A posteriori error estimates for variational inequalities : application to a two-phase flow in porous media

Jad Dabaghi

THÈSE

PRÉSENTÉE À

SORBONNE UNIVERSITÉ

ÉCOLE DOCTORALE : Sciences Mathématiques de Paris Centre (ED 386)

Par **Jad Dabaghi**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : Mathématiques Appliquées

Estimations d'erreur a posteriori pour des inégalités variationnelles : application à un écoulement diphasique en milieu poreux

---

Directeur de thèse : Martin Vohralík

Co-directeur de thèse : Vincent Martin

Soutenue le : 3 Juin 2019

Devant la commission d'examen formée de :

M. Faker Ben Belgacem	UTC Compiègne	Examinateur
M. Franz Chouly	Université de Bourgogne	Rapporteur
M. Patrick Hild	Université Toulouse III	Président du jury
M. Vincent Martin	UTC Compiègne	Co-directeur de Thèse
M. Pascal Omnes	CEA Saclay	Examinateur
Mme. Béatrice Rivière	Université Rice Houston	Rapporteur
Mme. Jean Roberts	INRIA Paris	Examinatrice
M. Martin Vohralík	INRIA Paris	Directeur de Thèse

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

informatics mathematics

A THESIS

PRESENTED AT

SORBONNE UNIVERSITY

DOCTORAL SCHOOL: Mathematical Sciences of Central Paris (ED 386)

by **Jad Dabaghi**

TO OBTAIN THE DEGREE OF
DOCTOR OF PHILOSOPHY
SPECIALITY: Applied Mathematics

A posteriori error estimates for variational inequalities: application to a two-phase flow in porous media

---

Thesis advisor: Martin Vohralík
Thesis co-advisor: Vincent Martin

Defended on: June 3rd 2019
In front of the examination committee consisting of:

M. Faker Ben Belgacem	UTC Compiègne	Examiner
M. Franz Chouly	University of Bourgogne	Reviewer
M. Patrick Hild	University of Toulouse III	Chairman
M. Vincent Martin	UTC Compiègne	Thesis co-advisor
M. Pascal Omnes	CEA Saclay	Examiner
Mme. Béatrice Rivière	Rice University Houston	Reviewer
Mme. Jean Roberts	INRIA Paris	Examiner
M. Martin Vohralík	INRIA Paris	Thesis advisor

A ma mère, mon père,

mon frère et ma soeur.

A toute ma famille.

Résumé

Dans cette thèse, nous considérons des inégalités variationnelles qui s'interprètent comme des équations aux dérivées partielles avec contraintes de complémentarité. Nous construisons des estimateurs d'erreur a posteriori pour des discrétisations utilisant la méthode des éléments finis et volumes finis, et des linéarisations inexactes faisant appel aux méthodes de Newton semi-lisse et à des solveurs algébriques quelconques. Nous considérons tout d'abord un problème modèle de contact entre deux membranes, puis une inégalité variationnelle parabolique et enfin un écoulement diphasique compositionnel avec changement de phases comme application industrielle.

Dans le premier chapitre, nous considérons un problème stationnaire de contact entre deux membranes. Ce problème s'inscrit dans la large gamme des inégalités variationnelles de première espèce. Nous discrétisons notre modèle par la méthode des éléments finis conformes d'ordre $p \geq 1$ et nous proposons deux formulations discrètes équivalentes : la première sous la forme d'une inégalité variationnelle et la seconde sous la forme d'un problème de type point-selle. Nous introduisons la différentiabilité au sens de Clarke pour traiter les non linéarités non différentiables. Cela permet d'utiliser des algorithmes de linéarisation de type Newton semi-lisse. Ensuite, un solveur itératif algébrique quelconque est utilisé pour le système linéaire obtenu. En utilisant la méthodologie de la reconstruction des flux équilibrés dans l'espace $\mathbf{H}(\mathrm{div}, \Omega)$, nous obtenons une borne supérieure de l'erreur totale dans la semi-norme d'énergie sur l'espace $H_0^1(\Omega)$. Cette borne est entièrement calculable à chaque pas du solveur de linéarisation semi-lisse et à chaque pas du solveur d'algèbre linéaire. Notre estimation d'erreur distingue en particulier les trois composantes de l'erreur, à savoir l'erreur de discrétisation (éléments finis), l'erreur de linéarisation (algorithme de Newton semi-lisse) et l'erreur d'algèbre linéaire (algorithme GMRES). Nous formulons ensuite des critères d'arrêts adaptatifs pour chaque solveur utilisé dans le but de réduire le nombre d'itérations. Nous prouvons également l'efficacité locale de nos estimateurs dans le contexte semi-lisse inexact modulo un terme de contact qui s'avère négligeable. Nos essais numériques illustrent la précision de nos estimations et le gain en terme de nombre d'itérations et témoignent de la performance de notre méthode adaptative semi-lisse inexacte.

Dans le second chapitre, nous nous intéressons à construire des estimations d'erreur a posteriori pour une inégalité variationnelle parabolique comme extension du premier chapitre au cas instationnaire. Nous discrétisons notre modèle en utilisant la méthode des éléments finis conformes d'ordre $p \geq 1$ en espace et le schéma d'Euler rétrograde en temps. Pour traiter les non linéarités, nous utilisons à nouveau des algorithmes de linéarisation de type Newton semi-lisse et nous employons également un solveur itératif algébrique quelconque pour le système linéaire obtenu. En utilisant la méthodologie de la reconstruction des flux équilibrés dans l'espace $\mathbf{H}(\mathrm{div}, \Omega)$, nous obtenons, quand $p = 1$, et à convergence

du solveur de linéarisation semi-lisse et d'algèbre linéaire, une borne supérieure de l'erreur totale dans la norme d'énergie sur l'espace $L^2(0,T;H_0^1(\Omega))$. De plus, nous estimons dans ce cas du mieux possible l'erreur en dérivée temporelle dans la norme d'énergie $L^2(0,T;H^{-1}(\Omega))$. Dans le cas $p \geq 1$, et à un pas quelconque des solveurs linéaires et non linéaires, nous présentons une estimation d'erreur a posteriori dans la norme d'énergie $L^2(0,T;H_0^1(\Omega))$. Nous distinguons dans ce cas les composantes de l'erreur totale, à savoir l'erreur de discrétisation, l'erreur de linéarisation et l'erreur d'algèbre linéaire. Cela permet en particulier de formuler des critères d'arrêts adaptatifs dans le but de réduire le nombre d'itérations.

Dans le troisième chapitre, nous abordons un problème d'écoulement diphasique compositionnel en milieu poreux liquide-gaz (eau-hydrogène) avec échange d'hydrogène entre les deux phases. Il s'agit d'un système d'équations aux dérivées partielles non linéaires avec contraintes de complémentarité non linéaires qui se réécrit sous la forme d'une inégalité variationnelle évolutive non linéaire. Nous employons la méthode des volumes finis centrés par maille pour la discrétisation en espace, et le schéma d'Euler rétrograde pour la discrétisation en temps. La discrétisation numérique engendre à chaque pas de temps un système non linéaire et non différentiable qui s'interprète comme une inégalité variationnelle. Comme dans les chapitres précédents, nous approchons la solution du système non linéaire obtenu par un algorithme de Newton semi-lisse inexact. Nous reconstruisons les flux équilibrés dans l'espace $\mathbf{H}(\mathrm{div},\Omega)$ en utilisant les flux numériques aux interfaces issus de la méthode des volumes finis, et les pressions de phases et la fraction molaire dans l'espace $H^1(\Omega)$ afin d'obtenir une borne supérieure de l'erreur entièrement calculable à chaque pas de temps, de Newton semi-lisse et du solveur d'algèbre linéaire. La mesure de l'erreur entre la solution exacte et la solution approchée est constituée ici de la norme duale du résidu complétée par un résidu défini sur les contraintes de complémentarité et par des termes exprimant la non conformité de la discrétisation en espace. Nous obtenons finalement des estimations d'erreur a posteriori distinguant les différentes composantes d'erreur, à savoir l'erreur de discrétisation, l'erreur de linéarisation et l'erreur d'algèbre linéaire. Ainsi, nous formulons des critères d'arrêt adaptatifs pour nos solveurs afin de réduire le nombre d'itérations. Les essais numériques réalisés corroborent les bénéfices de notre méthode adaptative semi-lisse inexacte.

Mots-clés

inégalité variationnelle elliptique - inégalité variationnelle parabolique - condition de complémentarité - problème de contact - écoulement diphasique - changement de phase - stockage des déchets radioactifs - méthode des éléments finis - méthode des éléments finis mixtes - méthode des volumes finis - non conformité - algorithme de Newton-min - algorithme de Newton–Fischer–Burmeister - semi lissité - estimation d'erreur a posteriori - flux équilibré - composantes d'erreur - critère d'arrêt - efficacité

Abstract

In this thesis, we consider variational inequalities in the form of partial differential equations with complementarity constraints. We construct a posteriori error estimates for discretizations using the finite element method and the finite volume method, for inexact linearizations employing any semismooth Newton solver and any iterative linear algebraic solver. First, we consider the model problem of contact between two membranes, next we consider its extension into a parabolic variational inequality, and to finish we treat a two-phase compositional flow with phase transition as an industrial application.

In the first chapter, we consider the stationnary problem of contact between two membranes. This problem belongs to the wide range of variational inequalities of the first kind. Our discretization is based on the finite element method with polynomials of order $p \geq 1$, and we propose two discrete equivalent formulations: the first one as a variational inequality, and the second one as a saddle-point-type problem. We employ the Clarke differential so as to treat the nondifferentiable nonlinearities. It enables us to use semismooth Newton algorithms. Next, any iterative linear algebraic solver is used for the linear system stemming from the discretization. Employing the methodology of equilibrated flux reconstructions in the space $\mathbf{H}(\mathrm{div}, \Omega)$, we get an upper bound on the total error in the energy norm $H_0^1(\Omega)$. This bound is fully computable at each semismooth Newton step and at each linear algebraic step. Our estimation distinguishes in particular the three components of the error, namely the discretization error (finite elements), the linearization error (semismooth Newton method), and the algebraic error (GMRES algorithm). We then formulate adaptive stopping criteria for our solvers to ultimately reduce the number of iterations. We also prove, in the inexact semismooth context, the local efficiency property of our estimators, up to a contact term that appears negligeable in numerics. Our numerical experiments illustrate the accuracy of our estimates and the reduction of the number of necessary iterations. They also show the performance of our adaptive inexacte semismooth Newton method.

In the second chapter, we are interested in deriving a posteriori error estimates for a parabolic variational inequality and we consider the extension of the model of the first chapter to the unsteady case. We discretize our model using the finite element method of order $p \geq 1$ in space and the backward Euler scheme in time. To treat the nonlinearities, we use again semismooth Newton algorithms, and we also employ an iterative algebraic solver for the linear system stemming from the discretization. Using the methodology of equilibrated flux reconstructions in the space $\mathbf{H}(\mathrm{div}, \Omega)$, we obtain, when $p = 1$ and at convergence of the semismooth solver and the algebraic solver, an upper bound for the total error in the energy norm $L^2(0, T; H_0^1(\Omega))$. Furthermore, we estimate in this case the time derivative error in a norm close to the energy norm $L^2(0, T; H^{-1}(\Omega))$. In the case $p \geq 1$, we present an a posteriori error estimate valid at each semismooth Newton step and at each linear algebraic step in the norm $L^2(0, T; H_0^1(\Omega))$. We distinguish in

this case the components of the total error, namely the discretization error, the linearization error, and the algebraic error. In particular, it enables us to devise adaptive stopping criteria for our solvers which reduces the number of iterations.

In the third chapter, we consider a two-phase liquid-gas compositional (water-hydrogen) flow with hydrogen mass exchange between the phases in porous media. It is a nonlinear system of partial differential equations with nonlinear complementarity constraints which can be interpreted as an evolutive in-time nonlinear variational inequality. We employ the cell-centered finite volume method for the space discretization and the backward Euler scheme for the time discretization. The discretization generates at each time step a nondifferentiable and nonlinear system. As for the previous chapters, we approximate the solution of the nonlinear system stemming from the discretization by an inexact semismooth Newton algorithm. The equilibrated flux reconstructions are obtained in the space $\mathbf{H}(\mathrm{div}, \Omega)$ using the numerical fluxes at the interfaces stemming from the finite volume method, and we reconstruct the phase pressures and the molar fraction in the space $H^1(\Omega)$ in order to obtain fully computable upper bound at each time step, semismooth Newton step, and algebraic step. The error measure between the exact solution and the approximate solution is made of the dual norm of the residual supplemented by a residual defined of the complementarity constraints and nonconforming space terms. We finally obtain a posteriori error estimate distinguishing the different error components, namely the discretization error, the linearization error, and the algebraic error. Thus, we formulate adaptive stopping criteria for our solvers in order to reduce the number of iterations. The numerical experiments confirms the benefits of our adaptive inexact semismooth approach.

Keywords: elliptic variational inequality - parabolic variational inequality - complementarity condition - contact problem - - two-phase flow - phase transition - storage of radioactive waste - finite elements - mixed finite elements - finite volumes - nonconformity - Newton-min algorithm - Newton–Fischer–Burmeister algorithm - semismoothness - a posteriori error estimate - equilibrated flux - error components - stopping criteria - efficiency

Remerciements

Nombreuses sont les personnes que j'aimerais remercier. Tout d'abord je tiens à adresser mes plus sincères remerciements à mon directeur de thèse, Monsieur Martin Vohralík, directeur de recherche et chef de l'équipe SERENA à l'INRIA Paris pour son soutien et sa gentillesse tout le long de la thèse. Je te suis très reconnaissant, Martin, de m'avoir donné la chance de faire une thèse et de m'avoir fait confiance. Je remercie également mon co-directeur de thèse, Monsieur Vincent Martin, maître de conférences à l'UTC de Compiègne et collaborateur externe de l'équipe SERENA, pour sa gentillesse, son aide et son écoute lors des moments difficiles. Vous avez tous les deux insisté sur l'importance de la rigueur et de la précision dans la préparation de mes travaux et je vous suis très reconnaissant. J'espère être capable à présent de prendre mon envol.

Ma gratitude va également à Ibtihel Ben Gharbia, ingénieure de recherche à l'IFP Energies nouvelles qui a accepté d'intevenir dans ma thèse dans le cadre de mon troisième chapitre. Du fond du cœur merci beaucoup Ibtihel pour ton aide, ton soutien et ta compréhension. Tes qualités humaines ne font pas débat et j'espère de tout cœur continuer à travailler avec toi par la suite. Ma reconnaissance va également à Soleiman Yousef, ingénieur de recherche à l'IFP Energies nouvelles qui a accepté de m'aider et répondre à beaucoup de mes questions quand je travaillais le dernier volet de ma thèse. J'ai bien conscience que je t'ai beaucoup sollicité Soleiman et je te remercie d'avoir trouvé le temps de répondre à mes questions avec beaucoup de clarté et de pédagogie.

Je remercie Béatrice Rivière, professeur à l'Université de Rice à Houston ainsi que Franz Chouly, professeur à l'Université de Bourgogne, d'avoir accepté d'être rapporteur de ma thèse. Merci également à Jean Roberts, directrice de recherche de l'INRIA Paris, Faker Ben Belgacem, professeur à l'UTC de Compiègne, Patrick Hild, professeur à l'Université Paul-Sabatier de Toulouse et Pascal Omnes, directeur de recherche au CEA Saclay qui ont accepté d'être des membres du jury.

Bien évidemment, je n'oublierai pas mes camarades doctorants et post-doctorants qui ont partagé cette aventure à mes côtés : Patrik Daniel, Fabien Wahl, Karol Cascavita, Amina Benaceur, Sarah Ali Hassan, Fatma Cheikh, Ani Miraçi, Nicolas Pignet, Mohammad Zakerzadeh, Matteo Cicuttin et Simon Legrand avec une attention toute particulière pour Patrik, Fabien, Karol, Amina, Sarah, Fatma, et Ani. J'ai eu de la chance de bénéficier d'un environnement scientifique à la fois très riche et convivial durant ces trois années et je tiens à exprimer toute ma sympathie à Michel Kern, Francois Clément, Pierre Weis, Géraldine Pichot, Alexandre Ern, Caroline Japhet, Jean Roberts et Jérôme Jaffré. Je n'oublierai pas la gentillesse de Jérôme et Jean qui m'ont accueilli avec beaucoup de chaleur lors de mon stage de Master 2 à l'INRIA Rocquencourt (ex projet POMDAPI 2). Michel, je te remercie pour tes conseils, ton éclairage, les longues conversations que nous avons partagées

ainsi que les excellentes références bibliographiques que tu m'as données. Merci à vous, Alexandre, pour votre aide lorsque j'en ai eu besoin, les conversations que nous avons partagées et votre attention.

Je remercie également Jan Papež, post-doctorant à l'INRIA Paris pour son aide précieuse et nos conversations très intéressantes sur les techniques d'implémentation sous Matlab. Ma gratitude va également à Monsieur Jean-Charles Gilbert, directeur de recherche à l'INRIA Paris, pour son aide et sa disponibilité lorsque j'avais des questions sur l'optimisation. J'ai également été très marqué par sa vision du monde de la recherche. Je souhaite comme lui que les scientifiques pourront un jour se réunir autour d'une table pour discuter des problématiques de manière désintéressée avec pour seul but la science, son essor et son développement. Merci également à Adel Blouza, maître de conférences à l'Université de Rouen qui a accepté de me recevoir à deux reprises à l'UPMC lorsque j'avais des questions sur le premier chapitre de ma thèse.

Une pensée toute particulière pour Nathalie Bonte qui a été l'assistante de notre équipe lors de mon arrivée à l'INRIA Rocquencourt, puis Cindy Croussouard qui a pris la relève après le départ de Nathalie, puis Kevin Bonny et Virgine Collette qui ont repris le relais par la suite et enfin Derya Gok qui a repris le flambeau suite au départ de Virginie. Je remercie beaucoup également Martine Girardot qui s'est toujours montrée très disponible lorsque j'avais des difficultés avec les tâches administratives. Vous êtes des personnes formidables pleines de gentillesse, de joie et de bonne humeur. Je ne vous oublierai pas.

Sur un plan plus personnel, j'espère maintenir l'amitié et les échanges scientifiques de qualité avec l'ensemble des membres de l'équipe SERENA. Je serai très honoré de pouvoir poursuivre quelques travaux de recherche avec mes encadrants de thèse et les membres de l'équipe.

Enfin, je dirai que tout ce travail n'aurait jamais vu le jour sans le soutien sans faille, indéfectible et inconditionnel de ma famille : ma mère, mon père, mon frère et ma sœur. Les mots ne suffiraient pas à exprimer toute la gratitude que j'ai pour toi maman. Tu as comme toujours été compréhensive, à l'écoute, généreuse et d'une gentillesse sans limite. Merci infiniment.

Contents

List of Figures

List of Tables

Prologue

Nombreuses furent les questions que je me posais avant d'intégrer l'équipe SERENA (Simulation for the Environment: Reliable and Efficient Numerical Algorithms) à l'INRIA Paris. Le projet SERENA s'est fixé pour objectif la construction et l'analyse d'outils de simulation pour des modèles mathématiques basés sur des équations aux dérivées partielles dans l'optique de traiter les problèmes liés à l'environnement et aux énergies. Parmi les multiples interrogations que j'avais, je citerais : Qu'est ce que l'INRIA m'apporterait pour vivre et satisfaire ma passion d'acquérir de la connaissance ? Comment assouvir mon désir d'apprendre dans un domaine aussi transversal que vaste, pour ne pas dire infini, qu'est les mathématiques appliquées ? Pourquoi les équations aux dérivées partielles (EDP) sont-elles si importantes et communément utilisées dans le domaine de la modélisation en général, et de celui des phénomènes physiques en particulier ? Quelles sont les limites de ces EDP dans la modélisation de ces phénomènes physiques et dans la modélisation numérique permettant de les simuler ? Quelle place occupent ces équations dans les défis majeurs du XXIe siècle et quel est le rôle du numérique, si répandu de nos jours, dans le traitement de ces équations ? Il est clair que c'est aussi compliqué et délicat de répondre à ces questions que de se les poser.
En fait, cette période passionnante, que fut ces 3 années passées à l'INRIA, m'a permis d'avoir quelques éléments de réponse. L'INRIA est certainement parmi les meilleurs endroits pour accéder à la connaissance et à la culture multidisciplinaire des mathématiques appliquées. Certes, mon parcours académique a été un précurseur favorable pour me situer, me positionner et être conscient des perspectives prometteuses que me présente cette aventure riche d'opportunités et de découvertes. En le vivant de l'intérieur pendant cette durée de thèse, je réponds affirmativement et sans hésitation que l'INRIA m'a apporté dans l'ensemble un excellent environnement cohérent pour réaliser ma recherche, alimenter ma culture mathématique et numérique et sans doute améliorer mes connaissances dans ce domaine si générique.

Pour finir, je propose de méditer longuement sur la pensée de Platon qui a toujours clamé que les mathématiques permettent d'accéder au monde des idées, et pour preuve, il a fait graver au portail de son Académie : "Que nul n'entre ici s'il n'est géomètre".

Introduction

I Préambule

Depuis toujours, l'homme s'est intéressé à comprendre primitivement et intuitivement par observation les nombreux phénomènes physiques l'environnant. Citons par exemple, à la période du Paléolithique, la découverte du feu en frottant deux pierres, le flottement du bois sur la surface de l'écoulement d'une rivière, le ruissellement et l'infiltration de l'eau dans le sol, le mouvement apparent des astres dans le ciel. Il a fallu attendre plusieurs millénaires pour que des scientifiques apportent un éclairage, une vision pérenne et structurée pour tenter de comprendre et d'analyser ces observations.

Au XXIe siècle, nous pouvons mentionner à titre d'exemple les défis relatifs au réchauffement climatique, à l'exploitation des ressources en sous sol comme le forage d'exploration pétrolière et gazière (voir Figure 1 gauche), au stockage des déchets radioactifs, à l'impact de l'activité humaine sur la pollution de l'eau, du sol et de l'air avec pour corollaire immédiat l'altération des nappes phréatiques (voir Figure 1 droite) et enfin au traitement de l'image pour les applications médicales.

Après cette brève revue sur l'impact et l'importance des mathématiques dans la compréhension de ce qui nous entoure, il devient plus naturel de parler des équations aux dérivées partielles (EDP). Les EDP ont pour vocation de décrire mathématiquement le comportement de nombreux phénomènes physiques mentionnés ci-dessus. Il existe plusieurs familles d'EDP. A titre d'exemple citons : l'équation de la chaleur, l'équation de la corde vibrante, l'équation des plaques, les équations de Navier–Stokes ou les équations de Maxwell. Malheureusement, les résoudre analytiquement est limité à des cas académiques. Dans la plupart des cas, nous sommes amenés à

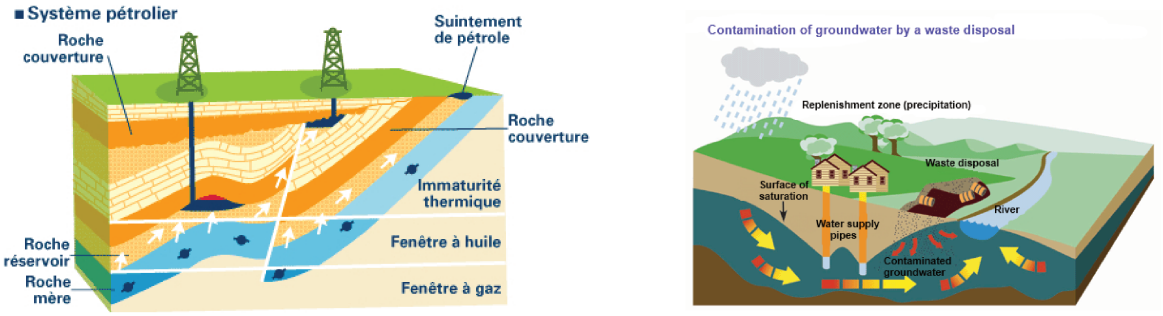

Figure 1: Extraction du pétrole à travers les roches (gauche), pollution de l'eau et contamination souterraine (droite). http://www.ec.gc.ca/eau-water.

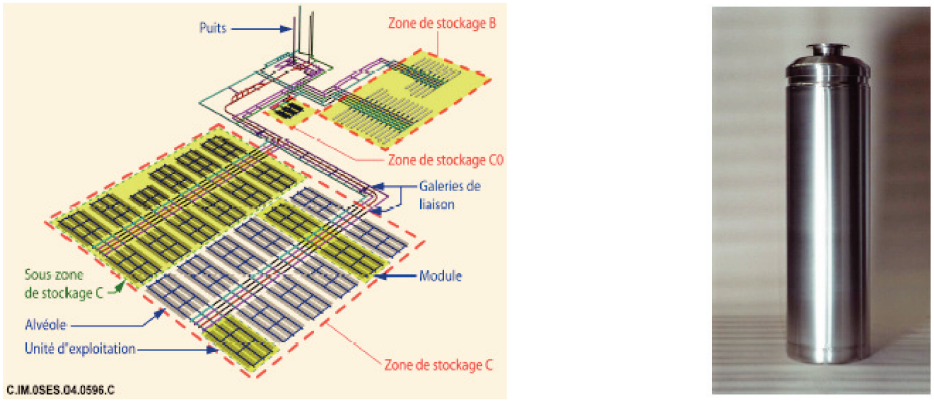

Figure 2: Solution de stockage des déchets radioactifs proposée par l'Andra (Agence nationale pour la gestion des déchets radioactifs). https://www.andra.fr/documents-et-ressources.

les approximer pour trouver des solutions numériques ayant un comportement aussi proche que possible du comportement des solutions analytiques inconnues et par conséquent des phénomènes modélisés observés. La tâche est évidemment ardue car il faut dans un premier temps cerner la phénoménologie du processus à modéliser tant par les aspects multi-échelles en temps qu'en espace, poser un modèle mathématique et le valider avec son étude théorique, choisir des données initiales à l'aide d'une connaissance conjecturale du comportement de la solution (par analyse phénoménologique et extrapolations des données existantes à petites échelles de temps par exemple) et utiliser un algorithme robuste de résolution.

Une catégorie spécifique d'EDP très étudiée est celle contenant des contraintes inégalités. On les retrouve dans de nombreux domaines scientifiques : l'économie (théorie des jeux) et la finance (modèle de Black–Scholes), mécanique des solides (problème de l'obstacle ou problème de contact entre plusieurs corps) et la mécanique des fluides (écoulement multiphasique avec transition de phase). L'analyse mathématique de ces EDP avec contraintes inégalités repose sur leur écriture sous forme "faible" entrant ainsi dans le cadre des inégalités variationnelles. Dans cette thèse, nous choisissons comme axe principal la modélisation mathématique et numérique des inéquations variationnelles avec comme objectif final l'étude des problèmes géophysiques en sous-sol. Plus précisément, notre étude est consacrée à un problème de contact stationnaire entre deux membranes, à une inégalité variationnelle parabolique qui peut être vue comme une extension du premier modèle, puis à un écoulement diphasique instationnaire avec transition de phase en milieu poreux. Le second modèle est un modèle intermédiaire entre le modèle de contact stationnaire, où l'unique non-linéarité réside dans le contact, et le modèle instationnaire diphasique avec transition de phase, qui est complètement non-linéaire. On voit bien le caractère transverse de ce sujet de thèse et sa portée dans le large éventail des problématiques environnementales actuelles. En somme, dans le dernier chapitre de cette thèse, nous nous intéressons aux conséquences de l'enfouissement et du stockage des déchets radioactifs dans les couches géologiques profondes (voir Figure 2). Ce stockage provoque entre autres une production d'hydrogène issue de la dégradation chimique de certains matériaux de stockage. L'objectif à terme pour les

industriels est de pouvoir contrôler cette émanation de gaz, responsable important de la dégradation de l'écosystème, en comprenant son évolution par la simulation numérique. On verra plus tard l'importance cruciale du contrôle de l'erreur de la solution approchée numérique, **a posteriori** en particulier, pour définir des seuils ou des critères d'arrêt permettant de décider si la solution numérique obtenue est acceptable ou non.

II Approximation numérique des inégalités variationnelles

II.I Modèle continu

Pour fixer les idées sur la démarche générale adoptée, on va considérer le modèle stationnaire à résoudre dans un domaine borné $\Omega \subset \mathbb{R}^d$ avec $d = 1, 2$ ou 3 : Etant donnés un espace de Hilbert V dont le dual topologique est noté V^*, un ensemble convexe fermé non vide $\mathcal{K} \subset V$ et une fonction $\mathcal{A} : \mathcal{K} \to V^*$, un problème d'inégalité variationnelle consiste à trouver un vecteur u tel que

$$\begin{aligned} &u \in \mathcal{K}, \\ &\langle \mathcal{A}(u), v - u \rangle \geq 0 \quad \forall v \in \mathcal{K}. \end{aligned}$$

Ici $\langle \cdot, \cdot \rangle$ désigne le crochet de dualité entre les espaces V^* et V. Ce modèle est muni de conditions aux limites pour la solution $u \in V$. On suppose que l'étude théorique nous permet de montrer l'existence de la solution u dans l'ensemble convexe $\mathcal{K}$.

II.II Modèle discret

Dans un premier temps, on approche le domaine Ω par un maillage $\mathcal{T}_h$ sur lequel notre inéquation variationnelle est vérifiée. Cette triangulation ou maillage est généralement réalisé à l'aide de segments en 1D, de triangles, rectangles ou hexagones en 2D et de tétraèdres ou hexaèdres en 3D (voir Figure 3). Ensuite, dépendant de la méthode numérique utilisée, on construit un espace d'approximation de dimension finie noté $\mathcal{K}_h$, où $\mathcal{K}_h$ est un convexe fermé non vide de V, non forcément inclus dans $\mathcal{K}$, dans lequel vit une solution approchée de u que l'on note u_h. La plupart du temps, les méthodes numériques employées sont : les méthodes d'éléments finis (voir Ciarlet [57], Brenner et Scott [35], Quarteroni et Valli [140], Maday, Bernardi et Rapetti [27], Ern et Guermond [77]), de volumes finis (voir Eymard, Gallouët et Herbin [84], Godlewski et Raviart [97]), d'éléments finis mixtes (voir Roberts et Thomas [146], Brezzi et Fortin [38], Chen [50], Boffi, Brezzi et Fortin [28]) et de Galerkine discontinues (voir Di Pietro et Ern [68] et Rivière [145]). La littérature citée n'est pas exhaustive, ces méthodes étant très largement étudiées. Chacune de ces méthodes présente des avantages comme des inconvénients et il n'existe pas de règles générales à leur utilisation, mais à titre d'exemple : la méthode des éléments finis est privilégiée dans les calculs des structures en mécanique des solides, la méthode des volumes finis et la méthode des éléments finis mixtes sont couramment utilisées dans les écoulements multiphasiques car elles vérifient la propriété

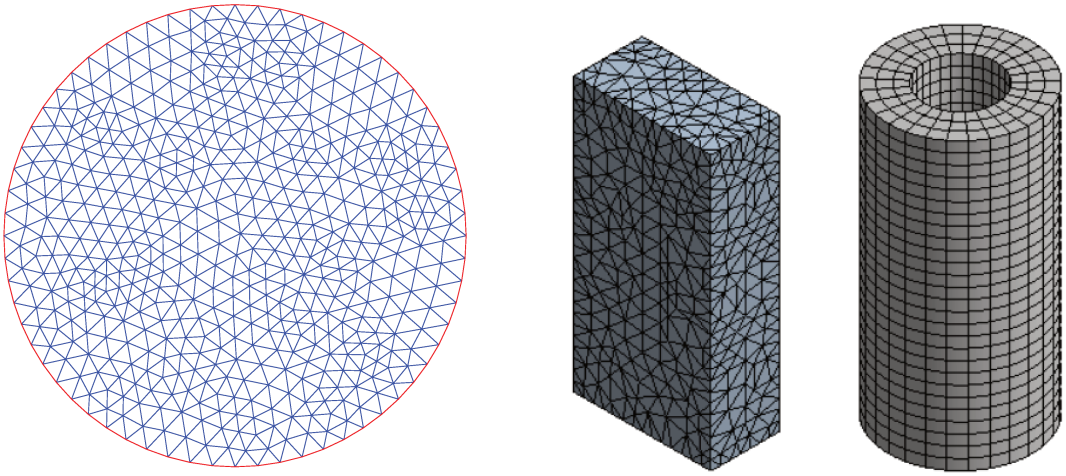

Figure 3: maillage 2D triangulaire non structuré (Matlab, gauche), maillage 3D tétraédrique non structuré (Wikipedia, milieu), maillage 3D hexaédrique structuré (Wikipedia, droite).

de conservation des flux et la méthode de Galerkine discontinue est intéressante car elle utilise des espaces d'approximation qui sont discontinus entre les mailles individuelles.

La solution numérique u_h recherchée vérifie dans le cas de la méthode des éléments finis le système non linéaire suivant :

$$\begin{aligned} &u_h \in \mathcal{K}_h, \\ &\langle \mathcal{A}(u_h), v_h - u_h \rangle \geq 0 \quad \forall v_h \in \mathcal{K}_h. \end{aligned} \tag{1}$$

Une manière de résoudre le problème (1) est de le reformuler en conditions de complémentarité. Dans ce cas, on identifie des opérateurs (non) linéaires $\mathcal{F} : \mathbb{R}^N \to \mathbb{R}^M$, $\mathcal{G} : \mathbb{R}^N \to \mathbb{R}^{N-M}$, $\mathcal{H} : \mathbb{R}^N \to \mathbb{R}^{N-M}$, $N > M > 1$ et le problème (1) se réécrit : trouver $\boldsymbol{X}_h \in \mathbb{R}^N$ vérifiant

$$\begin{aligned} &\mathcal{F}(\boldsymbol{X}_h) = 0, \\ &\mathcal{G}(\boldsymbol{X}_h) \geq 0, \quad \mathcal{H}(\boldsymbol{X}_h) \geq 0, \quad \mathcal{G}(\boldsymbol{X}_h) \cdot \mathcal{H}(\boldsymbol{X}_h) = 0, \end{aligned} \tag{2}$$

où $\mathcal{G}(\boldsymbol{X}_h) \geq 0$ signifie que les composantes du vecteur $\mathcal{G}(\boldsymbol{X}_h)$ sont positives et $\mathcal{G}(\boldsymbol{X}_h) \cdot \mathcal{H}(\boldsymbol{X}_h) = 0$ signifie l'orthogonalité de ces deux vecteurs.

II.III Linéarisation par les méthodes semi-lisses

La prochaine étape est dédiée à la résolution du système non linéaire (2). Nous reformulons les contraintes de complémentarité données par la seconde ligne de (2) à l'aide de C-fonctions ("C" comme complémentarité). Par définition, une fonction $f : \mathbb{R}^{N-M} \times \mathbb{R}^{N-M} \to \mathbb{R}^{N-M}$ est une C-fonction si

$$f(\boldsymbol{a}, \boldsymbol{b}) = 0 \iff \boldsymbol{a} \geq 0, \quad \boldsymbol{b} \geq 0, \quad \boldsymbol{a} \cdot \boldsymbol{b} = 0 \quad \forall\, (\boldsymbol{a}, \boldsymbol{b}) \in \mathbb{R}^{N-M} \times \mathbb{R}^{N-M}.$$

Les C-fonctions couramment utilisées sont la fonction minimum "min"

$$f(\boldsymbol{a}, \boldsymbol{b}) = \min(\boldsymbol{a}, \boldsymbol{b}) \tag{3}$$

et la fonction de Fischer–Burmeister

$$f_{\mathrm{FB}}(\boldsymbol{a},\boldsymbol{b}) = \sqrt{\boldsymbol{a}^2+\boldsymbol{b}^2} - (\boldsymbol{a}+\boldsymbol{b}) \tag{4}$$

(se référer aux ouvrages de Facchinei et Pang [88, 89] et Bonnans, Gilbert, Lemaréchal et Sagastizábal [29] pour plus de détails). Ces C-fonctions ne sont pas différentiables au sens de Fréchet puisque la fonction "min" n'est pas différentiable lorsque $\boldsymbol{a} = \boldsymbol{b}$ et la fonction de Fischer–Burmeister n'est pas différentiable en $\mathbf{0}$. Néanmoins, ces fonctions sont différentiables au sens de Clarke (voir l'ouvrage de Clarke [58]). Introduisons une C-fonction que l'on note $\tilde{\mathcal{C}} : \mathbb{R}^{N-M} \times \mathbb{R}^{N-M} \to \mathbb{R}^{N-M}$, définie par $\tilde{\mathcal{C}}(\mathcal{G}(\boldsymbol{X}_h), \mathcal{H}(\boldsymbol{X}_h)) = 0 \iff \mathcal{G}(\boldsymbol{X}_h) \geq 0,\ \mathcal{H}(\boldsymbol{X}_h) \geq 0,\ \mathcal{G}(\boldsymbol{X}_h) \cdot \mathcal{H}(\boldsymbol{X}_h) = 0$. Ainsi, en notant $\mathcal{C} : \mathbb{R}^N \to \mathbb{R}^{N-M}$ la fonction définie par $\mathcal{C}(\boldsymbol{X}_h) := \tilde{\mathcal{C}}(\mathcal{G}(\boldsymbol{X}_h), \mathcal{H}(\boldsymbol{X}_h))$, le problème (2) se réécrit

$$\begin{aligned} \mathcal{F}(\boldsymbol{X}_h) &= 0, \\ \mathcal{C}(\boldsymbol{X}_h) &= 0, \end{aligned}$$

ou, sous forme compacte,

$$\mathcal{S}(\boldsymbol{X}_h) = 0. \tag{5}$$

Ensuite, nous linéarisons (5) à l'aide de la méthode de Newton semi-lisse suivante : En considérant une donnée initiale $\boldsymbol{X}_h^0 \in \mathbb{R}^N$, l'algorithme de Newton semi-lisse génère à chaque pas de linéarisation $k \geq 1$ un système linéaire

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^k = \boldsymbol{F}^{k-1} \tag{6}$$

avec $\mathbb{A}^{k-1} \in \mathbb{R}^{N\times N}$ une matrice, $\boldsymbol{F}^{k-1} \in \mathbb{R}^N$ un vecteur et $\boldsymbol{X}_h^k \in \mathbb{R}^N$ la solution linéarisée. Habituellement, on arrête les itérations de Newton dès que le critère d'arrêt suivant est satisfait :

$$\left\|\mathcal{S}(\boldsymbol{X}_h^k)\right\| \leq \varepsilon_{\mathrm{lin}} \left\|\mathcal{S}(\boldsymbol{X}_h^0)\right\|, \tag{7}$$

avec $\varepsilon_{\mathrm{lin}}$ une tolérance fixée par l'utilisateur. Les méthodes de Newton sont étudiées depuis plusieurs décennies et pour trouver des études approfondies sur le sujet se référer aux ouvrages de Kantorovich [110], Kelley [114, 115] et Deuflhard [67], puis l'article d'Ortega [133].

