. 4924–4937.

[29] M. A. Biot. "Theory of elasticity and consolidation for a porous anisotropic solid". In: *J. Appl. Phys.* 26.2 (1955), pp. 182–185.

[30] D. Boffi, M. Botti, and D. A. Di Pietro. "A nonconforming high-order method for the Biot problem on general meshes". In: *SIAM J. Sci. Comput.* 38.3 (2016), A1508–A1537. DOI: 10.1137/15M1025505.

[31] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*. Vol. 44. Springer Series in Computational Mathematics. Berlin Heidelberg: Springer, 2013. DOI: 10.1007/978-3-642-36519-5.

[32] L. Botti, D. A. Di Pietro, and J. Droniou. *A Hybrid High-Order discretisation of the Brinkman problem robust in the Darcy and Stokes limits*. Submitted. Mar. 2018. URL: http://hal.archives-ouvertes.fr/hal-01746367.

[33] M. Botti, D. A. Di Pietro, and P. Sochala. "A Hybrid High-Order method for nonlinear elasticity". In: *SIAM J. Numer. Anal.* 55.6 (2017), pp. 2687–2717. DOI: 10.1137/16M1105943.

[34] M. Botti, D. A. Di Pietro, and P. Sochala. "A nonconforming high-order method for nonlinear poroelasticity". In: *Finite Volumes for Complex Applications VIII – Hyperbolic, Elliptic and Parabolic Problems*. 2017, pp. 537–545.

[35] M. Botti, D. A. Di Pietro, and P. Sochala. *Analysis of a Hybrid High-Order–discontinuous Galerkin discretization method for nonlinear poroelasticity*. Submitted. May 2018. URL: http://hal.archives-ouvertes.fr/hal-01785810.

[36] M. Botti and R. Riedlbeck. "Equilibrated stress tensor reconstruction and a posteriori error estimation for nonlinear elasticity". In: *Comput. Methods Appl. Math.* (June 2018). DOI: 10.1515/cmam-2018-0012.

[37] D. Braess, V. Pillwein, and J. Schöberl. "Equilibrated residual error estimates are p-robust". In: *Comput. Methods Appl. Mech. Engrg.* 198 (2009), pp. 1189–1197.

[38] D. Braess and J. Schöberl. "Equilibrated residual error estimator for edge elements". In: *Math. Comp.* 77.262 (2008), pp. 651–672.

[39] S. C. Brenner. "Korn's inequalities for piecewise H^1 vector fields". In: *Math. Comp.* 73.247 (2004), pp. 1067–1087. ISSN: 0025-5718.

[40] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, 2010. ISBN: 9780387709130.

[41] F. Brezzi, J. Douglas, and L. D. Marini. "Recent results on mixed finite element methods for second order elliptic problems". In: *Vistas in applied mathematics. Numerical analysis, atmospheric sciences, immunology*. Ed. by Dorodnitsyn Balakrishnan and Lions Eds. Optimization Software Inc., Publications Division, New York, 1986, pp. 25–43.

[42] F. Brezzi, G. Manzini, L. D. Marini, P. Pietra, and A. Russo. "Discontinuous Galerkin approximations for elliptic problems". In: *Numer. Methods Partial Differential Equations* 16.4 (2000), pp. 365–378. ISSN: 0749-159X.

[43] M. K. Brun, I. Berre, J. M. Nordbotten, and F. A. Radu. "Upscaling of the Coupling of Hydromechanical and Thermal Processes in a Quasi-static Poroelastic Medium". In: *Transport in Porous Media* 124.1 (2018), pp. 137–158. ISSN: 1573-1634. DOI: 10.1007/s11242-018-1056-8.

[44] H. Bryne and L. Preziosi. "Modeling tumour growth using the theory of mixtures". In: *Mathematical Medicine and Biology* 20.4 (2003), pp. 341–366.

[45] R. Burridge and Keller J. B. "Poroelasticity equations derived from microstructure". In: *J. Acoust. Soc. Am.* 70.4 (1981), pp. 1140–1146. DOI: 10.1121/1.386945.

[46] R. H. Cameron and W. T. Martin. "The orthogonal development of nonlinear functionals in series of Fourier–Hermite functionals". In: *Ann. Math.* 48 (1947), pp. 385–392.

[47] A. Cangiani, E. H. Georgoulis, and P. Houston. "*hp*-version discontinuous Galerkin methods on polygonal and polyhedral meshes". In: *Math. Models Methods Appl. Sci.* 24.10 (2014), pp. 2009–2041. ISSN: 0218-2025. DOI: 10.1142/S0218202514500146.

[48] Y. Cao, S. Chen, and A. J. Mier. "Analysis and numerical approximations of equations of nonlinear poroelasticity". In: *Discrete & Continuous Dynamical Systems - B* 18 (2013), pp. 1253–1273. DOI: 10.3924/dcdsb.2013.18.1253.

[49] M. Čermák, F. Hecht, Z. Tang, and M. Vohralík. "Adaptive inexact iterative algorithms based on polynomial-degree-robust a posteriori estimates for the Stokes problem". In: *Numer. Math.* 138.4 (2018), pp. 1027–1065. DOI: 10.1007//s00211-017-0925-3.

[50] M. Cervera, M. Chiumenti, and R. Codina. "Mixed stabilized finite element methods in nonlinear solid mechanics: Part II: Strain localization". In: *Comput. Methods in Appl. Mech. and Engrg.* 199.37–40 (2010), pp. 2571–2589. DOI: 10.1016/j.cma.2010.04.005.