II.IV Méthodes inexactes

La résolution du système (6) par une méthode directe peut s'avérer très coûteuse si on considère par exemple des maillages très fins conduisant à de trop grandes valeurs de N. On préfère employer des méthodes itératives entrant ainsi dans le cadre des méthodes de Newton inexactes. Parmi les méthodes itératives linéaires qui ont fait leurs preuves, citons la famille des algorithmes multigrilles (voir les ouvrages de Briggs [41] et Hackbusch [102] pour une étude détaillée) et la méthode du gradient conjugué lorsque la matrice $\mathbb{A}^{k-1}$ est symétrique définie positive (voir Olshanskii et Tyrtyshnikov [132], Kelley [114], Saad [149] et Liesen et Strakoš [120]). Sinon, on peut minimiser le résidu obtenu à l'aide de l'algorithme de GMRES (se référer à Saad et Schultz [150] et Brown et Saad [42]). Ainsi, à chaque itération $i \geq 0$ du

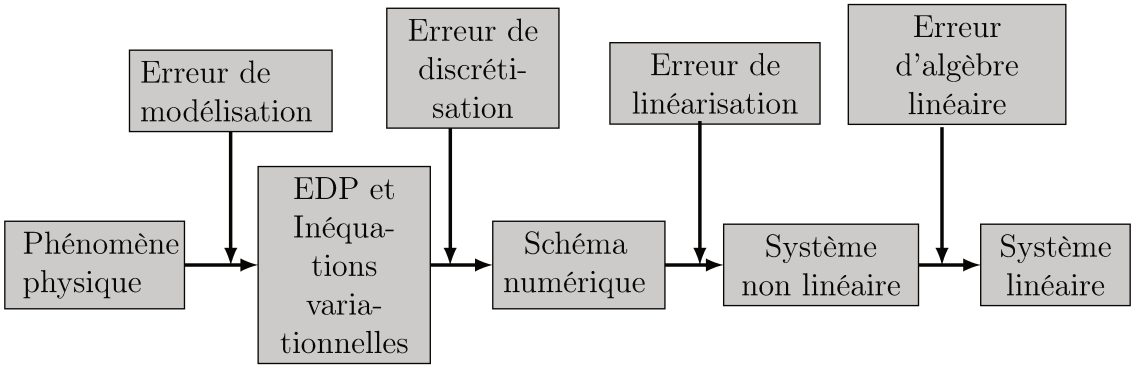

Figure 4: Illustration du processus allant de la modélisation du phénomène physique à sa résolution numérique approchée, avec différents types d'erreurs.

solveur d'algèbre linéaire au sein de chaque itération $k \geq 1$ du solveur non linéaire, on dispose d'une approximation $\boldsymbol{X}_h^{k,i} \in \mathbb{R}^N$ et on calcule un résidu défini par

$$\boldsymbol{R}_h^{k,i} := \boldsymbol{F}^{k-1} - \mathbb{A}^{k-1} \boldsymbol{X}_h^{k,i}. \tag{8}$$

Ici, le critère d'arrêt habituel pour le solveur itératif prend la forme

$$\frac{\left\| \boldsymbol{R}_h^{k,i} \right\|}{\left\| \boldsymbol{F}^{k-1} - \mathbb{A}^{k-1} \boldsymbol{X}_h^{k,0} \right\|} \leq \delta_k, \tag{9}$$

où $\boldsymbol{X}_h^{k,0}$ est le vecteur initial du solveur linéaire. La valeur δ_k, appelée "forcing term", peut varier à chaque itération k. Lorsque cette valeur est petite, on se rapproche de la méthode de Newton semi-lisse "exacte" et on effectuera alors beaucoup d'itérations en i du solveur d'algèbre linéaire, alors que si cette valeur est trop grande, on risque de faire trop d'itérations k du solveur de Newton semi-lisse. Il faut donc trouver un bon compromis entre résoudre rapidement les systèmes linéaires et ne pas effectuer trop d'itérations de Newton. Plusieurs études existent sur ce sujet. On peut mentionner Dembo, Eisenstat, Steihaug [65] et Eisenstat et Walker [72,73] pour les méthodes de Newton inexactes.

II.V Erreur

Naturellement, la résolution numérique décrite ci-dessus génère des erreurs qui sont typiquement de 3 natures : l'erreur liée au schéma numérique utilisé et au maillage choisi, l'erreur liée à l'emploi du solveur non linéaire et l'erreur liée à l'emploi du solveur algébrique linéaire. Lorsque ces erreurs sont très petites, cela veut dire qu'on a résolu de manière quasi exacte le système. Nous pouvons synthétiser de manière schématique notre propos (voir Figure 4).

On ne s'intéresse pas ici à l'erreur de modélisation (on suppose donc que l'EDP décrit suffisamment bien le phénomène physique). On va chercher à estimer les trois sources d'erreur causée par la simulation numérique. Cette estimation des composantes de l'erreur permettra en particulier de réduire le nombre d'itérations de la simulation numérique sans affecter ni dénaturer la solution obtenue, tout en garantissant à la fin de la simulation l'erreur totale maximale commise.

III Estimation d'erreur a posteriori

III.I Revue des méthodes a posteriori

Souvent, on mesure l'erreur commise à l'aide d'estimations dites a priori. Dans une estimation d'erreur a priori, la solution exacte notée u d'une EDP et sa solution approchée notée u_h sont liées par la relation suivante :

$$|||u - u_h||| \leq C(u)h^{\alpha}\left(\frac{1}{p}\right)^{\beta}$$

où C est une fonction qui dépend de u, α et β des paramètres strictement positifs, $|||\cdot|||$ une norme, h le pas du maillage et p le degré polynomial. Ce type d'estimation apporte des informations importantes sur la convergence de la méthode numérique employée car l'erreur diminue en raffinant le maillage et en augmentant le dégré polynomial p. Pour des exemples d'estimations d'erreur a priori, se référer à Strang [152], Johnson [109], Ern et Guermond [77] ou Maday, Bernardi, Rapetti [27]. Dans le cadre des éléments finis conformes, les estimations d'erreur a priori s'appuient fortement sur le fameux lemme de Céa [48].

L'inconvénient principal de ces estimations a priori est que la borne supérieure de l'erreur n'est pas calculable en pratique vu qu'elle dépend de la solution exacte u, qui est inconnue. A partir des années 1970, avec les travaux de Ladevèze [118], eux-mêmes inspirés de l'identité de Prager et Synge [138], des estimations sous le nom d'estimations d'erreur a posteriori se développent. Par "a posteriori" on entend évidemment que l'estimation a lieu après avoir calculé la solution approchée u_h. Parmi la longue énumération de travaux sur le sujet on peut citer les ouvrages de Babuška et Rheinboldt [11], Ainsworth et Oden [5], Repin [143], Verfürth [155] puis les contributions de Destyunder et Métivet [66], Braess et Schöberl [33], Vohralík [157], Ern et Vohralík [81] pour les problèmes elliptiques. Pour les problèmes paraboliques, on peut mentionner les apports de Verfürth [154], Nicaise et Soualem [130], Bergam, Bernardi et Mghazli [26], Lozinski, Picasso et Prachittham [125], Ern et Vohralík [80] et Ern, Smears et Vohralík [78]. Ces estimations prennent la forme suivante :

$$|||u - u_h||| \leq \eta(u_h), \tag{10}$$

où $\eta(u_h)$ est une quantité calculable dépendant seulement de la solution approchée u_h calculée. On distingue plusieurs types d'estimations d'erreur a posteriori :

- *Les estimations par résidus* : l'idée est de borner l'erreur par la norme duale du résidu, puis de borner cette norme duale par des estimateurs exprimant les résidus par éléments et faces. Une constante C_{rel} appelée communément constante de fiabilité dépendant des paramètres du problème apparaît :

  $$|||u - u_h||| \leq C_{\text{rel}}\eta(u_h).$$

  De plus, la propriété d'efficacité globale se traduit par :

  $$\eta(u_h) \leq C_{\text{eff}} |||u - u_h|||.$$

Pour cette méthode, les estimateurs locaux $\eta_K(u_h)$ sont peu coûteux numériquement. Le talon d'Achille est le calcul des constantes C_{eff} et C_{rel} qui dépendent d'un nombre conséquent de paramètres et ne sont pas connus en général. La littérature est très vaste pour cette méthode et on renvoit le lecteur aux travaux suivants [5, 26, 143, 154, 155].

- *Les estimations utilisant la technique de reconstruction des flux équilibrés* : l'idée est de reconstruire dans chaque patch (union de simplexes partageant un sommet commun) des flux qui sont réguliers $\mathbf{H}(\text{div})$. Cette méthode s'applique à toutes les méthodes numériques (éléments finis, volumes finis, Galerkine discontinue, etc...). Les estimateurs d'erreur calculés reflètent certaines propriétés physiques violées par la solution numérique. Cette résolution peut être plus coûteuse numériquement que pour les estimations par résidu mais elle a l'avantage de fournir une estimation d'erreur a posteriori où la constante C_{rel} devant les estimateurs est égale à 1. Cela se traduit par :

$$|||u - u_h||| \leq \eta(u_h) \quad \text{et} \quad \eta(u_h) \leq C_{\text{eff}} \, |||u - u_h||| \, .$$

  Par ailleurs, cette méthode permet de distinguer les différentes composantes d'erreur de la simulation numérique. Sur ce sujet plus récent, on renvoit le lecteur aux articles suivants [33, 66, 78, 81, 82, 126].

- *Les estimations hiérarchiques* : l'idée ici est de calculer une autre solution approchée $v_h \in \hat{V}_h$ plus précise que $u_h \in V_h$ de telle sorte que $|||u_h - v_h|||$ soit un estimateur d'erreur pour $|||u - u_h|||$. En fait, $\hat{V}_h$ est un espace d'approximation plus fin que V_h et vérifie $\hat{V}_h = Z_h \bigoplus V_h$. L'espace Z_h est le supplémentaire orthogonal de V_h. En utilisant la décomposition $u_h - v_h = w_h + z_h$ avec $w_h \in V_h$ et $z_h \in Z_h$ et en supposant l'inégalité de Cauchy–Schwarz forte $(\boldsymbol{\nabla} w_h, \boldsymbol{\nabla} z_h)_\Omega \leq \beta \, |||w_h||| \, |||z_h||| \;\; \forall w_h \in V_h, \;\; \forall z_h \in Z_h$, où $\beta < 1$ et $(\cdot, \cdot)_\Omega$ est le produit scalaire associé à la norme $|||\cdot|||$, on parvient à montrer que $|||z_h|||$ est un estimateur d'erreur pour $|||u - u_h|||$. Ainsi, nous obtenons à l'instar des méthodes précédentes

$$|||u - u_h||| \leq C_{\text{rel}} \eta(u_h) \quad \text{et} \quad \eta(u_h) \leq C_{\text{eff}} \, |||u - u_h||| \, .$$

  Nous renvoyons le lecteur aux contributions suivantes [8, 46, 74, 155].

- *Les estimations par les méthodes de la moyenne* : Cette méthode s'applique particulièrement lorsque l'on considère une norme d'énergie du type $|||u - u_h||| = \|\boldsymbol{\nabla}(u - u_h)\|$. Ici, on approche le gradient discret $\boldsymbol{\nabla} u_h$ par une quantité que l'on note $\widetilde{\boldsymbol{\nabla}} u_h$ de telle sorte qu'on ait la majoration $\|\boldsymbol{\nabla}(u - u_h)\| \leq C_{\text{rel}} \left\| \widetilde{\boldsymbol{\nabla}} u_h - \boldsymbol{\nabla} u_h \right\|$. Dans le cas particulier des éléments finis linéaires $\mathbb{P}_1$, le terme $\widetilde{\boldsymbol{\nabla}} u_h$ se calcule par la somme des moyennes locales du gradient $\boldsymbol{\nabla} u_h$:

$$\widetilde{\boldsymbol{\nabla}} u_h(\boldsymbol{a}) = \sum_{\substack{K \in \mathcal{T}_h \\ \boldsymbol{a} \in \mathcal{V}_K}} \frac{|K|}{|\omega_{\boldsymbol{a}}|} \boldsymbol{\nabla} u_h|_K .$$

Ici, $\mathcal{V}_K$ désigne l'ensemble des sommets du triangle $K \in \mathcal{T}_h$, $\omega_{\boldsymbol{a}}$ désigne le patch centré en $\boldsymbol{a}$ et $|K|$ la mesure au sens de Lebesgue de l'élément K. On retrouve également la propriété suivante :

$$|||u - u_h||| \leq C_{\text{rel}}\eta(u_h) \quad \text{et} \quad \eta(u_h) \leq C_{\text{eff}} |||u - u_h||| .$$

Pour des études plus détaillées sur le sujet, voir [12, 47, 61, 91, 155].

Dans toute cette thèse, nous utilisons la technique de reconstruction des flux équilibrés pour obtenir des estimations d'erreur a posteriori garanties avec $C_{\text{rel}} = 1$. Notons qu'un estimateur d'erreur a posteriori est optimal s'il satisfait les propriétés suivantes :

1. *Borne supérieure de l'erreur garantie* : la condition (10) doit être vérifiée et l'estimateur $\eta(u_h)$ doit être complètement calculable en fonction de u_h.

2. *Efficacité* : pour les problèmes stationnaires, pour chaque élément K du maillage $\mathcal{T}_h$, l'estimateur local $\eta_K(u_h)$ ($\eta(u_h)$ se décompose généralement comme une somme Euclidienne des termes $\eta_K(u_h)$) doit représenter une borne inférieure de l'erreur dans un voisinage de K (noté ζ_K) à une constante près ; ce qui se traduit par
$$\eta_K(u_h) \leq C_{\text{eff}} |||u - u_h|||_{\zeta_K} . \tag{11}$$
Pour les problèmes non-stationnaires, à chaque pas de temps $1 \leq n \leq N_{\text{t}}$, cela se traduit par
$$\eta_K^n \leq C_{\text{eff}} |||u - u_h|||_{\zeta_K \times (t^{n-1}, t^n]} . \tag{12}$$

3. *Exactitude asymptotique* : l'indice d'efficacité I_{eff} défini par
$$\text{I}_{\text{eff}} := \frac{\eta(u_h)}{|||u - u_h|||} \quad \text{pour les problèmes stationnaires}$$
ou
$$\text{I}_{\text{eff}} := \frac{\left\{ \sum_{n=1}^{N_{\text{t}}} \sum_{K \in \mathcal{T}_h} (\eta_K^n)^2 \right\}^{\frac{1}{2}}}{|||u - u_{h\tau}|||} \quad \text{pour les problèmes instationnaires}$$
tend vers 1 quand on raffine le maillage $\mathcal{T}_h$ et le pas de temps.

4. *Robustesse* : il faut garantir que la constante C_{eff} est indépendante des paramètres du problème (et de leurs variations).

5. *Faible coût numérique* : il faut si possible garantir un moindre coût numérique pour le calcul des estimateurs locaux η_K.

6. *Distinction des composantes d'erreur* : on doit être capable d'estimer toutes les composantes d'erreur issues de la simulation numérique.

III.II Critères d'arrêts adaptatifs

L'estimation des différentes sources de l'erreur permet de formuler des critères d'arrêts dits adaptatifs pour nos solveurs. On rappelle que les critères d'arrêts standards pour une méthode de Newton exacte et inexacte sont donnés respectivement par (7) et (9). En notant $u_h^{k,i}$ la représentation fonctionnelle du vecteur $\boldsymbol{X}_h^{k,i}$ issue de (8), puis $\eta_{\text{disc}}^{k,i}$ l'estimation de l'erreur de discrétisation, $\eta_{\text{lin}}^{k,i}$ l'estimation de l'erreur de linéarisation et $\eta_{\text{alg}}^{k,i}$ l'estimation de l'erreur d'algèbre linéaire pour un problème stationnaire, notre estimation d'erreur a posteriori prend la forme

$$\left|\!\left|\!\left| u - u_h^{k,i} \right|\!\right|\!\right| \leq \eta_{\text{disc}}^{k,i} + \eta_{\text{lin}}^{k,i} + \eta_{\text{alg}}^{k,i}.$$

Nos critères d'arrêts adaptatifs seront définis par :

$$\eta_{\text{alg}}^{k,i} \leq \gamma_{\text{alg}} \max\left\{\eta_{\text{lin}}^{k,i}, \eta_{\text{disc}}^{k,i}\right\}, \quad \eta_{\text{lin}}^{k,i} \leq \gamma_{\text{lin}} \eta_{\text{disc}}^{k,i}. \tag{13}$$

Ici γ_{alg} et γ_{lin} sont des poids fixés par l'utilisateur (généralement de l'odre de 0.1) et (13) exprime que les composantes d'erreur non principales ne doivent pas avoir une influence supérieure à 10% sur la composante principale. Les propriétés d'efficacité locale ou globale de la forme (11) se démontrent lorsque les critères d'arrêts adaptatifs sont satisfaits. Notons que ces estimations des composantes d'erreur sont heuristiques dans le sens où elles approchent et imitent le comportement des composantes exactes dans nos essais numériques mais nous n'apportons pas de preuves cependant qu'elles bornent ces composantes exactes. On a

$$\lim_{\substack{k\to\infty \\ i\to\infty}} \eta_{\text{lin}}^{k,i} = \lim_{\substack{k\to\infty \\ i\to\infty}} \eta_{\text{alg}}^{k,i} = 0. \tag{14}$$

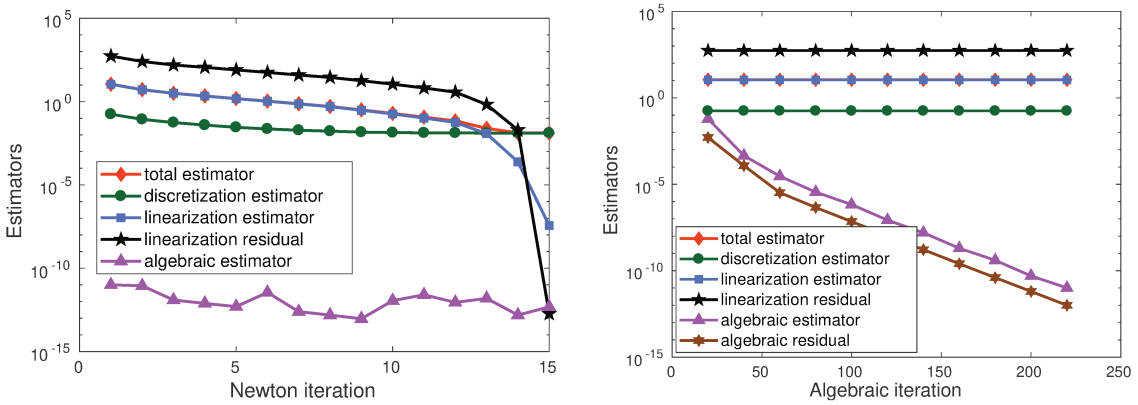

Figure 5: Illustration de la notion de critères d'arrêts adaptatifs. Critère d'arrêt adaptatif pour un solveur de linéarisation semi-lisse Newton-min (figure de gauche). Critère d'arrêt adaptatif pour un solveur d'algèbre linéaire GMRES (figure de droite). Chapitre 1 Figure 1.8, Figure 1.9 et https://hal.archives-ouvertes.fr/hal-01666845.

On observe dans l'exemple concret de la Figure 5 issue du Chapitre 1 que dans une telle configuration, il est raisonnable d'arrêter les itérations du solveur non

linéaire de Newton dès la 14^{eme} itération car l'estimateur de linéarisation (courbe bleue) n'influence plus le comportement de l'erreur total (donné par l'estimateur de discrétisation) qui commence à stagner. De même, la figure de droite montre que pour une itération de Newton-min fixée, l'estimateur d'algèbre linéaire est très petit devant l'estimateur de linéarisation dès le début des itérations ($i = 10$). Pour une résolution exacte il aurait fallu attendre l'itération $i = 220$ avant d'arrêter l'algorithme.

On voit désormais que notre travail peut se synthétiser de la manière illustrée dans la Figure 6.

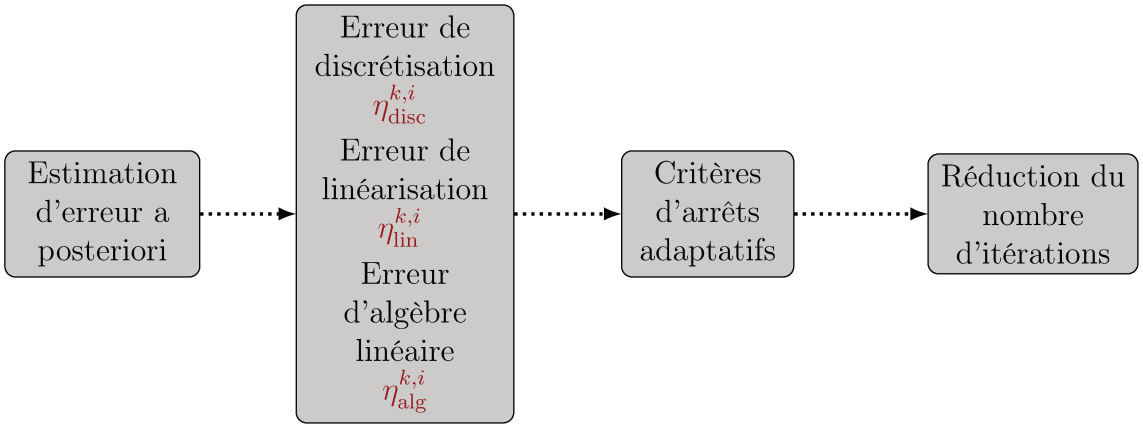

Figure 6: Interactions entre les diverses composantes d'erreur au sein d'une simulation numérique.

Au passage, on mentionne que nos estimateurs d'erreur permettent de raffiner le maillage aux endroits où l'erreur est importante même si cela n'est pas effectué dans ce travail.

IV Description des problèmes et état de l'art

Le but principal de cette thèse est de concevoir des estimations d'erreur a posteriori pour des équations aux dérivées partielles avec contraintes de complémentarité et de proposer des critères d'arrêts adaptatifs pour chaque solveur utilisé. Ceci permet de réduire le nombre d'itérations tout en garantissant la précision des résultats. Nous abordons trois problèmes : au Chapitre 1, nous traitons une inégalité variationnelle linéaire stationnaire décrivant un problème de contact entre deux membranes. Au Chapitre 2, nous étudions une inégalité variationnelle linéaire parabolique instationnaire comme extension du modèle du chapitre 1 . Enfin, au Chapitre 3, nous présentons une inégalité variationnelle non linéaire parabolique décrivant un écoulement diphasique avec transition de phase en milieu poreux.

IV.I Chapitre 1 : problème de contact entre deux membranes comme une inégalité variationnelle stationnaire

Problème

Nous abordons dans un premier temps un problème d'inégalité variationnelle stationnaire. Soit $\Omega \subset \mathbb{R}^2$ un domaine polygonal de $\mathbb{R}^2$. On considère le problème : trouver u_1, u_2 et λ tel que

$$\begin{cases} -\mu_1 \Delta u_1 - \lambda = f_1 & \text{dans} \quad \Omega, \\ -\mu_2 \Delta u_2 + \lambda = f_2 & \text{dans} \quad \Omega, \\ (u_1 - u_2)\lambda = 0, \quad u_1 - u_2 \geq 0, \quad \lambda \geq 0 & \text{dans} \quad \Omega, \\ u_1 = g & \text{sur} \quad \partial\Omega, \\ u_2 = 0 & \text{sur} \quad \partial\Omega. \end{cases} \tag{15}$$

Ici, u_1 est le déplacement de la première membrane, u_2 est le déplacement de la seconde membrane (qui se situe sous la première) et λ un multiplicateur de Lagrange traduisant l'action de la première membrane sur la seconde. Les coefficients μ_1 et μ_2 strictement positifs sont les tensions de chaque membrane. Les termes source $f_1 \in L^2(\Omega)$ et $f_2 \in L^2(\Omega)$ sont des forces extérieures. Les deux premières lignes de (15) décrivent le comportement cinématique de chaque membrane et la troisième ligne de (15) les conditions de complémentarité linéaires. Celles-ci permettent de gérer deux situations physiques distinctes : soit les membranes sont en contact, auquel cas $u_1 = u_2$ et $\lambda > 0$, soit les membranes sont séparées, ce qui se traduit par $u_1 - u_2 > 0$ et $\lambda = 0$. Pour simplifier l'étude, on supposera que $g > 0$ est une constante. Ce problème s'inscrit dans la gamme des inégalités variationnelles de première espèce : Trouver $\boldsymbol{u} = (u_1, u_2) \in \boldsymbol{\mathcal{K}}_g$ tel que $\forall \boldsymbol{v} = (v_1, v_2) \in \boldsymbol{\mathcal{K}}_g$

$$a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq l(\boldsymbol{v} - \boldsymbol{u}), \tag{16}$$

où $\boldsymbol{\mathcal{K}}_g := \left\{(v_1, v_2) \in H^1_g(\Omega) \times H^1_0(\Omega),\ v_1 - v_2 \geq 0 \text{ presque partout dans } \Omega\right\}$, est un convexe fermé non vide de l'espace produit $H^1_g(\Omega) \times H^1_0(\Omega)$. Ici, $H^1_g(\Omega)$ est l'espace affine des fonctions $H^1(\Omega)$ à valeur g au bord $\partial\Omega$, a est une forme bilinéaire continue sur $\left[H^1(\Omega)\right]^2 \times \left[H^1(\Omega)\right]^2$ et coercive sur $\left[H^1_0(\Omega)\right]^2 \times \left[H^1_0(\Omega)\right]^2$ et l une forme linéaire continue sur $\left[H^1(\Omega)\right]^2$. Tous les détails sont donnés dans le Chapitre 1. Ce problème est similaire, quoique plus complexe, au problème d'obstacle et la littérature existente sur le sujet est très vaste. Nous le considérons ici en raison de sa structure d'inégalité variationnelle (16).

Etude bibliographique

La communauté scientifique s'est penchée depuis plusieurs décennies sur l'étude des estimations d'erreurs a posteriori pour les inégalités variationnelles. La liste des travaux est trop longue pour être intégralement détaillée. Nous ne mentionnons qu'une partie et dans l'odre chronologique : nous avons les travaux pionniers de Brezzi, Hager et Raviart [39,40] puis les travaux de Ainsworth, Oden et Lee [6], Kornhuber [116], Chen et Nochetto [53], Coorevits, Hild et Pelle [60], Veeser [153], Louf, Combe et Pelle [123], Belhachmi et Ben Belgacem [16], Bartels et Carstensen [12],

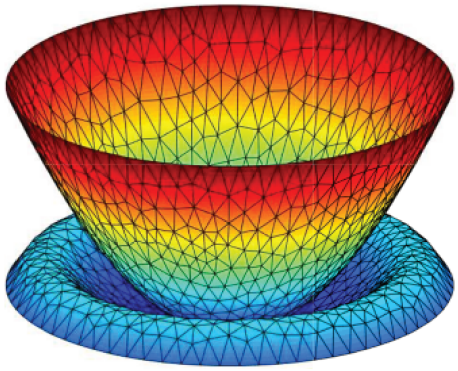

Figure 7: Contact entre deux membranes. source: Matlab.

Braess [31], Repin [144], Ben Belgacem, Bernardi, Blouza et Vohralík [19], Auliac, Belhachmi, Ben Belgacem et Hecht [10], Gudi et Porwal [99–101], Bürg et Schröder [43], Chouly, Fabre, Hild, Pousin et Renard [55].

Généralement, c'est le problème elliptique d'obstacle qui est considéré. Ce dernier consiste à trouver la position d'équilibre d'une membrane élastique dont les bords sont fixés et devant passer au-dessus d'un obstacle donné. On peut citer les ouvrages de Rodrigues [148] et Hlaváček, Haslinger, Nečas et Lovíšek [105] pour avoir une vision à large spectre sur ce problème. Avant d'aller plus loin, nous présentons les idées principales de certains travaux cités plus haut.

Chen et Nochetto [53] s'intéressent à concevoir des estimations d'erreur a posteriori pour le problème de l'obstacle dans le cas d'une discrétisation par la méthode des éléments finis $\mathbb{P}_1$. Ils utilisent un multiplicateur de Lagrange pour traiter la condition de contact et introduisent un opérateur d'interpolation préservant la positivité. Ils considèrent la méthode des estimateurs par résidu pour obtenir une borne supérieure de l'erreur de discrétisation. Plus précisément, ils construisent une mesure de l'erreur entre la solution exacte et la solution approchée à partir de la norme d'énergie sur l'espace $H^1(\Omega)$. Ils obtiennent une borne supérieure pour cette norme d'énergie à une constante C_{rel} près constituée de termes entièrement calculables localement qui sont de 4 natures : un estimateur de résidu par face, un estimateur de résidu par élément et deux estimateurs liés à l'obstacle faisant intervenir l'opérateur d'interpolation. Les auteurs de ce papier ne prouvent qu'un résultat partiel d'efficacité locale. En effet, ils obtiennent une borne inférieure pour l'erreur d'énergie constituée de l'estimateur de résidu par face et par élément ainsi que d'une mesure de Radon continue notée μ.

Veeser [153] propose des estimations a posteriori dans le cas d'une discrétisation par éléments finis $\mathbb{P}_1$ en s'affranchissant de l'opérateur d'interpolation de Chen et Nochetto. Il considère en premier le cas particulier des obstacles affines et ensuite le cas des obstacles quelconques en utilisant également un multiplicateur de Lagrange. Il utilise la méthode des estimations d'erreurs par résidu pour obtenir des estimations d'erreur de l'erreur de discrétisation. Plus précisément, il considère une mesure de l'erreur constituée de deux termes : le premier étant l'erreur d'énergie au sens $H^1_0(\Omega)$ entre la solution exacte et la solution approchée et le second l'erreur au sens $H^{-1}(\Omega)$ entre le multiplicateur de Lagrange exact et sa version discrète. La borne supérieure de cette mesure de l'erreur se décompose localement et ne contient que des termes

intégralement calculables à une constante C_{rel} près, qui sont de trois natures : un estimateur de résidu par face, un estimateur de résidu par élément et un estimateur des contraintes. La propriété d'efficacité locale est également démontrée pour chaque estimateur.

Braess [31] synthétise les approches précédentes en apportant une vision d'ensemble et un cadre quasi-générique à la construction d'estimateurs a posteriori pour le problème d'obstacle.

Gudi et Porwal [99, 100] traitent le problème elliptique d'obstacle en utilisant les méthodes de Galerkine discontinues. Ils utilisent un opérateur d'interpolation $\mathcal{I}_h$ introduit par Brenner [34]. En notant u la solution exacte et u_h la solution approchée, leur idée consiste à utiliser une inégalité triangulaire $\|\boldsymbol{\nabla}(u-u_h)\|_\Omega \leq \|\boldsymbol{\nabla}(u-\mathcal{I}_h(u_h))\|_\Omega + \|\boldsymbol{\nabla}(\mathcal{I}_h(u_h)-u_h)\|_\Omega$ puis d'estimer chacun de termes de cette inégalité. Ils obtiennent ainsi une borne supérieure pour l'erreur de discrétisation entièrement calculable. Les estimateurs fournis ressemblent à ceux présentés par Veeser [153]. Par ailleurs, ils prouvent l'efficacité locale de leurs estimateurs en s'inspirant du travail de Veeser [153].

Un autre problème étudié par Chouly et Hild [56], entrant dans la gamme des inégalités variationnelles, est le problème de contact sans frottements en élasticité linéaire où les contraintes sont traitées avec une méthode à la Nitsche. Cette dernière consiste à obtenir une discrétisation du modèle par la méthode des éléments finis sans passer par l'emploi d'un multiplicateur de Lagrange. Concernant l'analyse a posteriori du problème introduit dans [56], Chouly, Fabre, Hild, Pousin et Renard (voir [55]) proposent d'estimer l'erreur de discrétisation en employant les estimations par résidu. En particulier, ils obtiennent une borne supérieure de l'erreur dans la norme d'énergie $H_0^1(\Omega)$ complétée par un terme issu de la méthode de "Nitsche", puis ils démontrent l'efficacité locale pour chaque estimateur.

Nous nous intéressons ici au problème de contact entre deux membranes (15) étudié par Ben Belgacem, Bernardi, Blouza et Vohralík [17]. Les auteurs ont utilisé une discrétisation par la méthode des éléments finis $\mathbb{P}_1$ conforme et la résolution numérique a été effectuée grâce à la méthode appelée "primal dual active set strategy" proposée par Hintermüller, Ito et Kunisch [104]. Au niveau de l'analyse a posteriori, les mêmes auteurs ont proposé quelques années plus tard dans [19], une reconstruction des flux dans l'espace $\mathbf{H}(\text{div},\Omega)$ en utilisant l'espace de Raviart–Thomas de plus bas degré $\mathbf{RT}_0$. Ils estiment l'erreur de discrétisation via des estimateurs d'erreur qui s'expriment localement. Ces estimateurs sont de trois natures : un estimateur de résidu, un estimateur de flux et un estimateur des contraintes. Notons que leur estimation d'erreur est garantie avec la constante $C_{\text{rel}} = 1$. La propriété d'efficacité locale est également démontrée pour chaque estimateur. En somme, ils obtiennent une estimation d'erreur a posteriori optimale.

Contribution principale

Dans le cadre du premier chapitre de cette thèse, nous proposons une discrétisation de (15) par la méthode des éléments finis de degré $p \geq 1$ quelconque. Cette nouvelle discrétisation génère un système linéaire d'équations complétés par des contraintes de complémentarité linéaire du type (2). Nous utilisons une approche originale pour traiter la non linéarité, plus précisément nous nous servons des algorithmes de New-

ton semi-lisse inexacts (voir [64, 86, 112, 127]). Ainsi, à chaque pas de linéarisation semi-lisse $k \geq 1$ et chaque pas d'algèbre linéaire $i \geq 0$, nous étudions le résidu $\boldsymbol{R}_h^{k,i}$ défini par

$$\boldsymbol{R}_h^{k,i} := \boldsymbol{F}^{k-1} - \mathbb{A}^{k-1} \boldsymbol{X}_h^{k,i}, \quad \text{à l'instar de (8).} \tag{17}$$

Notons que d'autres méthodes de résolution sont souvent employées dans la littérature. On peut mentionner la méthode des points intérieurs de Wright [160], la méthode active-set de Kanzow [111] et la méthode "primal dual active set strategy" proposée par Hintermüller, Ito et Kunisch [104]. Pour une étude plus approfondie et détaillée sur l'ensemble des techniques de résolution nous référons aux livres de Facchinei et Pang [88, 89]).

Nous disposons dans (17) d'une solution numérique linéarisée $\boldsymbol{u}_h^{k,i}$ représentée par le vecteur algébrique $\boldsymbol{X}_h^{k,i}$. Une difficulté centrale à traiter est la violation des contraintes de complémentarités $\mathcal{F}(\boldsymbol{X}_h^{k,i}) \not\geq 0$ et $\mathcal{G}(\boldsymbol{X}_h^{k,i}) \not\geq 0$ (voir (2)) au cours des itérations k du solveur semi-lisse et i du solveur d'algèbre linéaire.

Au niveau de l'analyse a posteriori, nous adoptons pour la construction des flux de discrétisation la méthode des flux équilibrés chère à Destyunder et Métivet [66], Braess et Schöberl [33] et Ern et Vohralík [81]. Pour la construction des flux qui permettent d'estimer l'erreur d'algèbre linéaire, nous utilisons la procédure introduite récemment dans Papež, Rüde, Vohralík et Wohlmuth [136]. Cette procédure utilise une reconstruction par niveaux au sein d'une hiérarchie de maillages. Ensuite, nous fournissons une estimation d'erreur a posteriori valide à chaque pas $k \geq 1$ et chaque pas $i \geq 0$, avant de distinguer et estimer chaque composante d'erreur issue de la simulation numérique, à savoir l'erreur de discrétisation par éléments finis, l'erreur de linéarisation par la méthode de Newton semi-lisse et enfin l'erreur d'algèbre linéaire liée à l'emploi d'un solveur itératif algébrique. Cela n'a pas été proposé par l'ensemble des travaux cités plus haut. Notre estimation prend la forme

$$\left|\!\left|\!\left| \boldsymbol{u} - \boldsymbol{u}_h^{k,i} \right|\!\right|\!\right| \leq \eta_{\text{disc}}^{k,i} + \eta_{\text{lin}}^{k,i} + \eta_{\text{alg}}^{k,i}.$$

Nous proposons ensuite un algorithme adaptatif de Newton semi-lisse inexact au sein duquel les critères d'arrêt pour chacun de nos solveurs sont basés sur les estimateurs d'erreur préalablement établis :

$$\text{(a) } \eta_{\text{alg}}^{k,i} \leq \gamma_{\text{alg}} \max\left\{\eta_{\text{disc}}^{k,i}, \eta_{\text{lin}}^{k,i}\right\}, \qquad \text{(b) } \eta_{\text{lin}}^{k,i} \leq \gamma_{\text{lin}} \eta_{\text{disc}}^{k,i}. \tag{18}$$

Nous verrons que sous ces critères d'arrêt, nous pouvons réduire de manière importante le nombre d'itérations (voir Figure 8 gauche). En effet, nous avons voulu montrer par cet exemple concret que la méthode de Newton exacte basée sur les critères d'arrêts (9) et (7) où $\delta_k = 2 \cdot 10^{-12}$ et $\varepsilon_{\text{lin}} = 10^{-10}$ requiert un grand nombre d'itérations total d'algèbre linéaire pour converger. La méthode de Newton inexacte pour δ_k variant en fonction de k (voir Chapitre 1, Section 1.8) et $\varepsilon_{\text{lin}} = 10^{-10}$ requiert moins d'itérations pour converger que la méthode de Newton exacte. Néanmois, la méthode de Newton semi-lisse inexacte adaptative utilisant les critères d'arrêts (18) avec $\gamma_{\text{alg}} = \gamma_{\text{lin}} = 0.3$ ne requiert que très peu d'itérations au total. De plus, la Figure 8 (droite) apporte des informations importantes. En effet, l'indice d'efficacité diminue au cours des itérations de Newton semi-lisse jusqu'à atteindre la valeur optimale proche de 1. Ainsi, nous avons à chaque itération k un contrôle sur l'erreur,

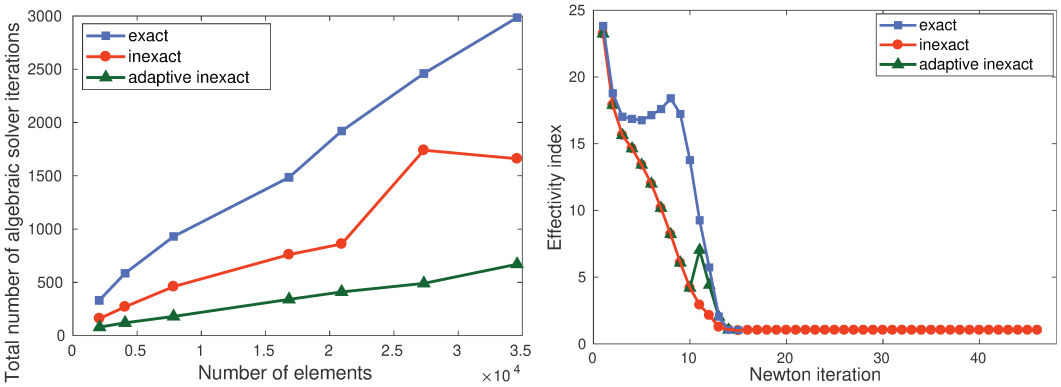

Figure 8: Nombre total cumulé d'itérations du solveur d'algèbre linéaire en fonction du nombre d'éléments du maillage pour trois méthodes : Newton semi-lisse exacte, Newton semi-lisse inexacte et Newton semi-lisse inexacte adaptatif (gauche). Indice d'efficacité en fonction du nombre d'itérations de Newton semi-lisse (droite). Source : Chapter 1 Figure 1.10 et https://hal.archives-ouvertes.fr/hal-01666845.

et plus on avance dans les itérations k, plus notre estimation d'erreur a posteriori est proche de l'erreur exacte.

Ensuite, nous montrons que lorsque nos critères d'arrêts adaptatifs (18) sont vérifiés, nos estimateurs vérifient la propriété d'efficacité locale :

$$\eta_K(\boldsymbol{u}_h^{k,i}) \leq C_{\text{eff}} \left|\!\left|\!\left| \boldsymbol{u} - \boldsymbol{u}_h^{k,i} \right|\!\right|\!\right|_{\zeta_K},$$

sauf pour un terme de contact numériquement négligeable. Pour finir, nous testons numériquement dans le cas des éléments finis $\mathbb{P}_1$ notre approche en utilisant la fonction semi-lisse min (3) combinée avec le solveur itératif GMRES. Nous montrons en particulier que notre approche réduit significativement le nombre total cumulé d'itérations (voir Figure 8 gauche) et que nous obtenons une estimation d'erreur a posteriori optimale.

IV.II Chapitre 2 : Problème évolutif en temps du contact entre deux membranes comme une inégalité variationnelle linéaire parabolique

Après avoir étudié une inégalité variationnelle stationnaire, nous nous intéressons aux inégalités variationnelles paraboliques comme une étape intermédiaire afin de nous rapprocher le plus possible du modèle industriel d'écoulement diphasique avec transition de phase (voir Chapitre 3).

Problème

Il s'agit ainsi d'étudier le problème suivant : Soit $\Omega \subset \mathbb{R}^2$ un domaine polygonal de $\mathbb{R}^2$ et $T > 0$. On considère le problème évolutif en temps : Trouver u_1, u_2 et λ tel

que

$$\begin{cases} \partial_t u_1 - \mu_1 \Delta u_1 - \lambda = f_1 & \text{dans} \quad \Omega \times]0,T[\,, \\ \partial_t u_2 - \mu_2 \Delta u_2 + \lambda = f_2 & \text{dans} \quad \Omega \times]0,T[\,, \\ (u_1 - u_2)\lambda = 0, \quad u_1 - u_2 \geq 0, \quad \lambda \geq 0 & \text{dans} \quad \Omega \times]0,T[\,, \\ u_1 = g & \text{sur} \quad \partial\Omega \times]0,T[\,, \\ u_2 = 0 & \text{sur} \quad \partial\Omega \times]0,T[\,, \\ u_1(\boldsymbol{x},0) = u_1^0(\boldsymbol{x}),\ u_2(\boldsymbol{x},0) = u_2^0(\boldsymbol{x}),\ u_1^0(\boldsymbol{x}) - u_2^0(\boldsymbol{x}) \geq 0 & \text{dans} \quad \Omega. \end{cases} \tag{19}$$

Les deux premières lignes de (19) sont des équations de type parabolique linéaire alors que la troisième ligne de (19) représente les conditions de complémentarités linéaires permettant de traiter deux situations disctinctes: $u_1(\boldsymbol{x},t) - u_2(\boldsymbol{x},t) > 0,\ \lambda(\boldsymbol{x},t) = 0$ ou $u_1(\boldsymbol{x},t) = u_2(\boldsymbol{x},t),\ \lambda(\boldsymbol{x},t) > 0$. Les termes sources $f_1 \in L^2(0,T;L^2(\Omega))$ et $f_2 \in L^2(0,T;L^2(\Omega))$ sont des forces surfaciques et μ_1 et μ_2 sont des contantes strictement positives. A l'instant $t = 0$, les données initiales appartiennent à l'espace $H^1(\Omega)$ et à chaque instant $t > 0$, les conditions de Dirichlet sont imposées pour chaque inconnue. On supposera pour simplifier notre étude que la condition de Dirichlet $g > 0$ est constante en temps et en espace.Ici, les fonctions espace-temps u_1, u_2 décrivent le comportement de la première membrane, respectivement de la seconde membrane et λ est une fonction espace-temps traduisant l'action de la première membrane sur la seconde. Les termes sources $f_1 \in L^2(0,T;L^2(\Omega))$ et $f_2 \in L^2(0,T;L^2(\Omega))$ sont des forces surfaciques et les constantes $\mu_1 > 0$ et $\mu_2 > 0$ les tensions de chaque membrane. A l'instant $t = 0$, les données initiales appartiennent à l'espace $H^1(\Omega)$ et à chaque instant $t > 0$, les conditions de Dirichlet sont imposées pour chaque membrane. On supposera pour simplifier notre étude que $g > 0$ est une constante. Les conditions de complémentarités données par la troisième ligne de (19) indiquent que soit les membranes sont séparées, $u_1(\boldsymbol{x},t) - u_2(\boldsymbol{x},t) > 0,\ \lambda(\boldsymbol{x},t) = 0$, soit elles sont en contact, $u_1(\boldsymbol{x},t) = u_2(\boldsymbol{x},t),\ \lambda(\boldsymbol{x},t) > 0$. Pour l'écriture de la formulation variationnelle de notre problème, nous avons employé deux formulations. La première, de type point-selle, consiste à chercher pour presque tout instant $t \in]0,T[$ les inconnues $u_1 \in L^2(0,T;H_g^1(\Omega))$, $u_2 \in L^2(0,T;H_0^1(\Omega))$ et $\lambda \in L^2(0,T;\Lambda)$ tel que $\partial_t u_\alpha \in L^2(0,T;H^{-1}(\Omega))$, $\alpha = \{1,2\}$, et $\forall (v_1,v_2,\chi) \in H_0^1(\Omega) \times H_0^1(\Omega) \times \Lambda$,

$$\begin{aligned} &\sum_{\alpha=1}^{2} \langle \partial_t u_\alpha(t), v_\alpha \rangle + \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_\alpha(t), \boldsymbol{\nabla} v_\alpha \right)_\Omega - \left(\lambda(t), v_1 - v_2 \right)_\Omega = \sum_{\alpha=1}^{2} \langle f_\alpha, v_\alpha \rangle, \\ &\left(\chi - \lambda(t), u_1(t) - u_2(t) \right)_\Omega \geq 0. \end{aligned} \tag{20}$$

Ici, $\Lambda := \{\lambda \in L^2(\Omega),\ \lambda \geq 0 \quad \text{dans} \quad \Omega\}$. Cette formulation est commode pour discrétiser notre problème et pour l'implémentation numérique qui découle. La deuxième formulation que nous utilisons reviendra à transformer le modèle continu (19) en une inégalité variationnelle parabolique : chercher $\boldsymbol{u} := (u_1,u_2) \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ tel que $\partial_t u_\alpha \in L^2(0,T,H^{-1}(\Omega))$ et $\forall \boldsymbol{v} := (v_1,v_2) \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$

$$\int_0^T \langle \partial_t \boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u} \rangle(t)\,\mathrm{d}t + \int_0^T a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u})(t)\,\mathrm{d}t \geq \int_0^T l(\boldsymbol{v} - \boldsymbol{u})(t)\,\mathrm{d}t, \tag{21}$$

où

$$\boldsymbol{\mathcal{K}}_g^{\mathrm{t}} := \left\{ (v_1,v_2) \in L^2(0,T;H_g^1(\Omega)) \times L^2(0,T;H_0^1(\Omega)),\ v_1 - v_2 \geq 0 \text{ p.p. dans} \, \Omega \times]0,T[\right\}.$$

Dans notre travail, nous proposons une discrétisation en espace faisant appel à la méthode des éléments finis de degré quelconque $p \geq 1$ et une discrétisation en temps utilisant le schéma d'Euler rétrograde. Ensuite, nous construisons des estimations d'erreur a posteriori en suivant la technique de reconstruction des flux équilibrés et la démarche adoptée dans le Chapitre 1.

Etude bibliographique

Notons d'abord que notre formulation ressemble fortement au problème classique de l'obstacle parabolique. Ce type de formulation est richement étudié. On peut mentionner depuis la fin des années 1960 le livre de Lions [121] et plus tard au début des années 1980 l'ouvrage de Glowinski, Lions et Trémolières [96] puis celui de Bensoussan et Lions [25]. On cite également les monographies de Puel [139] et Brezis [36]. En général, la discrétisation du problème (21) repose sur le schéma d'Euler implicite en temps et sur la méthode des éléments finis de Lagrange d'ordre 1 pour la partie spatiale.

Au niveau des méthodes de résolution du système discrétisé (s'écrivant à chaque pas de temps $1 \leq n \leq N_{\mathrm{t}}$ comme le système (2)), on retrouve toutes les techniques utilisées pour la résolution d'une inégalité variationnelle stationnaire : voir [64, 86, 88, 89, 104, 111, 112, 127, 160].

Au niveau de l'analyse a posteriori des inéquations variationnelles paraboliques, le chantier est toujours ouvert et peu de travaux on été effectués à notre connaissance sur le sujet, comparé aux inéquations variationnelles stationnaires. Moon, Nochetto, Petersdorff et Zhang [129] ont étudié un système d'inégalités paraboliques modélisant un problème financier. Le modèle utilisé est celui de Black-Scholes avec contraintes (obstacle). Ils utilisent le schéma d'Euler rétrograde pour discrétiser la dérivée temporelle et la méthode des éléments finis $\mathbb{P}_1$ pour les termes en espace. Ils établissent des estimations d'erreur a posteriori pour la norme d'énergie $L^2(0,T;H^1(\Omega))$. Une discussion a lieu concernant l'efficacité des estimateurs de l'erreur spatiale dans la région de non contact et sur l'efficacité de l'estimateur de l'erreur temporelle.

Nous citons également la contribution d'Achdou, Hecht et Pommier [3] dans laquelle les auteurs étudient un problème d'obstacle parabolique. Comme pour les inégalités variationnelles stationnaires, l'introduction d'un multiplicateur de Lagrange s'avère très utile pour obtenir une équation parabolique avec des contraintes de type complémentarité. Les auteurs utilisent au niveau de la discrétisation temporelle le schéma d'Euler rétrogade et pour les termes en espace la méthode des éléments finis de Lagrange conforme d'ordre 1. Pour la construction du multiplicateur de Lagrange discret, ils utilisent l'opérateur d'interpolation introduit par Chen et Nochetto [53]. Enfin, pour l'analyse a posteriori, la méthode des résidus est employée. En utilisant la norme d'énergie $L^2(0,T;H^1(\Omega))$, ils obtiennent des estimations d'erreur a posteriori sur l'erreur de discrétisation.

Les travaux cités précédemment sur les estimations d'erreur a posteriori pour les inégalités variationnelles paraboliques s'appuient essentiellement sur deux contributions : les travaux effectués sur les estimations d'erreur a posteriori pour les inégalités variationnelles elliptiques stationnaires et les estimations d'erreur a posteriori pour les équations variationnelles paraboliques. La communauté scientifique s'est penchée relativement tardivement sur l'analyse a posteriori de ces problèmes

avec comme problème prototype l'équation de la chaleur. On cite les apports de Verfürth [154], Nicaise et Soualem [130], Bernardi, Bergham et Mghzali [26], Ern et Vohralík [80] et Ern, Smears et Vohralík [78, 79]. La méthode des estimations d'erreur a posteriori par la technique des résidus est utilisée dans [26, 130, 154] alors que dans [78–80] les auteurs utilisent la technique de reconstruction des flux équilibrés. Par ailleurs, l'efficacité locale espace-temps des estimateurs sur l'erreur de discrétisation (12) est prouvée récemment dans [79] ce qui constitue une avancée assez importante. On voit donc l'effet vase communiquant essentiel entre ces différents travaux pour réaliser notre tâche.

Contribution principale

A notre connaissance, le problème (19) n'a jamais été étudié et nous proposons donc d'établir des estimations d'erreur a posteriori pour ce dernier. Nous utilisons la méthodologie introduite dans le Chapitre 1, à savoir la reconstruction des flux équilibrés de discrétisation et d'algèbre linéaire, estimation d'erreur a posteriori valide à chaque pas $k \geq 1$ du solveur de linéarisation semi-lisse et chaque pas $i \geq 0$ du solveur d'algèbre linéaire, distinction des composantes d'erreur et enfin conception d'un algorithme de Newton semi-lisse inexact adaptatif. Une difficulté importante apparaît dans notre analyse a posteriori. En notant $\boldsymbol{u}_{h\tau}^{k,i}$, la solution numérique espace-temps, avec les indices k, i indiquant les solveurs inexacts, il semble très délicat de trouver une borne supérieure entièrement calculable pour le terme $\left\| \partial_t \left(\boldsymbol{u} - \boldsymbol{u}_{h\tau}^{k,i} \right) \right\|_{[L^2(0,T;H^{-1}(\Omega))]^2}$. Nous procédons en plusieurs étapes. Tout d'abord, dans le cadre des éléments finis $\mathbb{P}_1$ et des solveurs exacts, nous fournissons une estimation d'erreur a posteriori dans la norme d'énergie $L^2(0,T;H_0^1(\Omega))$:

$$\|\boldsymbol{u} - \boldsymbol{u}_{h\tau}\|_{[L^2(0,T;H_0^1(\Omega))]^2} \leq \left\{ \sum_{n=1}^{N_\mathrm{t}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\eta_K^n(\boldsymbol{u}_{h\tau}) \right)^2 (t) \, \mathrm{dt} \right\}^{\frac{1}{2}}.$$

Ensuite, nous approchons la norme d'énergie $H^{-1}(\Omega)$ en dérivée temporelle par la norme d'énergie dans l'espace $[L^2(0,T;H_0^1(\Omega))]^2$ d'un objet auxilliaire issu d'une inégalité variationnelle auxilière. Enfin, nous proposons une estimation d'erreur a posteriori valide à chaque pas $k \geq 1$ et $i \geq 0$ et pour des degrés polynomiaux quelconques $p \geq 1$ de la norme d'énergie sous la forme :

$$\left\| \boldsymbol{u} - \boldsymbol{u}_{h\tau}^{k,i} \right\|_{[L^2(0,T;H_0^1(\Omega))]^2} \leq \left\{ \sum_{n=1}^{N_\mathrm{t}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\eta_K^n(\boldsymbol{u}_{h\tau}^{k,i}) \right)^2 (t) \, \mathrm{dt} \right\}^{\frac{1}{2}}.$$

Cela constitue une autre nouveauté de notre travail car les travaux cités plus haut n'estiment que l'erreur de discrétisation.

IV.III Chapitre 3 : Ecoulement diphasique avec transition de phase comme une inégalité variationnelle non linéaire parabolique

Problème

Dans ce dernier chapitre, nous nous intéressons à la problématique du stockage des déchets radioactifs dans les couches géologiques profondes. Comme expliqué au préalable, ce stockage provoque entre autre une éjection de gaz responsable en partie de la dégradation de l'environnement. Pour décrire ce phénomène physique, nous considérons une situation simplifiée décrite par un écoulement diphasique compositionel liquide-gaz où les deux composants présents sont l'eau et l'hydrogène. On suppose que l'eau n'est que présente dans la phase liquide. L'hydrogène peut exister sous deux formes : liquide et gaz.

Pour illustrer notre propos, notons $\mathcal{P}$ l'ensemble des phases présentes et $\mathcal{C}$ l'ensemble des composants présents. De même, notons $\mathcal{P}_c$ l'ensemble des phases contenant le composant c et $\mathcal{C}^p$ l'ensemble des composants présents dans la phase $p \in \mathcal{P}$. Le symbole "w" est associé à l'eau et le symbole "h" à l'hydrogène. De même, le symbole "l" est associé à la phase liquide alors que le symbole "g" est associé à la phase gazeuse. Ainsi, nous avons $\mathcal{P} = \{\mathrm{l}, \mathrm{g}\}$, $\mathcal{C} = \{\mathrm{w}, \mathrm{h}\}$, $\mathcal{C}^{\mathrm{l}} = \{\mathrm{w}, \mathrm{h}\}$, $\mathcal{C}^{\mathrm{g}} = \{\mathrm{h}\}$. En fait, au fil des années, les matériaux stockés en sous-sol vont dégager de l'hydrogène gazeux ce qui correspondra à la période diphasique. Au début de la simulation, l'hydrogène sera présent en trop faible quantité et sera complètement dissous dans la phase liquide ; cela correspond à un régime monophasique liquide où l'hydrogène gazeux n'a pas encore apparu. On choisit comme inconnues principales la saturation liquide S^{l}, la pression liquide P^{l} et la fraction molaire d'hydrogène liquide $\chi^{\mathrm{l}}_{\mathrm{h}}$. Le système étudié est le suivant:

$$
\begin{aligned}
&\partial_t(\phi\rho^{\mathrm{l}}_{\mathrm{w}}S^{\mathrm{l}} + \phi\rho^{\mathrm{g}}_{\mathrm{w}}S^{\mathrm{g}}) + \boldsymbol{\nabla}\cdot(\rho^{\mathrm{l}}_{\mathrm{w}}\mathbf{q}^{\mathrm{l}} + \rho^{\mathrm{g}}_{\mathrm{w}}\mathbf{q}^{\mathrm{g}} + \mathbf{J}^{\mathrm{l}}_{\mathrm{w}} + \mathbf{J}^{\mathrm{g}}_{\mathrm{w}}) = Q_{\mathrm{w}},\\
&\partial_t(\phi\rho^{\mathrm{l}}_{\mathrm{h}}S^{\mathrm{l}} + \phi\rho^{\mathrm{g}}_{\mathrm{h}}S^{\mathrm{g}}) + \boldsymbol{\nabla}\cdot(\rho^{\mathrm{l}}_{\mathrm{h}}\mathbf{q}^{\mathrm{l}} + \rho^{\mathrm{g}}_{\mathrm{h}}\mathbf{q}^{\mathrm{g}} + \mathbf{J}^{\mathrm{l}}_{\mathrm{h}} + \mathbf{J}^{\mathrm{g}}_{\mathrm{h}}) = Q_{\mathrm{h}},\\
&1 - S^{\mathrm{l}} \geq 0,\ H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}}\chi^{\mathrm{l}}_{\mathrm{h}} \geq 0,\ [1 - S^{\mathrm{l}}]\cdot\left[H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}}\chi^{\mathrm{l}}_{\mathrm{h}}\right] = 0.
\end{aligned}
\tag{22}
$$

Tous les détails seront donnés dans le Chapitre 3, mais notons que les deux premières lignes de (22) forment les équations de conservation pour chaque composant : eau et hydrogène. Les termes associés aux dérivées temporelles décrivent l'accumulation des composants, alors que ceux associés à l'opérateur divergence correspondent aux flux des composants. Les flux pour chaque composant sont définis par la somme des vitesses de Darcy $(\mathbf{q}^p)_{p\in\mathcal{P}_c}$ pour chaque phase contenant le composant considéré et des flux de Fick $(\mathbf{J}^p_c)_{c\in\mathcal{P}_c}$ associés au composant considéré. La troisième ligne de (22) correspond aux contraintes de complémentarité non linéaires traduisant la transition de phase : si la contrainte $1 - S^{\mathrm{l}} > 0$ est active, on se situe dans un régime diphasique et donc $H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}}\chi^{\mathrm{l}}_{\mathrm{h}} = 0$. Si la contrainte $1 - S^{\mathrm{l}} = 0$ est active, l'écoulement est monophasique liquide et $H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}}\chi^{\mathrm{l}}_{\mathrm{h}} > 0$. Ici H désigne la constante de Henry. Dans notre travail, nous proposons pour (22), une discrétisation spatiale par la méthode des volumes finis centrés sur les mailles et une discrétisation temporelle utilisant le schéma d'Euler rétrograde. Ensuite,

comme pour les chapitres précédents, nous utilisons des solveurs semi lisses inexacts pour traiter les non linéarités. Nous construisons enfin des estimations d'erreur a posteriori dans le but d'estimer chaque composante d'erreur issue de la simulation numérique.

Etude bibliographique

Le modèle mathématique (22) a été introduit par Jaffré et Sboui [107]. Ils utilisent un schéma volume fini 1D pour la discrétisation en espace et le schéma d'Euler rétrograde pour la discrétisation en temps. A l'issue de la discrétisation, ils obtiennent en particulier le problème discret sous la forme (2) avec $N = 3N_{\text{sp}}$ et $M = N_{\text{sp}}$, où N_{sp} désigne le nombre de cellules. L'étude de (22) a été repris ensuite par Ben Gharbia et Jaffré [24]. Ils réécrivent les contraintes de complémentarités non linéaires données par la troisième ligne de (22) via des fonctions semi-lisses. Comme expliqué au préalable, une fonction semi-lisse possède un caractère différentiable plus faible que celui de Fréchet. Les auteurs de [24] utilisent plus précisément la fonction semi-lisse "min" (3) générant ainsi un algorithme de Newton-min. Les résultats numériques de ce travail montrent que sous le choix de conditions initiales "raisonnables", ce modèle est assez proche de la réalité physique observée. Dans la même catégorie, on peut mentionner l'article de Lauser, Hager, Helmig et Wohlmuth [119] adoptant une approche très générale de résolution d'un système d'équations à plusieurs phases et transitions de phases en milieu poreux.

Dans la littérature, un modèle également très étudié est le modèle multiphasique sans composant :

$$\partial_t \left(\phi S^p\right) + \boldsymbol{\nabla} \cdot \mathbf{q}^p = Q^p \quad \forall p \in \mathcal{P}, \tag{23}$$

où S^p est la saturation de la phase p, $\mathbf{q}^p$ est la vitesse de Darcy de la phase p et ϕ la porosité du milieu poreux. Les inconnues pour cette formulation sont la saturation de la phase p notée S^p et la pression de la phase p notée P^p. Ce problème très général a été étudié depuis plusieurs décennies et la littérature sur ce sujet est très riche et vaste. Nous mentionnons les ouvrages de Chavent et Jaffré [49] pour la réécriture de (23) via la pression globale, Chen, Huan et Ma [52], Chen [51], Gross et Reusken [98] et Sha [151] pour une liste exhaustive des diverses méthodes numériques applicables au problème (23). On peut citer également l'article d'Epshteyn et Rivière [75] s'appuyant sur la formulation pression globale introduite par Chavent et Jaffré [49]. Toujours dans [75], les auteurs utilisent pour la discrétisation en temps le schéma d'Euler rétrograde et pour la discrétisation en espace la méthode de Galerkine discontinue. Le degré d'approximation polynomial de la pression globale P_{glob} n'est pas le même que celui de la saturation de la phase non mouillante S_{n} (deuxième inconnue). Ensuite, des estimations d'erreur a priori pour chaque inconnue sont fournies.

Quelques années plus tard, Rankin et Rivière [141] étudient le modèle "black-oil". Il s'agit d'un modèle à trois phases (liquide, vapeur et aqueuse) et trois composants (l'huile, le gaz et l'eau). Plus précisément, le gaz existe sous forme liquide et vapeur alors que l'eau (respectivement l'huile) existe dans la phase aqueuse (respectivement dans la phase liquide). Les auteurs utilisent le schéma d'Euler rétrograde pour discrétiser la dérivée temporelle, alors que la discrétisation de la partie spatiale

repose sur la méthode de Galerkine discontinue.

Un volet très important dans l'étude des formulations multiphasiques est l'implémentation des équations discrétisées. On imagine bien que pour un maillage très fin, la taille des matrices va rapidement devenir gigantesque surtout si le modèle en question possède beaucoup de phases et composants. Des travaux ont été proposés pour affiner et réduire le coût de la résolution numérique. On peut mentionner l'article de Yotov [161] qui transforme la résolution du problème très coûteux en un problème d'interface. Ce problème d'interface est résolu via une méthode de Newton-GMRES inexacte et un algorithme multigrille (Cycle V) avec comme lisseur un solveur de type Newton-GMRES. Cette approche est originale car généralement les lisseurs utilisés dans le cadre de l'algorithme multigrille sont les algorithmes de Jacobi et Gauss-Seidel.

Une difficulté majeure de notre formulation diphasique (22) et plus généralement des formulation multiphasiques est de prouver un théorème d'existence et unicité d'une solution faible. En conséquence, pour l'analyse a posteriori, il est très délicat de définir une norme d'erreur afin d'obtenir des estimations du type:

$$\left|\!\left|\!\left| P^{\mathrm{l}} - P^{\mathrm{l}}_{h\tau} \right|\!\right|\!\right| + \left|\!\left|\!\left| S^{\mathrm{l}} - S^{\mathrm{l}}_{h\tau} \right|\!\right|\!\right| + \left|\!\left|\!\left| \chi^{\mathrm{l}}_{\mathrm{h}} - \chi^{\mathrm{l}}_{\mathrm{h},h\tau} \right|\!\right|\!\right| \leq \eta(P^{\mathrm{l}}_{h\tau}, S^{\mathrm{l}}_{h\tau}, \chi^{\mathrm{l}}_{\mathrm{h},h\tau}) \tag{24}$$

où $|||\cdot|||$ est une norme et $S^{\mathrm{l}}_{h\tau}$, $P^{\mathrm{l}}_{h\tau}$, respectivement $\chi^{\mathrm{l}}_{\mathrm{h},h\tau}$ des fonctions espaces-temps approximant les inconnues S^{l}, P^{l}, respectivement $\chi^{\mathrm{l}}_{\mathrm{h}}$.

Les estimations d'erreur a posteriori pour les modèles multiphasiques ont fait l'objet d'une attention toute particulière au cours de cette dernière décennie. Nous renvoyons les lecteurs aux travaux de Vohralík et Wheeler [158] pour des estimations d'erreur a posteriori utilisant un modèle diphasique incompressible reformulé par la pression globale. Ce travail inclut une distinction des diverses composantes d'erreur et des critères d'arrêts adaptatifs pour chaque solveur utilisé. On cite également Cancès, Pop et Vohralík [44] pour une étude théorique traitant un cadre où l'on bénéficie de l'existence et unicité d'une solution faible pour un écoulement diphasique avec un composant par phase, et [69, 71] pour une généralisation à des écoulements multiphasiques compositionnels.

Contribution principale

Les auteurs des différents travaux cités ci-dessus utilisent comme mesure de l'erreur la norme duale du résidu complétée par des termes décrivant la non conformité des inconnues. Cependant, ils ne considèrent pas la transition de phase. C'est ce que nous développons en détail dans ce chapitre, en introduisant un estimateur de contraintes ou de transition de phase détectant dans quelles cellules le gaz apparaît (voir Figure 9 gauche). Dans le même esprit que dans les chapitres précédents, nous utilisons des algorithmes de Newton semi-lisse conduisant à résoudre à chaque pas de temps $1 \leq n \leq N_{\mathrm{t}}$ le problème linéaire

$$\mathbb{A}^{n,k-1} \boldsymbol{X}_h^{n,k} = \boldsymbol{F}^{n,k-1}. \tag{25}$$

Nous privilégions à nouveau l'emploi des solveurs itératifs à chaque pas $k \geq 1$ de Newton semi-lisse, donnant lieu à étudier le résidu $\boldsymbol{R}_h^{n,k,i}$ défini par

$$\boldsymbol{R}_h^{n,k,i} := \boldsymbol{F}^{n,k-1} - \mathbb{A}^{n,k-1} \boldsymbol{X}_h^{n,k,i}.$$

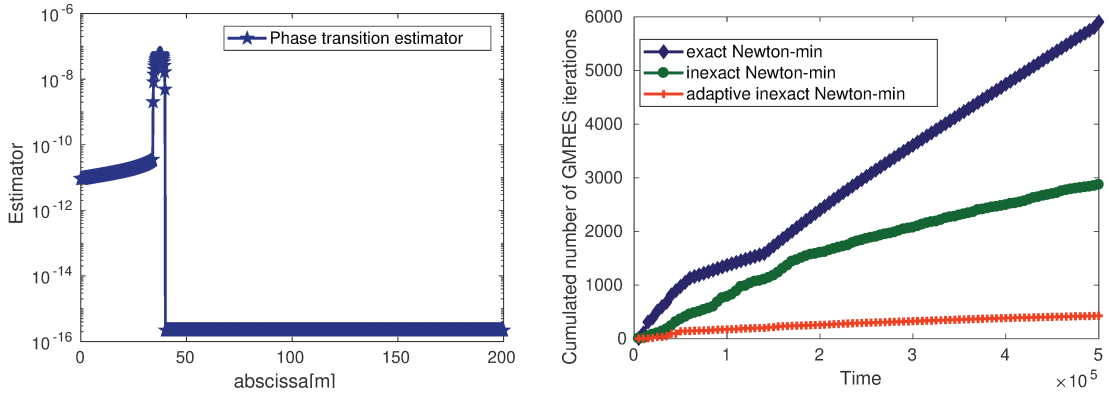

Figure 9: Estimateur de transition de phase (gauche). L'estimateur s'active dans les cellulles ou le gaz apparaît. Nombre cumulé d'itérations de GMRES en fonction du temps (droite). Source : Chapitre 3 Figure 3.8, Figure 3.12 et https://hal.archives-ouvertes.fr/hal-01919067.

Notons également que dans [24], les auteurs ont fait appel au différenciateur automatique ADMAT pour calculer les matrices Jacobiennes à chaque pas $k \geq 1$ de Newton. Dans cette thèse, nous avons proposé une implémentation numérique ne faisant pas appel à cet outil. L'analyse a posteriori de notre modèle repose sur l'hypothèse d'existence d'une solution faible. De plus, nous ne pouvons pas, contrairement aux chapitres précédents, obtenir une borne supérieure d'une norme d'énergie sous forme (24) . Nous définissons en revanche une mesure de l'erreur égale à une norme duale de résidu complétée par un résidu des contraintes de complémentarité et par des termes évaluant la non conformité des pressions et de la fraction molaire. Nous obtenons ainsi une estimation d'erreur a posteriori valide à chaque pas de temps $1 \leq n \leq N_t$, à chaque pas $k \geq 1$ d'un solveur de Newton semi-lisse et à chaque pas $i \geq 0$ d'un solveur itératif algébrique. En particulier nous parvenons de nouveau à estimer les différentes composantes d'erreur et proposer un algorithme de Newton semi-lisse inexact adaptatif.

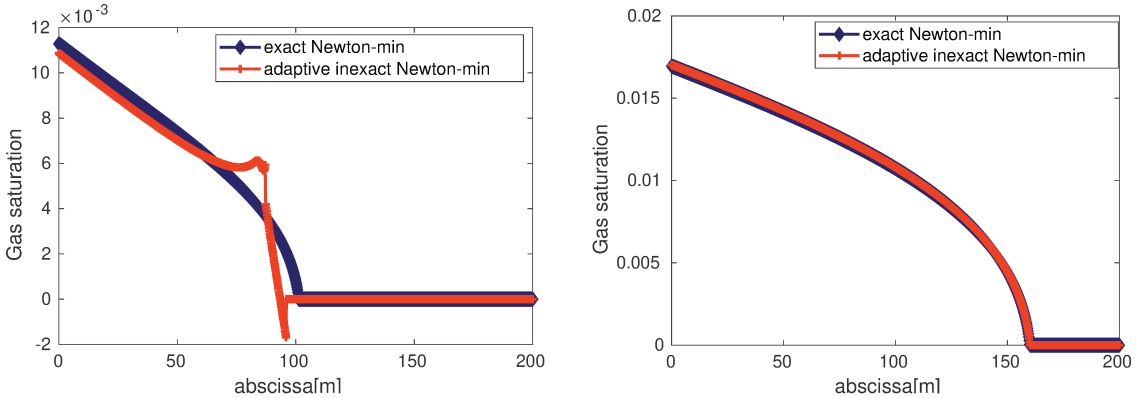

Figure 10: Saturation du gaz au début du régime diphasique (gauche), saturation du gaz vers la fin du régime diphasique (droite). Source : Chapitre 3 Figure 3.13 et https://hal.archives-ouvertes.fr/hal-01919067.

Les simulations sont effectuées en Matlab en dimension une d'espace en utilisant

les fonctions semi-lisse min (3) et Fischer–Burmeister (4) et l'algorithme de GMRES. Nous montrons en particulier que notre démarche réduit significativement le nombre total cumulé d'itérations du solveur algébrique de GMRES (voir Figure 9 droite) et qu'elle préserve la précision de la solution numérique (voir Figure 10).

Chapter 1

Problem of contact between two membranes as a stationnary variational inequality

Abstract

We propose an adaptive inexact version of a class of semismooth Newton methods. As a model problem, we study the system of variational inequalities describing the contact between two membranes. This problem is discretized with conforming finite elements of order $p \geq 1$, yielding a nonlinear algebraic system of PDE'S with complementarity constraints. We consider any iterative semismooth linearization algorithm like the Newton-min or the Newton–Fischer–Burmeister which we complement by any iterative linear algebraic solver. We then derive an a posteriori estimate on the error between the exact solution and the approximate solution which is valid at any step of the linearization and algebraic resolutions. Our estimate is based on flux reconstructions in discrete subspaces of $\mathbf{H}(\mathrm{div}, \Omega)$ and on potential reconstructions in discrete subspaces of $H^1(\Omega)$ satisfying the constraints. It distinguishes the discretization, linearization, and algebraic components of the error. Consequently, we can formulate adaptive stopping criteria for both solvers, giving rise to an adaptive version of the considered inexact semismooth Newton algorithm. Under these criteria, the efficiency of our estimates is also established, meaning that we prove them equivalent with the error up to a generic constant, except for a typically small contact term. Numerical experiments for the Newton-min algorithm in combination with the GMRES algebraic solver confirm the efficiency of our adaptive method.

Keywords: variational inequality, complementarity condition, contact problem, semismooth Newton method, a posteriori estimate, adaptivity, stopping criterion

In this first chapter, we study the stationnary problem of contact between two membranes. The contents presented is taken from the article [62] submitted for publication, with some additional results. We provide in particular more details concerning the continuous problem (Section 1.2.1) and we give the expression of the discrete complementarity constraints in the Lagrange basis as well as its dual basis (Section 1.2.3) for any polynomial degree $p \geq 1$.

1.1 Introduction

Consider a system of algebraic inequalities written in the following form: find a vector $\boldsymbol{X}_h \in \mathbb{R}^n$, such that

$$\begin{array}{rll} \mathbb{E}\boldsymbol{X}_h & = \boldsymbol{F}, & \\ \boldsymbol{K}(\boldsymbol{X}_h) & \geq \boldsymbol{0}, \quad \boldsymbol{G}(\boldsymbol{X}_h) \geq \boldsymbol{0}, & \quad \boldsymbol{K}(\boldsymbol{X}_h) \cdot \boldsymbol{G}(\boldsymbol{X}_h) = 0, \end{array} \tag{1.1}$$

where, for some integers $n > 1$ and $0 < m < n$, $\mathbb{E} \in \mathbb{R}^{n-m,n}$ is a matrix, $\boldsymbol{K} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{R}^m$ are affine operators, and $\boldsymbol{F} \in \mathbb{R}^{n-m}$ is a given vector. The first line of (1.1) typically represents the discretization of a linear partial differential equation (PDE) (the model example for this study is described further in (1.5)). The second line of (1.1) represents linear complementarity constraints (called linear even if the Euclidean scalar product $\boldsymbol{K}(\boldsymbol{X}_h)\cdot\boldsymbol{G}(\boldsymbol{X}_h)$ is not linear) and states that the vectors $\boldsymbol{K}(\boldsymbol{X}_h)$ and $\boldsymbol{G}(\boldsymbol{X}_h)$ have non-negative components and are orthogonal, *i.e.* $\boldsymbol{G}(\boldsymbol{X}_h)\cdot\boldsymbol{K}(\boldsymbol{X}_h) = 0$. Numerous algorithms have been developed in the past for approximate solution of (1.1), see for example the overview of Facchinei and Pang [88, 89], the book of Bonnans *et al.* [29], and the survey of Aganagić [4]. In particular, we mention the approach by interior point method of Wright [160], the active set strategy by Kanzow [111], and the primal-dual active set strategy by Hintermüler *et al.* [104]. Another approach that will be used here is to rewrite the complementarity conditions as a system of nonsmooth nonlinear equations by the means of C-functions, see for instance [64, 87–89]. The C-functions are not smooth in the classical sense (Fréchet-differentiable), but admit a weaker smoothness called the Clarke derivative, see [58]. Semismooth Newton algorithms like the Newton-min are often used in practice as they show local quadratic convergence properties, see [21–23, 88, 89].

The goal of the present study is to perform an a posteriori analysis of problem (1.1), where $\mathbb{E}$ is given by a discretization of a PDE and the complementarity constraints are treated using semismooth C-functions, and to derive an inexact semismooth Newton algorithm with adaptive stopping criteria. Assume that any C-function is applied to (1.1). This yields an equivalent formulation of (1.1) that requests to find a vector $\boldsymbol{X}_h \in \mathbb{R}^n$ such that

$$\mathcal{S}(\boldsymbol{X}_h) = \boldsymbol{0}, \tag{1.2}$$

where $\mathcal{S} : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear non-differentiable Fréchet functional. Next, let any semismooth nonlinear solver be applied to system (1.2), yielding at a semismooth step $k \geq 1$ a linear system

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^k = \boldsymbol{B}^{k-1}, \tag{1.3}$$

where $\mathbb{A}^{k-1} \in \mathbb{R}^{n,n}$ is a matrix and $\boldsymbol{B}^{k-1} \in \mathbb{R}^n$ is a vector. Finally, let any iterative algebraic solver be applied to (1.3), yielding at step $i \geq 1$ an approximation $\boldsymbol{X}_h^{k,i}$ to $\boldsymbol{X}_h$. Note that $\boldsymbol{X}_h^{k,i}$ does not solve (1.3) but only

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^{k,i} = \boldsymbol{B}^{k-1} - \boldsymbol{R}^{k,i},$$

where $\boldsymbol{R}^{k,i} = \boldsymbol{B}^{k-1} - \mathbb{A}^{k-1}\boldsymbol{X}_h^{k,i} \in \mathbb{R}^n$ is the algebraic residual vector of (1.3). Similarly, $\boldsymbol{X}_h^{k,i}$ does not solve (1.2) as $\mathcal{S}(\boldsymbol{X}_h^{k,i}) \neq 0$ in general.