[51] L. Chamoin, P. Ladevèze, and F. Pled. "On the techniques for constructing admissible stress fields in model verification: Performances on engineering examples". In: *Internat. J. Numer. Methods Engrg.* 88.5 (2011), pp. 409–441. DOI: 10.1002/nme.3180.

[52] C.S. Chang. "Uncertainty in one-dimensional consolidation analysis". In: *J. Geothec. Eng.* 111.12 (1985), pp. 1411–1424.

[53] J. Charrier. "Strong and weak error estimates for elliptic partial differential equations with random coefficients". In: *SIAM J. Numer. Anal.* 50.1 (2012), pp. 216–246. DOI: 10.1137/100800531.

[54] H. Chi, L. Beirão da Veiga, and G.H. Paulino. "Some basic formulations of the virtual element method (VEM) for finite deformations". In: *Computer Methods in Applied Mechanics and Engineering* 318 (2017), pp. 148–192. ISSN: 0045-7825. DOI: 10.1016/j.cma.2016.12.020.

[55] H. Chi, C. Talischi, O. Lopez-Pamies, and G. H. Paulino. "Polygonal finite elements for finite elasticity". In: *Int. J. Numer. Methods Eng.* 101.4 (2015), pp. 305–328. ISSN: 1097-0207. DOI: 10.1002/nme.4802.

[56] P. G. Ciarlet and P. Ciarlet. "Another approach to linearized elasticity and a new proof of Korn's inequality". In: *Math. Models Methods Appl. Sci.* 15.02 (2005), pp. 259–271. DOI: 10.1142/S0218202505000352.

[57] B. Cockburn, D. A. Di Pietro, and A. Ern. "Bridging the Hybrid High-Order and Hybridizable Discontinuous Galerkin Methods". In: *M2AN Math. Model. Numer. Anal.* (2015). Published online. DOI 10.1051/m2an/2015051.

[58] P. R. Conrad and Y. M. Marzouk. "Adaptive Smolyak Pseudospectral Approximations". In: *SIAM J. Sci. Comp.* 35.6 (2013), A2643–A2670. DOI: 10.1137/120890715.

[59] P. G. Constantine, M. S. Eldred, and E. T. Phipps. "Sparse pseudospectral approximation method". In: *Comput. Methods Appl. Mech. Engrg.* 229 (2012), pp. 1–12.

[60] P. Cosenza, M. Ghoreychi, G. De Marsily, G. Vasseur, and S. Violette. "Theoretical prediction of poroelastic properties of argillaceous rocks from in situ specific storage coefficient". In: *Water Resour. Res.* 38.10 (2002), p. 1207. DOI: 10.1029/2001WR001201.

[61] O. Coussy. *Poromechanics*. J. Wiley and Sons, ltd., 2004. DOI: 10.1002/0470092718.

[62] A.A. Darrag and M.A Tawil. "The consolidation of soils under stochastic initial excess pore pressure". In: *Applied Mathematical Modelling* 17.11 (1993), pp. 609–612. DOI: 10.1016/0307-904X(93)90069-S.

[63] R. Davies, G. Foulger, A. Bindley, and P. Styles. "Induced seismicity and hydraulic fracturing for the recovery of hydrocarbons". In: *Mar. Petrol. Geol.* 45.1 (2013), pp. 171–185. DOI: 10.1016/j.marpetgeo.2013.03.016.

[64] K. Deimling. "Nonlinear functional analysis". In: *Springer-Verlag, Berlin* (1985). DOI: 10.1007/978-3-662-00547-7.

[65] P. Delgado and V. Kumar. "A stochastic Galerkin approach to uncertainty quantification in poroelastic media". In: *Applied Mathematics and Computation* 266.1 (2015), pp. 328–338. DOI: 10.1016/j.amc.2015.04.127.

[66] M. Destrade and R. W. Ogden. "On the third- and fourth-order constants of incompressible isotropic elasticity". In: *The Journal of the Acoustical Society of America* 128.6 (2010), pp. 3334–3343. DOI: 10.1121/1.3505102.

[67] P. Destuynder and B. Métivet. "Explicit error bounds in a conforming finite element method". In: *Math. Comp.* 68.228 (1999), pp. 1379–1396.

[68] E. Detournay and A. H. D. Cheng. "Fundamentals of poroelasticity". In: *Comprehensive rock engineering. Vol. 2*. Ed. by J.A. Hudson. Pergamon, 1993, pp. 113–171.

[69] D. A. Di Pietro and J. Droniou. "A Hybrid High-Order method for Leray–Lions elliptic equations on general meshes". In: *Math. Comp.* 86.307 (2017), pp. 2159–2191. DOI: 10.1142/S0218202517500191.

[70] D. A. Di Pietro and J. Droniou. "$W^{s,p}$-approximation properties of elliptic projectors on polynomial spaces, with application to the error analysis of a Hybrid High-Order discretization of Leray–Lions problems". In: *Math. Models Methods Appl. Sci.* 27.5 (2017), pp. 879–908. DOI: 10.1142/S0218202517500191.

[71] D. A. Di Pietro, J. Droniou, and A. Ern. "A Discontinuous-Skeletal Method for Advection-Diffusion-Reaction on General Meshes". In: *SIAM J. Numer. Anal.* 53.5 (2015), pp. 2135–2157. DOI: 10.1137/140993971.

[72] D. A. Di Pietro, J. Droniou, and G. Manzini. "Discontinuous Skeletal Gradient Discretisation methods on polytopal meshes". In: *J. Comput. Phys.* 355 (2018), pp. 397–425. DOI: 10.1016/j.jcp.2017.11.018.