An important amount of work has been performed in the last years on a posteriori analysis of partial differential equations (see for instance the contributions of Prager and Synge [138] and Ladevèze [118], and the books of Verfürth [155], Ainsworth [5] and Repin [143] for a general introduction). Concerning general error estimates for variational inequalities discretized as in (1.1) or (1.2), let us mention the pioneering work of Brezzi, Hager, and Raviart [39, 40]. Next, for reliable a posteriori error estimates we mention Ainsworth, Oden, and Lee [6], Kornhuber [116], Repin [144] and Bürg and Schröder [43]. For the elliptic obstacle problem we can more precisely mention the papers of Chen and Nochetto [53], see also the references therein, Veeser [153], Bartels and Carstensen [12], and Braess [31] for linear finite elements, Gudi and Porwal [99–101] for discontinuous Galerkin methods, and Chouly, Fabre, Hild, Pousin and Renard [55] for Nitsche's method in elasticity with constraint. Not to solve (1.3) exactly or with a high precision leads to the concept of an inexact semismooth Newton method. Such approaches are heavily used in practice and theoretical foundations can be found in [42,65,72,92,114] for the case of inexact Newton methods and in [86, 88, 89, 112, 127] for inexact semismooth Newton methods. All these approaches do not take into account the discretization error of the PDE by the given numerical scheme. In this work, we focus on inexact semismooth solutions of (1.3) and (1.2). We follow the approach by equilibrated flux reconstructions, solving auxiliary local problems (see Destuynder and Métivet [66] and Braess and Schöberl [33]). A reconstruction of the primal variable satisfying the constraints on the given step $k \geq 1$, $i \geq 1$, will also be performed. More precisely, following the concepts from Becker, Johnson, Rannacher [15], Coorevits, Hild, Pelle [60], Louf, Combe, Pelle [123], Jiránek *et al.* [108], Ern and Vohralík [81], Arioli, Georgoulis, Loghin [9], and Papež *et al.* [136], our estimate takes the form

$$e(\boldsymbol{X}_h^{k,i}) \leq \eta(\boldsymbol{X}_h^{k,i}) = \eta_{\mathrm{disc}}^{k,i} + \eta_{\mathrm{lin}}^{k,i} + \eta_{\mathrm{alg}}^{k,i}, \tag{1.4}$$

where $e(\boldsymbol{X}_h^{k,i})$ stands for the error between the approximation corresponding to the algebraic vector $\boldsymbol{X}_h^{k,i}$ and the unknown exact solution of the continuous problem. The a posteriori error estimate $\eta(\boldsymbol{X}_h^{k,i})$, fully computable from $\boldsymbol{X}_h^{k,i}$, enables to distinguish the components of the error, caused by the discretization, the linearization, and the algebraic resolution. The proposed criteria then request to stop the algebraic (respectively linearization) solver whenever the algebraic estimator $\eta_{\mathrm{alg}}^{k,i}$ (respectively the linearization estimator $\eta_{\mathrm{lin}}^{k,i}$) does not contribute significantly to the overall estimator $\eta(\boldsymbol{X}_h^{k,i})$. Local stopping criteria are derived as well. We thus conceive an adaptive inexact version of an important class of semismooth Newton methods and answer the practical questions: 1) to which precision should (1.3) and (1.2) be solved? 2) what is the error in $\boldsymbol{X}_h^{k,i}$?

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain. We exemplify the above approach with the following problem that models the contact between two membranes: find u_1, u_2, and λ such that

$$\left\{\begin{array}{llll} -\mu_1 \Delta u_1 - \lambda = f_1 & \text{in} & \Omega, & \\ -\mu_2 \Delta u_2 + \lambda = f_2 & \text{in} & \Omega, & \\ (u_1 - u_2)\lambda = 0, \quad u_1 - u_2 \geq 0, \quad \lambda \geq 0 & \text{in} & \Omega, & \\ u_1 = g & \text{on} & \partial\Omega, & \\ u_2 = 0 & \text{on} & \partial\Omega, & \end{array}\right. \tag{1.5}$$

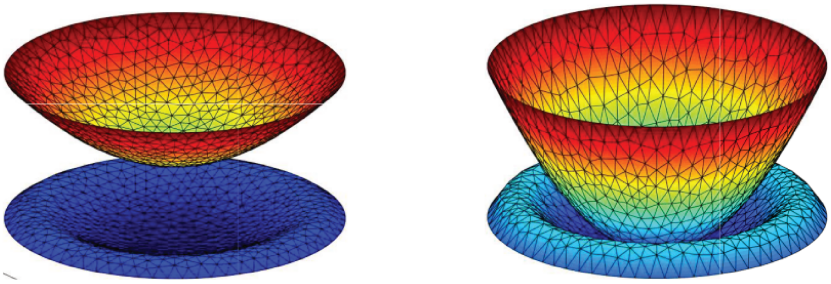

Figure 1.1: Position of the two membranes in two different configurations. Left figure: the membranes are separated. Right figure: The membranes are in contact.

where u_1 and u_2 represent vertical displacements of the two membranes and λ is a Lagrange multiplier characterizing the action of the second membrane on the first one. The constant parameters $\mu_1, \mu_2 > 0$ correspond to the tension of the membranes, and $f_1, f_2 \in L^2(\Omega)$ are given external forces. The boundary condition prescribed by a constant $g > 0$ ensures that the first membrane is above the second one on $\partial\Omega$. In (1.5), the two first equations represent the kinematic behavior of the membranes, and the third one represents the linear complementarity conditions saying that either the membranes are separated ($u_1 > u_2$, $\lambda = 0$), or they are in contact ($u_1 = u_2$, $\lambda \geq 0$), see Figure 1.1. This kind of formulation, and the closely related but simpler elliptic obstacle problems where the goal is to find the equilibrium position of a single elastic membrane constrained to lie below or above some given obstacle, is well understood today, see, *e.g.*, Hlaváček *et al.* [105] and Rodrigues [148] for general concepts. For the quadratic finite element approximation of the Signorini problem refer to [10, 16]. The existence and uniqueness of a weak solution of (1.5) follows by Lions and Stampacchia [122], an a priori analysis for linear finite elements was performed in [17, 18], and an a posteriori analysis was undertaken in [19]. Therein, however, it was supposed that the discrete system (1.1) is solved exactly for continuous and piecewise linear elements (polynomial degree $p = 1$). In the present paper, two new difficulties are treated: first, the inexact solve leads to approximate solutions that do not fulfill the constraints even when $p = 1$. The second difficulty is caused by the nonconformity of the method when $p > 1$. Indeed, the approximate solution is sought in a convex set which is not a subset of the continuous convex set. For this reason, the analysis in the literature is often performed for linear finite elements only which skirts this difficulty.

This contribution is organized as follows. In Section 1.2, we give details on the model problem (1.5) and its finite element discretization for all polynomial degrees $p \geq 1$, which is new to the best of our knowledge when $p > 1$. In contrast to [62], we provide in this thesis the expression of the discrete complementarity constraints in the Lagrange basis and its dual basis (see Section 1.2.3) for any polynomial degree $p \geq 1$. This leads to algebraic systems of the form (1.1). In Section 1.3, we present the concept of the inexact semismooth Newton method giving rise to systems (1.2)–(1.3). The various flux reconstructions, in particular employing a multilevel mesh hierarchy for the algebraic error components following Papež *et al.* [136], are described in Section 1.4. Next, Section 1.5 is dedicated to the construction of the a posteriori error estimate of the form (1.4). In Section 1.6, we present the

adaptive inexact semismooth algorithm and in Section 1.7, we prove the converse inequality to (1.4) up to a generic constant and up to a typically small contact term, assessing the quality of our estimates. Finally, Section 1.8 is devoted, for $p = 1$, to numerical experiments that confirm the theoretical results and Section 1.10 summarizes our conclusions.

1.2 Model problem and its finite element discretization

In this section, we set up the notation, describe in details the model problem (1.5), and introduce its finite element discretization for all polynomial degrees $p \geq 1$. We first recall the definition of some functional spaces. Let $\mathcal{D}(\Omega)$ be the space of $\mathcal{C}^\infty$ functions with compact support on Ω and $\mathcal{D}'(\Omega)$ the dual space of $\mathcal{D}(\Omega)$. Let $H^1(\Omega)$ be the space of L^2 functions on the domain Ω which admit a weak gradient in $[L^2(\Omega)]^2$ and $H^1_0(\Omega)$ its zero-trace subspace. Similarly, $\mathbf{H}(\mathrm{div}, \Omega)$ stands for the space of $[L^2(\Omega)]^2$ functions having a weak divergence in $L^2(\Omega)$. Moreover, we define the sets

$$H^1_g(\Omega) := \left\{v \in H^1(\Omega),\ v = g \text{ on } \partial\Omega\right\} \quad \text{and} \quad \Lambda := \left\{\chi \in L^2(\Omega),\ \chi \geq 0 \text{ a.e. in } \Omega\right\}.$$

The standard notations $\boldsymbol{\nabla}$ and $\boldsymbol{\nabla}\cdot$ are used respectively for the weak gradient and divergence operators. For a nonempty set $\mathcal{O}$ of $\mathbb{R}^2$, we denote its Lebesgue measure by $|\mathcal{O}|$ and the $L^2(\mathcal{O})$ scalar product by $(u, v)_\mathcal{O} := \int_\mathcal{O} uv \,\mathrm{dx}$ for $u, v \in L^2(\mathcal{O})$. We also use the following notations: $\|v\|^2_\mathcal{O} := (v, v)_\mathcal{O}$, and $\|\boldsymbol{\nabla} v\|^2_\mathcal{O} := (\boldsymbol{\nabla} v, \boldsymbol{\nabla} v)_\mathcal{O}$. Besides, the Poincaré–Friedrichs and the Poincaré–Wirtinger inequalities, see [14, 137], state that if $\overline{v}_\mathcal{O}$ denotes the mean value of v and $h_\mathcal{O}$ the diameter of $\mathcal{O}$, then

$$\|v\|_\mathcal{O} \leq C_{\mathrm{PF}} h_\mathcal{O} \|\boldsymbol{\nabla} v\|_\mathcal{O} \qquad \forall v \in H^1_0(\mathcal{O}), \tag{1.6a}$$
$$\|v - \overline{v}_\mathcal{O}\|_\mathcal{O} \leq C_{\mathrm{PW}} h_\mathcal{O} \|\boldsymbol{\nabla} v\|_\mathcal{O} \qquad \forall v \in H^1(\mathcal{O}). \tag{1.6b}$$

The constants C_{PF} and C_{PW} can be precisely estimated in many cases. In particular, if $\mathcal{O}$ is convex, C_{PW} can be taken as $\frac{1}{\pi}$, see [14, 137] whereas C_{PF} is at most 1. Then, we define the energy norm:

$$\forall \boldsymbol{v} = (v_1, v_2) \in [H^1_0(\mathcal{O})]^2, \quad |||\boldsymbol{v}|||_\mathcal{O} := \left\{\sum_{\alpha=1}^{2} \mu_\alpha \|\boldsymbol{\nabla} v_\alpha\|^2_\mathcal{O}\right\}^{\frac{1}{2}}. \tag{1.7}$$

When $\mathcal{O} = \Omega$, we use the shorthand notation $|||\boldsymbol{v}||| := |||\boldsymbol{v}|||_\mathcal{O}$. We also define the following rescaling of the $H^{-1}(\mathcal{O})$ norm:

$$\forall v \in H^{-1}(\mathcal{O}), \quad |||v|||_{H^{-1}_*(\mathcal{O})} := \sup_{\psi \in H^1_0(\mathcal{O}),\ \max(\mu_1^{\frac{1}{2}}, \mu_2^{\frac{1}{2}}) \|\boldsymbol{\nabla}\psi\|_\mathcal{O} = 1} \langle v, \psi \rangle. \tag{1.8}$$

1.2.1 Continuous and reduced variational problem

For $(f_1, f_2) \in [L^2(\Omega)]^2$ and g a positive constant, the weak formulation corresponding to (1.5) consists in: find $(u_1, u_2, \lambda) \in H^1_g(\Omega) \times H^1_0(\Omega) \times \Lambda$ such that

$$\begin{cases} \displaystyle\sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_\alpha, \boldsymbol{\nabla} v_\alpha\right)_\Omega - \left(\lambda, v_1 - v_2\right)_\Omega = \sum_{\alpha=1}^{2} \left(f_\alpha, v_\alpha\right)_\Omega \quad \forall\, (v_1, v_2) \in [H^1_0(\Omega)]^2, \\ \left(\chi - \lambda, u_1 - u_2\right)_\Omega \geq 0 \quad \forall \chi \in \Lambda. \end{cases} \tag{1.9}$$

Setting $\boldsymbol{u} := (u_1, u_2)$, $\boldsymbol{v} := (v_1, v_2)$, and defining

$$a(\boldsymbol{u}, \boldsymbol{v}) := \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_\alpha, \boldsymbol{\nabla} v_\alpha\right)_\Omega, \; b(\boldsymbol{v}, \chi) := \left(\chi, v_1 - v_2\right)_\Omega, \; l(\boldsymbol{v}) := \sum_{\alpha=1}^{2} \left(f_\alpha, v_\alpha\right)_\Omega, \tag{1.10}$$

problem (1.9) rewrites: find $\boldsymbol{u} \in H^1_g(\Omega) \times H^1_0(\Omega)$ and $\lambda \in \Lambda$ such that

$$\begin{cases} a(\boldsymbol{u}, \boldsymbol{v}) - b(\boldsymbol{v}, \lambda) &= l(\boldsymbol{v}) & \forall \boldsymbol{v} \in [H^1_0(\Omega)]^2, \\ b\left(\boldsymbol{u}, \chi - \lambda\right) &\geq 0 & \forall \chi \in \Lambda. \end{cases}$$

Now, we prove the equivalence between the continuous problem (1.5) and its variational expression (1.9). Note that the proof is already available in [17, Proposition 1], but we add more details to facilitate the reading.

Proposition 1.2.1. *Problems (1.5) and (1.9) are equivalent in the sense that any triple $(u_1, u_2, \lambda) \in H^1_g(\Omega) \times H^1_0(\Omega) \times \Lambda$ is a solution of (1.5) if and only if it is a solution of (1.9).*

Proof. The proof is standard and follows the lines of [17, Proposition 1]. Assume that $(u_1, u_2, \lambda) \in H^1_g(\Omega) \times H^1_0(\Omega) \times \Lambda$ is a solution of (1.5). Multiplying the first line of (1.5) by a test function $v_1 \in H^1_0(\Omega)$ and the second line of (1.5) by a test function $v_2 \in H^1_0(\Omega)$, summing these two equations and employing the Green formula gives the first line of (1.9). Furthermore, as $\lambda\,(u_1 - u_2) = 0$, for any $\chi \in \Lambda$ we have

$$\left(\chi - \lambda, u_1 - u_2\right)_\Omega = \left(\chi, u_1 - u_2\right)_\Omega - \left(\lambda, u_1 - u_2\right)_\Omega = \left(\chi, u_1 - u_2\right) \geq 0.$$

Therefore, (u_1, u_2, λ) is a solution to (1.9). Conversely, assume that (u_1, u_2, λ) is a solution to (1.9). Taking $v_1 \in \mathcal{D}(\Omega) \subset H^1_0(\Omega)$ and taking $v_2 = 0$ in (1.9) we get

$$\mu_1 \left(\boldsymbol{\nabla} u_1, \boldsymbol{\nabla} v_1\right)_\Omega - \left(\lambda, v_1\right)_\Omega = \langle f_1, v_1 \rangle. \tag{1.11}$$

In the distributional sense (1.11) reads

$$-\mu_1 \sum_{j=1}^{2} \left\langle \frac{\partial^2 u_1}{\partial x_j^2}, v_1 \right\rangle_{\mathcal{D}'(\Omega), \mathcal{D}(\Omega)} = \langle \lambda + f_1, v_1 \rangle_{\mathcal{D}'(\Omega), \mathcal{D}(\Omega)}$$

Then,

$$-\mu_1 \Delta u_1 - \lambda = f_1 \quad \text{in} \quad \mathcal{D}'(\Omega).$$

Next, taking $v_2 \in \mathcal{D}(\Omega) \subset H_0^1(\Omega)$ and $v_1 = 0$ and employing the same methodology we get

$$-\mu_2 \Delta u_2 + \lambda = f_2 \quad \text{in} \quad \mathcal{D}'(\Omega).$$

Furthermore, if the second line of (1.9) is satisfied, taking $\chi = \lambda + \mathbb{1}_{\mathcal{O}} \in \Lambda$ with $\mathcal{O}$ any measurable subset of Ω, and next $\chi = 0$ we obtain

$$u_1 - u_2 \geq 0, \quad \text{and} \quad (\lambda, u_1 - u_2)_\Omega \leq 0.$$

As $\lambda \geq 0$ it yields $\lambda\,(u_1 - u_2) = 0$. Thus, (u_1, u_2, λ) is a solution to (1.5) and this concludes the proof. □

The formulation (1.9) provides a charaterization of the weak solution. This saddle-point type formulation can also be rewritten as a variational inequality: Find $\boldsymbol{u} = (u_1, u_2) \in \boldsymbol{\mathcal{K}}_g$ such that $\forall \boldsymbol{v} = (v_1, v_2) \in \boldsymbol{\mathcal{K}}_g$,

$$a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq l(\boldsymbol{v} - \boldsymbol{u}). \tag{1.12}$$

where $\boldsymbol{\mathcal{K}}_g$ is a convex set defined by

$$\boldsymbol{\mathcal{K}}_g := \left\{(v_1, v_2) \in H_g^1(\Omega) \times H_0^1(\Omega),\ v_1 - v_2 \geq 0 \text{ a.e. in } \Omega\right\}.$$

For a proof see [17, Lemma 2]. Existence and uniqueness of a weak solution for (1.12) follows by the Lions–Stampacchia theorem (see [37]).

We now present the finite element discretization of problems (1.12) and (1.9).

1.2.2 Discretization by finite elements

Let $\mathcal{T}_h$ be a conforming simplicial mesh of Ω, *i.e.* $\mathcal{T}_h$ is a set of triangles verifying

$$\bigcup_{K \in \mathcal{T}_h} \overline{K} = \overline{\Omega}$$

where the intersection of the closure of two elements of $\mathcal{T}_h$ is either an empty set, a vertex, or an edge. The set of vertices of $\mathcal{T}_h$ is denoted by $\mathcal{V}_h$ and is partitionned into the interior vertices $\mathcal{V}_h^{\text{int}}$ and the boundary vertices $\mathcal{V}_h^{\text{ext}}$. We denote by $\mathcal{N}_h^{\text{int}}$ the number of interior vertices and by $\mathcal{N}_h^{\text{ext}}$ the number of boundary vertices. The vertices of an element $K \in \mathcal{T}_h$ are collected in the set $\mathcal{V}_K$. Denote by h_K the diameter of a triangle K and $h := \max_{K \in \mathcal{T}_h} h_K$. Furthermore, for $\boldsymbol{a} \in \mathcal{V}_h$, let the patch $\omega_h^{\boldsymbol{a}} \subset \Omega$ be the domain made up of the elements of $\mathcal{T}_h$ that share $\boldsymbol{a}$. The vector $\mathbf{n}_{\omega_h^{\boldsymbol{a}}}$ stands for its outward unit normal. In the sequel, we use the discrete conforming space of piecewise polynomial functions

$$X_h^p := \left\{v_h \in \mathcal{C}^0(\overline{\Omega});\ v_h|_K \in \mathbb{P}_p(K)\ \ \forall K \in \mathcal{T}_h\right\} \subset H^1(\Omega),$$

where $\mathbb{P}_p(K)$ stands for the set of polynomials of total degree less than or equal to $p \geq 1$ on the element $K \in \mathcal{T}_h$. We also denote by $\mathcal{V}_{\text{d}}^p$ the set of the Lagrange nodes (points $\boldsymbol{x}_l$) of the space X_h^p and by $\mathcal{N}_{\text{d}}^p$ its cardinality. The internal Lagrange nodes are collected in the set $\mathcal{V}_{\text{d}}^{p,\text{int}}$ (with $\mathcal{N}_{\text{d}}^{p,\text{int}}$ its cardinality) and the external ones are

collected in the set $\mathcal{V}_{\rm d}^{p,\rm ext}$ (with $\mathcal{N}_{\rm d}^{p,\rm ext}$ its cardinality). The Lagrange basis functions of X_h^p are denoted by $(\psi_{h,\boldsymbol{x}_l})_{1\le l\le \mathcal{N}_{\rm d}^p}$ for $\boldsymbol{x}_l \in \mathcal{V}_{\rm d}^p$. We recall that $\psi_{h,\boldsymbol{x}_l}(\boldsymbol{x}_l) = 1$ and $\psi_{h,\boldsymbol{x}_l}(\boldsymbol{x}_{l'}) = 0$ for all $(\boldsymbol{x}_{l'})_{1\le l'\neq l\le \mathcal{N}_{\rm d}^p} \in \mathcal{V}_{\rm d}^p$. In the particular case $p = 1$, the set $\mathcal{V}_{\rm d}^1$ coincides with the mesh vertices $\mathcal{V}_h$ and the Lagrange basis functions are the "hat" basis functions and are denoted by $\psi_{h,\boldsymbol{a}}$, $\boldsymbol{a} \in \mathcal{V}_h$. Still in this case, we denote

$$M_{\boldsymbol{a}} := (\psi_{h,\boldsymbol{a}}, 1)_{\omega_h^{\boldsymbol{a}}} = \frac{|\omega_h^{\boldsymbol{a}}|}{3}.$$

We also introduce the boundary aware set and space

$$X_{gh}^p := \{v_h \in X_h^p,\ v_h = g \text{ on } \partial\Omega\} \subset H_g^1(\Omega), \quad \text{and} \quad X_{0h}^p := X_h^p \cap H_0^1(\Omega),$$

as well as the convex set

$$\boldsymbol{\mathcal{K}}_{gh}^p := \left\{(v_{1h}, v_{2h}) \in X_{gh}^p \times X_{0h}^p,\ v_{1h}(\boldsymbol{x}_l) - v_{2h}(\boldsymbol{x}_l) \ge 0\ \ \forall\, (\boldsymbol{x}_l)_{1\le l\le \mathcal{N}_{\rm d}^{p,\rm int}} \in \mathcal{V}_{\rm d}^{p,\rm int}\right\}. \tag{1.13}$$

Observe that $\boldsymbol{\mathcal{K}}_{gh}^1 \subset \boldsymbol{\mathcal{K}}_g$ holds but $\boldsymbol{\mathcal{K}}_{gh}^p \not\subset \boldsymbol{\mathcal{K}}_g$ when $p > 1$, see [17, 53, 153]. The discrete counterpart to (1.12) consists in: find $\boldsymbol{u}_h = (u_{1h}, u_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ such that

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) \ge l(\boldsymbol{v}_h - \boldsymbol{u}_h) \qquad \forall \boldsymbol{v}_h = (v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p. \tag{1.14}$$

As a result of the Lions–Stampacchia theorem, problem (1.14) admits a unique solution that is nonconforming ($\boldsymbol{u}_h \notin \boldsymbol{\mathcal{K}}_g$) when $p \ge 2$. Moreover, following the methodology of [18, equation (4.5)] or [43] we define the functions λ_{1h} and λ_{2h} in X_h^p by

$$\left\{\begin{array}{rcll} \langle \lambda_{1h}, z_{1h}\rangle_h & := & \mu_1\,(\boldsymbol{\nabla} u_{1h}, \boldsymbol{\nabla} z_{1h})_\Omega - (f_1, z_{1h})_\Omega & \forall z_{1h} \in X_{0h}^p, \\ \langle \lambda_{2h}, z_{2h}\rangle_h & := & -\mu_2\,(\boldsymbol{\nabla} u_{2h}, \boldsymbol{\nabla} z_{2h})_\Omega + (f_2, z_{2h})_\Omega & \forall z_{2h} \in X_{0h}^p, \\ \langle \lambda_{1h}, \psi_{h,\boldsymbol{x}_l}\rangle_h & := & 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm ext}, & \\ \langle \lambda_{2h}, \psi_{h,\boldsymbol{x}_l}\rangle_h & := & 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm ext}, & \end{array}\right. \tag{1.15}$$

where for all $(w_h, v_h) \in X_h^p \times X_h^p$

$$\langle w_h, v_h\rangle_h := \left\{\begin{array}{ll} \displaystyle\sum_{\boldsymbol{a}\in\mathcal{V}_h} w_h(\boldsymbol{a}) v_h(\boldsymbol{a}) M_{\boldsymbol{a}} & \text{if} \quad p = 1, \qquad (1.16) \\ (w_h, v_h)_\Omega & \text{if} \quad p \ge 2. \qquad (1.17) \end{array}\right.$$

Note that (1.16) corresponds to the use of a mass lumping and will lead to particular properties in the case $p = 1$. Extending [17, Proposition 12] to the case $p \ge 2$ we have:

Lemma 1.2.2. *Let $(u_{1h}, u_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ be the solution of the reduced discrete problem (1.14). Then, the functions λ_{1h} and λ_{2h} defined by (1.15) coincide and we set $\lambda_h := \lambda_{1h} = \lambda_{2h} \in X_h^p$.*

Proof. We substract the first two equations of (1.15) taking $z_{1h} = z_{2h} = \psi_{h,\boldsymbol{x}_l}$ with $\boldsymbol{x}_l$ any internal Lagrange node to get $\forall l = 1 \ldots \mathcal{N}_{\rm d}^{p,\rm int}$

$$\langle \lambda_{1h} - \lambda_{2h}, \psi_{h,\boldsymbol{x}_l}\rangle_h = (\mu_1 \boldsymbol{\nabla} u_{1h} + \mu_2 \boldsymbol{\nabla} u_{2h}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{x}_l})_\Omega - (f_1 + f_2, \psi_{h,\boldsymbol{x}_l})_\Omega\,.$$

Taking $v_{1h} := u_{1h} + \psi_{h,\boldsymbol{x}_l}$ and $v_{2h} := u_{2h} + \psi_{h,\boldsymbol{x}_l}$ in (1.14) and noting that $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}^p_{gh}$ where $\psi_{h,\boldsymbol{x}_l}$ is the Lagrange basis function associated to the internal node $\boldsymbol{x}_l$, we get

$$\langle \lambda_{1h} - \lambda_{2h}, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}. \tag{1.18}$$

In the same way, considering $v_{1h} := u_{1h} - \psi_{h,\boldsymbol{x}_l}$ and $v_{2h} := u_{2h} - \psi_{h,\boldsymbol{x}_l}$ we get

$$\langle \lambda_{1h} - \lambda_{2h}, \psi_{h,\boldsymbol{x}_l} \rangle_h \leq 0 \quad \forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}. \tag{1.19}$$

Finally, combining (1.18) and (1.19) with the last two equations of (1.15) provides

$$\langle \lambda_{1h} - \lambda_{2h}, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0 \quad \forall 1 \leq l \leq \mathcal{N}_{\mathrm{d}}^{p}.$$

For $p \geq 2$, $\lambda_{1h} - \lambda_{2h} \in X_h^p$ is L^2-orthogonal to all test functions in the space X_h^p, which implies $\lambda_{1h} = \lambda_{2h}$. For $p = 1$, $\lambda_{1h} = \lambda_{2h}$ holds true because $M_{\boldsymbol{a}} > 0$. □

Furthermore, $\lambda_h \in X_h^p$ satisfies the following property:

Lemma 1.2.3. *Let $(u_{1h}, u_{2h}) \in \boldsymbol{\mathcal{K}}^p_{gh}$ be the solution of the reduced problem (1.14) and let λ_h be defined by (1.15). Then, there holds*

$$\langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}. \tag{1.20}$$

Proof. Let $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$ be an internal node. First observe that $(v_{1h}, v_{2h}) := (u_{1h} + \psi_{h,\boldsymbol{x}_l}, u_{2h}) \in \boldsymbol{\mathcal{K}}^p_{gh}$. The conclusion follows from the reduced problem (1.14), the characterization (1.15), and Lemma 1.2.2 giving $\forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$

$$\mu_1 \left(\boldsymbol{\nabla} u_{1h}, \boldsymbol{\nabla} \psi_{h,\boldsymbol{x}_l} \right)_\Omega - \left(f_1, \psi_{h,\boldsymbol{x}_l} \right)_\Omega = \langle \lambda_{1h}, \psi_{h,\boldsymbol{x}_l} \rangle_h = \langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0.$$

□

Following Lemma 1.2.3 we suggest to define the discrete convex set for λ_h by

$$\Lambda_h^p := \left\{ v_h \in X_h^p;\ \langle v_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0\ \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}},\ \langle v_h, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0\ \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{ext}} \right\}. \tag{1.21}$$

Observe that $\Lambda_h^p \not\subset \Lambda$ for $p \geq 2$ and in the case $p = 1$, Λ_h^p reduces to

$$\Lambda_h^1 = \left\{ v_h \in X_{0h}^1;\ v_h(\boldsymbol{a}) \geq 0 \quad \forall \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}} \right\} \subset \Lambda, \tag{1.22}$$

which is the same as in [17, Section 4]. Note that when $\chi_h \in \Lambda_h^p$ and $\boldsymbol{v}_h \in \boldsymbol{\mathcal{K}}^p_{gh}$,

$$\langle \chi_h, v_{1h} - v_{2h} \rangle_h = \sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}} \left(v_{1h} - v_{2h} \right) (\boldsymbol{x}_l) \, \langle \chi_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0. \tag{1.23}$$

A discrete formulation, built by the Galerkin method corresponding to problem (1.9) consists in: find $(u_{1h}, u_{2h}, \lambda_h) \in X^p_{gh} \times X^p_{0h} \times \Lambda^p_h$ such that

$$\begin{aligned} &\sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_{\alpha h}, \boldsymbol{\nabla} z_{\alpha h} \right)_\Omega - \langle \lambda_h, z_{1h} - z_{2h} \rangle_h = \sum_{\alpha=1}^{2} \left(f_\alpha, z_{\alpha h} \right)_\Omega \quad \forall (z_{1h}, z_{2h}) \in [X^p_{0h}]^2, \\ &\langle \chi_h - \lambda_h, u_{1h} - u_{2h} \rangle_h \geq 0 \quad \forall \chi_h \in \Lambda^p_h. \end{aligned} \tag{1.24}$$

Lemma 1.2.4. *For any solution $(u_{1h}, u_{2h}, \lambda_h)$ of problem (1.24), the pair (u_{1h}, u_{2h}) is a solution of problem (1.14). Conversely, for any solution (u_{1h}, u_{2h}) of problem (1.14), defining the function $\lambda_h = \lambda_{\alpha h}$, $\alpha = 1, 2$ by (1.15), the triple $(u_{1h}, u_{2h}, \lambda_h)$ is a solution to problem (1.24).*

Proof. For the case $p = 1$, the proof is given in [18, Lemma 13]. Let $p \geq 2$ and let $(u_{1h}, u_{2h}, \lambda_h)$ be the solution of problem (1.24). Decomposing $u_{1h} - u_{2h}$ in the Lagrange basis we obtain

$$u_{1h} - u_{2h} = \sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p} (u_{1h} - u_{2h})(\boldsymbol{x}_l)\psi_{h,\boldsymbol{x}_l}.$$

Next, note that for $\chi_h, \lambda_h \in \Lambda_h^p$, $\chi_h + \lambda_h \in \Lambda_h^p$. Take $\chi_h + \lambda_h$ as a test function in (1.24), we get

$$\langle \chi_h, u_{1h} - u_{2h} \rangle_h \geq 0 \quad \forall \chi_h \in \Lambda_h^p,$$

and then

$$\sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p} (u_{1h} - u_{2h})(\boldsymbol{x}_l) \langle \chi_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall \chi_h \in \Lambda_h^p. \tag{1.25}$$

We construct the basis $(\Theta_{h,\boldsymbol{x}_l})_{1 \leq l \leq \mathcal{N}_{\mathrm{d}}^p}$ of X_h^p, dual to $(\psi_{h,\boldsymbol{x}_l})_{1 \leq l \leq \mathcal{N}_{\mathrm{d}}^p}$, satisfying

$$\begin{aligned} &\langle \Theta_{h,\boldsymbol{x}_l}, \psi_{h,\boldsymbol{x}_l} \rangle_h = 1 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p, \\ &\langle \Theta_{h,\boldsymbol{x}_l}, \psi_{h,\boldsymbol{x}_l^*} \rangle_h = 0 \quad \forall 1 \leq l^* \neq l \leq \mathcal{N}_{\mathrm{d}}^p. \end{aligned} \tag{1.26}$$

Note that each vector of the dual basis $\Theta_{h,\boldsymbol{x}_l}$ can be determined by inverting a diagonal (lumped mass) matrix for $p = 1$ and the finite element mass matrix for $p \geq 2$; importantly, all $\Theta_{h,\boldsymbol{x}_l}$, $1 \leq l \leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ belong to Λ_h^p. Note, however, that the support of $\Theta_{h,\boldsymbol{x}_l}$ is typically not local. Finally, taking in (1.25) $\chi_h = \Theta_{h,\boldsymbol{x}_l^*}$, for all $\boldsymbol{x}_l^* \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$, we obtain

$$\sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p} (u_{1h} - u_{2h})(\boldsymbol{x}_l) \langle \Theta_{h,\boldsymbol{x}_l^*}, \psi_{h,\boldsymbol{x}_l} \rangle_h = (u_{1h} - u_{2h})(\boldsymbol{x}_l^*) \geq 0 \quad \forall \boldsymbol{x}_l^* \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}.$$

Therefore, $\boldsymbol{u}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$. Now, we prove (1.14). Let $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$. Taking $z_{1h} := v_{1h} - u_{1h} \in X_{0h}^p$ and $z_{2h} := v_{2h} - u_{2h} \in X_{0h}^p$ as test functions in (1.24) provides

$$\langle \lambda_h, v_{1h} - v_{2h} \rangle_h - \langle \lambda_h, u_{1h} - u_{2h} \rangle_h = a(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) - l(\boldsymbol{v}_h - \boldsymbol{u}_h). \tag{1.27}$$

Using (1.23) with $\lambda_h \in \Lambda_h^p$ and $\boldsymbol{v}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$ and taking $\chi_h = 0$ in (1.24) gives

$$\langle \lambda_h, v_{1h} - v_{2h} \rangle_h \geq 0, \quad \langle -\lambda_h, u_{1h} - u_{2h} \rangle_h \geq 0. \tag{1.28}$$

Combining (1.27), and (1.28) provides (1.14).