[73] D. A. Di Pietro and A. Ern. "A family of arbitrary order mixed methods for heterogeneous anisotropic diffusion on general meshes". In: *IMA J. Numer. Anal.* (2016). Published online. DOI: 10.1093/imanum/drw003.

[74] D. A. Di Pietro and A. Ern. "A hybrid high-order locking-free method for linear elasticity on general meshes". In: *Comput. Meth. Appl. Mech. Engrg.* 283 (2015), pp. 1–21. DOI: 10.1016/j.cma.2014.09.009.

[75] D. A. Di Pietro and A. Ern. "Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier–Stokes equations". In: *Math. Comp.* 79 (2010), pp. 1303–1330. DOI: 10.1090/S0025-5718-10-02333-1.

[76] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*. Vol. 69. Mathématiques & Applications. Berlin: Springer-Verlag, 2012. DOI: 10.1007/978-3-642-22980-0.

[77] D. A. Di Pietro, A. Ern, and J.-L. Guermond. "Discontinuous Galerkin methods for anisotropic semi-definite diffusion with advection". In: *SIAM J. Numer. Anal.* 46.2 (2008), pp. 805–831. DOI: 10.1137/060676106.

[78] D. A. Di Pietro, A. Ern, and S. Lemaire. "An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators". In: *Comput. Methods Appl. Math.* 14.4 (2014), pp. 461–472.

[79] D. A. Di Pietro, B. Kapidani, R. Specogna, and F. Trevisan. "An Arbitrary-Order Discontinuous Skeletal Method for Solving Electrostatics on General Polyhedral Meshes". In: *IEEE Transactions on Magnetics* 53.6 (June 2017), pp. 1–4. ISSN: 0018-9464. DOI: 10.1109/TMAG.2017.2666546.

[80] D. A. Di Pietro and S. Lemaire. "An extension of the Crouzeix–Raviart space to general meshes with application to quasi-incompressible linear elasticity and Stokes flow". In: *Math. Comp.* 84.291 (2015), pp. 1–31.

[81] D. A. Di Pietro and S. Nicaise. "A locking-free discontinuous Galerkin method for linear elasticity in locally nearly incompressible heterogeneous media". In: *Appl. Numer. Math.* 63 (2013), pp. 105–116. ISSN: 0168-9274. DOI: 10.1016/j.apnum.2012.09.009.

[82] D. A. Di Pietro and R. Tittarelli. "Numerical Methods for PDEs. State of the Art Techniques". In: ed. by L. Formaggia D. A. Di Pietro A. Ern. SEMA-SIMAI 15. Springer, 2018. Chap. An introduction to Hybrid High-Order methods. URL: http://arxiv.org/abs/1703.05136.

[83] P. Dörsek and J. Melenk. "Symmetry-free, *p*-robust equilibrated error indication for the *hp*-version of the FEM in nearly incompressible linear elasticity". In: *Comput. Methods Appl. Math.* 13 (2013), pp. 291–304.

[84] J. Droniou. "Finite volume schemes for fully non-linear elliptic equations in divergence form". In: *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique* 40.6 (2006), pp. 1069–1100.

[85] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. *The gradient discretisation method*. Preprint hal-01366646. Nov. 2016.

[86] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. *The gradient discretisation method: A framework for the discretisation and numerical analysis of linear and nonlinear elliptic and parabolic problems*. Maths & Applications. To appear. Springer, 2017, 509p. URL: https://hal.archives-ouvertes.fr/hal-01382358.

[87] J. Droniou and B. P. Lamichhane. "Gradient Schemes for Linear and Non-linear Elasticity Equations". In: *Numer. Math.* 129.2 (2015), pp. 251–277. ISSN: 0029-599X. DOI: 10.1007/s00211-014-0636-y.

[88] T. Dupont and R. Scott. "Polynomial approximation of functions in Sobolev spaces". In: *Math. Comp.* 34.150 (1980), pp. 441–463.

[89] G. Duvaut and Lions J.L. *Inequalities in Mechanics and Physics*. 1st ed. Grundlehren der mathematischen Wissenschaften 219. Springer-Verlag Berlin Heidelberg, 1976. ISBN: 978-3-642-66167-9,978-3-642-66165-5.

[90] W. Ehlers and G. Eipper. "Finite Elastic Deformations in Liquid-Saturated and Empty Porous Solids". In: *Transport in Porous Media* 34.1 (1999), pp. 179–191. ISSN: 1573-1634. DOI: 10.1023/A:1006565509095.

[91] L. El Alaoui, A. Ern, and M. Vohralík. "Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems". In: *Comput. Methods Appl. Mech. Engrg.* 200 (2011), pp. 2782–2795.

[92] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*. Vol. 159. Applied Mathematical Sciences. New York, NY: Springer-Verlag, 2004.

[93] A. Ern, A. F. Stephansen, and P. Zunino. "A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally small and anisotropic diffusivity". In: *IMA J. Numer. Anal.* 29.2 (2009), pp. 235–256. ISSN: 0272-4979. DOI: 10.1093/imanum/drm050.

[94] A. Ern and M. Vohralík. "A posteriori error estimation based on potential and flux reconstruction for the heat equation". In: *SIAM J. Numer. Anal.* 48.1 (2010), pp. 198–223.

[95] A. Ern and M. Vohralík. "Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs". In: *SIAM J. Sci. Comput.* 35.4 (2013), A1761–A1791.

[96] A. Ern and M. Vohralík. *Broken stable H^1 and $H(\mathrm{div})$ polynomial extensions for polynomial-degree-robust potential and flux reconstruction in three space dimensions*. preprint hal-01422204. 2017.

[97] A. Ern and M. Vohralík. "Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations". In: *SIAM J. Numer. Anal.* 53.2 (2015), pp. 1058–1081.

[98] O. G. Ernst, A. Mugler, H-J. Starkloff, and E. Ullmann. "On the convergence of generalized polynomial chaos expansions". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 46 (02 Mar. 2012), pp. 317–339. ISSN: 1290-3841. DOI: 10.1051/m2an/2011045.

[99] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in advanced mathematics. Boca Raton (Fla.): CRC Press, 1992. URL: http://opac.inria.fr/record=b1089059.

[100] D.G. Frias, M.A. Murad, and F. Pereira. "Stochastic computational modelling of highly heterogeneous poroelastic media with long-range correlations". In: *Int. J. Numer. Anal. Meth. Geomech.* 28.1 (2004), pp. 1–32. DOI: 10.1002/nag.323.

[101] G. Fu. *A high-order HDG method for the Biot's consolidation model*. Submitted. 2018.

[102] F.J. Gaspar, F.J. Lisbona, C.W. Oosterlee, and P.N. Vabishchevich. "An efficient multigrid solver for a reformulated version of the poroelasticity system". In: *Comput. Methods Appl. Mech. and Engrg.* 196 (2007), pp. 1447–1457. DOI: 10.1016/j.cma.2006.03.020.

[103] F. Gassmann. "On Elasticity of Porous Media". In: *Vierteljahresheft der Naturforschenden Gesselschaft* 96 (1951), pp. 1–23.

[104] G. N. Gatica, A. Márquez, and W. Rudolph. "A priori and a posteriori error analyses of augmented twofold saddle point formulations for nonlinear elasticity problems". In: *Comput. Meth. Appl. Mech. Engrg.* 264 (2013), pp. 23–48. DOI: 10.1016/j.cma.2013.05.010.

[105] G. N. Gatica and E. P. Stephan. "A mixed-FEM formulation for nonlinear incompressible elasticity in the plane". In: *Numerical Methods for Partial Differential Equations* 18.1 (2002), pp. 105–128. ISSN: 1098-2426. DOI: 10.1002/num.1046.

[106] G. N. Gatica and W. L. Wendland. "Coupling of Mixed Finite Elements and Boundary Elements for A Hyperelastic Interface Problem". In: *SIAM J. on Numer. Anal.* 34.6 (1997), pp. 2335–2356. DOI: 10.1137/S0036142995291317.

[107] B. Gatmiri and P. Delage. "A formulation of fully coupled thermal-hydraulic-mechanical behavior of saturated porous media–numerical approach". In: *Int. J. Numerical and Anal. Meth. Geomech.* 21.3 (1997), pp. 199–225. DOI: 10.1002/(SICI)1096-9853(199703)21:3<199::AID-NAG865>3.0.CO;2-M.

[108] T. Gerstner and M. Griebel. "Dimension-Adaptive Tensor-Product Quadrature". In: *Computing* 71.1 (2003), pp. 65–87.

[109] T. Gerstner and M. Griebel. "Numerical Integration Using Sparse Grids". In: *Numerical Algorithms* (1998), pp. 209–232.

[110] R. G. Ghanem. "Probabilistic characterization of transport in heterogeneous media". In: *Comput. Meth. Appl. Mech. Eng.* 158.3 (1998), pp. 199–220. ISSN: 0045-7825. DOI: 10.1016/S0045-7825(97)00250-8.

[111] R. G. Ghanem and S. Dham. "Stochastic Finite Element Analysis for Multiphase Flow in Heterogeneous Porous Media". In: *Transport in Porous Media* 32.3 (1998), pp. 239–262. ISSN: 1573-1634. DOI: 10.1023/A:1006514109327.

[112] R. G. Ghanem and S. D. Spanos. *Stochastic Finite Elements: a Spectral Approach*. Springer Verlag, 1991.

[113] S.M. Gorelick and M. D. Zoback. "Earthquake triggering and large-scale geologic storage of carbon dioxide". In: *Proc. Natl. Acad. Sci. U.S.A.* 109.26 (2012), pp. 10164–10168. DOI: 10.1073/pnas.1202473109.

[114] D. H. Green and H. F. Wang. "Specific storage as a poroelastic coefficient". In: *Water Resour. Res.* 26.7 (1990), pp. 1631–1637. DOI: 10.1029/WR026i007p01631.

[115] P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman Advanced Pub. Program. London, 1985.

[116] G. Guennebaud and B. Jacob. *Eigen v3*. http://eigen.tuxfamily.org. 2010.

[117] J. B. Haga, H. Osnes, and H. P. Langtangen. "On the causes of pressure oscillations in low-permeable and low-compressible porous media". In: *Int. J. Numer. Anal. Methods Geomech.* 36.12 (2012), pp. 1507–1522. DOI: 10.1002/nag.1062.

[118] A. Hannukainen, R. Stenberg, and M. Vohralík. "A unified framework for a posteriori error estimation for the Stokes problem". In: *Numer. Math.* 122 (2012), pp. 725–769.

[119] T. D. Hein and M. Kleiber. "Stochastic Finite Element modeling in linear transient heat transfer". In: *Comput. Meth. Appl. Mech. Eng.* 144 (1997), pp. 111–124.