Conversely, let $(u_{1h}, u_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ be the solution of the reduced problem (1.14) and let $(z_{1h}, z_{2h}) \in X_{0h}^p \times X_{0h}^p$ be arbitrary. The characterization (1.15) combined with Lemma 1.2.2 and Lemma 1.2.3 yields $\lambda_h \in \Lambda_h^p$. Next, subtracting the two equations of (1.15) gives the first line of (1.24). It remains to prove the second line

of (1.24). Let now $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}^p_{gh}$. The first line in (1.24) now implies (1.27) and the reduced problem (1.14) yields

$$-\langle \lambda_h, u_{1h} - u_{2h} \rangle_h + \langle \lambda_h, v_{1h} - v_{2h} \rangle_h \geq 0 \quad \forall\, (v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}^p_{gh}. \tag{1.29}$$

For $v_{1h} := u_{1h} - \displaystyle\sum_{\boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm int}} u_{1h}(\boldsymbol{x}_l)\psi_{h,\boldsymbol{x}_l} \in X^p_{gh}$ and $v_{2h} := 0 \in X^p_{0h}$, $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}^p_{gh}$, and using the definition of Λ^p_h, we have $\langle \lambda_h, v_{1h} - v_{2h} \rangle_h = 0$ and the inequality (1.29) yields

$$-\langle \lambda_h, u_{1h} - u_{2h} \rangle_h \geq 0.$$

To conclude the proof, we use (1.23) with $\boldsymbol{u}_h \in \boldsymbol{\mathcal{K}}^p_{gh}$ and for any $\chi_h \in \Lambda^p_h$. □

Remark 1.2.5. *Taking $\chi_h = 0$ and then $\chi_h = 2\lambda_h \in \Lambda^p_h$ in the second line of (1.24) yields*

$$\langle \lambda_h, u_{1h} - u_{2h} \rangle_h = 0. \tag{1.30}$$

Combining (1.30) with the definition of Λ^p_h and the fact that the solution $\boldsymbol{u}_h \in \boldsymbol{\mathcal{K}}^p_{gh}$ give the complementarity conditions

$$\begin{aligned} &(u_{1h} - u_{2h})\,(\boldsymbol{x}_l) \geq 0 \quad && \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm int}, \\ &\langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0, && \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm int}, \\ &\langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0 && \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm ext}, \\ &\langle \lambda_h, u_{1h} - u_{2h} \rangle_h = 0. \end{aligned} \tag{1.31}$$

Finally, Problem (1.24) reads: find $(u_{1h}, u_{2h}, \lambda_h) \in X^p_{gh} \times X^p_{0h} \times \Lambda^p_h$ such that

$$\begin{aligned} &\sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_{\alpha h}, \boldsymbol{\nabla} z_{\alpha h}\right)_\Omega - \langle \lambda_h, z_{1h} - z_{2h} \rangle_h = \sum_{\alpha=1}^{2} (f_\alpha, z_{\alpha h})_\Omega \quad \forall (z_{1h}, z_{2h}) \in [X^p_{0h}]^2, \\ &(u_{1h} - u_{2h})\,(\boldsymbol{x}_l) \geq 0 \ \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm int},\ \langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \ \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm int},\ \langle \lambda_h, u_{1h} - u_{2h} \rangle_h = 0. \end{aligned} \tag{1.32}$$

From the previous analysis, we deduce

Lemma 1.2.6. *The formulation (1.24) is well posed for any polynomial degree $p \geq 1$.*

Remark 1.2.7. *Problems (1.14) and (1.24) are equivalent in the sense of Lemma 1.2.4. To our knowledge it is the first time that a formulation for the membranes problem is proposed for $p \geq 2$, although no proof of convergence is provided. Moreover, when $p \geq 2$, the solution is nonconforming: the constraints $u_{1h} \geq u_{2h}$ and $\lambda_h \geq 0$ do not hold in general. However, the following a posteriori analysis remains valid.*

1.2.3 Numerical resolution and discrete complementarity problems

We now express the discrete problem (1.32) under an algebraic form. We use for this purpose first the Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1 \leq l \leq \mathcal{N}_{\rm d}^p}$ of Λ^p_h for the variable λ_h. Second,

we prefer employing the subfamily $(\Theta_{h,\boldsymbol{x}_l})_{1\le l\le \mathcal{N}_{\rm d}^{p,\rm int}}$ of the basis $(\Theta_{h,\boldsymbol{x}_l})_{1\le l\le \mathcal{N}_{\rm d}^{p}}$ of Λ_h^p, dual to $(\psi_{h,\boldsymbol{x}_l})_{1\le l\le \mathcal{N}_{\rm d}^{p}}$. Concerning the discrete solution $u_{1h}\in X_{gh}^p$ we use a lifting as follows: $u_{1h}=u_{1h}^*+g$ where $u_{1h}^*\in X_{0h}^p$ and $g>0$ is the constant boundary value so that $\boldsymbol{\nabla} u_{1h}=\boldsymbol{\nabla} u_{1h}^*$. The matricial representation of the lifting is denoted by $\boldsymbol{X}_{1h}\in\mathbb{R}^{\mathcal{N}_{\rm d}^{p,\rm int}}$ and satisfies

$$u_{1h}=\sum_{l=1}^{\mathcal{N}_{\rm d}^{p,\rm int}}(\boldsymbol{X}_{1h})_l\,\psi_{h,\boldsymbol{x}_l}+g,\quad\text{where}\quad(\boldsymbol{X}_{1h})_l=u_{1h}^*(\boldsymbol{x}_l). \tag{1.33}$$

The discrete functional u_{2h} is expressed in the Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1\le l\le \mathcal{N}_{\rm d}^{p,\rm int}}$ as

$$u_{2h}=\sum_{l=1}^{\mathcal{N}_{\rm d}^{p,\rm int}}(\boldsymbol{X}_{2h})_l\,\psi_{h,\boldsymbol{x}_l},\quad\text{where}\quad\boldsymbol{X}_{2h}\in\mathbb{R}^{\mathcal{N}_{\rm d}^{p,\rm int}}\quad\text{and}\quad(\boldsymbol{X}_{2h})_l=u_{2h}(\boldsymbol{x}_l). \tag{1.34}$$

Case of Lagrange basis for λ_h, $p=1$

In this case, $\mathcal{N}_{\rm d}^{p,\rm int}=\mathcal{N}_h^{\rm int}$ and the discrete Lagrange multiplier λ_h is decomposed in the basis $(\psi_{h,\boldsymbol{a}_l})_{1\le l\le \mathcal{N}_h^{\rm int}}$ as

$$\lambda_h=\sum_{l=1}^{\mathcal{N}_h^{\rm int}}(\boldsymbol{X}_{3h})_l\,\psi_{h,\boldsymbol{a}_l},\quad\text{where}\quad\boldsymbol{X}_{3h}\in\mathbb{R}^{\mathcal{N}_h^{\rm int}}\quad\text{and}\quad(\boldsymbol{X}_{3h})_l=\lambda_h(\boldsymbol{a}_l). \tag{1.35}$$

Taking successively $z_{1h}=\psi_{h,\boldsymbol{a}_l}$, $z_{2h}=0$ and $z_{1h}=0$, $z_{2h}=\psi_{h,\boldsymbol{a}_l}$ for all $\boldsymbol{a}_l\in\mathcal{V}_h^{\rm int}$, the first line of (1.32) reads

$$\mathbb{E}_1\boldsymbol{X}_h=\boldsymbol{F},$$

where $\boldsymbol{X}_h^T:=[\boldsymbol{X}_{1h},\boldsymbol{X}_{2h},\boldsymbol{X}_{3h}]^T$ is the unknown vector, $\mathbb{E}_1\in\mathbb{R}^{2\mathcal{N}_h^{\rm int},3\mathcal{N}_h^{\rm int}}$ is a rectangular matrix defined by

$$\mathbb{E}_1:=\begin{bmatrix}\mu_1\mathbb{S} & \mathbf{0} & -\mathbb{D}\\ \mathbf{0} & \mu_2\mathbb{S} & +\mathbb{D}\end{bmatrix},$$

where the stiffness matrix $\mathbb{S}\in\mathbb{R}^{\mathcal{N}_h^{\rm int},\mathcal{N}_h^{\rm int}}$ is defined by

$$\mathbb{S}_{l,m}:=(\boldsymbol{\nabla}\psi_{h,\boldsymbol{a}_l},\boldsymbol{\nabla}\psi_{h,\boldsymbol{a}_m})_\Omega\quad 1\le l,m\le\mathcal{N}_h^{\rm int}, \tag{1.36}$$

and the diagonal mass lumped matrix $\mathbb{D}$ defined by

$$\mathbb{D}_{l,l}:=M_{\boldsymbol{a}_l}:=(\psi_{h,\boldsymbol{a}_l},1)_\Omega=\frac{|\omega_h^{\boldsymbol{a}_l}|}{3}. \tag{1.37}$$

The right-hand side vector $\boldsymbol{F}$ is defined by blocks ($\boldsymbol{F}^T:=(\boldsymbol{F}_1,\boldsymbol{F}_2)^T$) by

$$\begin{aligned}(\boldsymbol{F}_1)_l&:=(f_1,\psi_{h,\boldsymbol{a}_l})_\Omega\quad 1\le l\le\mathcal{N}_h^{\rm int},\\ (\boldsymbol{F}_2)_l&:=(f_2,\psi_{h,\boldsymbol{a}_l})_\Omega\quad 1\le l\le\mathcal{N}_h^{\rm int}.\end{aligned}$$

The linear complementarity constraints are easily expressed as

$$\boldsymbol{X}_{1h}+g\mathbf{1}-\boldsymbol{X}_{2h}\ge 0,\quad\boldsymbol{X}_{3h}\ge 0,\quad(\boldsymbol{X}_{1h}+g\mathbf{1}-\boldsymbol{X}_{2h})\cdot\boldsymbol{X}_{3h}=0, \tag{1.38}$$

where $\mathbf{1}=(1,1,\cdots,1)^T\in\mathbb{R}^{\mathcal{N}_h^{\rm int}}$. Thus, when $p=1$, problem (1.32) is equivalent to search $\boldsymbol{X}_h\in\mathbb{R}^{3\mathcal{N}_h^{\rm int}}$ such that

$$\begin{aligned}&\mathbb{E}_1\boldsymbol{X}_h=\boldsymbol{F},\\ &\boldsymbol{X}_{1h}+g\mathbf{1}-\boldsymbol{X}_{2h}\ge 0,\quad\boldsymbol{X}_{3h}\ge 0,\quad(\boldsymbol{X}_{1h}+g\mathbf{1}-\boldsymbol{X}_{2h})\cdot\boldsymbol{X}_{3h}=0.\end{aligned} \tag{1.39}$$

Case of Lagrange basis for λ_h, case $p \geq 2$

We now express problem (1.32) under an algebraic form, using the Lagrange basis for λ_h when $p \geq 2$. We first start by defining the finite element mass matrix $\mathbb{M}$ as

$$\mathbb{M}_{l,l'} := \left(\psi_{h,\boldsymbol{x}_l}, \psi_{h,\boldsymbol{x}_{l'}}\right)_\Omega, \quad 1 \leq l, l' \leq \mathcal{N}_{\mathrm{d}}^p. \tag{1.40}$$

For $\boldsymbol{x}_{l'}$ an internal node, we extract from $\mathbb{M}$ the rectangular matrix $\widehat{\mathbb{M}}$ defined as

$$\left(\widehat{\mathbb{M}}_{l,l'}\right)_{\substack{1 \leq l \leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} \\ 1 \leq l' \leq \mathcal{N}_{\mathrm{d}}^p}} := \left(\mathbb{M}_{j,j'}\right)_{\substack{\boldsymbol{x}_j \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}} \\ \boldsymbol{x}_{j'} \in \mathcal{V}_{\mathrm{d}}^p}}. \tag{1.41}$$

The discrete lagrange multiplier λ_h is decomposed in the full space X_h^p as

$$\lambda_h = \sum_{l=1}^{\mathcal{N}_{\mathrm{d}}^p} \left(\widetilde{\boldsymbol{X}}_{3h}\right)_l \psi_{h,\boldsymbol{x}_l} \quad \text{with} \quad \widetilde{\boldsymbol{X}}_{3h} \in \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^p}. \tag{1.42}$$

Taking successively $z_{1h} = \psi_{h,\boldsymbol{x}_l}$, $z_{2h} = 0$ and $z_{1h} = 0$, $z_{2h} = \psi_{h,\boldsymbol{x}_l}$ for all $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$, the first line of (1.32) reads

$$\widetilde{\mathbb{E}}_p \boldsymbol{X}_h = \boldsymbol{F},$$

where $\boldsymbol{X}_h^T := \left[\boldsymbol{X}_{1h}, \boldsymbol{X}_{2h}, \widetilde{\boldsymbol{X}}_{3h}\right]^T \in \mathbb{R}^{2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} + \mathcal{N}_{\mathrm{d}}^p}$ is the unknown vector, $\widetilde{\mathbb{E}}_p \in \mathbb{R}^{2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}, 2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} + \mathcal{N}_{\mathrm{d}}^p}$ is a rectangular matrix defined by

$$\widetilde{\mathbb{E}}_p := \begin{bmatrix} \mu_1 \mathbb{S} & \mathbf{0} & -\widehat{\mathbb{M}} \\ \mathbf{0} & \mu_2 \mathbb{S} & +\widehat{\mathbb{M}} \end{bmatrix},$$

where the stiffness matrix $\mathbb{S} \in \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}, \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ is defined by

$$\mathbb{S}_{l,m} := \left(\boldsymbol{\nabla} \psi_{h,\boldsymbol{x}_l}, \boldsymbol{\nabla} \psi_{h,\boldsymbol{x}_m}\right)_\Omega \quad 1 \leq l, m \leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}. \tag{1.43}$$

The right-hand side vector $\boldsymbol{F}$ is defined by blocks ($\boldsymbol{F}^T := (\boldsymbol{F}_1, \boldsymbol{F}_2)^T$) by

$$\begin{aligned} (\boldsymbol{F}_1)_l &:= (f_1, \psi_{h,\boldsymbol{x}_l})_\Omega \quad 1 \leq l \leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}, \\ (\boldsymbol{F}_2)_l &:= (f_2, \psi_{h,\boldsymbol{x}_l})_\Omega \quad 1 \leq l \leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}. \end{aligned}$$

The first complementarity constraint of (1.32) is expressed in the Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1 \leq l \leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ for u_{1h} and u_{2h} as

$$\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h} \geq 0. \tag{1.44}$$

Next, observe that for any $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$

$$\langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h = \sum_{l'=1}^{\mathcal{N}_{\mathrm{d}}^p} \left(\widetilde{\boldsymbol{X}}_{3h}\right)_{l'} \left(\psi_{h,\boldsymbol{x}_{l'}}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega = \left(\widehat{\mathbb{M}} \widetilde{\boldsymbol{X}}_{3h}\right)_l.$$

Then, the second complementarity constraint of (1.32) is expressed in the Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1 \leq l \leq \mathcal{N}_{\mathrm{d}}^p}$ for λ_h as

$$\widehat{\mathbb{M}} \widetilde{\boldsymbol{X}}_{3h} \geq 0. \tag{1.45}$$

Furthermore, as $\lambda_h \in \Lambda_h^p$ we have

$$\langle \lambda_h, u_{1h} - u_{2h} \rangle_h = \sum_{l=1}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} \sum_{l'=1}^{\mathcal{N}_{\mathrm{d}}^{p}} ((\boldsymbol{X}_{1h})_l - (\boldsymbol{X}_{2h})_l) \left(\widetilde{\boldsymbol{X}}_{3h}\right)_{l'} \left(\psi_{h,\boldsymbol{x}_{l'}}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega + (\lambda_h, g)_\Omega \,.$$

Decomposing the constant function g in the complete Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\mathrm{d}}^p}$ and employing the definition of the space Λ_h^p given by (1.21) we get

$$(\lambda_h, g)_\Omega = \sum_{l'=1}^{\mathcal{N}_{\mathrm{d}}^{p}} \sum_{l=1}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} \left(\widetilde{\boldsymbol{X}}_{3h}\right)_{l'} g \left(\psi_{h,\boldsymbol{x}_{l'}}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega = g\mathbf{1} \cdot \widehat{\mathbb{M}}\widetilde{\boldsymbol{X}}_{3h},$$

where $\mathbf{1} \in \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$. Finally, the last complementarity constraint of (1.32) is expressed in the Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ as

$$(\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}) \cdot \widehat{\mathbb{M}}\widetilde{\boldsymbol{X}}_{3h} = 0. \tag{1.46}$$

Thus, when $p \geq 2$, problem (1.32) is equivalent to search $\boldsymbol{X}_h \in \mathbb{R}^{2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}+\mathcal{N}_{\mathrm{d}}^{p}}$ such that

$$\begin{aligned} &\widetilde{\mathbb{E}}_p \boldsymbol{X}_h = \boldsymbol{F}, \\ &\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h} \geq 0, \quad \widehat{\mathbb{M}}\widetilde{\boldsymbol{X}}_{3h} \geq 0, \quad (\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}) \cdot \widehat{\mathbb{M}}\widetilde{\boldsymbol{X}}_{3h} = 0. \end{aligned} \tag{1.47}$$

Case of dual basis for λ_h, case $p \geq 2$

To finish we provide a characterization of problem (1.32) under an algebraic form in the basis $(\Theta_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\mathrm{d}}^p}$ dual to $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\mathrm{d}}^p}$ when $p \geq 2$. This time, the discrete Lagrange multiplier λ_h is decomposed in the basis $\Theta_{h,\boldsymbol{x}_l}$ as

$$\lambda_h = \sum_{l=1}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} (\boldsymbol{X}_{3h})_l \, \Theta_{h,\boldsymbol{x}_l}, \quad \text{with} \quad \boldsymbol{X}_{3h} \in \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}, \tag{1.48}$$

because $\lambda_h \in \Lambda_h^p$ and thus the components on $\Theta_{h,\boldsymbol{x}_l}$ for $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{ext}}$ are 0. Taking successively $z_{1h} = \psi_{h,\boldsymbol{x}_l}$, $z_{2h} = 0$ and $z_{1h} = 0$, $z_{2h} = \psi_{h,\boldsymbol{x}_l}$ for all $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$, the first line of (1.32) reads

$$\mathbb{E}_p \boldsymbol{X}_h = \boldsymbol{F},$$

where $\boldsymbol{X}_h^T := [\boldsymbol{X}_{1h}, \boldsymbol{X}_{2h}, \boldsymbol{X}_{3h}]^T \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ is the unknown vector, and $\mathbb{E}_p \in \mathbb{R}^{2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}},3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ is a rectangular matrix defined by

$$\mathbb{E}_p := \begin{bmatrix} \mu_1\mathbb{S} & \mathbf{0} & -\mathbb{I}_{\mathrm{d}} \\ \mathbf{0} & \mu_2\mathbb{S} & +\mathbb{I}_{\mathrm{d}} \end{bmatrix}$$

with $\mathbb{I}_{\mathrm{d}} \in \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}},\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ the identity matrix. The right-hand side vector $\boldsymbol{F} \in \mathbb{R}^{2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ and the stiffness matrix $\mathbb{S} \in \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}},\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ are defined above for the Lagrange basis. As therein, the first complementarity constraint of (1.32) is unchanged

$$\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h} \geq 0. \tag{1.49}$$

Next, using (1.26) we have for any $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$

$$\langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h = \sum_{l'=1}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} (\boldsymbol{X}_{3h})_{l'} \left(\Theta_{h,\boldsymbol{x}_{l'}}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega = (\boldsymbol{X}_{3h})_l \,.$$

Therefore, the second complementarity constraint of (1.32) is expressed in the dual basis $(\Theta_{h,\boldsymbol{x}_l})_{1\le l\le \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ as

$$\boldsymbol{X}_{3h} \ge 0. \tag{1.50}$$

Furthermore,

$$\begin{aligned}\langle \lambda_h, u_{1h} - u_{2h} \rangle_h &= \sum_{l=1}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} \sum_{l'=1}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} \left((\boldsymbol{X}_{1h})_l - (\boldsymbol{X}_{2h})_l\right) (\boldsymbol{X}_{3h})_{l'} \left(\Theta_{h,\boldsymbol{x}_{l'}}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega + (\lambda_h, g)_\Omega \\ &= (\boldsymbol{X}_{1h} - \boldsymbol{X}_{2h}) \cdot \boldsymbol{X}_{3h} + (\lambda_h, g)_\Omega \,.\end{aligned} \tag{1.51}$$

Next,

$$(\lambda_h, g)_\Omega = \sum_{l'=1}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} \sum_{l=1}^{\mathcal{N}_{\mathrm{d}}^{p}} g\, (\boldsymbol{X}_{3h})_{l'} \left(\Theta_{h,\boldsymbol{x}_{l'}}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega = g\mathbf{1} \cdot \boldsymbol{X}_{3h}, \tag{1.52}$$

where $\mathbf{1} \in \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$. Finally, combining (1.51) and (1.52), the last complementarity constraint is expressed in the dual basis $(\Theta_{h,\boldsymbol{x}_l})_{1\le l\le \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ as

$$(\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}) \cdot \boldsymbol{X}_{3h} = 0. \tag{1.53}$$

Thus, when $p \ge 2$, problem (1.32) is equivalent to search $\boldsymbol{X}_h \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ such that

$$\begin{aligned}&\mathbb{E}_p \boldsymbol{X}_h = \boldsymbol{F}, \\ &\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h} \ge 0, \quad \boldsymbol{X}_{3h} \ge 0, \quad (\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}) \cdot \boldsymbol{X}_{3h} = 0.\end{aligned} \tag{1.54}$$

Comments

We expressed problem (1.32) for $p \ge 1$ with λ_h in the Lagrange basis and for $p \ge 2$ with λ_h in its dual basis. When $p = 1$, we obtain discrete complementarity constraints in internal vertices of the mesh and the system (1.39) is practical for implementation. When $p \ge 2$, we obtained a discrete complementarity problem (1.47) expressed in the Lagrange nodes. The construction of the rectangular matrix $\widetilde{\mathbb{E}}_p$ requires to compute the finite element mass matrix and the complementarity constraints (1.45) and (1.46) also contain the mass matrix. The characterization of problem (1.32) in the basis $\Theta_{h,\boldsymbol{x}_l}$, when $p \ge 2$ shows that the construction of the rectangular matrix $\mathbb{E}_p$ only depends on the identity matrix and the constraints are expressed in a very convenient manner. Nevertheless, in this case, the unknown vector $\boldsymbol{X}_h$ has $\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ coordinates expressed in the basis $\Theta_{h,\boldsymbol{x}_l}$. So, to express it afterwards in the Lagrange basis, (when we want to work with the function $\lambda_h^{k,i} \in \Lambda_h^p$) it requires to invert the finite element mass matrix which has some cost. In the sequel, we consider the case $p \ge 1$ and the expression of (1.32) written in the basis $(\Theta_{h,\boldsymbol{x}_l})_{1\le l\le \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$.

1.2.4 C-functions

We now express the complementarity constraints given by the second line of (1.54) via non-differentiable equations. Let us recall that a function $f : (\mathbb{R}^m)^2 \to \mathbb{R}^m$ ($m \geq 1$) is a C-function or a complementarity function if

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in (\mathbb{R}^m)^2 \qquad f(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{0} \quad \Longleftrightarrow \quad \boldsymbol{x} \geq \mathbf{0}, \quad \boldsymbol{y} \geq \mathbf{0}, \quad \boldsymbol{x} \cdot \boldsymbol{y} = 0.$$

Examples of C-functions are the min function

$$(\min\{\boldsymbol{x}, \boldsymbol{y}\})_l := \min\{\boldsymbol{x}_l, \boldsymbol{y}_l\} \qquad l = 1, \ldots, m, \tag{1.55}$$

the Fischer–Burmeister function

$$(f_{\mathrm{FB}}(\boldsymbol{x}, \boldsymbol{y}))_l := \sqrt{\boldsymbol{x}_l^2 + \boldsymbol{y}_l^2} - (\boldsymbol{x}_l + \boldsymbol{y}_l) \quad l = 1, \ldots, m, \tag{1.56}$$

or the Mangasarian function

$$(f_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{y}))_l := \xi(|\boldsymbol{x}_l - \boldsymbol{y}_l|) - \xi(\boldsymbol{y}_l) - \xi(\boldsymbol{x}_l) \quad l = 1, \ldots, m,$$

where $\xi : \mathbb{R} \mapsto \mathbb{R}$ is an increasing function satisfying $\xi(\mathbf{0}) = \mathbf{0}$. The Fischer–Burmeister function is not differentiable in $(\mathbf{0}, \mathbf{0})$ whereas the min function and the Mangasarian function are not differentiable when $\boldsymbol{x} = \boldsymbol{y}$. For more details on C-functions see [88, 89]. Let $\tilde{\mathbf{C}}$ be any C-function, *i.e.*, satisfying (for $m = \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$) $\tilde{\mathbf{C}}(\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}, \boldsymbol{X}_{3h}) = 0 \iff \boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h} \geq 0$, $\boldsymbol{X}_{3h} \geq 0$, and $(\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}) \cdot \boldsymbol{X}_{3h} = 0$. Then, introducing the function $\mathbf{C} : \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} \to \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ defined as $\mathbf{C}(\boldsymbol{X}_h) = \tilde{\mathbf{C}}(\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}, \boldsymbol{X}_{3h})$, the problem (1.54) can be equivalently rewritten as

$$\begin{cases} \mathbb{E}_p \boldsymbol{X}_h & = \boldsymbol{F}, \\ \mathbf{C}(\boldsymbol{X}_h) & = \mathbf{0}. \end{cases} \tag{1.57}$$

The C-functions that are commonly used are locally Lipschitz and continuous, thus differentiable almost everywhere as a result of the Rademacher Theorem (see [58, 88]). Thus, it is possible to weaken the $\mathcal{C}^1$ assumption that would be necessary for the Newton algorithm by constructing a semismooth Newton scheme (see [29, 88, 89]). For instance we can employ the Newton-min, the Newton–Fischer–Burmeister or the Newton–Mangasarian algorithms.

1.3 Inexact semismooth Newton methods

We address in this section the numerical solution of the system of nonlinear algebraic inequalities corresponding to (1.24). We assume that an iterative linearization procedure is applied such that for a given initial vector $\boldsymbol{X}_h^0 := (\boldsymbol{X}_{1h}^0, \boldsymbol{X}_{2h}^0, \boldsymbol{X}_{3h}^0)^T \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$, on step $k \geq 1$, one looks for $\boldsymbol{X}_h^k \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ such that

$$\mathbb{A}^{k-1} \boldsymbol{X}_h^k = \boldsymbol{B}^{k-1}, \tag{1.58}$$

where the square matrix $\mathbb{A}^{k-1}$ and the right-hand side vector $\boldsymbol{B}^{k-1}$ are respectively defined by

$$\mathbb{A}^{k-1} := \begin{bmatrix} \mathbb{E}_p \\ \mathbf{J}_{\mathbf{C}}(\boldsymbol{X}_h^{k-1}) \end{bmatrix}, \qquad \boldsymbol{B}^{k-1} := \begin{bmatrix} \boldsymbol{F} \\ \mathbf{J}_{\mathbf{C}}(\boldsymbol{X}_h^{k-1}) \boldsymbol{X}_h^{k-1} - \mathbf{C}(\boldsymbol{X}_h^{k-1}) \end{bmatrix}. \tag{1.59}$$

Note that since the first line of (1.57) is linear, the corresponding Jacobian is constant and equal to $\mathbb{E}_p$. The semismooth nonlinearity occurs in the second line of (1.57). The Clarke subdifferential of the semismooth C-function $\mathbf{C}$ at $\boldsymbol{X}_h^{k-1}$ is a set composed of $2^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ Jacobians (*cf.* [88, 89]).

1.3.1 Example of the semismooth method

We explicit in this Section the semismooth Newton scheme associated to the min function for problem (1.39) and for problem (1.54).

Lagrange basis, $p = 1$

If we consider the semismooth function min (1.55),

$$\min\{\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}, \boldsymbol{X}_{3h}\} = \min\left\{\begin{pmatrix} u_{1h}(\boldsymbol{a}_1) - u_{2h}(\boldsymbol{a}_1) \\ \vdots \\ u_{1h}(\boldsymbol{a}_{\mathcal{N}_h^{\mathrm{int}}}) - u_{2h}(\boldsymbol{a}_{\mathcal{N}_h^{\mathrm{int}}}) \end{pmatrix}, \begin{pmatrix} \lambda_h(\boldsymbol{a}_1) \\ \vdots \\ \lambda_h(\boldsymbol{a}_{\mathcal{N}_h^{\mathrm{int}}}) \end{pmatrix}\right\},$$

and if we define the block matrices $\mathbb{K}$ and $\mathbb{G}$ in $\mathbb{R}^{\mathcal{N}_h^{\mathrm{int}}, 3\mathcal{N}_h^{\mathrm{int}}}$ respectively by

$$\mathbb{K} := \left[\mathbb{I}_{\mathrm{d}\mathcal{N}_h^{\mathrm{int}} \times \mathcal{N}_h^{\mathrm{int}}}, -\mathbb{I}_{\mathrm{d}\mathcal{N}_h^{\mathrm{int}} \times \mathcal{N}_h^{\mathrm{int}}}, \mathbf{0}_{\mathcal{N}_h^{\mathrm{int}} \times \mathcal{N}_h^{\mathrm{int}}}\right],$$
$$\mathbb{G} := \left[\mathbf{0}_{\mathcal{N}_h^{\mathrm{int}} \times \mathcal{N}_h^{\mathrm{int}}}, \mathbf{0}_{\mathcal{N}_h^{\mathrm{int}} \times \mathcal{N}_h^{\mathrm{int}}}, \mathbb{I}_{\mathrm{d}\mathcal{N}_h^{\mathrm{int}} \times \mathcal{N}_h^{\mathrm{int}}}\right],$$

the l^{th} row of the Jacobian matrix in the sense of Clarke $\mathbf{J}_{\mathbf{C}}(\boldsymbol{X}_h^{k-1})$ is either given by the l^{th} row of $\mathbb{K}$ if $u_{1h}^{k-1}(\boldsymbol{a}_l) - u_{2h}^{k-1}(\boldsymbol{a}_l) \leq \lambda_h^{k-1}(\boldsymbol{a}_l)$, or by the l^{th} row of $\mathbb{G}$ if $u_{1h}^{k-1}(\boldsymbol{a}_l) - u_{2h}^{k-1}(\boldsymbol{a}_l) > \lambda_h^{k-1}(\boldsymbol{a}_l)$.

Dual basis, $p \geq 2$

Considering the semismooth function min (1.55), we have for $p \geq 2$

$$\min\{\boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}, \boldsymbol{X}_{3h}\} = \min\left\{\begin{pmatrix} u_{1h}(\boldsymbol{x}_1) - u_{2h}(\boldsymbol{x}_1) \\ \vdots \\ u_{1h}(\boldsymbol{x}_{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}) - u_{2h}(\boldsymbol{x}_{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}) \end{pmatrix}, \begin{pmatrix} (\boldsymbol{X}_{3h})_1 \\ \vdots \\ (\boldsymbol{X}_{3h})_{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} \end{pmatrix}\right\},$$

and if we define the block matrices $\mathbb{K}$ and $\mathbb{G}$ in $\mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}, 3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ respectively by

$$\mathbb{K} := \left[\mathbb{I}_{\mathrm{d}\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} \times \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}, -\mathbb{I}_{\mathrm{d}\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} \times \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}, \mathbf{0}_{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} \times \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}\right],$$
$$\mathbb{G} := \left[\mathbf{0}_{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} \times \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}, \mathbf{0}_{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} \times \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}, \mathbb{I}_{\mathrm{d}\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} \times \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}\right],$$

the l^{th} row of the Jacobian matrix in the sense of Clarke $\mathbf{J}_{\mathbf{C}}(\boldsymbol{X}_h^{k-1})$ is either given by the l^{th} row of $\mathbb{K}$ if $u_{1h}^{k-1}(\boldsymbol{x}_l) - u_{2h}^{k-1}(\boldsymbol{x}_l) \leq \left(\boldsymbol{X}_{3h}^{k-1}\right)_l$, or by the l^{th} row of $\mathbb{G}$ if $u_{1h}^{k-1}(\boldsymbol{x}_l) - u_{2h}^{k-1}(\boldsymbol{x}_l) > \left(\boldsymbol{X}_{3h}^{k-1}\right)_l$.

1.3.2 Algebraic resolution (general case $p \geq 1$)

It is reasonable to consider a semismooth solver that converges, although we would like to stress that it is not necessary for the validity of the a posteriori estimate we derive. Suppose now that some iterative algebraic solver is applied to the linear system (1.58). Given an initial vector $\boldsymbol{X}_h^{k,0} \in \mathbb{R}^{3\mathcal{N}_{\rm d}^{p,\rm int}}$, often taken as $\boldsymbol{X}_h^{k,0} = \boldsymbol{X}_h^{k-1}$, this yields on step $i \geq 1$ an approximation $\boldsymbol{X}_h^{k,i}$ to $\boldsymbol{X}_h^k$ satisfying

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^{k,i} = \boldsymbol{B}^{k-1} - \boldsymbol{R}_h^{k,i}, \tag{1.60}$$

where $\boldsymbol{R}_h^{k,i} := \boldsymbol{B}^{k-1} - \mathbb{A}^{k-1}\boldsymbol{X}_h^{k,i} \in \mathbb{R}^{3\mathcal{N}_{\rm d}^{p,\rm int}}$ is the algebraic residual vector. Note that $\boldsymbol{R}_h^{k,i}$ has a block structure of the form $\left(\boldsymbol{R}_h^{k,i}\right)^T := \left(\boldsymbol{R}_{1h}^{k,i}, \boldsymbol{R}_{2h}^{k,i}, \boldsymbol{R}_{3h}^{k,i}\right)^T$, with $\boldsymbol{R}_{1h}^{k,i} \in \mathcal{N}_{\rm d}^{p,\rm int}$ corresponds to the test functions v_{1h} in the first line of (1.24) (with $v_{2h} = 0$), $\boldsymbol{R}_{2h}^{k,i}$ corresponds to the test functions v_{2h} in the first line of (1.24) (with $v_{1h} = 0$), and issues from the second line of (1.24) the complementarity constraints (1.31). Following [136], we associate respectively with $\boldsymbol{R}_{1h}^{k,i}$ and $\boldsymbol{R}_{2h}^{k,i}$ elementwise discontinuous polynomials $r_{1h}^{k,i}$ and $r_{2h}^{k,i}$ of degree $p \geq 1$ that vanish on the boundary of Ω. These can be easily computed solving on each element $K \in \mathcal{T}_h$ a small problem with an element mass matrix given as follows. For $\boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm int}$, denote by $N_{h,\boldsymbol{x}_l}$ the number of mesh elements forming the support of the basis function $\psi_{h,\boldsymbol{x}_l}$. Then, $\forall K \in \mathcal{T}_h$, $\forall \alpha \in \{1,2\}$, define $r_{\alpha h}^{k,i}|_K \in \mathbb{P}_p(K)$ by

$$(r_{\alpha h}^{k,i}, \psi_{h,\boldsymbol{x}_l})_K := \frac{(\boldsymbol{R}_{\alpha h}^{k,i})_l}{N_{h,\boldsymbol{x}_l}} \quad \text{and} \quad r_{\alpha h}^{k,i}|_{\partial K \cap \partial\Omega} := 0$$

for all basis functions $\psi_{h,\boldsymbol{x}_l}, \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm int}$ nonzero on K. It is easily seen that the first $2\mathcal{N}_{\rm d}^{p,\rm int}$ lines of (1.60) read, *cf.*(1.24) and (1.15),

$$\begin{aligned} \mu_1 \left(\boldsymbol{\nabla} u_{1h}^{k,i}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{x}_l}\right)_\Omega &= \left(f_1 + \tilde{\lambda}_{h,l}^{k,i} - r_{1h}^{k,i}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega \quad \forall l = 1, \ldots, \mathcal{N}_{\rm d}^{p,\rm int}, \\ \mu_2 \left(\boldsymbol{\nabla} u_{2h}^{k,i}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{x}_l}\right)_\Omega &= \left(f_2 - \tilde{\lambda}_{h,l}^{k,i} - r_{2h}^{k,i}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega \quad \forall l = 1, \ldots, \mathcal{N}_{\rm d}^{p,\rm int}, \end{aligned} \tag{1.61}$$

where

$$\tilde{\lambda}_{h,l}^{k,i} = \begin{cases} \lambda_h^{k,i}(\boldsymbol{x}_l) \text{ (real number given by the vertex value of } \lambda_h^{k,i}) \text{ if } p = 1 \text{ and } \mathcal{N}_{\rm d}^{p,\rm int} = \mathcal{V}_h^{\rm int} \\ \lambda_h^{k,i} \text{ (function } \lambda_h^{k,i}, \text{ the index } l \text{ is discarded) if } p \geq 2. \end{cases} \tag{1.62}$$

In the sequel, we also use the shorthand notation

$$\tilde{\lambda}_{h,\boldsymbol{a}}^{k,i} = \begin{cases} \lambda_h^{k,i}(\boldsymbol{a}) & \text{if} \quad p = 1 \\ \lambda_h^{k,i} & \text{if} \quad p \geq 2 \end{cases}. \tag{1.63}$$

Using the representations $r_{\alpha h}^{k,i}$ of the algebraic vectors and (1.61) will be useful to formulate our a posteriori estimators in Section 1.5.

1.4 Flux reconstructions

This section introduces flux reconstructions that will be central in our a posteriori error analysis, following some general concepts in [33, 66, 81], see also the references therein. Let $k \geq 1$ be a semismooth linearization step and $i \geq 1$ be a linear solver step. Denote by $\mathbf{\Pi}_{\mathbb{P}_p}$ the L^2-orthogonal projection onto the space $\mathbb{P}_p(\mathcal{T}_h)$ of piecewise discontinuous polynomials of order $p \geq 1$. Our first goal will be to fulfill:

Assumption 1.4.1. *There exist* $\boldsymbol{\sigma}_{\alpha h}^{k,i} \in \mathbf{H}(\mathrm{div}, \Omega)$, $\alpha \in \{1, 2\}$, *such that*

$$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i} = \mathbf{\Pi}_{\mathbb{P}_p}(f_\alpha) - (-1)^\alpha \lambda_h^{k,i} \quad \in \mathbb{P}_p(\mathcal{T}_h). \tag{1.64}$$

Recall that by definition, the strong form (1.64) implies

$$\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i} + (-1)^\alpha \lambda_h^{k,i}, q_h\right)_K = (f_\alpha, q_h)_K \quad \forall q_h \in \mathbb{P}_p(K), \forall K \in \mathcal{T}_h,$$

where, recall, $\mathbb{P}_p(K)$ stands for the set of polynomials of degree at most p on the element $K \in \mathcal{T}_h$. Second, following [81], in order to distinguish the algebraic and discretization error components, we make:

Assumption 1.4.2. *There exist* $(\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}, \boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i}) \in [\mathbf{H}(\mathrm{div}, \Omega)]^2$, *such that*

$$\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} + \boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i} = \boldsymbol{\sigma}_{\alpha h}^{k,i} \quad \textit{and} \quad \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} = r_{\alpha h}^{k,i} \quad \forall \alpha \in \{1, 2\}\,.$$

Remark 1.4.3. *The construction of the fluxes is based on the first two diffusion equations in* (1.5) *that are linear. Thus we do not need to construct any linearization fluxes as in [81]. These reconstructed fluxes* $\boldsymbol{\sigma}_{\alpha h}^{k,i}$ *are an approximation in* $\mathbf{H}(\mathrm{div}, \Omega)$ *to the opposite of the gradient of* $u_{\alpha h}^{k,i}$ *multiplied by* μ_α *and are supposed to be separated into two contributions: one essentially lifts the algebraic residual, while the other is assumed to deal with the discretization error.*

1.4.1 Discretization flux reconstruction

We now provide a way to obtain the discretization flux reconstructions $(\boldsymbol{\sigma}_{1h,\mathrm{disc}}^{k,i}, \boldsymbol{\sigma}_{2h,\mathrm{disc}}^{k,i}) \in [\mathbf{H}(\mathrm{div}, \Omega)]^2$. This is done via solution of local mixed systems, on the patches $\omega_h^{\boldsymbol{a}}$ around the mesh vertices $\boldsymbol{a} \in \mathcal{V}_h$ of the mesh $\mathcal{T}_h$. The Raviart–Thomas spaces of order $p \geq 1$ [38, 50, 142, 146] are defined by

$$\mathbf{RT}_p(\Omega) := \{\boldsymbol{\tau}_h \in \mathbf{H}(\mathrm{div}, \Omega),\ \boldsymbol{\tau}_h|_K \in \mathbf{RT}_p(K)\ \ \forall K \in \mathcal{T}_h\}\,,$$

where $\mathbf{RT}_p(K) := [\mathbb{P}_p(K)]^2 + \vec{\boldsymbol{x}}\mathbb{P}_p(K)$, with $\vec{\boldsymbol{x}} = [x_1, x_2]^T$. For $\boldsymbol{a} \in \mathcal{V}_h$, let

$$\mathbf{RT}_p(\omega_h^{\boldsymbol{a}}) := \{\boldsymbol{\tau}_h \in \mathbf{H}(\mathrm{div}, \omega_h^{\boldsymbol{a}}),\ \boldsymbol{\tau}_h|_K \in \mathbf{RT}_p(K),\ \forall K \in \mathcal{T}_h \text{ such that } K \subset \omega_h^{\boldsymbol{a}}\}\,,$$

and let $\mathbb{P}_p(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}})$ stand for piecewise discontinuous polynomials of order $p \geq 1$ in the patch $\omega_h^{\boldsymbol{a}}$. Define consequently the spaces $\mathbf{V}_h^{\boldsymbol{a}}$ and $Q_h^{\boldsymbol{a}}$ when $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$ by

$$\mathbf{V}_h^{\boldsymbol{a}} := \left\{\boldsymbol{\tau}_h \in \mathbf{RT}_p(\omega_h^{\boldsymbol{a}}),\ \boldsymbol{\tau}_h\cdot\mathbf{n}_{\omega_h^{\boldsymbol{a}}} = 0 \text{ on } \partial\omega_h^{\boldsymbol{a}}\right\},\quad Q_h^{\boldsymbol{a}} := \left\{q_h \in \mathbb{P}_p(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}}),\ (q_h, 1)_{\omega_h^{\boldsymbol{a}}} = 0\right\} \tag{1.65}$$

and when $\boldsymbol{a} \in \mathcal{V}_h^{\text{ext}}$ by

$$\mathbf{V}_h^{\boldsymbol{a}} := \left\{ \boldsymbol{\tau}_h \in \mathbf{RT}_p(\omega_h^{\boldsymbol{a}}),\ \boldsymbol{\tau}_h \cdot \mathbf{n}_{\omega_h^{\boldsymbol{a}}} = 0 \ \text{ on } \ \partial\omega_h^{\boldsymbol{a}} \backslash \partial\Omega \right\}, \quad Q_h^{\boldsymbol{a}} := \mathbb{P}_p(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}}). \tag{1.66}$$

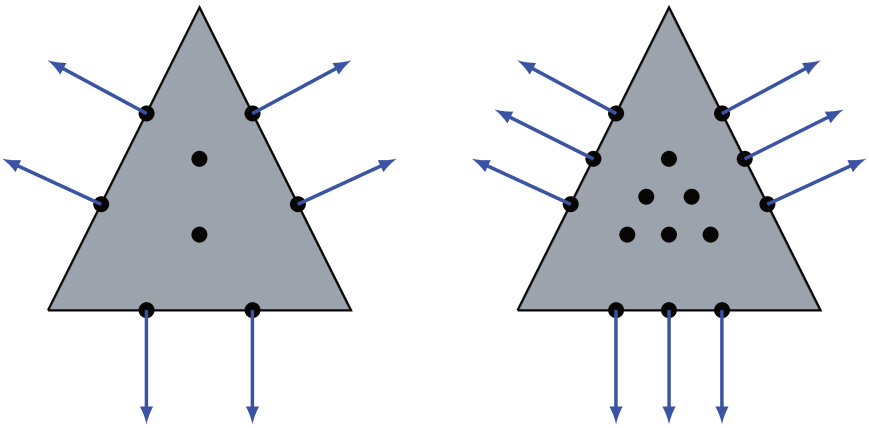

Figure 1.2: Left: Degrees of freedom for the $\mathbf{RT}_1$ space in a triangle. Right: Degrees of freedom for the $\mathbf{RT}_2$ space in a triangle.

For a triangle K the dimension of $\mathbf{RT}_p(K)$ is equal to $(p+1)(p+3)$ (see also Figure 1.2 for a quick illustration) and for a function $\boldsymbol{v} \in \mathbf{RT}_p(K)$ one has

$$\boldsymbol{v} = \sum_{j=1}^{(p+1)(p+3)} N_j(\boldsymbol{v}) \cdot \boldsymbol{\Phi}_j$$

where $(\boldsymbol{\Phi}_j)_{1 \le j \le (p+1)(p+3)}$ are the Raviart–Thomas basis functions and $N_j(\boldsymbol{v})$ the degrees of freedom. Furthermore the degrees of freedom are given by

$$\begin{cases} (\boldsymbol{v} \cdot \boldsymbol{n}_K, w)_e & \forall w \in \mathbb{P}_p(e), \quad e \in \partial K, \\ (\boldsymbol{v}, \boldsymbol{w})_K & \forall \boldsymbol{w} \in (\mathbb{P}_{p-1}(K))^2 \end{cases} \tag{1.67}$$

where $\boldsymbol{n}_K$ is the outward unit normal to K. Therefore, for each edge $e \in \partial K$ we have $3\,(p+1)$ degrees of freedom given by the first line of (1.67), and for each element $K \in \mathcal{T}_h$, we have $p\,(p+1)$ degrees of freedom given by the second line of (1.67).

Definition 1.4.4. *Let* $\left(u_{1h}^{k,i}, u_{2h}^{k,i}, \lambda_h^{k,i}\right)$ *be the approximate solution given by* (1.60)*, verifying in particular* (1.61)*. For each vertex* $\boldsymbol{a} \in \mathcal{V}_h$*, define* $\boldsymbol{\sigma}_{\alpha h,\text{disc}}^{k,i,\boldsymbol{a}} \in \mathbf{V}_h^{\boldsymbol{a}}$ *and* $\gamma_{\alpha h}^{k,i,\boldsymbol{a}} \in Q_h^{\boldsymbol{a}}$*, by solving:*

$$\begin{aligned} \left(\boldsymbol{\sigma}_{\alpha h,\text{disc}}^{k,i,\boldsymbol{a}}, \boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} - \left(\gamma_{\alpha h}^{k,i,\boldsymbol{a}}, \boldsymbol{\nabla}\cdot\boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} &= -\left(\mu_\alpha \psi_{h,\boldsymbol{a}} \boldsymbol{\nabla} u_{\alpha h}^{k,i}, \boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} && \forall \boldsymbol{\tau}_h \in \mathbf{V}_h^{\boldsymbol{a}}, \\ \left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\text{disc}}^{k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} &= \left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} && \forall q_h \in Q_h^{\boldsymbol{a}}, \end{aligned} \tag{1.68}$$

where the right-hand sides are defined by

$$\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}} := \left(f_\alpha - (-1)^\alpha \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i} - r_{\alpha h}^{k,i}\right)\psi_{h,\boldsymbol{a}} - \mu_\alpha \boldsymbol{\nabla} u_{\alpha h}^{k,i} \cdot \boldsymbol{\nabla}\psi_{h,\boldsymbol{a}} \quad \forall \boldsymbol{a} \in \mathcal{V}_h. \tag{1.69}$$

where we recall the notation (1.62)*. Then set*

$$\boldsymbol{\sigma}_{1h,\mathrm{disc}}^{k,i} := \sum_{\boldsymbol{a}\in\mathcal{V}_h} \boldsymbol{\sigma}_{1h,\mathrm{disc}}^{k,i,\boldsymbol{a}} \quad \textit{and} \quad \boldsymbol{\sigma}_{2h,\mathrm{disc}}^{k,i} := \sum_{\boldsymbol{a}\in\mathcal{V}_h} \boldsymbol{\sigma}_{2h,\mathrm{disc}}^{k,i,\boldsymbol{a}}. \tag{1.70}$$

Remark 1.4.5. *Using* (1.61)*, we have*

$$\begin{aligned}\left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, 1\right)_{\omega_h^{\boldsymbol{a}}} =& \left(f_\alpha, \psi_{h,\boldsymbol{a}}\right)_{\omega_h^{\boldsymbol{a}}} - (-1)^\alpha \left(\tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}, \psi_{h,\boldsymbol{a}}\right)_{\omega_h^{\boldsymbol{a}}} - \left(r_{\alpha h}^{k,i}, \psi_{h,\boldsymbol{a}}\right)_{\omega_h^{\boldsymbol{a}}} \\ & - \mu_\alpha \left(\boldsymbol{\nabla} u_{\alpha h}^{k,i}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{a}}\right)_{\omega_h^{\boldsymbol{a}}} \\ =& 0.\end{aligned} \tag{1.71}$$

Then, the Neumann compatibility condition is satisfied for (1.68)*.*

Proposition 1.4.6. *The flux reconstruction* $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i} \in \mathbf{H}(\mathrm{div},\Omega)$ *and satisfies the equilibration*

$$\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i}, q_h\right)_K = \left(f_\alpha - (-1)^\alpha \lambda_h^{k,i} - r_{\alpha h}^{k,i}, q_h\right)_K \quad \forall q_h \in \mathbb{P}_p(K). \tag{1.72}$$

Proof. First, we have $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}} \in \mathbf{H}(\mathrm{div},\omega_h^{\boldsymbol{a}})$. Noting that $\Omega = \omega_h^{\boldsymbol{a}} \cup (\Omega\backslash\omega_h^{\boldsymbol{a}})$ (see Figure 1.3) and extending $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}$ by zero outside $\omega_h^{\boldsymbol{a}}$ we obtain (thanks to (1.65)-(1.66)) that $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}} \in \mathbf{H}(\mathrm{div},\Omega)$ and thus $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i} \in \mathbf{H}(\mathrm{div},\Omega)$.

Next, for $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$, defining $\widetilde{Q}_h^{\boldsymbol{a}} := \mathbb{P}_p(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}})$ we have obviously the inclusion $Q_h^{\boldsymbol{a}} \subset \widetilde{Q}_h^{\boldsymbol{a}}$ and the decomposition $q_h = q_h^* + c^*$ where $q_h \in \widetilde{Q}_h^{\boldsymbol{a}}$, $q_h^* \in Q_h^{\boldsymbol{a}}$, and $c^* = \frac{1}{|\omega_h^{\boldsymbol{a}}|}(q_h, 1)_{\omega_h^{\boldsymbol{a}}}$ a constant. Note that the definition of the spaces $\mathbf{V}_h^{\boldsymbol{a}}$ (1.65)-(1.66) in combination with the Green formula give $\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, 1\right)_{\omega_h^{\boldsymbol{a}}} = \left(\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}\cdot\boldsymbol{n}_{\omega_h^{\boldsymbol{a}}}, 1\right)_{\partial\omega_h^{\boldsymbol{a}}} = 0$. Thus,

$$\begin{aligned}\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} &= \left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, q_h^* + c^*\right)_{\omega_h^{\boldsymbol{a}}}, \\ &= \left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, q_h^*\right)_{\omega_h^{\boldsymbol{a}}} + c^* \left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, 1\right)_{\omega_h^{\boldsymbol{a}}}, \\ &= \left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, q_h^*\right)_{\omega_h^{\boldsymbol{a}}}, \\ &= \left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, q_h^*\right)_{\omega_h^{\boldsymbol{a}}}.\end{aligned}$$

Employing the Neumann compatibility property (1.71) we get

$$\left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, q_h^*\right)_{\omega_h^{\boldsymbol{a}}} = \left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, q_h - c^*\right)_{\omega_h^{\boldsymbol{a}}} = \left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}}. \tag{1.73}$$

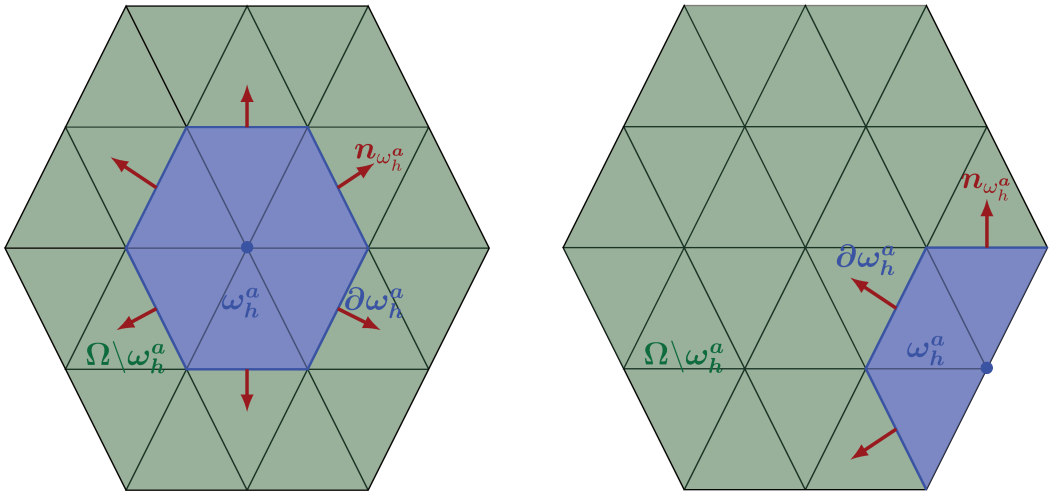

Figure 1.3: Left: internal patch (blue). Right: boundary patch (blue).

Finally,

$$\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}},q_h\right)_{\omega_h^{\boldsymbol{a}}}=\left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}},q_h\right)_{\omega_h^{\boldsymbol{a}}}\quad\forall q_h\in\widetilde{Q}_h^{\boldsymbol{a}}. \tag{1.74}$$

We thus observe that the condition (1.74) holds for all piecewise polynomials belonging to $\mathbb{P}_p(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}})$ and not only for those with zero mean value. Next, crucial argument is that the polynomials in $\mathbb{P}_p(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}})$ are discontinuous, so that we can restrict q_h to any element $K\in\mathcal{T}_h$. If $\boldsymbol{a}\in\mathcal{V}_h^{\mathrm{ext}}$, this is immediate. Then, for any $q_h\in\mathbb{P}_p(K)$ we get

$$\begin{aligned}\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i},q_h\right)_K&=\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}},q_h\right)_K=\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}},q_h\right)_K\\&=\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\left(f_\alpha-(-1)^\alpha\tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}-r_{\alpha h}^{k,i}\right)\psi_{h,\boldsymbol{a}}-\mu_\alpha\boldsymbol{\nabla}u_{\alpha h}^{n,k,i}\cdot\boldsymbol{\nabla}\psi_{h,\boldsymbol{a}},q_h\right)_K.\end{aligned} \tag{1.75}$$

Remarking that $\sum_{\boldsymbol{a}\in\mathcal{V}_K}\psi_{h,\boldsymbol{a}}|_K=1$, $\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\lambda_h^{k,i}\psi_{h,\boldsymbol{a}}\right)|_K=\lambda_h^{k,i}|_K$ for $p\geq 2$, and $\left(\sum_{\boldsymbol{a}\in\mathcal{V}_K}\tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\psi_{h,\boldsymbol{a}}\right)|_K=\left(\sum_{\boldsymbol{a}\in\mathcal{V}_K}\lambda_h^{k,i}(\boldsymbol{a})\psi_{h,\boldsymbol{a}}\right)|_K=\lambda_h^{k,i}|_K$ if $p=1$, we get

$$\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i},q_h\right)_K=\left(f_\alpha-(-1)^\alpha\lambda_h^{k,i}-r_{\alpha h}^{k,i},q_h\right)_K. \tag{1.76}$$

□

Remark 1.4.7. *When the source terms f_α are polynomials, (1.72) reads*

$$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i}=f_\alpha-(-1)^\alpha\lambda_h^{k,i}-r_{\alpha h}^{k,i}. \tag{1.77}$$

Furthermore, (1.77) implies that Definition 1.4.4 combined with the rest of Assumption 1.4.2 concerning $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}$.

1.4.2 Algebraic flux reconstruction via a multilevel approach

In this section, we describe the algebraic flux reconstructions $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}$ following [136]. Unlike the reconstruction of the discretization flux, which works on the given mesh $\mathcal{T}_h$ only, we need to suppose here the existence of a multilevel hierarchy of meshes $\mathcal{T}_j$ that are nested in the sense that $\mathcal{T}_j$ is a refinement of $\mathcal{T}_{j-1}, 1 \le j \le J$, $\mathcal{T}_h = \mathcal{T}_J$, see Figure 1.4. The set of vertices of $\mathcal{T}_j$ is denoted by $\mathcal{V}_j$ and it is partitioned into interior vertices $\mathcal{V}_j^{\mathrm{int}}$ and boundary vertices $\mathcal{V}_j^{\mathrm{ext}}$. For each vertex $\boldsymbol{a} \in \mathcal{V}_j$, there is one hat basis function denoted by $\psi_{j,\boldsymbol{a}}$, with support $\omega_j^{\boldsymbol{a}}$ (so $\psi_{J,\boldsymbol{a}} = \psi_{h,\boldsymbol{a}}$ and $\omega_J^{\boldsymbol{a}} = \omega_h^{\boldsymbol{a}}$). Let X_0^0 be the space of continuous piecewise affine polynomials on $\mathcal{T}_0$. Therein, we

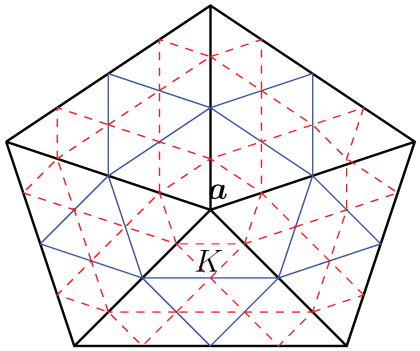

Figure 1.4: Example of nested meshes with $J = 2$. Coarsest mesh $\mathcal{T}_0$ with $\boldsymbol{a} \in \mathcal{V}_0$ and $\omega_0^{\boldsymbol{a}}$ constituted by 5 elements (black, thick), first refined mesh $\mathcal{T}_1$ (blue, thin), and second refined mesh $\mathcal{T}_2 = \mathcal{T}_h$ (red, dashed). $\mathbf{V}_{1,0}^{\boldsymbol{a}}$ consists of $\mathbf{RT}_p$ functions associated with blue (thin) elements and edges that lie inside $\omega_0^{\boldsymbol{a}}$.

construct two Riesz representers $\rho_{\alpha 0}^{k,i}$ of the algebraic residuals $r_{\alpha h}^{k,i}$, $\alpha = 1, 2$, by

$$\left(\boldsymbol{\nabla}\rho_{10}^{k,i}, \boldsymbol{\nabla}v_0\right) = \left(r_{1h}^{k,i}, v_0\right) \ \forall v_0 \in X_0^0, \quad \left(\boldsymbol{\nabla}\rho_{20}^{k,i}, \boldsymbol{\nabla}v_0\right) = \left(r_{2h}^{k,i}, v_0\right) \ \forall v_0 \in X_0^0.$$

Next, for $\boldsymbol{a} \in \mathcal{V}_{j-1}$, $1 \le j \le J$, let

$$\begin{aligned}
\mathbf{RT}_p(\omega_{j-1}^{\boldsymbol{a}}) &:= \left\{\boldsymbol{\tau}_j \in \mathbf{H}(\mathrm{div}, \omega_{j-1}^{\boldsymbol{a}}),\ \boldsymbol{\tau}_j|_K \in \mathbf{RT}_p(K),\ \forall K \in \mathcal{T}_j \text{ such that } K \subset \omega_{j-1}^{\boldsymbol{a}}\right\},\\
\mathbb{P}_p(\mathcal{T}_j|_{\omega_{j-1}^{\boldsymbol{a}}}) &:= \left\{q_j \in L^2(\omega_{j-1}^{\boldsymbol{a}}),\ q_j|_K \in \mathbb{P}_p(K),\ \forall K \in \mathcal{T}_j \text{ such that } K \subset \omega_{j-1}^{\boldsymbol{a}}\right\}.
\end{aligned}$$

We then define

$$\begin{aligned}
\mathbf{V}_{j,j-1}^{\boldsymbol{a}} &:= \left\{\boldsymbol{\tau}_j \in \mathbf{RT}_p(\omega_{j-1}^{\boldsymbol{a}}),\ \boldsymbol{v}_j{\cdot}\mathbf{n}_{\omega_{j-1}^{\boldsymbol{a}}} = 0 \text{ on } \partial\omega_{j-1}^{\boldsymbol{a}}\right\},\\
Q_{j,j-1}^{\boldsymbol{a}} &:= \left\{q_j \in \mathbb{P}_p(\mathcal{T}_j|_{\omega_{j-1}^{\boldsymbol{a}}}),\ (q_j, 1)_{\omega_{j-1}^{\boldsymbol{a}}} = 0\right\},
\end{aligned} \tag{1.78}$$

when $\boldsymbol{a} \in \mathcal{V}_{j-1}^{\mathrm{int}}$ and

$$\mathbf{V}_{j,j-1}^{\boldsymbol{a}} := \left\{\boldsymbol{\tau}_j \in \mathbf{RT}_p(\omega_{j-1}^{\boldsymbol{a}}),\ \boldsymbol{\tau}_j{\cdot}\mathbf{n}_{\omega_{j-1}^{\boldsymbol{a}}} = 0 \text{ on } \partial\omega_{j-1}^{\boldsymbol{a}}\backslash\partial\Omega\right\}, \quad Q_{j,j-1}^{\boldsymbol{a}} := \mathbb{P}_p(\mathcal{T}_j|_{\omega_{j-1}^{\boldsymbol{a}}}), \tag{1.79}$$

when $\boldsymbol{a} \in \mathcal{V}_{j-1}^{\mathrm{ext}}$. Denote also by Π_{j-1} the $L^2(\Omega)$-orthogonal projection onto the broken space $\mathbb{P}_p(\mathcal{T}_{j-1})$. With an abuse of notation, we set $\Pi_0 := 0$. Following [136], the reconstructions $\boldsymbol{\sigma}_{1j,\mathrm{alg}}^{k,i}$ and $\boldsymbol{\sigma}_{2j,\mathrm{alg}}^{k,i}$ is obtained as follows:

Definition 1.4.8. *Let $(u_{1h}^{k,i}, u_{2h}^{k,i}, \lambda_h^{k,i})$ be the approximate solution given by (1.60), verifying in particular (1.61). Let $1 \le j \le J$. For any $\boldsymbol{a} \in \mathcal{V}_{j-1}$, we prescribe $\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}} \in \mathbf{V}_{j,j-1}^{\boldsymbol{a}}$ and $\gamma_{\alpha j}^{k,i,\boldsymbol{a}} \in Q_{j,j-1}^{\boldsymbol{a}}$ by solving*

$$\begin{aligned} \left(\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}}, \boldsymbol{\tau}_j\right)_{\omega_{j-1}^{\boldsymbol{a}}} - \left(\gamma_{\alpha j}^{k,i,\boldsymbol{a}}, \boldsymbol{\nabla}\cdot\boldsymbol{\tau}_j\right)_{\omega_{j-1}^{\boldsymbol{a}}} &= 0 && \forall \boldsymbol{\tau}_j \in \mathbf{V}_{j,j-1}^{\boldsymbol{a}}, \\ \left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}}, q_j\right)_{\omega_{j-1}^{\boldsymbol{a}}} &= \left(\tilde{g}_{\alpha j}^{k,i,\boldsymbol{a}}, q_j\right)_{\omega_{j-1}^{\boldsymbol{a}}} && \forall q_j \in Q_{j,j-1}^{\boldsymbol{a}}, \end{aligned} \tag{1.80}$$

where the right-hand sides are defined by

$$\tilde{g}_{\alpha j}^{k,i,\boldsymbol{a}} := (\mathrm{Id} - \Pi_{j-1}) \left(r_{\alpha h}^{k,i} \psi_{j-1,\boldsymbol{a}} - \boldsymbol{\nabla}\rho_{\alpha 0}^{k,i} \cdot \boldsymbol{\nabla}\psi_{j-1,\boldsymbol{a}}\right) \quad \forall \boldsymbol{a} \in \mathcal{V}_{j-1}^{\mathrm{int}}.$$

Then set

$$\boldsymbol{\sigma}_{1h,\mathrm{alg}}^{k,i} := \sum_{j=1}^{J} \sum_{\boldsymbol{a}\in\mathcal{V}_{j-1}} \boldsymbol{\sigma}_{1j,\mathrm{alg}}^{k,i,\boldsymbol{a}} \quad \textit{and} \quad \boldsymbol{\sigma}_{2h,\mathrm{alg}}^{k,i} := \sum_{j=1}^{J} \sum_{\boldsymbol{a}\in\mathcal{V}_{j-1}} \boldsymbol{\sigma}_{2j,\mathrm{alg}}^{k,i,\boldsymbol{a}}.$$

Remark 1.4.9. *Observe that*

$$\begin{aligned} \left(\tilde{g}_{\alpha j}^{k,i,\boldsymbol{a}}, 1\right)_{\omega_{j-1}^{\boldsymbol{a}}} &= \left(r_{\alpha h}^{k,i}\psi_{j-1,\boldsymbol{a}}, 1\right)_{\omega_{j-1}^{\boldsymbol{a}}} - \left(\boldsymbol{\nabla}\rho_{\alpha 0}^{k,i} \cdot \boldsymbol{\nabla}\psi_{j-1,\boldsymbol{a}}, 1\right)_{\omega_{j-1}^{\boldsymbol{a}}} \\ &\quad - \left(\Pi_{j-1}\left(r_{\alpha h}^{k,i}\psi_{j-1,\boldsymbol{a}}\right), 1\right)_{\omega_{j-1}^{\boldsymbol{a}}} - \left(\Pi_{j-1}\left(\boldsymbol{\nabla}\rho_{\alpha 0}^{k,i} \cdot \boldsymbol{\nabla}\psi_{j-1,\boldsymbol{a}}\right), 1\right)_{\omega_{j-1}^{\boldsymbol{a}}}, \\ &= 0. \end{aligned}$$

Then, the Neumann compatibility condition is satisfied for problems (1.80).

In this definition, each flux $\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}}$ is computed on a "coarse" patch $\omega_{j-1}^{\boldsymbol{a}}$ (level $j-1$) using $\mathbf{RT}_p$ functions defined on the "fine" mesh $\mathcal{T}_j$ (level j). The source term $\tilde{g}_{\alpha j}^{k,i,\boldsymbol{a}}$ is designed to pass the residual error from one level to the next. Once the coarsest Riesz representers $\rho_{\alpha 0}^{k,i}$ are computed, the flux reconstructions are computed starting from level $j = 1$ up to level J. Crucially, the algebraic flux reconstruction $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}$ satisfy the following equilibration property. The proof is already available in [136] and we report it here for the sake of self-containedness.

Proposition 1.4.10. *For a semismooth iteration $k \ge 1$ and an algebraic solver iteration $i \ge 0$, the algebraic flux $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}$ satisfies*

$$\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} \in \mathbf{H}(\mathrm{div},\Omega) \quad \textit{and} \quad \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} = r_{\alpha h}^{k,i}. \tag{1.81}$$

Proof. Let $\boldsymbol{a} \in \mathcal{V}_{j-1}$. Extending $\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}}$ by zero outside $\omega_{j-1}^{\boldsymbol{a}}$ we obtain using (1.78) and (1.79) that $\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}} \in \mathbf{H}(\mathrm{div},\Omega)$ and thus $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} \in \mathbf{H}(\mathrm{div},\Omega)$. Next,

$$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} = \sum_{j=1}^{J} \sum_{\boldsymbol{a}\in\mathcal{V}_{j-1}} \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}}.$$

The second line of the mixed system (1.80) means that $\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}}$ is the $L^2(\omega_{j-1}^{\boldsymbol{a}})$-orthogonal projection of $\tilde{g}_{\alpha j}^{k,i,\boldsymbol{a}}$ onto $\mathbb{P}_p(\mathcal{T}_j|_{\omega_{j-1}^{\boldsymbol{a}}})$: $\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}} = \Pi_j \tilde{g}_{\alpha j}^{k,i,\boldsymbol{a}}$. Next,

as the meshes are nested, $\Pi_{j-1}v \in \mathbb{P}_p(\mathcal{T}_{j-1}) \subset \mathbb{P}_p(\mathcal{T}_j)\ \forall v \in L^2(\omega_{j-1}^{\boldsymbol{a}})$. Then, $(\Pi_j \circ \Pi_{j-1})(v) = \Pi_j(\Pi_{j-1}(v)) = \Pi_{j-1}(v)$. Finally,

$$\begin{aligned}\Pi_j \tilde{g}_{\alpha j}^{k,i,\boldsymbol{a}} &= \Pi_j\left((\mathrm{Id} - \Pi_{j-1})\left(r_{\alpha h}^{k,i}\psi_{j-1,\boldsymbol{a}} - \boldsymbol{\nabla}\rho_{\alpha 0}^{k,i}\cdot\boldsymbol{\nabla}\psi_{j-1,\boldsymbol{a}}\right)\right)\\ &= \Pi_j\left(r_{\alpha h}^{k,i}\psi_{j-1,\boldsymbol{a}} - \boldsymbol{\nabla}\rho_{\alpha 0}^{k,i}\cdot\boldsymbol{\nabla}\psi_{j-1,\boldsymbol{a}}\right) - \Pi_{j-1}\left(r_{\alpha h}^{k,i}\psi_{j-1,\boldsymbol{a}} - \boldsymbol{\nabla}\rho_{\alpha 0}^{k,i}\cdot\boldsymbol{\nabla}\psi_{j-1,\boldsymbol{a}}\right).\end{aligned}$$

The partition of unity $\displaystyle\sum_{\boldsymbol{a}\in\mathcal{V}_{j-1}} \psi_{j-1,\boldsymbol{a}} = 1$ gives

$$\sum_{\boldsymbol{a}\in\mathcal{V}_{j-1}} \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha j,\mathrm{alg}}^{k,i,\boldsymbol{a}} = \Pi_J r_{\alpha h}^{k,i} - \Pi_{j-1} r_{\alpha h}^{k,i}.$$

Finally,

$$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} = \Pi_J r_{\alpha h}^{k,i} - \Pi_0 r_{\alpha h}^{k,i} = r_{\alpha h}^{k,i}$$

where we recall that $\Pi_0 = 0$ by definition. □

Consequently, it is immediate to check that the flux reconstructions Definitions 1.4.4 and 1.4.8 satisfy Assumption 1.4.2.

1.5 A posteriori error estimates

We derive in this section, for any polynomial degree $p \geq 1$, an a posteriori estimate on the error between the exact solution $\boldsymbol{u}$ and the approximate solution $\boldsymbol{u}_h^{k,i}$ valid at each iteration $k \geq 1$ of a linearization solver and each iteration $i \geq 1$ of the iterative algebraic solver satisfying (1.60), (1.61). The main difficulty is located in the treatment of the constraints (1.31). Indeed, for $p = 1$, $k \geq 1$, $i \geq 0$, in general $(u_{1h}^{k,i} - u_{2h}^{k,i})(\boldsymbol{a}) \not\geq 0$, $\lambda_h^{k,i}(\boldsymbol{a}) \not\geq 0$, $\lambda_h^{k,i}(\boldsymbol{a})\cdot(u_{1h}^{k,i} - u_{2h}^{k,i})(\boldsymbol{a}) \neq 0\ \forall \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$. At convergence, $k \to \infty$, $i \to \infty$, we have $u_{1h}^{k,i} - u_{2h}^{k,i} \geq 0$, $\lambda_h^{k,i} \geq 0$, but still $\lambda_h^{k,i}\cdot\left(u_{1h}^{k,i} - u_{2h}^{k,i}\right) \neq 0$. For $p \geq 2$, $k \geq 1$, $i \geq 0$ and for the decomposition of the complementarity constraints in the Lagrange basis and the dual basis, we have $(u_{1h}^{k,i} - u_{2h}^{k,i})(\boldsymbol{x}_l) \not\geq 0$, $\lambda_h^{k,i}(\boldsymbol{x}_l) \not\geq 0$, $\left(\lambda_h^{k,i}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega \not\geq 0\ \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$, and $\left(\lambda_h^{k,i}, u_{1h}^{k,i} - u_{2h}^{k,i}\right)_\Omega \not\geq 0$ in general. Note that even at convergence, in general $u_{1h}^{k,i} - u_{2h}^{k,i} \not\geq 0$, $\lambda_h^{k,i} \not\geq 0$, and $\lambda_h^{k,i}\cdot\left(u_{1h}^{k,i} - u_{2h}^{k,i}\right) \neq 0$.

In this respect, it will be useful to introduce positive and negative parts of $\lambda_h^{k,i}$,

$$\lambda_h^{k,i} := \lambda_h^{k,i,\mathrm{pos}} + \lambda_h^{k,i,\mathrm{neg}}, \quad \lambda_h^{k,i,\mathrm{pos}} := \max\{\lambda_h^{k,i}, 0\}, \quad \lambda_h^{k,i,\mathrm{neg}} := \min\{\lambda_h^{k,i}, 0\},$$

and the convex set $\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p$ defined by

$$\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p := \left\{(v_{1h}, v_{2h}) \in X_{gh}^p \times X_{0h}^p,\ v_{1h} - v_{2h} \geq 0\right\} \subset \boldsymbol{\mathcal{K}}_g. \tag{1.82}$$

Note that $\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^1 = \boldsymbol{\mathcal{K}}_{gh}^1$ but only $\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p \subset \boldsymbol{\mathcal{K}}_{gh}^p$ for $p \geq 2$ only. In what follows, we introduce local elementwise estimators in the form $\eta_{\cdot,K}^{k,i}$, $K \in \mathcal{T}_h$, and global estimators by $\eta_\cdot^{k,i} := \left\{\displaystyle\sum_{K\in\mathcal{T}_h}\left(\eta_{\cdot,K}^{k,i}\right)^2\right\}^{\frac{1}{2}}$. The first main result of this article is:

Theorem 1.5.1. *Let $\boldsymbol{u} = (u_1, u_2) \in \boldsymbol{\mathcal{K}}_g$ be the solution of the continuous reduced problem (1.12). Let $\boldsymbol{u}_h^{k,i} = \left(u_{1h}^{k,i}, u_{2h}^{k,i}\right) \in X_{gh}^p \times X_{0h}^p$ and $\lambda_h^{k,i} \in X_h^p$ be the approximation given by (1.60) for any $p \geq 1$, verifying in particular (1.61). Let $\boldsymbol{\sigma}_{1h}^{k,i}$ and $\boldsymbol{\sigma}_{2h}^{k,i}$ be equilibrated flux reconstructions satisfying Assumptions 1.4.1 and 1.4.2. Let $\boldsymbol{s}_h^{k,i} \in \widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p$ be arbitrary. For $\alpha \in \{1,2\}$, define the estimators*

$$\eta_{\mathrm{F},K,\alpha}^{k,i} := \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} u_{\alpha h}^{k,i} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h}^{k,i} \right\|_K, \quad \eta_{\mathrm{osc},K,\alpha} := \frac{h_K}{\pi} \mu_\alpha^{-\frac{1}{2}} \left\| f_\alpha - \boldsymbol{\Pi}_{\mathbb{P}_p}(f_\alpha) \right\|_K,$$

$$\eta_{\mathrm{C},K}^{k,i,\mathrm{pos}} := 2 \left(\lambda_h^{k,i,\mathrm{pos}}, u_{1h}^{k,i} - u_{2h}^{k,i} \right)_K, \quad \eta_1^{k,i} := \left(\sum_{K \in \mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{F},K,\alpha}^{k,i} + \eta_{\mathrm{osc},K,\alpha} \right)^2 \right)^{\frac{1}{2}},$$

$$\eta_{\mathrm{nonc},1,K}^{k,i} := \left| \left| \left| \boldsymbol{s}_h^{k,i} - \boldsymbol{u}_h^{k,i} \right| \right| \right|_K, \qquad \eta_{\mathrm{nonc},2,K}^{k,i} := C_{\mathrm{PF}} h_\Omega \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} \right)^{\frac{1}{2}} \left\| \lambda_h^{k,i,\mathrm{neg}} \right\|_K,$$

$$\eta_{\mathrm{nonc},3,K}^{k,i} := 2 C_{\mathrm{PF}} h_\Omega \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} \right)^{\frac{1}{2}} \left\| \lambda_h^{k,i,\mathrm{pos}} \right\|_\Omega \left| \left| \left| \boldsymbol{s}_h^{k,i} - \boldsymbol{u}_h^{k,i} \right| \right| \right|_K.$$

Then, the following a posteriori error estimate holds:

$$\left| \left| \left| \boldsymbol{u} - \boldsymbol{u}_h^{k,i} \right| \right| \right| \leq \eta^{k,i} := \left\{ \left(\eta_1^{k,i} + \eta_{\mathrm{nonc},1}^{k,i} + \eta_{\mathrm{nonc},2}^{k,i} \right)^2 + \eta_{\mathrm{nonc},3}^{k,i} + \sum_{K \in \mathcal{T}_h} \eta_{\mathrm{C},K}^{k,i,\mathrm{pos}} \right\}^{\frac{1}{2}}. \tag{1.83}$$

Remark 1.5.2. *The estimate (1.83) gives a practical way to bound the energy error between the exact solution $\boldsymbol{u}$ and the approximation $\boldsymbol{u}_h^{k,i}$ on each linearization step $k \geq 1$ and on each linear solver step $i \geq 1$. The estimators of Theorem 1.5.1 reflect various violations of physical properties of the approximate solution $\boldsymbol{u}_h^{k,i}$: $\eta_{\mathrm{F},K,\alpha}^{k,i}$ and $\eta_{\mathrm{osc},K,\alpha}$ represent the nonconformity of the flux,* i.e., *the fact that $-\mu_\alpha \boldsymbol{\nabla} u_{\alpha h}^{k,i} \notin \mathbf{H}(\mathrm{div}, \Omega)$; $\eta_{\mathrm{C},K}^{k,i,\mathrm{pos}}$ reflects inconsistencies in the contact conditions at the discrete level,* i.e., *the fact that $(u_{1h}^{k,i} - u_{2h}^{k,i})\lambda_h^{k,i} \neq 0$; the estimators $\eta_{\mathrm{nonc},1,K}^{k,i}$, $\eta_{\mathrm{nonc},2,K}^{k,i}$, and $\eta_{\mathrm{nonc},3,K}^{k,i}$ stem from the possible departure of the discrete solution $\boldsymbol{u}_h^{k,i}$ from the convex set $\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p$ and the possible negativity of the discrete Lagrange multiplier $\lambda_h^{k,i}$ because of the inexact semismooth linearization ($p \geq 1$) and high-order nonconformity ($p \geq 2$). More precisely, in the case $p = 1$ these three local estimators are nonzero whenever $\boldsymbol{u}_h^{k,i} \notin \boldsymbol{\mathcal{K}}_{gh}^1$ or $\lambda_h^{k,i} \notin \Lambda_h^1$ (recall the respective definitions (1.13) and (1.22)). Here $\boldsymbol{s}_h^{k,i}$ is designed to be an approximation of $\boldsymbol{u}_h^{k,i}$ that lies inside $\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p$, see the possible definition 1.98 below. In Corollary 1.5.6, the estimators $\eta_{\mathrm{F},K,\alpha}^{k,i}$ will be divided in three parts to exhibit the errors contributions that come from from the discretization, semismooth linearization, and linear algebra for the case $p = 1$.*

Remark 1.5.3. *In [19], an a posteriori error estimate between the exact solution $\boldsymbol{u}$ and the discrete $\mathbb{P}_1$ finite element solution $\boldsymbol{u}_h$ given by (1.14) for $p = 1$ not taking into account inexact nonlinear and linear solvers was derived. The estimate of [19, Lemma 3.3] has the form*

$$|||\boldsymbol{u} - \boldsymbol{u}_h||| \leq \left\{ \sum_{K \in \mathcal{T}_h} \left\{ \sum_{\alpha=1}^{2} (\eta_{\mathrm{F},K,\alpha}^{\infty,\infty} + \eta_{\mathrm{osc},K,\alpha})^2 + \eta_{\mathrm{C},K}^{\infty,\infty,\mathrm{pos}} \right\} \right\}^{\frac{1}{2}}, \tag{1.84}$$

where the variables at convergence are denoted with indices $(k,i)=(\infty,\infty)$. *Supposing convergence,* $\boldsymbol{u}_h^{\infty,\infty}=\boldsymbol{u}_h\in\boldsymbol{\mathcal{K}}_{gh}^1$ *(so that one can take* $\boldsymbol{s}_h^{\infty,\infty}=\boldsymbol{u}_h$*),* $\lambda_h^{\infty,\infty}=\lambda_h\in\Lambda_h^1$ *(so that* $\lambda_h^{\infty,\infty,\mathrm{neg}}=0$*),* $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{\infty,\infty}=0$*, and* $\boldsymbol{\sigma}_{\alpha h}^{\infty,\infty}=\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i}$*. Thus* $\eta_{\mathrm{nonc},1,K}^{\infty,\infty}=\eta_{\mathrm{nonc},2,K}^{\infty,\infty}=\eta_{\mathrm{nonc},3,K}^{\infty,\infty}=0$ *and* $\eta_{\mathrm{F},K,\alpha}^{\infty,\infty}$ *does not contain the algebraic flux contribution. Therefore, estimate* (1.83) *for* $p=1$ *takes the same form as* (1.84) *at convergence, which shows the consistency of our approach.*

Proof of Theorem 1.5.1. First, as $\boldsymbol{u}_h^{k,i}$ does not belong to $\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p$ in general, we define the $a-$orthogonal projection $\boldsymbol{s}$ of $\boldsymbol{u}_h^{k,i}$ to the nonempty closed convex set $\boldsymbol{\mathcal{K}}_g$ by

$$a(\boldsymbol{s},\boldsymbol{v}-\boldsymbol{s})\geq a(\boldsymbol{u}_h^{k,i},\boldsymbol{v}-\boldsymbol{s})\quad\forall\boldsymbol{v}\in\boldsymbol{\mathcal{K}}_g, \tag{1.85}$$

where we recall that the bilinear symmetric form a was defined in (1.10). Problem (1.85) is well-posed thanks to the Lions–Stampacchia theorem [122], because a defines a scalar product on $[H_0^1(\Omega)]^2$. Developing the square, the projection $\boldsymbol{s}$ satisfies for each $\boldsymbol{v}\in\boldsymbol{\mathcal{K}}_g$

$$\left|\!\left|\!\left|\boldsymbol{v}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|^2=|\!|\!|\boldsymbol{v}-\boldsymbol{s}|\!|\!|^2+2a(\boldsymbol{v}-\boldsymbol{s},\boldsymbol{s}-\boldsymbol{u}_h^{k,i})+\left|\!\left|\!\left|\boldsymbol{s}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|^2. \tag{1.86}$$

Since $a(\boldsymbol{v}-\boldsymbol{s},\boldsymbol{s}-\boldsymbol{u}_h^{k,i})\geq 0$ from (1.85), taking successively in (1.86) $\boldsymbol{v}=\boldsymbol{u}$ and $\boldsymbol{v}=\boldsymbol{s}_h^{k,i}$ for any $\boldsymbol{s}_h^{k,i}\in\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p\subset\boldsymbol{\mathcal{K}}_g$, we obtain

$$|\!|\!|\boldsymbol{u}-\boldsymbol{s}|\!|\!| \leq \left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|, \tag{1.87}$$

$$\left|\!\left|\!\left|\boldsymbol{s}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right| \leq \left|\!\left|\!\left|\boldsymbol{s}_h^{k,i}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|=\eta_{\mathrm{nonc},1}^{k,i}. \tag{1.88}$$

Second, the energy norm of the error is decomposed as

$$\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|^2=a(\boldsymbol{u}-\boldsymbol{u}_h^{k,i},\boldsymbol{u}-\boldsymbol{u}_h^{k,i})=a(\boldsymbol{u}-\boldsymbol{u}_h^{k,i},\boldsymbol{u}-\boldsymbol{s})+a(\boldsymbol{u}-\boldsymbol{u}_h^{k,i},\boldsymbol{s}-\boldsymbol{u}_h^{k,i}). \tag{1.89}$$

We estimate both terms in (1.89) separately. The second one is bounded by the Cauchy–Schwarz inequality and (1.88),

$$a(\boldsymbol{u}-\boldsymbol{u}_h^{k,i},\boldsymbol{s}-\boldsymbol{u}_h^{k,i})\leq\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|\ \left|\!\left|\!\left|\boldsymbol{s}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|\leq\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|\ \eta_{\mathrm{nonc},1}^{k,i}. \tag{1.90}$$

The rest of the proof is dedicated to bounding the first one.