[120] R. Herbin and F. Hubert. "Benchmark on discretization schemes for anisotropic diffusion problems on general grids". In: *Finite Volumes for Complex Applications V*. Ed. by R. Eymard and J.-M. Hérard. John Wiley & Sons, 2008, pp. 659–692.

[121] J. G. Heywood and R. Rannacher. "Finite-element approximation of the nonstationary Navier–Stokes problem. Part IV: error analysis for second-order time discretization". In: *SIAM J. Numer. Anal.* 27.2 (1990), pp. 353–384.

[122] H. P. Hong. "One-dimensional consolidation with uncertain properties". In: *Canadian Geotechnical journal* 29 (1991), pp. 161–165.

[123] L. Hu, P. H. Winterfield, P. Fakcharoenphol, and Wu Y. S. "A novel fully-coupled flow and geomechanics model in enhanced geothermal reservoirs". In: *J. Pet. Sci. Eng.* 107 (2013), pp. 1–11. DOI: 10.1016/j.petrol.2013.04.005.

[124] D. S. Hughes and J. L. Kelly. "Second-Order Elastic Deformation of Solids". In: *Phys. Rev.* 92 (5 1953), pp. 1145–1149.

[125] M. Iskandarani, S. Wang, A. Srinivasan, W. C. Thacker., J. Winokur, and O. Knio. "An overview of uncertainty quantification techniques with application to oceanic and oil-spill simulations". In: *J. Geophys. Res.: Oceans* 121.4 (2016), pp. 2789–2808. ISSN: 2169-9291. DOI: 10.1002/2015JC011366.

[126] B. Jah and R. Juanes. "Coupled multiphase flow and poromechanics: A computational model of pore pressure effects on fault slip and earthquake triggering". In: *Water Resour. Res.* 50.5 (2014), pp. 3776–3808. DOI: 10.1002/2013WR015175.

[127] L. Jin and M. D. Zoback. "Fully coupled nonlinear fluid flow and poroelasticity in arbitrarily fractured porous media: A hybrid-dimensional computational model". In: *Journal Geophysical Research: Solid Earth* 122 (2017), pp. 7626–7658. DOI: 10.1002/2017JB014892.

[128] A. Khan, C. E. Powell, and D. J. Silvester. *Robust preconditioning for stochastic Galerkin formulations of parameter-dependent linear elasticity equations*. Submitted. Mar. 2018.

[129] Kwang-Yeon Kim. "Guaranteed a posteriori error estimator for mixed finite element methods of linear elasticity with weak stress symmetry". In: *SIAM J. Numer. Anal.* 48.6 (2011), pp. 2364–2385.

[130] A.E. Kolesov, P.N. Vabishchevich, and M.V. Vasilyeva. "Splitting scheme for poroelasticity and thermoelasticity problems". In: *Comput. and Math. with Appl.* 67 (2014), pp. 2185–2198. DOI: 10.1016/j.camwa.2014.02.005.

[131] J. Korsawe, G. Starke, W. Wang, and O. Kolditz. "Finite element analysis of poro-elastic consolidation in porous media: Standard and mixed approaches". In: *Comput. Methods Appl. Mech. Engrg.* 195 (2006), pp. 1096–1115. DOI: 10.1016/j.cma.2005.04.011.

[132] P. Ladevèze. "Comparaison de modèles de milieux continus". PhD thesis. Université Pierre et Marie Curie (Paris 6), 1975.

[133] P. Ladevèze and D. Leguillon. "Error estimate procedure in the finite element method and applications". In: *SIAM J. Numer. Anal.* 20 (1983), pp. 485–509.

[134] P. Ladevèze, J. P. Pelle, and P. Rougeot. "Error estimation and mesh optimization for classical finite elements". In: *Engrg. Comp.* 8.1 (1991), pp. 69–80.

[135] L. D. Landau and E. M. Lifshitz. *Theory of elasticity*. Pergamon London, 1959, 134p.

[136] O. P. Le Maître and O. M. Knio. *Spectral Methods for Uncertainty Quantification*. Scientific Computation. Springer, 2010.

[137] O. Le Maître, M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio. "A Stochastic Projection Method for Fluid Flow II.: Random Process". In: *J. Comput. Phys.* 181.1 (2002), pp. 9–44. ISSN: 0021-9991. DOI: 10.1006/jcph.2002.7104.

[138] S. Lemaire. "Discrétisations non-conformes d'un modèle poromécanique sur maillages généraux". PhD thesis. Université Paris-Est, Dec. 2013.

[139] S. E. Leon, D. W. Spring, and G. H. Paulino. "Reduction in mesh bias for dynamic fracture using adaptive splitting of polygonal finite elements". In: *Int. J. Numer. Methods Eng.* 100.8 (2014), pp. 555–576. ISSN: 1097-0207. DOI: 10.1002/nme.4744.

[140] R. W. Lewis and B. A. Schrefler. *The finite element method in the static and dynamic deformation and consolidation of porous media*. Numerical methods in engineering. John Wiley, 1998. ISBN: 9780471928096.

[141] H. Liu, B. Hu, and Z. W. Yu. "Stochastic finite element method for random temperature inconcrete structures". In: *Int. J. Solids Struct.* 38.1 (2001), pp. 6965–6983.

[142] M. Loéve. *Probability theory*. Vol. 2. Springer-Verlang, New York, 1977.

[143] A. F. D. Loula and J. N. C. Guerreiro. "Finite element analysis of nonlinear creeping flows". In: *Comput. Meth. Appl. Mech. Engrg.* 79.1 (1990), pp. 87–109.

[144] R. Luce and B. I. Wohlmuth. "A local a posteriori error estimator based on equilibrated fluxes". In: *SIAM J. Numer. Anal.* 42 (2004), pp. 1394–1414.

[145] P. Luo, C. Rodrigo, F. J. Gaspar, and C. W. Oosterlee. "Multigrid method for nonlinear poroelasticity equations". In: *Comput. Visual. Sci.* 17 (2015), pp. 255–265. DOI: 10.1007/s00791-016-0260-8.

[146] Bowen R. M. "Compressible porous media models by use of the theory of mixtures". In: *Int. J. Eng. Sci.* 20.6 (1982), pp. 697–735. DOI: 10.1016/0020-7225(82)90082-9.

[147] Bowen R. M. "Incompressible porous media models by use of the theory of mixtures". In: *Int. J. Eng. Sci.* 18.9 (1980), pp. 1129–1148. DOI: 10.1016/0020-7225(80)90114-7.

[148] H. G. Matthies and A. Keese. "Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations". In: *Comput. Methods Appl. Mech. Engrg.* 194.12-16 (2005), pp. 1295–1331.

[149] V. Maz'ya and J. Rossman. *Elliptic equations in polyhedral domains*. Vol. 162. AMS: Mathematical Surveys and Monographs, 2010.

[150] A. Mehrabian and Y. N. Abousleiman. "Gassmann equation and the constitutive relations for multiple-porosity and multiple-permeability poroelasticity with applications to oil and gas shale". In: *Int. J. Numer. Anal. Meth. Geomech.* 39 (2015), pp. 1547–1569. DOI: 10.1002/nag.2399.

[151] S. E. Minkoff, C. M. Stone, S. Bryant, M. Peszynsak, and M. F. Wheeler. "Coupled fluid flow and geomechanical deformation modeling". In: *J. Pet. Sci. Eng.* 38 (2003), pp. 37–56. DOI: 10.1016/S0920-4105(03)00021-4.

[152] G. J. Minty. "On a "monotonicity" method for the solution of nonlinear equations in Banach spaces". In: *Proceedings of the National Academy of Sciences of the United States of America* 50.6 (1963), pp. 1038–1041.

[153] D. Moos, P. Peska, T. Finkbeiner, and M.D. Zoback. "Comprehensive Wellbore Stability Analysis Utilizing Quantitative Risk Assessment". In: *J. Pet. Sci. Eng.* 38.3 (2003), pp. 97–109. DOI: 10.1016/S0920-4105(03)00024-X.

[154] M. A. Murad and F. D. Loula. "Improved accuracy in finite element analysis of Biot's consolidation problem". In: *Comput. Methods Appl. Mech. Engrg.* 93.3 (1992), pp. 359–382.

[155] M. A. Murad and F. D. Loula. "On stability and convergence of finite element approximations of Biot's consolidation problem". In: *Interat. J. Numer. Methods Engrg.* 37.4 (1994). DOI: 10.1002/nme.1620370407.

[156] A. Naumovich. "On finite volume discretization of the three-dimensional Biot poroelasticity system in multilayer domains". In: *Comput. Meth. App. Math.* 6.3 (2006), pp. 306–325.

[157] J. Nečas. *Introduction to the theory of nonlinear elliptic equations*. Chichester: A Wiley-Interscience Publication. John Wiley & Sons Ltd., 1986. Reprint of the 1983 edition.

[158] S. Nicaise, K. Witowski, and B. I. Wohlmuth. "An a posteriori error estimator for the Lamé equation based on H(div)-conforming stress approximations". In: *IMA J. Numer. Anal.* 28.2 (2008), pp. 331–353. DOI: 10.1093/imanum/drm008.

[159] S. Nishimura, K. Shimada, and H. Fujii. "Consolidation inverse analysis considering spacial variability and non-linearity of soil parameters". In: *Soils and Foundations* 42.3 (2002), pp. 45–61. DOI: 10.3208/sandf.42.3_45.

[160] J. M. Nordbotten. "Cell-centered finite volume discretizations for deformable porous media". In: *Int. J. Numer. Meth. Eng.* 100.6 (2014), pp. 399–418. DOI: 10.1002/nme.4734.

[161] S. Ohnimus, E. Stein, and E. Walhorn. "Local error estimates of FEM for displacements and stresses in linear elasticity by solving local Neumann problems". In: *Int. J. Numer. Meth. Engng.* 52 (2001), pp. 727–746.

[162] C.M. Oldenburg. "The risk of induced seismicity: Is cap-rock on shaky ground ?" In: *Greenh. Gases: Sci. Technol.* 2.4 (2012), pp. 217–218. DOI: 10.1002/ghg.1299.

[163] C. Ortner and E. Süli. "Discontinuous Galerkin Finite Element Approximation of Nonlinear Second-Order Elliptic and Hyperbolic Systems". In: *SIAM J. on Numer. Anal.* 45.4 (2007), pp. 1370–1397. DOI: 10.1137/06067119X.

[164] P. J. Phillips. "Finite element methods in linear poroelasticity: Theoretical and computational results". PhD thesis. University of Texas at Austin, Dec. 2005.

[165] P. J. Phillips and M. F. Wheeler. "A coupling of mixed and continuous Galerkin finite element methods for poroelasticity I: the continuous in time case". In: *Comput. Geosci.* 11 (2007), pp. 131–144.