The reduced problem (1.12) for $\boldsymbol{v}=\boldsymbol{s}\in\boldsymbol{\mathcal{K}}_g$ yields

$$a(\boldsymbol{u},\boldsymbol{u}-\boldsymbol{s})\leq l(\boldsymbol{u}-\boldsymbol{s}). \tag{1.91}$$

Setting $\boldsymbol{w}=\boldsymbol{u}-\boldsymbol{s}$, we estimate the first term in (1.89) using (1.91) and adding and substracting $b(\boldsymbol{w},\lambda_h^{k,i})$ and employing the definitions of b and l of (1.10)

$$\begin{aligned}a(\boldsymbol{u}-\boldsymbol{u}_h^{k,i},\boldsymbol{w})&\leq l(\boldsymbol{w})+b(\boldsymbol{w},\lambda_h^{k,i})-a(\boldsymbol{u}_h^{k,i},\boldsymbol{w})-b(\boldsymbol{w},\lambda_h^{k,i}),\\&=\sum_{\alpha=1}^{2}\left(f_\alpha-(-1)^\alpha\lambda_h^{k,i},w_\alpha\right)_\Omega-\sum_{\alpha=1}^{2}\left(\mu_\alpha\boldsymbol{\nabla}u_{\alpha h}^{k,i},\boldsymbol{\nabla}w_\alpha\right)_\Omega-b(\boldsymbol{w},\lambda_h^{k,i}).\end{aligned} \tag{1.92}$$

Besides, as $\boldsymbol{\sigma}_{\alpha h}^{k,i} \in \mathbf{H}(\mathrm{div},\Omega)$ by Assumption 1.4.1 and since $w_\alpha \in H_0^1(\Omega)$, the Green formula gives

$$\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i}, w_\alpha\right)_\Omega = -\left(\boldsymbol{\sigma}_{\alpha h}^{k,i}, \boldsymbol{\nabla} w_\alpha\right)_\Omega \quad \forall \alpha \in \{1,2\}\,. \tag{1.93}$$

Then, using (1.93) in (1.92), one has

$$\begin{aligned} a(\boldsymbol{u}-\boldsymbol{u}_h^{k,i},\boldsymbol{w}) \le \sum_{\alpha=1}^{2}\sum_{K\in\mathcal{T}_h}\Big\{&\left(f_\alpha-(-1)^\alpha\lambda_h^{k,i}-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i}, w_\alpha\right)_K \\ &-\left(\mu_\alpha^{\frac12}\boldsymbol{\nabla}u_{\alpha h}^{k,i}+\mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h}^{k,i}, \mu_\alpha^{\frac12}\boldsymbol{\nabla}w_\alpha\right)_K\Big\} - b(\boldsymbol{w},\lambda_h^{k,i}). \end{aligned} \tag{1.94}$$

It remains to bound each of the three terms in (1.94). Using the divergence property (1.64) of Assumption 1.4.1, the Cauchy–Schwarz and Poincaré–Wirtinger (1.6b) inequalities, since $w_\alpha \in H^1(K)$, and denoting by $\overline{w}_{\alpha,K}$ the mean of w_α over K, we have for $\alpha = 1,2$

$$\begin{aligned} \left(f_\alpha-\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i}-(-1)^\alpha\lambda_h^{k,i}, w_\alpha\right)_K &= \left(f_\alpha-\boldsymbol{\Pi}_{\mathbb{P}_p}(f_\alpha), w_\alpha-\overline{w}_{\alpha,K}\right)_K, \\ &\le \eta_{\mathrm{osc},K,\alpha}\left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla}w_\alpha\right\|_K. \end{aligned} \tag{1.95}$$

Furthermore, by the Cauchy–Schwarz inequality

$$-\left(\mu_\alpha^{\frac12}\boldsymbol{\nabla}u_{\alpha h}^{k,i}+\mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h}^{k,i}, \mu_\alpha^{\frac12}\boldsymbol{\nabla}w_\alpha\right)_K \le \eta_{\mathrm{F},K,\alpha}^{k,i}\left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla}w_\alpha\right\|_K. \tag{1.96}$$

Next, as $\boldsymbol{u}\in\boldsymbol{\mathcal{K}}_g$, $\boldsymbol{w}=\boldsymbol{u}-\boldsymbol{s}$, and $-b(\boldsymbol{u},\lambda_h^{k,i,\mathrm{pos}}) \le 0$ we have

$$\begin{aligned} -b(\boldsymbol{w},\lambda_h^{k,i}) &\le -b(\boldsymbol{w},\lambda_h^{k,i,\mathrm{neg}}) + b(\boldsymbol{s}-\boldsymbol{u}_h^{k,i},\lambda_h^{k,i,\mathrm{pos}}) + b(\boldsymbol{u}_h^{k,i},\lambda_h^{k,i,\mathrm{pos}}) \\ &\le -\left(\lambda_h^{k,i,\mathrm{neg}}, w_1-w_2\right)_\Omega + \left(\lambda_h^{k,i,\mathrm{pos}}, (s_1-u_{1h}^{k,i})-(s_2-u_{2h}^{k,i})\right)_\Omega \\ &\quad + \frac12\sum_{K\in\mathcal{T}_h} 2\left(\lambda_h^{k,i,\mathrm{pos}}, u_{1h}^{k,i}-u_{2h}^{k,i}\right)_K. \end{aligned}$$

Using (1.7), we see

$$\|\boldsymbol{\nabla}(w_1-w_2)\|_\Omega \le \sum_{\alpha=1}^{2}\mu_\alpha^{-\frac12}\left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla}w_\alpha\right\|_\Omega \le \left(\frac{1}{\mu_1}+\frac{1}{\mu_2}\right)^{\frac12}|||\boldsymbol{w}|||\,.$$

Thus, the Cauchy–Schwarz and Poincaré–Friedrichs (1.6a) inequalities noting that both w_α and $(s_\alpha - u_{\alpha h}^{k,i})$ belong to $H_0^1(\Omega)$, and also employing (1.88) give

$$-b(\boldsymbol{w},\lambda_h^{k,i}) \le \eta_{\mathrm{nonc},2}^{k,i}|||\boldsymbol{w}||| + \frac12\eta_{\mathrm{nonc},3}^{k,i} + \frac12\sum_{K\in\mathcal{T}_h}\eta_{\mathrm{C},K}^{k,i,\mathrm{pos}}. \tag{1.97}$$

Therefore, combining (1.89), (1.90), (1.94), (1.95), (1.96), (1.97), and (1.87), we get

$$\left|\left|\left|\boldsymbol{u}-\boldsymbol{u}_h^{k,i}\right|\right|\right|^2 \le \left(\eta_{\mathrm{nonc},1}^{k,i}+\eta_1^{k,i}+\eta_{\mathrm{nonc},2}^{k,i}\right)\left|\left|\left|\boldsymbol{u}-\boldsymbol{u}_h^{k,i}\right|\right|\right| + \frac12\eta_{\mathrm{nonc},3}^{k,i} + \frac12\sum_{K\in\mathcal{T}_h}\eta_{\mathrm{C},K}^{k,i,\mathrm{pos}}.$$

To conclude, the inequality $AB \le \frac12(A^2+B^2)$ gives the result (1.83). □

The construction of $\boldsymbol{s}_h^{k,i} \in \widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p \subset \boldsymbol{\mathcal{K}}_g$ when the polynomial degree $p \geq 2$ is not easy for implementation. When the polynomial degree $p = 1$, any reasonable definition of $\boldsymbol{s}_h^{k,i}$ should lead to vanishing $\eta_{\mathrm{nonc},1,K}^{k,i}$ and $\eta_{\mathrm{nonc},3,K}^{k,i}$ when the constraint $\boldsymbol{u}_h^{k,i} \in \boldsymbol{\mathcal{K}}_{gh}^1$ is satisfied. A possibility that we will use below in Section 1.8 for numerical experiments is to define $\boldsymbol{s}_h^{k,i} \in \boldsymbol{\mathcal{K}}_{gh}^1 = \widetilde{\boldsymbol{\mathcal{K}}}_{gh}^1 \subset \boldsymbol{\mathcal{K}}_g$ such that for each $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$

$$\boldsymbol{s}_h^{k,i}(\boldsymbol{a}) := \begin{cases} \boldsymbol{u}_h^{k,i}(\boldsymbol{a}) = \left(u_{1h}^{k,i}(\boldsymbol{a}), u_{2h}^{k,i}(\boldsymbol{a})\right) & \text{if } u_{1h}^{k,i}(\boldsymbol{a}) \geq u_{2h}^{k,i}(\boldsymbol{a}), \\ \left(\frac{1}{2}\left(u_{1h}^{k,i}(\boldsymbol{a}) + u_{2h}^{k,i}(\boldsymbol{a})\right), \frac{1}{2}\left(u_{1h}^{k,i}(\boldsymbol{a}) + u_{2h}^{k,i}(\boldsymbol{a})\right)\right) & \text{if } u_{1h}^{k,i}(\boldsymbol{a}) < u_{2h}^{k,i}(\boldsymbol{a}). \end{cases} \tag{1.98}$$

Concerning $\lambda_h^{k,i}$, an estimate is provided in (recall the definition given in (1.8)):

Theorem 1.5.4. *Assume the hypotheses and notations of Theorem 1.5.1 and let $\lambda \in \Lambda$ be the solution of problem (1.9). Then the following a posteriori estimate holds:*

$$\left|\!\left|\!\left|\lambda - \lambda_h^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\Omega)} \leq \eta^{k,i} + \eta_1^{k,i}. \tag{1.99}$$

Proof. The proof follows the one in [19, Corollary 3.5]. We only give the essential elements. Let $\mu_{\mathrm{m}} := \max(\mu_1, \mu_2)$. Employing (1.8) and extending appropriately b,

$$\left|\!\left|\!\left|\lambda - \lambda_h^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\Omega)} = \sup_{\substack{\psi \in H_0^1(\Omega) \\ \mu_{\mathrm{m}}\|\boldsymbol{\nabla}\psi\|_\Omega^2 = 1}} \langle \lambda_h^{k,i} - \lambda, \psi \rangle = \sup_{\substack{\boldsymbol{\phi} \in [H_0^1(\Omega)]^2 \\ \mu_{\mathrm{m}} \sum_{\alpha=1}^2 \|\boldsymbol{\nabla}\phi_\alpha\|^2 = 1}} b(\boldsymbol{\phi}, \lambda_h^{k,i} - \lambda).$$

Fix $\boldsymbol{\phi} \in [H_0^1(\Omega)]^2$ such that $\mu_{\mathrm{m}} \sum_{\alpha=1}^2 \|\boldsymbol{\nabla}\phi_\alpha\|^2 = 1$. Invoking (1.9), we have

$$-b(\boldsymbol{\phi}, \lambda - \lambda_h^{k,i}) = l(\boldsymbol{\phi}) + b(\boldsymbol{\phi}, \lambda_h^{k,i}) - a(\boldsymbol{u}_h^{k,i}, \boldsymbol{\phi}) - a(\boldsymbol{u} - \boldsymbol{u}_h^{k,i}, \boldsymbol{\phi}).$$

The last term is estimated as $-a(\boldsymbol{u} - \boldsymbol{u}_h^{k,i}, \boldsymbol{\phi}) \leq \left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|$, since $|||\boldsymbol{\phi}||| \leq 1$. The first three terms are identical to the first three terms of (1.92) but with $\boldsymbol{\phi} \in [H_0^1(\Omega)]^2$ instead of $\boldsymbol{w}$. Thus, using the estimates (1.95) and (1.96), one gets

$$-b(\boldsymbol{\phi}, \lambda - \lambda_h^{k,i}) \leq \eta_1^{k,i} + \left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|,$$

which combined with (1.83) gives the result. □

Remark 1.5.5. *At convergence, for $\mathbb{P}_1$ finite elements, estimate (1.99) reduces to (3.30) in [19] with a slightly sharper treatment of the oscillation in f_α.*

So far, we have established a posteriori estimates between the exact and approximate solution. When $p = 1$, the nonconformity estimators can be interpreted as semismooth linearization estimators so that we set

$$\eta_{\mathrm{lin},1,K}^{k,i} := \eta_{\mathrm{nonc},1,K}^{k,i}, \quad \eta_{\mathrm{lin},2,K}^{k,i} := \eta_{\mathrm{nonc},2,K}^{k,i}, \quad \eta_{\mathrm{lin},3,K}^{k,i} := \eta_{\mathrm{nonc},3,K}^{k,i}. \tag{1.100}$$

We now provide an estimate distinguishing the different error components, namely the finite element discretization error, the semismooth linearization error, and the linear algebra error for the case $p = 1$. This distinction is heuristic, based on the property that $\eta_{\mathrm{lin}}^{k,i} \to 0$ and $\eta_{\mathrm{alg}}^{k,i} \to 0$ when $k \to 0$ and $i \to 0$. A similar distinction for the several case $p \geq 2$ is possible but a little longer to write down.

Corollary 1.5.6. *Consider the assumptions and notations of Theorem 1.5.1 in the case $p = 1$. Define for $\alpha \in \{1, 2\}$ and for $K \in \mathcal{T}_h$*

$$\eta^{k,i}_{\mathrm{alg},K,\alpha} := \left\| \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{alg}} \right\|_K, \quad \eta^{k,i}_{\mathrm{disc},K,\alpha} := \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} u^{k,i}_{\alpha h} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{disc}} \right\|_K, \tag{1.101a}$$

$$\eta^{k,i}_{\mathrm{disc}} := \left\{ \sum_{\alpha=1}^{2} \sum_{K \in \mathcal{T}_h} \left(\eta^{k,i}_{\mathrm{disc},K,\alpha} + \eta_{\mathrm{osc},K,\alpha} \right)^2 \right\}^{\frac{1}{2}} + \left\{ \sum_{K \in \mathcal{T}_h} \left| \eta^{k,i,\mathrm{pos}}_{\mathrm{C},K} \right| \right\}^{\frac{1}{2}}, \tag{1.101b}$$

$$\eta^{k,i}_{\mathrm{lin}} := \eta^{k,i}_{\mathrm{lin},1} + \eta^{k,i}_{\mathrm{lin},2} + \left(\eta^{k,i}_{\mathrm{lin},3} \right)^{\frac{1}{2}}, \quad \eta^{k,i}_{\mathrm{alg}} := \left\{ \sum_{\alpha=1}^{2} \sum_{K \in \mathcal{T}_h} (\eta^{k,i}_{\mathrm{alg},K,\alpha})^2 \right\}^{\frac{1}{2}}. \tag{1.101c}$$

Then,

$$\left| \left| \left| \boldsymbol{u} - \boldsymbol{u}^{k,i}_h \right| \right| \right| \leq \eta^{k,i}_{\mathrm{disc}} + \eta^{k,i}_{\mathrm{lin}} + \eta^{k,i}_{\mathrm{alg}}.$$

Proof. As for $(A, B) \in \mathbb{R}_+ \times \mathbb{R}_+$, $(A + B)^{\frac{1}{2}} \leq A^{\frac{1}{2}} + B^{\frac{1}{2}}$, we have

$$\eta^{k,i} \leq \eta^{k,i}_1 + \eta^{k,i}_{\mathrm{lin},1} + \eta^{k,i}_{\mathrm{lin},2} + \left(\eta^{k,i}_{\mathrm{lin},3} \right)^{\frac{1}{2}} + \left(\sum_K |\eta^{k,i,\mathrm{pos}}_{\mathrm{C},K}| \right)^{\frac{1}{2}}. \tag{1.102}$$

Next, the definition of $\eta^{k,i}_1$ combined with the triangle inequality to separate the algebraic estimators $\eta^{k,i}_{\mathrm{alg},K,\alpha}$ from the discretization estimators $\eta^{k,i}_{\mathrm{disc},K,\alpha}$ give

$$\eta^{k,i}_1 \leq \left(\sum_{K \in \mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta^{k,i}_{\mathrm{disc},K,\alpha} + \eta_{\mathrm{osc},K,\alpha} \right)^2 \right)^{\frac{1}{2}} + \left(\sum_{K \in \mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta^{k,i}_{\mathrm{alg},K,\alpha} \right)^2 \right)^{\frac{1}{2}}, \tag{1.103}$$

which concludes the proof. □

1.6 Adaptive inexact semismooth Newton method using a posteriori stopping criteria

We propose in this section an adaptive inexact semismooth Newton method. In the spirit of [9, 81, 108, 128], it is designed to only perform the linearization and algebraic resolution with minimal necessary precision and thus to avoid unnecessary iterations. We rely on Corollary 1.5.6 that estimates the different error components and design adaptive stopping criteria for both linearization and algebraic solvers. The results of this section are for simplicity presented for $p = 1$; extension to $p \geq 2$ is merely technical.

1.6.1 Stopping criteria

Recall that we employ a semismooth Newton method for the nonlinear problem (1.57), yielding on each step $k \geq 1$ and each step $i \geq 1$ the linear system (1.60). Let γ_{lin} and γ_{alg} be two positive parameters typically of order 0.1, representing the desired relative size of the algebraic and linearization errors. We propose the following

stopping criteria, balancing globally the algebraic, linearization, and discretization error components:

$$\text{(a) } \eta_{\mathrm{alg}}^{k,i} \leq \gamma_{\mathrm{alg}} \max\left\{\eta_{\mathrm{disc}}^{k,i}, \eta_{\mathrm{lin}}^{k,i}\right\}, \qquad \text{(b) } \eta_{\mathrm{lin}}^{k,i} \leq \gamma_{\mathrm{lin}} \eta_{\mathrm{disc}}^{k,i}. \tag{1.104}$$

Remark 1.6.1. *For $K \in \mathcal{T}_h$, let $\gamma_{\mathrm{lin},K}$, $\gamma_{\mathrm{alg},K}$ be two fixed parameters, typically of order 0.1, representing the desired local relative sizes of the linearization and algebraic errors components. Following [81, 108] and the references therein, one can aim at the balance of all error components in each mesh cell in place of (1.104), while simultaneously guaranteeing the global criteria (1.104). These local criteria read, with*

$$\eta_{\mathrm{lin},K}^{k,i} := \left(1 + \left(2h_\Omega C_{\mathrm{PF}} \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)^{\frac{1}{2}} \frac{\left\|\lambda_h^{k,i,\mathrm{pos}}\right\|_\Omega}{\left|\!\left|\!\left|\boldsymbol{s}_h^{k,i} - \boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|_\Omega}\right)^{\frac{1}{2}}\right) \eta_{\mathrm{lin},1,K}^{k,i} + \eta_{\mathrm{lin},2,K}^{k,i},$$

$$\eta_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha}^{k,i} \leq \min_{K \subset \omega_h^{\boldsymbol{a}}} \left\{\gamma_{\mathrm{alg},K} \max\left\{\eta_{\mathrm{disc},K,\alpha}^{k,i}, \eta_{\mathrm{lin},K}^{k,i}\right\}\right\} \qquad \forall \alpha \in \{1,2\}, \tag{1.105a}$$

$$\eta_{\mathrm{lin},K}^{k,i} \leq \min_{\alpha \in \{1,2\}} \left\{\gamma_{\mathrm{lin},K} \eta_{\mathrm{disc},K,\alpha}^{k,i}\right\}, \tag{1.105b}$$

where

$$\eta_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha}^{k,i} := \left\{\sum_{K \subset \omega_h^{\boldsymbol{a}}} \left(\eta_{\mathrm{alg},K,\alpha}^{k,i}\right)^2\right\}^{\frac{1}{2}} = \left\|\mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}}. \tag{1.106}$$

The (complicated) form of the term $\eta_{\mathrm{lin},K}^{k,i}$ ensures that local criteria (1.105) imply the global criteria (1.104), and stems from the different scalings of $\eta_{\mathrm{lin},1,K}^{k,i}$ and $\eta_{\mathrm{lin},2,K}^{k,i}$ with respect to $\eta_{\mathrm{lin},3,K}^{k,i}$ in Theorem 1.5.1. In particular, local efficiency will be proven below based on (1.105).

Remark 1.6.2. *When $p \geq 2$, we will prove below the local efficiency of the leading estimators from Theorem 1.5.1 directly (recall that we have only introduced Corollary 1.5.6 for $p = 1$). Then, the analogue of the local stopping criterion (1.105a) will be*

$$\eta_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha}^{k,i} \leq \min_{K \subset \omega_h^{\boldsymbol{a}}} \left\{\gamma_{\mathrm{alg},K} \eta_{\mathrm{disc},K,\alpha}^{k,i}\right\} \qquad \forall \alpha \in \{1,2\}, \tag{1.107}$$

where $\eta_{\mathrm{alg},K,\alpha}^{k,i}$ and $\eta_{\mathrm{disc},K,\alpha}^{k,i}$ are given by (1.101a) and $\eta_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha}^{k,i}$ is given by (1.106).

1.6.2 Adaptive inexact semismooth Newton algorithm

The adaptive inexact algorithm that we propose is as follows:

Algorithm 1 Adaptive inexact semismooth Newton algorithm

0. Choose an initial vector $\boldsymbol{X}_h^0 \in \mathbb{R}^{3\mathcal{N}_h^{\rm int}}$ and set $k=1$.

1. From $\boldsymbol{X}_h^{k-1}$ define $\mathbb{A}^{k-1} \in \mathbb{R}^{3\mathcal{N}_h^{\rm int},3\mathcal{N}_h^{\rm int}}$ and $\boldsymbol{B}^{k-1} \in \mathbb{R}^{3\mathcal{N}_h^{\rm int}}$ by (1.59).

2. Consider the linear system

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^k = \boldsymbol{B}^{k-1}. \tag{1.108}$$

3. Set $\boldsymbol{X}_h^{k,0} := \boldsymbol{X}_h^{k-1}$ as initial guess for the iterative linear solver, set $i := 0$.

4*a*. Perform $\nu \geq 1$ steps of a chosen linear solver for (1.108), starting from $\boldsymbol{X}_h^{k,i}$. This yields on step $i+\nu$ an approximation $\boldsymbol{X}_h^{k,i+\nu}$ to $\boldsymbol{X}_h^k$ satisfying

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^{k,i+\nu} = \boldsymbol{B}^{k-1} - \boldsymbol{R}_h^{k,i+\nu}.$$

4*b*. Compute the estimators of Corollary 1.5.6 and check the stopping criterion for the linear solver in the form (1.104)(a). Set $i := i+\nu$. If satisfied, set $\boldsymbol{X}_h^k = \boldsymbol{X}_h^{k,i}$. If not go back to **4*a***.

5. Check the stopping criterion for the nonlinear solver in the form (1.104)(b). If satisfied, return $\boldsymbol{X}_h = \boldsymbol{X}_h^k$. If not, set $k = k+1$ and go back to **1**.

1.7 Efficiency

We prove in this section local efficiency of the a posteriori error estimators of Corollary 1.5.6 for p=1, *i.e.*, we establish that the derived estimators form a local lower bound for the error, up to a generic constant, and up to data oscillation and a typically small contact term also with inexact linearization and algebraic solvers. As a particular consequence, the overall estimate is proven equivalent to the overall error. We proceed in several steps, following [32,81,82,136] and the references therein. First, we introduce primal continuous problems on patches of mesh elements which are such that the energy norms of their solutions represent lower bounds of the error in the patches. Next, we exploit the stability of the local mixed finite element problems in Definition 1.4.4. We finally bound all the estimators by the local discretization estimator up to a constant, relying on the imposed local stopping criteria of (1.105). In the generic case $p \geq 2$, we do not address the inexact linearization solver and show that the leading term in Theorem 1.5.1 is locally efficient. We assume in the sequel for simplicity that f_1 and f_2 are piecewise $\mathbb{P}_p$ polynomials. This obviously yields $\eta_{{\rm osc},K,\alpha} = 0$, $\forall \alpha \in \{1,2\}$. We do not treat here the "complementarity" estimators $\eta_{{\rm C},K}^{k,i,{\rm pos}}$ that are typically numerically very small. Their local efficiency could be proven, when $p=1$, along the lines of [19, Proposition 3.9].

1.7.1 Continuous-level problems with hat functions on patches

For each vertex $\boldsymbol{a} \in \mathcal{V}_h$, define the spaces

$$\begin{aligned} H_*^1(\omega_h^{\boldsymbol{a}}) &:= \left\{ v \in H^1(\omega_h^{\boldsymbol{a}});\ (v,1)_{\omega_h^{\boldsymbol{a}}} = 0 \right\} && \boldsymbol{a} \in \mathcal{V}_h^{\rm int}, \\ H_*^1(\omega_h^{\boldsymbol{a}}) &:= \{ v \in H^1(\omega_h^{\boldsymbol{a}});\ v = 0 \text{ on } \partial\omega_h^{\boldsymbol{a}} \cap \partial\Omega \} && \boldsymbol{a} \in \mathcal{V}_h^{\rm ext}. \end{aligned}$$

Then there exists a constant $C_{\mathrm{cont,PF}} > 0$ only depending on the shape regularity of the mesh $\mathcal{T}_h$ such that

$$\|\boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\|_{\omega_h^{\boldsymbol{a}}} \leq C_{\mathrm{cont,PF}} \|\boldsymbol{\nabla} v\|_{\omega_h^{\boldsymbol{a}}} \quad \forall v \in H^1_*(\omega_h^{\boldsymbol{a}}), \tag{1.109}$$

see Carstensen and Funken [45], Braess *et al.* [32], or Ern and Vohralík [82]. Then, for any $p \geq 1$ we have

Lemma 1.7.1. *Let (u_1, u_2, λ) be the solution of (1.9) and let $(u_{1h}^{k,i}, u_{2h}^{k,i}, \lambda_h^{k,i})$ be the approximation given by (1.60), verifying in particular (1.61). Let $\boldsymbol{a} \in \mathcal{V}_h$, and for $\alpha \in \{1,2\}$ let $\zeta_{\alpha,\boldsymbol{a}} \in H^1_*(\omega_h^{\boldsymbol{a}})$ be the solution of*

$$(\mu_\alpha \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}, \boldsymbol{\nabla} v)_{\omega_h^{\boldsymbol{a}}} = \left(-\mu_\alpha \psi_{h,\boldsymbol{a}} \boldsymbol{\nabla} u_{\alpha h}^{k,i}, \boldsymbol{\nabla} v\right)_{\omega_h^{\boldsymbol{a}}} + \left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, v\right)_{\omega_h^{\boldsymbol{a}}} \quad \forall v \in H^1_*(\omega_h^{\boldsymbol{a}}), \tag{1.110}$$

where $\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}$ is defined in (1.69). Let $\mu_{\mathrm{m}} := \max(\mu_1, \mu_2)$. Then, for $\alpha \in \{1,2\}$,

$$\begin{aligned}\left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}\right\|_{\omega_h^{\boldsymbol{a}}} \leq C_{\mathrm{cont,PF}} \Bigg(&\left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}} + \mu_{\mathrm{m}}^{\frac{1}{2}} \mu_\alpha^{-\frac{1}{2}} \left|\left|\left|\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right|\right|\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})} \\ &+ \left\|\mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}}\Bigg).\end{aligned} \tag{1.111}$$

Proof. Let $\alpha \in \{1,2\}$. There holds

$$\left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}\right\|_{\omega_h^{\boldsymbol{a}}} = \sup_{v \in H^1_*(\omega_h^{\boldsymbol{a}}), \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} v\right\|_{\omega_h^{\boldsymbol{a}}} = 1} \left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}, \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} v\right)_{\omega_h^{\boldsymbol{a}}}. \tag{1.112}$$

Consider $v \in H^1_*(\omega_h^{\boldsymbol{a}})$ with $\left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} v\right\|_{\omega_h^{\boldsymbol{a}}} = 1$. As $\zeta_{\alpha,\boldsymbol{a}}$ is the solution of (1.110), using the definition (1.69) and considering the test functions $(\psi_{h,\boldsymbol{a}} v, 0)$ and $(0, \psi_{h,\boldsymbol{a}} v) \in \left(H_0^1(\omega_h^{\boldsymbol{a}})\right)^2 \subset \left(H_0^1(\Omega)\right)^2$ in (1.9), we obtain

$$\begin{aligned}\left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}, \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} v\right)_{\omega_h^{\boldsymbol{a}}} &= \left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right), \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(\psi_{h,\boldsymbol{a}} v\right)\right)_{\omega_h^{\boldsymbol{a}}} \\ &\quad + \left((-1)^\alpha \left(\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right) - r_{\alpha h}^{k,i}, \psi_{h,\boldsymbol{a}} v\right)_{\omega_h^{\boldsymbol{a}}}.\end{aligned} \tag{1.113}$$

Moreover, as $\psi_{h,\boldsymbol{a}} v \in H_0^1(\omega_h^{\boldsymbol{a}})$, $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} \in \mathbf{H}(\mathrm{div}, \omega_h^{\boldsymbol{a}})$, and $\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} = r_{\alpha h}^{k,i}$ by Assumption 1.4.2, the Green formula and the Cauchy–Schwarz inequality give

$$\begin{aligned}\left|\left(r_{\alpha h}^{k,i}, \psi_{h,\boldsymbol{a}} v\right)_{\omega_h^{\boldsymbol{a}}}\right| &= \left|-\left(\mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}, \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(\psi_{h,\boldsymbol{a}} v\right)\right)_{\omega_h^{\boldsymbol{a}}}\right| \\ &\leq \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(\psi_{h,\boldsymbol{a}} v\right)\right\|_{\omega_h^{\boldsymbol{a}}} \left\|\mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}}.\end{aligned} \tag{1.114}$$

Multiplying and dividing $\left(\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}, \psi_{h,\boldsymbol{a}} v\right)_{\omega_h^{\boldsymbol{a}}}$ by $\left\|\mu_{\mathrm{m}}^{\frac{1}{2}} \boldsymbol{\nabla}\left(\psi_{h,\boldsymbol{a}} v\right)\right\|_{\omega_h^{\boldsymbol{a}}}$ and using that $\psi_{h,\boldsymbol{a}} v \in H_0^1(\omega_h^{\boldsymbol{a}})$ which allows us to employ the definition (1.8), we get

$$\left|\left(\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}, \psi_{h,\boldsymbol{a}} v\right)_{\omega_h^{\boldsymbol{a}}}\right| \leq \left|\left|\left|\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right|\right|\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})} \mu_{\mathrm{m}}^{\frac{1}{2}} \mu_\alpha^{-\frac{1}{2}} \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(\psi_{h,\boldsymbol{a}} v\right)\right\|_{\omega_h^{\boldsymbol{a}}}. \tag{1.115}$$

Finally, the Cauchy–Schwarz inequality leads to

$$\left(\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right), \mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(\psi_{h,\boldsymbol{a}} v\right)\right)_{\omega_h^{\boldsymbol{a}}} \leq \left\|\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}} \left\|\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(\psi_{h,\boldsymbol{a}} v\right)\right\|_{\omega_h^{\boldsymbol{a}}}. \tag{1.116}$$

The result now follows by combining (1.114), (1.115), and (1.116) with (1.109) together with (1.112). □

1.7.2 Local efficiency of the estimators

Recall the definition of $\zeta_{\alpha,\boldsymbol{a}}$ from (1.110) in Lemma 1.7.1. Following [32, 82], there exists a constant $C_{\rm st} > 0$ only depending on the shape regularity of the mesh $\mathcal{T}_h$ such that the discretization flux reconstructions $\boldsymbol{\sigma}_{\alpha h,\rm disc}^{k,i,\boldsymbol{a}}$ of Definition 1.4.4 satisfy

$$\left\|\mu_\alpha^{\frac{1}{2}}\psi_{h,\boldsymbol{a}}\boldsymbol{\nabla}u_{\alpha h}^{k,i} + \mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}_{\alpha h,\rm disc}^{k,i,\boldsymbol{a}}\right\|_{\omega_h^{\boldsymbol{a}}} \leq C_{\rm st}\left\|\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}\right\|_{\omega_h^{\boldsymbol{a}}}. \tag{1.117}$$

Our second main result is:

Theorem 1.7.2. *Let the flux reconstructions $\boldsymbol{\sigma}_{\alpha h,\rm alg}^{k,i}$ and $\boldsymbol{\sigma}_{\alpha h,\rm disc}^{k,i}$ be given respectively by Definitions 1.4.4 and 1.4.8 for $p \geq 1$. Let the local stopping criteria (1.105) be satisfied for the estimators of Corollary 1.5.6 for $p = 1$. Let also (1.107) holds for $p \geq 2$ for the estimators of Theorem 1.5.1. Let finally the algebraic parameters $\gamma_{{\rm alg},K}$ be such that*

$$\begin{aligned}
\gamma_{{\rm alg},K} &\leq \frac{1}{6C_{\rm st}C_{\rm cont,PF}\max\left\{1,\gamma_{{\rm lin},K}\right\}} \quad \text{if} \quad p = 1,\\
\gamma_{{\rm alg},K} &\leq \frac{1}{6C_{\rm st}C_{\rm cont,PF}} \quad \text{if} \quad p \geq 2.
\end{aligned} \tag{1.118}$$

Setting

$$\delta_K := 2C_{\rm st}C_{\rm cont,PF}\left(1 + \gamma_{{\rm lin},K} + \gamma_{{\rm alg},K}\max\left\{1,\gamma_{{\rm lin},K}\right\}\right) \quad \text{if} \quad p = 1,$$

and

$$\delta_K := 2C_{\rm st}C_{\rm cont,PF}\left(1 + \gamma_{{\rm alg},K}\right) \quad \text{if} \quad p \geq 2,$$

we have for $\alpha \in \{1,2\}$

$$\begin{aligned}
&\eta_{{\rm disc},K,\alpha}^{k,i} + \eta_{{\rm lin},K}^{k,i} + \eta_{{\rm alg},K,\alpha}^{k,i}\\
&\qquad \leq \delta_K \sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\left\|\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}} + \mu_{\rm m}^{\frac{1}{2}}\mu_\alpha^{-\frac{1}{2}}\left|\!\left|\!\left|\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})}\right)
\end{aligned}$$

if $p = 1$ and

$$\begin{aligned}
\eta_{{\rm F},K,\alpha}^{k,i} &= \left\|\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}u_{\alpha h}^{k,i} + \mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}_{\alpha h}^{k,i}\right\|_K\\
&\leq \eta_{{\rm disc},K,\alpha}^{k,i} + \eta_{{\rm alg},K,\alpha}^{k,i}\\
&\leq \delta_K \sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\left\|\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}} + \mu_{\rm m}^{\frac{1}{2}}\mu_\alpha^{-\frac{1}{2}}\left|\!\left|\!\left|\lambda - \tilde{\lambda}_h^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})}\right) \quad \text{if} \quad p \geq 2.
\end{aligned}$$

Proof. We first treat the case $p = 1$. Let $\alpha \in \{1, 2\}$. First, the local criteria (1.105a) and (1.105b) and the definition of δ_K yield

$$\eta^{k,i}_{\mathrm{disc},K,\alpha} + \eta^{k,i}_{\mathrm{lin},K} + \eta^{k,i}_{\mathrm{alg},K,\alpha} \leq \frac{\delta_K}{2C_{\mathrm{st}}C_{\mathrm{cont,PF}}}\eta^{k,i}_{\mathrm{disc},K,\alpha}. \tag{1.119}$$

Next, By the partition of unity $\sum_{\boldsymbol{a}\in\mathcal{V}_K}\psi_{h,\boldsymbol{a}}|_K = 1|_K$, definition (1.101a), (1.70) which implies $\boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{disc}}|_K = \sum_{\boldsymbol{a}\in\mathcal{V}_K}\boldsymbol{\sigma}^{k,i,\boldsymbol{a}}_{\alpha h,\mathrm{disc}}|_K$, stability (1.117), and energy lower bound (1.111), we have

$$\begin{aligned}\eta^{k,i}_{\mathrm{disc},K,\alpha} \leq C_{\mathrm{st}}C_{\mathrm{cont,PF}}\sum_{\boldsymbol{a}\in\mathcal{V}_K}\Bigg(&\left\|\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(u_\alpha - u^{k,i}_{\alpha h}\right)\right\|_{\omega_h^{\boldsymbol{a}}} + \mu_{\mathrm{m}}^{\frac{1}{2}}\mu_\alpha^{-\frac{1}{2}}\left|\!\left|\!\left|\lambda - \tilde{\lambda}^{k,i}_{h,\boldsymbol{a}}\right|\!\right|\!\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})}\\ &+ \left\|\mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{alg}}\right\|_{\omega_h^{\boldsymbol{a}}}\Bigg).\end{aligned} \tag{1.120}$$

Using successively the local criteria (1.105), and since any triangle has three vertices

$$\begin{aligned}\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left\|\mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{alg}}\right\|_{\omega_h^{\boldsymbol{a}}} &= \sum_{\boldsymbol{a}\in\mathcal{V}_K}\eta^{k,i}_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha} \leq 3\gamma_{\mathrm{alg},K}\max\left\{\eta^{k,i}_{\mathrm{disc},K,\alpha}, \eta^{k,i}_{\mathrm{lin},K}\right\}\\ &\leq 3\gamma_{\mathrm{alg},K}\max\left\{1, \gamma_{\mathrm{lin},K}\right\}\eta^{k,i}_{\mathrm{disc},K,\alpha}.\end{aligned} \tag{1.121}$$

Employing now crucially assumption (1.118), it follows that

$$C_{\mathrm{st}}C_{\mathrm{cont,PF}}\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left\|\mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{alg}}\right\|_{\omega_h^{\boldsymbol{a}}} \leq \frac{\eta^{k,i}_{\mathrm{disc},K,\alpha}}{2}. \tag{1.122}$$

Finally, combine (1.122) with (1.120) to bound $\eta^{k,i}_{\mathrm{disc},K,\alpha}$ without the term containing $\boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{alg}}$, and conclude using (1.119).