[166] P. J. Phillips and M. F. Wheeler. "A coupling of mixed and continuous Galerkin finite element methods for poroelasticity II: the discrete-in-time case". In: *Comput. Geosci.* 11 (2007), pp. 145–158.

[167] P. J. Phillips and M. F. Wheeler. "A coupling of mixed and discontinuous Galerkin finite-element methods for poroelasticity". In: *Comput. Geosci.* 12 (2008), pp. 417–435.

[168] P. J. Phillips and M. F. Wheeler. "Overcoming the problem of locking in linear elasticity and poroelasticity: An heuristic approach". In: *Comput. Geosci.* 13 (2009), pp. 5–12. DOI: 10.1007/s10596-008-9114-x.

[169] M. Pitteri and G. Zanotto. *Continuum models for phase transitions and twinning in crystals*. Chapman & Hall/CRC, 2003.

[170] W. Prager and J. L. Synge. "Approximations in elasticity based on the concept of function space". In: *Quart. Appl. Math.* 5 (1947), pp. 241–269.

[171] M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio. "Uncertainty quantification in reacting-flow simulations through non-intrusive spectral projection". In: *Combustion and Flame* 132.3 (2003), pp. 545–555.

[172] S. I. Repin. *A posteriori estimates for partial differential equations*. Vol. 4. Radon Series on Computational and Applied Mathematics. Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[173] J. R. Rice and M. P. Cleary. "Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents". In: *Rev. Geophys.* 14.2 (1976), pp. 227–241. DOI: 10.1029/RG014i002p00227.

[174] R. Riedlbeck, D. A. Di Pietro, A. Ern, S. Granet, and K. Kazymyrenko. "Stress and flux reconstruction in Biot's poro-elasticity problem with application to a posteriori analysis". In: *Comput. and Math. with Appl.* 73.7 (2017), pp. 1593–1610.

[175] R. Riedlbeck, D. A. Di Pietro, and E. Ern. "Equilibrated stress reconstructions for linear elasticity problems with application to a posteriori error analysis". In: *Finite Volumes for Complex Applications VIII – Methods and Theoretical Aspects*. 2017, pp. 293–301.

[176] C. Rodrigo, F.J. Gaspar, X. Hu, and L.T. Zikatanov. "Stability and monotonicity for some discretizations of the Biot's consolidation model". In: *Comput. Methods Appl. Mech. and Engrg.* 298 (2016), pp. 183–204. DOI: 10.1016/j.cma.2015.09.019.

[177] D. Sandri. "Sur l'approximation numérique des écoulements quasi-Newtoniens dont la viscosité suit la loi puissance ou la loi de Carreau". In: *Math. Modelling and Num. Anal."* 27.2 (1993), pp. 131–155.

[178] O.D. Schirra. "New Korn-type inequalities and regularity of solutions to linear elliptic systems and anisotropic variational problems involving the trace-free part of the symmetric gradient". In: *Calc. Var.* 43 (2012), pp. 147–172. DOI: 10.1007/s00526-011-0406-y.

[179] D. Schötzau and Schwab C. "Time Discretization of Parabolic Problems by the HP-Version of the Discontinuous Galerkin Finite Element Method". In: *SIAM J. Numer. Anal.* 38.3 (2000), pp. 837–875. DOI: 10.1137/S0036142999352394.

[180] C. Schwab and R. A. Todor. "Karhunen-Loève Approximation of Random Fields by Generalized Fast Multipole Methods". In: *J. Comput. Phys.* 217.1 (2006), pp. 100–122. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2006.01.048.

[181] R. E. Showalter. "Diffusion in poro-elastic media". In: *J. Math. Anal. Appl.* 251 (2000), pp. 310–340. DOI: 10.1006/jmaa.2000.7048.

[182] R. Shukla, P. Ranjith, S. Choi, and A. Haque. "Study of Caprock Integrity in Geosequestration of Carbon Dioxide". In: *Int. J. Geomech.* 11.4 (2010), pp. 294–301. DOI: 10.1061/(ASCE)GM.1943-5622.0000015.

[183] I. Smears. "Robust and efficient preconditioners for the discontinuous Galerkin time-stepping method". In: *IMA J. Numer. Anal.* 37.4 (2017), pp. 1961–1985. DOI: 10.1093/imanum/drw050.

[184] S.A. Smolyak. "Quadrature and interpolation formulas for tensor products of certain classes of functions". In: *Dokl. Akad. Nauk SSSR* 4.240-243 (1963), p. 123.

[185] S.-C. Soon, B. Cockburn, and H. K. Stolarski. "A hybridizable discontinuous Galerkin method for linear elasticity". In: *Internat. J. Numer. Methods Engrg.* 80.8 (2009), pp. 1058–1092.

[186] D. W. Spring, S. E. Leon, and G. H. Paulino. "Unstructured polygonal meshes with adaptive refinement for the numerical simulation of dynamic cohesive fracture". In: *Int. J. Fracture* 189.1 (2014), pp. 33–57. ISSN: 1573-2673. DOI: 10.1007/s10704-014-9961-5.

[187] C.C. Swan, R.S. Lakes, R.A. Brand, and K.J. Stewart. "Micromechanically based poroelastic modeling of fluid flow in haversian bone". In: *J. Biomech. Eng.* 125.1 (2003), pp. 25–37.

[188] C. Talischi, G. H. Paulino, A. Pereira, and I. F. M. Menezes. "PolyMesher: a general-purpose mesh generator for polygonal elements written in Matlab". In: *Structural and Multidisciplinary Optimization* 45.3 (2012), pp. 309–328. DOI: 10.1007/s00158-011-0706-z.

[189] K. Terzaghi. "Die berechnung der durchlassigkeitszier des tones aus dem verlauf der hydrodynemischen spannungserscheinungen". In: *Sitz. Akad. Wissen.* 132.2a (1923), pp. 105–124.

[190] K. Terzaghi. *Theoretical soil mechanics*. New York: Wiley, 1943.

[191] L. R. G. Treolar. *The Physics of Rubber Elasticity*. Oxford: Clarendon Press, 1975.

[192] J. Troyen, O. Le Maître, M. Ndjinga, and A. Ern. "Intrusive projection methods with upwinding for uncertain nonlinear hyperbolic systems". In: *J. Comput. Phys.* 229.18 (2010), pp. 6485–6511. DOI: 10.1016/j.cam.2010.05.043.

[193] B. Tully and Y. Ventikos. "Cerebral water transport using multiple-network poroelastic theory: application to normal pressure hydrocephalus". In: *J. Fluid Mech.* 667 (2011), pp. 188–215. DOI: 10.1017/S0022112010004428.

[194] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley-Teubner, Chichester, 1996.

[195] R. Verfürth. "A review of a posteriori error estimation techniques for elasticity problems". In: *Comput. Meth. Appl. Mech. Engrg.* 176 (1999), pp. 419–440.

[196] M. Vogelius. "An analysis of the p-version of the finite element method for nearly incompressible materials. Uniformly valid, optimal error estimates". In: *Numer. Math.* 41 (1983), pp. 39–53.

[197] C. Wang, J. Wang, R. Wang, and R. Zhang. "A Locking-free Weak Galerkin Finite Element Method for Elasticity Problems in the Primal Formulation". In: *J. Comput. Appl. Math.* 307.C (2016), pp. 346–366. ISSN: 0377-0427. DOI: 10.1016/j.cam.2015.12.015.

[198] H. F. Wang. *Theory of linear poroelasticity with applications to geomechanics and hydrogeology*. Princeton University Press, 2000.

[199] J. Wang and X. Ye. "A weak Galerkin mixed finite element method for second order elliptic problems". In: *Math. Comp.* 83.289 (2014), pp. 2101–2126.

[200] M. F. Wheeler, G. Xue, and I. Yotov. "Coupling multipoint flux mixed finite element methods with continuous Galerkin methods for poroelasticity". In: *Comput. Geosci.* 18 (2014), pp. 57–75.

[201] N. Wiener. "The Homogeneous Chaos". In: *Amer. J. Math.* 60 (1938), pp. 897–936.

[202] P. Wriggers, W. T. Rust, and B. D. Reddy. "A virtual element method for contact". In: *Computational Mechanics* 58.6 (2016), pp. 1039–1050. ISSN: 1432-0924. DOI: 10.1007/s00466-016-1331-x.

[203] D. Xiu and G.E. Karniadakis. "The Wiener-Askey Polynomial Chaos for Stochastic Differential Equations". In: *SIAM J. Sci. Comp.* 24.2 (2002), pp. 619–644.

[204] S.-Y. Yi. "Convergence analysis of a new mixed finite element method for Biot's consolidation model". In: *Numer. Methods Partial Differential Equations* 30.4 (2014), pp. 1189–1210. DOI: 10.1002/num.21865.

[205] A. Ženíšek. "The existence and uniqueness theorem in Biot's consolidation theory". In: *Aplikace Matematiky* 29 (1984), pp. 194–211.

[206] R. W. Zimmerman. "Coupling in poroelasticity and thermoelasticity". In: *Int. J. Rock Mech. Min. Sci.* 37.1–2 (Jan. 2000), pp. 79–87. ISSN: 0148-9062.

Résumé

Dans cette thèse, on s'intéresse à de nouveaux schémas de discrétisation afin de résoudre les équations couplées de la poroélasticité et nous présentons des résultats analytiques et numériques concernant des problèmes issus de la poromécanique. Nous proposons de résoudre ces problèmes en utilisant les méthodes Hybrid High-Order (HHO), une nouvelle classe de méthodes de discrétisation polyédriques d'ordre arbitraire. Cette thèse a été conjointement financé par le Bureau de Recherches Géologiques et Minières (BRGM) et le LabEx NUMEV. Le couplage entre l'écoulement souterrain et la déformation géomécanique est un sujet de recherche crucial pour les deux institutions de cofinancement.

Mots clés: méthodes Hybrides d'Ordre Élevé, méthodes Galerkin discontinues, maillages polyédriques, ordre d'approximation arbitraire, poromécanique, problème de Biot, élasticité non-linéaire, écoulement de Darcy, couplage hydraulique, formulation primale, analyse fonctionnelle discrète, quantification d'incertitude.

Abstract

In this manuscript we focus on novel discretization schemes for solving the coupled equations of poroelasticity and we present analytical and numerical results for poromechanics problems relevant to geoscience applications. We propose to solve these problems using Hybrid High-Order (HHO) methods, a new class of nonconforming high-order methods supporting general polyhedral meshes. This Ph.D. thesis was conjointly founded by the Bureau de recherches géologiques et minières (BRGM) and LabEx NUMEV. The coupling between subsurface flow and geomechanical deformation is a crucial research topic for both cofunding institutions.

Keywords: Hybrid High-Order methods, discontinuous Galerkin methods, arbitrary order, polyhedral meshes, poromechanics, Biot problem, nonlinear elasticity, Darcy flow, hydromechanical coupling, primal formulation, discrete functional analysis, uncertainty quantification.