If $p \geq 2$ the analogue of equation (1.119) reads

$$\eta^{k,i}_{\mathrm{disc},K,\alpha} + \eta^{k,i}_{\mathrm{alg},K,\alpha} \leq \frac{\delta_K}{2C_{\mathrm{st}}C_{\mathrm{cont,PF}}}\eta^{k,i}_{\mathrm{disc},K,\alpha}.$$

While, inequalities (1.120) and (1.122) remain the same, inequality (1.121) reads

$$\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left\|\mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{alg}}\right\|_{\omega_h^{\boldsymbol{a}}} \leq 3\gamma_{\mathrm{alg},K}\eta^{k,i}_{\mathrm{disc},K,\alpha}. \tag{1.123}$$

The conclusion follows immediately. □

1.8 Numerical experiments

This section illustrates numerically our theoretical developments in the case of linear finite elements $p = 1$. We consider the unit disk $\Omega := \{(r, \theta) \in [0, 1] \times [0, 2\pi]\}$ using

the polar coordinates, and an analytical solution given in [19] by, for all $(r,\theta) \in \Omega$,

$$u_1(r,\theta) := g(2r^2-1),$$
$$u_2(r,\theta) := \begin{cases} g(2r^2-1) & \text{if } r \le 1/\sqrt{2}, \\ g(1-r)(2r^2-1)\dfrac{\sqrt{2}}{\sqrt{2}-1} & \text{if } r \ge 1/\sqrt{2}, \end{cases} \quad \lambda(r,\theta) := \begin{cases} 2g & \text{if } r \le 1/\sqrt{2}, \\ 0 & \text{if } r \ge 1/\sqrt{2}. \end{cases}$$

This triple is the solution of the system (1.5) for the data f_1 and f_2 given by

$$f_1(r,\theta) := \begin{cases} -10g & \text{if } r \le 1/\sqrt{2}, \\ -8g & \text{if } r \ge 1/\sqrt{2}, \end{cases} \quad f_2(r,\theta) := \begin{cases} -6g & \text{if } r \le 1/\sqrt{2}, \\ -g\dfrac{1+8r-18r^2}{r}\dfrac{\sqrt{2}}{\sqrt{2}-1} & \text{if } r \ge 1/\sqrt{2}. \end{cases}$$

The parameters μ_1 and μ_2 are set to 1 and the boundary condition for the first membrane g is equal to 0.05. We use the semismooth Newton Algorithm 1 with the min function (1.55) combined with the GMRES linear solver for the system (1.58). For the computation of $\boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{alg}}$, $\alpha = 1,2$, following Section 1.4.2, we consider three levels of uniform mesh refinement ($J = 3$). We also define the linearization and algebraic residuals by

$$\boldsymbol{R}^{k,i}_{\mathrm{lin}} := \begin{pmatrix} \boldsymbol{F} - \mathbb{E}\boldsymbol{X}^{k,i}_h \\ -\mathbf{C}(\boldsymbol{X}^{k,i}_h) \end{pmatrix} \quad \text{and} \quad \boldsymbol{R}^{k,i}_{\mathrm{alg}} := \boldsymbol{B}^{k-1} - \mathbb{A}^{k-1}\boldsymbol{X}^{k,i}_h. \tag{1.124}$$

Three different approaches are tested: 1) The *exact* Newton-min method. Here both the linear and nonlinear solvers are iterated to "almost" convergence. More precisely, we take $\varepsilon_{\mathrm{alg}} := 2{\cdot}10^{-12}$ and $\varepsilon_{\mathrm{lin}} := 10^{-10}$ and replace respectively the stopping criteria **4*b*** and **5** of Algorithm 1 by criteria on the relative residuals,

$$\text{(a)} \ \frac{\left\|\boldsymbol{R}^{k,i}_{\mathrm{alg}}\right\|}{\|\boldsymbol{B}^{k-1}\|} \le \varepsilon_{\mathrm{alg}}, \qquad \text{(b)} \ \frac{\left\|\boldsymbol{R}^{k,i}_{\mathrm{lin}}\right\|}{\left\|\begin{pmatrix} \boldsymbol{F} \\ 0 \end{pmatrix}\right\|} \le \varepsilon_{\mathrm{lin}}. \tag{1.125}$$

2) The *inexact* Newton-min method. Here we consider $\alpha_{\mathrm{alg}} := 1$ and $\varepsilon_{\mathrm{lin}} := 10^{-10}$ in

$$\text{(a)} \ \frac{\left\|\boldsymbol{R}^{k,i}_{\mathrm{alg}}\right\|}{\|\boldsymbol{B}^{k-1}\|} \le \alpha_{\mathrm{alg}}\frac{\left\|\boldsymbol{R}^{k,i}_{\mathrm{lin}}\right\|}{\left\|\begin{pmatrix} \boldsymbol{F} \\ 0 \end{pmatrix}\right\|}, \qquad \text{(b)} \ \frac{\left\|\boldsymbol{R}^{k,i}_{\mathrm{lin}}\right\|}{\left\|\begin{pmatrix} \boldsymbol{F} \\ 0 \end{pmatrix}\right\|} \le \varepsilon_{\mathrm{lin}}. \tag{1.126}$$

3) The *adaptive inexact* Newton-min method (see Algorithm 1) that relies on the stopping criteria (1.104)(a) and (1.104)(b) with $\gamma_{\mathrm{alg}} := 0.3$ and $\gamma_{\mathrm{lin}} := 0.3$.
In the cases of inexact and adaptive inexact methods, the criteria are computed every $\nu := 10$ linear iterations. An ILU preconditionner is used to speed up the GMRES solver. The initial linearization guess $\boldsymbol{X}^0_h \in \mathbb{R}^{3\mathcal{N}^{\mathrm{int}}_h}$ has its first $\mathcal{N}^{\mathrm{int}}_h$ components equal to g and its next components equal to zero.

Figure 1.5 displays the behavior of the solution when the Newton-min and the GMRES solvers have converged. We observe a contact zone in the area $r \lesssim 1/\sqrt{2}$, where λ_h is positive. In the sequel, when the stopping criterion of the nonlinear

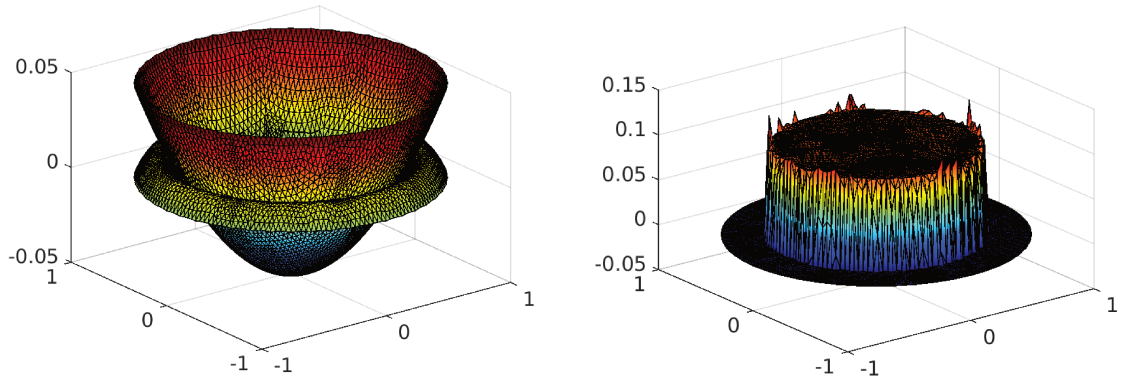

Figure 1.5: Solution at convergence for approximately 8000 elements. Left: position of the membranes (u_{1h}, u_{2h}). Right: discrete action (λ_h).

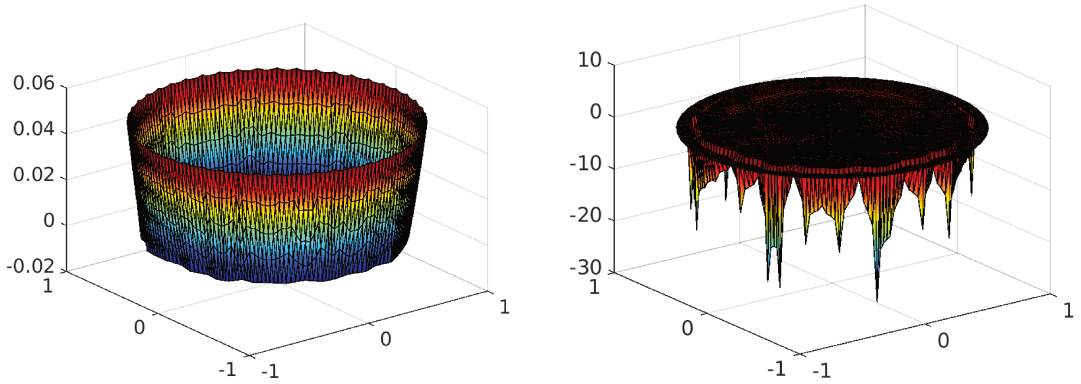

Figure 1.6: Left: $u_{1h}^{k,i} - u_{2h}^{k,i}$ at the second Newton-min step ($k = 2$, $i = 20$). Right: discrete action $\lambda_h^{k,i}$ for $k = 3$, $i = 20$.

solver is satisfied , the index k will be denoted by $\overline{k}$, and similarly the index i at the various stopping criteria will be denoted by $\overline{i}$.

Figure 1.6 then shows the possible violation of the physical constraints during the iterations, before convergence is reached, see Remark 1.5.2. The figure on the left shows that $u_{1h}^{k,i} < u_{2h}^{k,i}$ can appear and the one on the right shows that $\lambda_h^{k,i}$ can be negative.

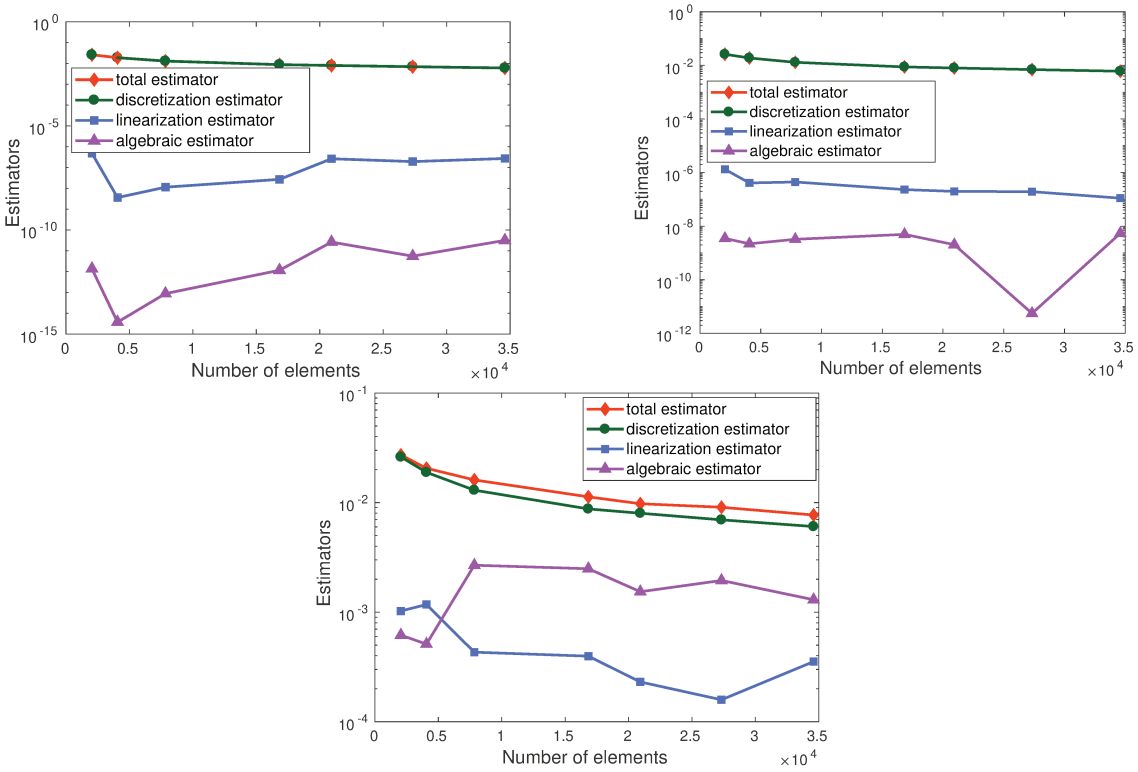

Figure 1.7: A posteriori estimators at convergence ($\eta^{\overline{k},\overline{i}}$, $\eta^{\overline{k},\overline{i}}_{\text{disc}}$, $\eta^{\overline{k},\overline{i}}_{\text{lin}}$, $\eta^{\overline{k},\overline{i}}_{\text{alg}}$) as a function of the number of mesh elements. Exact (left), inexact (middle), and adaptive inexact (right) Newton-min methods with respectively the stopping criteria (1.125),(1.126), and (1.105). The log scales are different in each graph.

Figure 1.7 displays the curves of the different estimators as a function of the number of mesh elements when the nonlinear and algebraic stopping criteria are satisfied. We observe that the total estimators $\eta^{\overline{k},\overline{i}}$ (1.83) are almost identical for the three methods (exact, inexact, and adaptive inexact). Moreover, they have values close to $\eta^{\overline{k},\overline{i}}_{\text{disc}}$, which is consistent with the fact that the error components from Newton-min and GMRES are relatively small. Next, $\eta^{\overline{k},\overline{i}}_{\text{alg}}$ takes values below 10^{-11} for the exact semismooth Newton and below 10^{-8} for the inexact semismooth Newton, whereas $\eta^{\overline{k},\overline{i}}_{\text{lin}}$ takes similar values in both cases (below 10^{-6}). The adaptive inexact Newton method proposed here shows a different behavior: both $\eta^{\overline{k},\overline{i}}_{\text{alg}}$ and $\eta^{\overline{k},\overline{i}}_{\text{lin}}$ take larger values that are just sufficiently small not to influence the overall error estimator. It is also interesting to note the following fact. Although the norm of the linearization residual vector $\boldsymbol{R}^{k,i}_{\text{lin}}$ is requested to lie below $\varepsilon_{\text{lin}} = 10^{-10}$ in both (1.125)(b) and (1.126)(b), the value of the present linearization estimator $\eta^{k,i}_{\text{lin}}$ still remains quite large, with values around 10^{-6} (see Figure 1.7, left and middle). Clearly, there is a huge difference between the l^2 size of the residual vector as expressed by $\left\|\boldsymbol{R}^{k,i}_{\text{lin}}\right\|$ and the size of its lifting back to the physical space expressed by $\eta^{k,i}_{\text{lin}}$.

Figure 1.8 shows the evolution of the various estimators and the behavior of $\|\boldsymbol{R}^{k,i}_{\text{lin}}\|$ and $\|\boldsymbol{R}^{k,i}_{\text{alg}}\|$ given by (1.124) during the algebraic iterations of the first

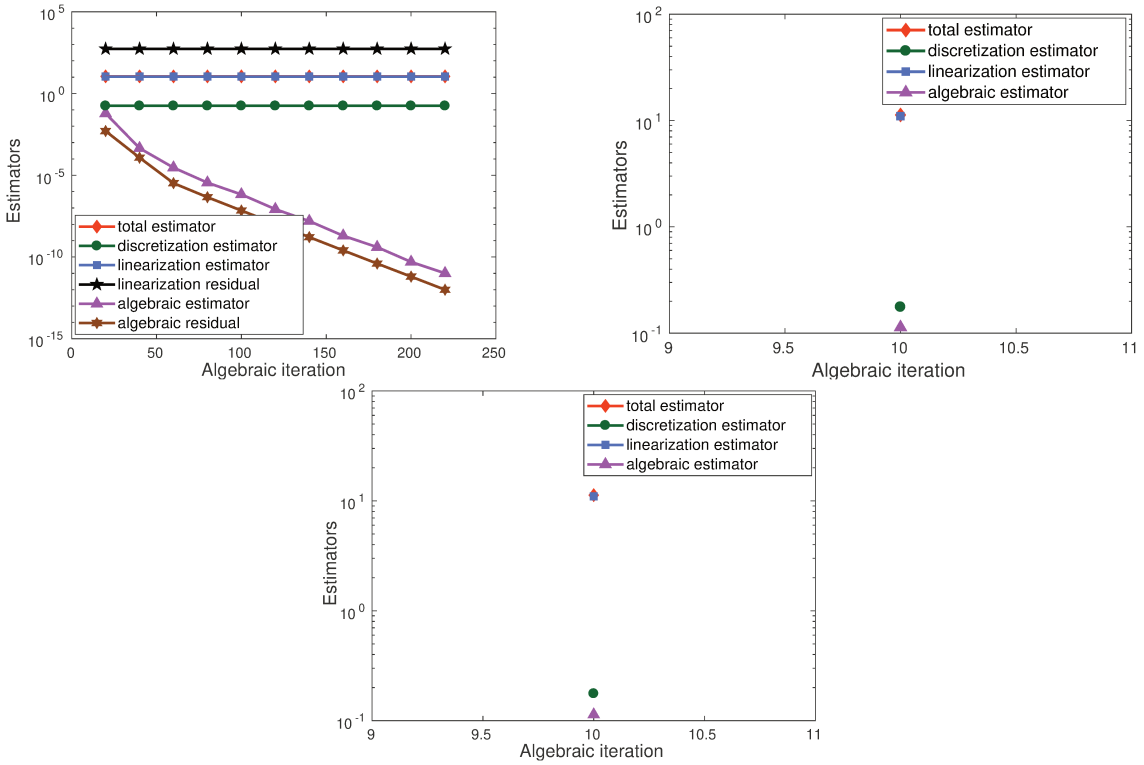

Figure 1.8: Estimators as a function of the algebraic iterations for $k = 1$. Exact (left), inexact (middle), and adaptive inexact (right) Newton-min methods.

Newton-min step (approximately 8000 elements, $k = 1$, i varies). In the exact resolution, we observe that aproximately 220 GMRES iterations are needed to achieve the criterion (1.125)(a), whereas in the inexact and adaptive inexact cases only 10 GMRES iterations are required to satisfy the stopping criteria (respectively (1.126)(a) and (1.104)(a)). In the inexact and adaptive inexact cases, the estimators are computed only once (every $\nu = 10$ iterations) and the total and linearization estimators are approximately equal.

Figure 1.9 represents the evolution of the various estimators as a function of the semismooth Newton iterations when the algebraic solver stopping criteria have been satisfied (approximately 8000 elements, k varies, $i = \bar{i}$). For the three methods, the linearization estimator dominates and is close to the total estimator until approximately the 14^{th} iteration. Next, one can observe that during the Newton-min iterations, the linearization estimator steadily decreases, whereas the discretization one roughly stagnates. The linearization iterations are then stopped in the adaptive inexact Newton-min case when the discretization error becomes dominant, whereas the inexact Newton-min performs many unnecessary additional iterations. This can in general also be the case for the exact Newton-min algorithm but one actually does not see many unnecessary additional iterations here since the convergence gets extremely fast at the end. In terms of numbers, the inexact Newton-min requires 46 iterations to satisfy the stopping criterion (1.126)(b), whereas the exact Newton-min and the adaptive inexact Newton-min methods require 15 and 14 iterations to achieve respectively the criteria (1.125)(b) and (1.104)(b).

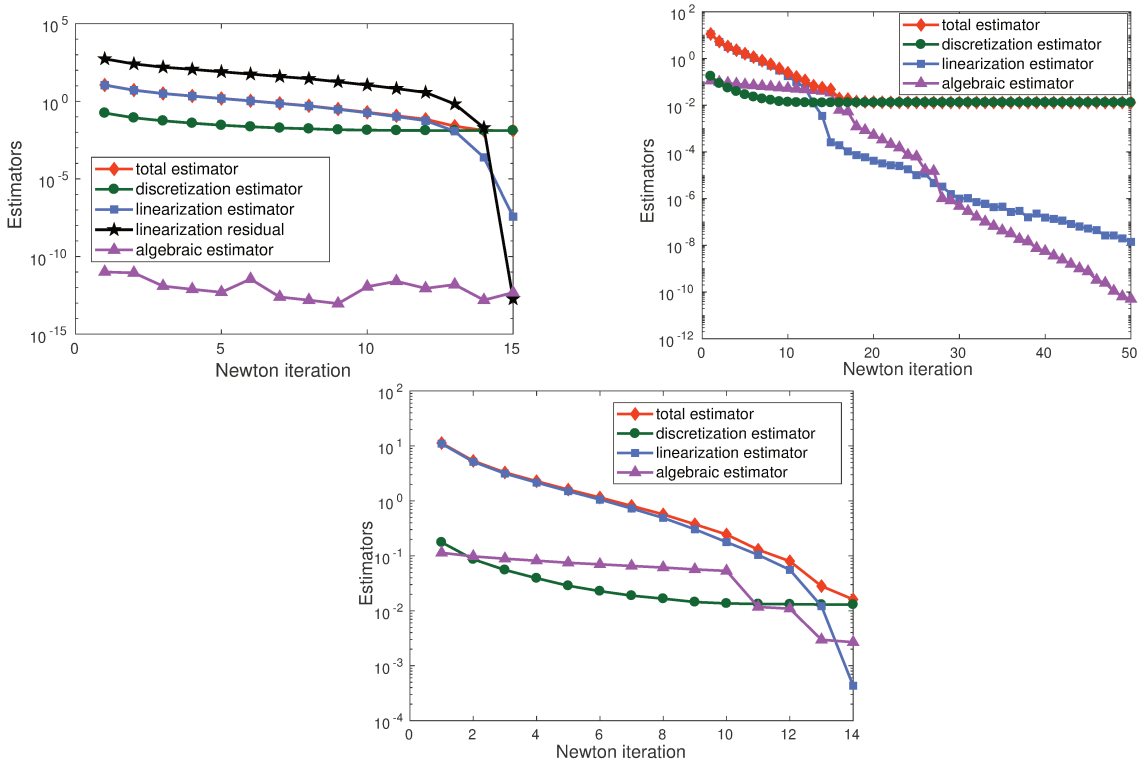

Figure 1.9: Estimators as a function of the Newton-min iterates k ($i = \bar{i}$). Exact (left), inexact (middle), and adaptive inexact (right) Newton-min methods.

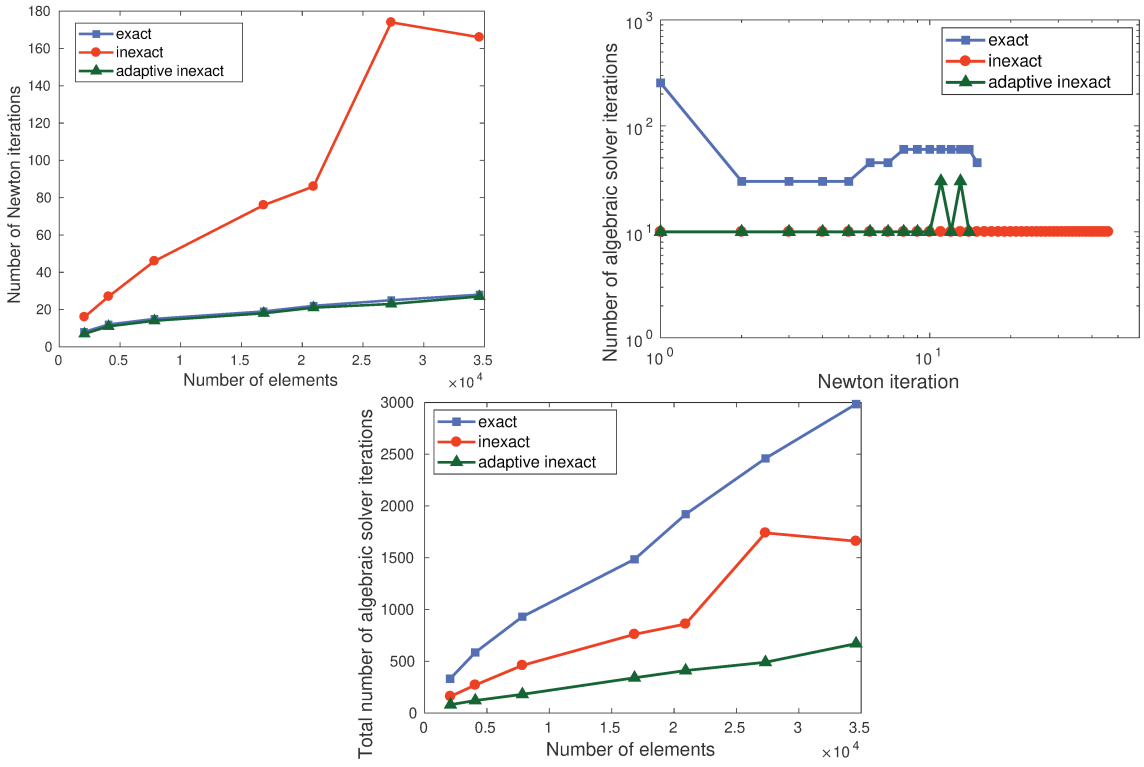

Figure 1.10: Number of Newton-min iterations per number of elements (left), number of algebraic solver iterations per Newton-min step for 8000 elements (middle), and total number of linear solver iterations per number of elements (right).

Figure 1.10 illustrates the overall performance of the three approaches. In the first graph, the behavior of the three methods is represented when the number of mesh elements is increased. In particular, the inexact Newton-min method requires many more semismooth iterations to converge in comparison with the other methods. The exact and the adaptive inexact Newton-min methods lead to approximately the same number of nonlinear iterations. The second graph of Figure 1.10 focuses on the required number of algebraic steps to satisfy the linear stopping criterion for each method at each Newton-min step for a mesh containing approximately 8000 elements. Many algebraic iterations are necessary in the exact Newton-min case, while in the inexact and adaptive inexact cases, the algebraic solver converges almost all the times in 10 iterations. The total number of algebraic iterations is displayed as a function of the number of elements in the right part of Figure 1.10. We observe that exact Newton-min is the most expensive method (3000 iterations for approximately 35000 elements), whereas inexact and adaptive inexact require respectively 1660 and 670 iterations. Thus, globally our approach yields an economy by a factor of roughly 2 with respect to inexact Newton-min and roughly 5 with respect to exact Newton-min in terms of total algebraic solver iterations.

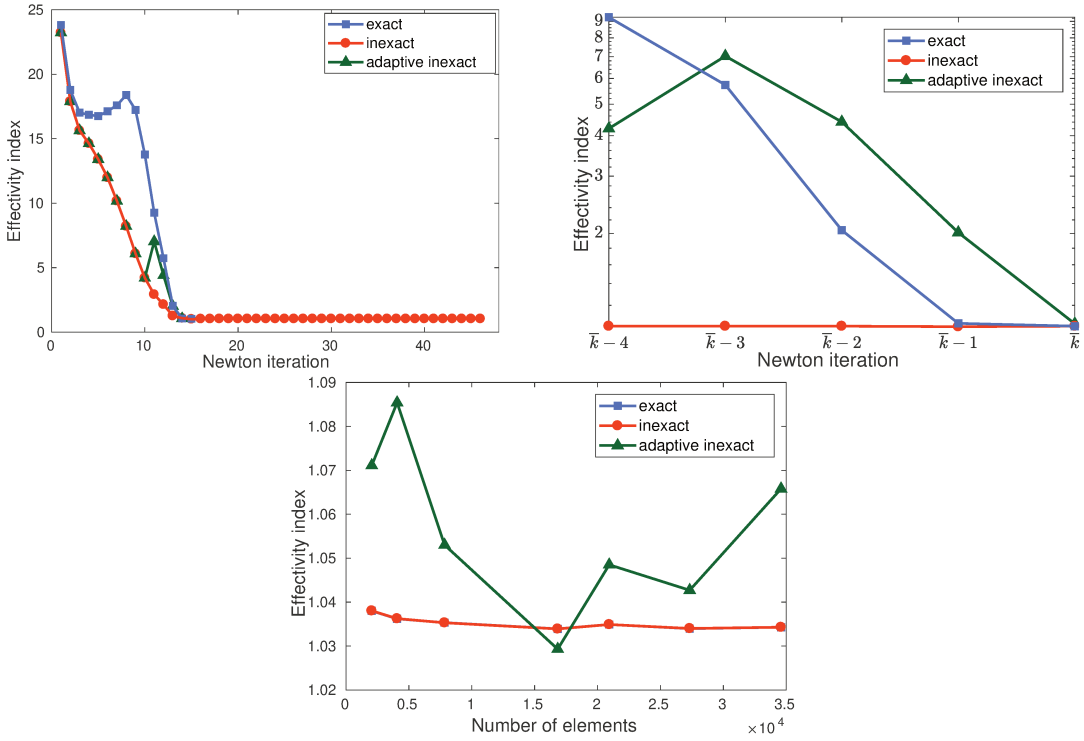

Figure 1.11: Effectivity index as a function of the Newton-min steps for three methods (left), effectivity indices for the last five semismooth Newton steps (middle), and effectivity indices as a function of the number of mesh elements (right); $\overline{k}$ stands for the last Newton-min step for each method ($\overline{k}$ = 15, 46, and 14 respectively for exact, inexact and adaptive inexact methods).

The effectivity indices, defined as the ratio of the total estimator $\eta^{k,\bar{i}}$ over the energy norm $\left|\!\left|\!\left| \boldsymbol{u} - \boldsymbol{u}_h^{k,\bar{i}} \right|\!\right|\!\right|$, are displayed in Figure 1.11 as a function of the Newton-min

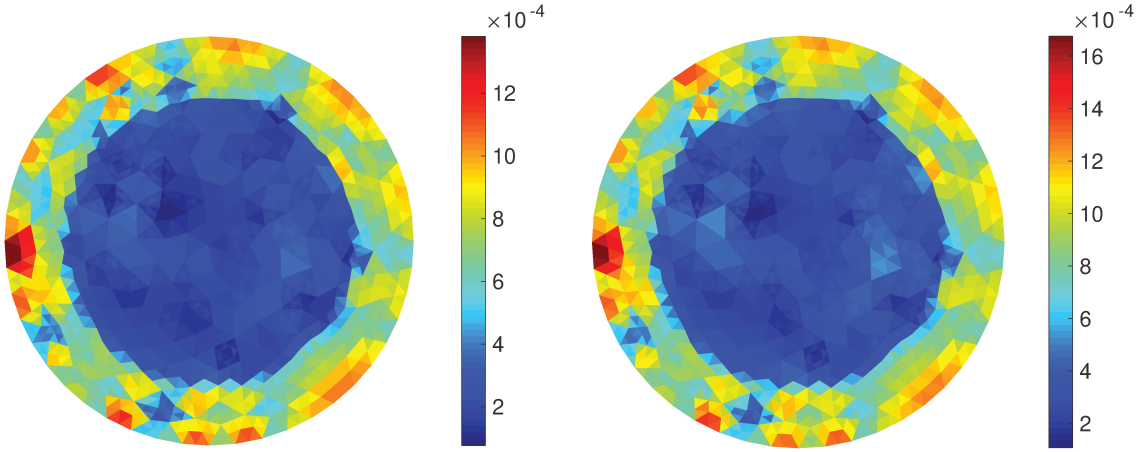

Figure 1.12: Error in energy norm (left) and total estimator (right), adaptive inexact Newton-min method, $p = 1$.

Table 1.1: Number of iterations for the adaptive inexact Newton-min method for several parameters γ_{alg} and γ_{lin}.

$(\gamma_{\text{alg}}, \gamma_{\text{lin}})$	$(0.3, 0.3)$	$(0.03, 0.3)$	$(0.3, 0.03)$	$(0.03, 0.03)$
Newton-min iterations	26	26	27	27
Average algebraic iterations	26	43	25	42
Total iterations	670	1130	680	1140

iterations for the three methods (approximately 8000 elements, k varies, $i = \bar{i}$.) We observe that they always decrease to the optimal value 1, when the computational effort grows. In the middle part of Figure 1.11, we zoom on the last five semismooth Newton iterations for all the methods. In the right part of Figure 1.11, we displayed the value of the effectivity indices for each method for several number of mesh elements when the Newton-min solver and the GMRES solver have converged ($k = \overline{k}$, $i = \bar{i}$). Note that the curves of inexact and adaptive inexact Newton-min are superimposed. We observe that increasing the mesh size will not influence the behavior of the effectivity indices. It is indeed still close to the optimal value of 1.

Figure 1.12 shows the local distribution of the total error estimator $\eta^{k,\bar{i}}$ and of the error in the energy norm $\left|\!\left|\!\left| \boldsymbol{u} - \boldsymbol{u}_h^{k,\bar{i}} \right|\!\right|\!\right|$ in the case of the adaptive inexact semismooth Newton method (approximately 8000 elements, $k = 3$, $i = \bar{i}$). We observe a very close agreement, even in the presence of algebraic and linearization errors.

Finally, Table 1.1 tests the dependency of our adaptive inexact methodology on the coefficients γ_{lin} and γ_{alg} in the algebraic and linearization stopping criteria (1.104)(a) and (1.104)(b) (on the finest mesh with 35000 elements). We represent on the first line the number of Newton-min iterations required to satisfy the stopping criterion (1.104)(b) and on the second one the number of algebraic iterations required to obtain the stopping criterion (1.104)(a), averaged over all Newton iterations. As the linearization convergence is fast, the choice of γ_{lin} has a very small impact on the number of linearization iterations, but choosing γ_{alg} small adds many additional iterations. In any case, however, the overall number of algebraic itera-

tions remains (much) smaller than for the exact and inexact semismooth Newton methods.

1.9 Numerical implementation

Here we give some details on the numerical implementation associated to the numerical resolution of system (1.57) by the Newton–Fischer–Burmeister algorithm. Recall that we provided in Section 1.3.1 the construction of the Jacobian matrix associated to the min function. Next, we briefly explain in the case $p = 1$ the construction of the equilibrated discretization flux reconstruction $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i}$.

Numerical resolution for the Newton–Fischer–Burmeister algorithm when $p = 1$

In this case, the C-function Fischer–Burmeister is defined by

$$(f_{\mathrm{FB}}(\boldsymbol{x},\boldsymbol{y}))_l := \sqrt{\boldsymbol{x}_l^2+\boldsymbol{y}_l^2} - (\boldsymbol{x}_l+\boldsymbol{y}_l) \quad l=1,\ldots,\mathcal{N}_h^{\mathrm{int}}.$$

Taking $\boldsymbol{x} = \boldsymbol{X}_{1h} + g\mathbf{1} - \boldsymbol{X}_{2h}$ where $(\boldsymbol{X}_{1h})_l = u_{1h}(\boldsymbol{a}_l)$ $\forall 1 \le l \le \mathcal{N}_h^{\mathrm{int}}$ and $(\boldsymbol{X}_{2h})_l = u_{2h}(\boldsymbol{a}_l)$ $\forall 1 \le l \le \mathcal{N}_h^{\mathrm{int}}$ and $\boldsymbol{y} = \boldsymbol{X}_{3h}$ where $(\boldsymbol{X}_{3h})_l = \lambda_h(\boldsymbol{a}_l)$ $\forall 1 \le l \le \mathcal{N}_h^{\mathrm{int}}$ (see Section 1.2.3) we have

$$\begin{aligned}(f_{\mathrm{FB}}(\boldsymbol{X}_{1h}+g\mathbf{1}-\boldsymbol{X}_{2h},\boldsymbol{X}_{3h}))_l :=& \sqrt{(u_{1h}(\boldsymbol{a}_l)+g-u_{2h}(\boldsymbol{a}_l))^2+(\lambda_h(\boldsymbol{a}_l))^2}\\ &-(u_{1h}(\boldsymbol{a}_l)+g-u_{2h}(\boldsymbol{a}_l)+\lambda_h(\boldsymbol{a}_l))\,.\end{aligned}$$

Next, denoting by $f_{\mathrm{FB}}^{\circ}(\boldsymbol{X}_h) := f_{\mathrm{FB}}(\boldsymbol{X}_{1h}+g\mathbf{1}-\boldsymbol{X}_{2h},\boldsymbol{X}_{3h})$, the Jacobian matrix $\mathbb{J}_{f_{\mathrm{FB}}^{\circ}}(\boldsymbol{X}_h^{k-1}) \in \mathbb{R}^{\mathcal{N}_h^{\mathrm{int}},3\mathcal{N}_h^{\mathrm{int}}}$ (here $\boldsymbol{X}_h^{k-1}$, $k \ge 1$ is the previous iterate from the semismooth Newton scheme) for the C-function f_{FB}° is defined by

$$\mathbb{J}_{f_{\mathrm{FB}}^{\circ}}(\boldsymbol{X}_h^{k-1}) := \left[\ \mathbb{J}_1(\boldsymbol{X}_h^{k-1}) \,\middle|\, \mathbb{J}_2(\boldsymbol{X}_h^{k-1}) \,\middle|\, \mathbb{J}_3(\boldsymbol{X}_h^{k-1})\ \right]$$

with the diagonal matrix $\mathbb{J}_1(\boldsymbol{X}_h^{k-1}) \in \mathbb{R}^{\mathcal{N}_h^{\mathrm{int}},\mathcal{N}_h^{\mathrm{int}}}$ defined by

$$\left(\mathbb{J}_1(\boldsymbol{X}_h^{k-1})\right)_{l,l} := \frac{u_{1h}^{k-1}(\boldsymbol{a}_l)-u_{2h}^{k-1}(\boldsymbol{a}_l)}{\sqrt{\left(u_{1h}^{k-1}(\boldsymbol{a}_l)-u_{2h}^{k-1}(\boldsymbol{a}_l)\right)^2+\left(\lambda_h^{k-1}(\boldsymbol{a}_l)\right)^2}}-1,$$

the diagonal matrix $\mathbb{J}_2(\boldsymbol{X}_h^{k-1}) \in \mathbb{R}^{\mathcal{N}_h^{\mathrm{int}},\mathcal{N}_h^{\mathrm{int}}}$ defined by

$$\left(\mathbb{J}_2(\boldsymbol{X}_h^{k-1})\right)_{l,l} := -\left(\mathbb{J}_1(\boldsymbol{X}_h^{k-1})\right)_{l,l},$$

and the diagonal matrix $\mathbb{J}_3(\boldsymbol{X}_h^{k-1}) \in \mathbb{R}^{\mathcal{N}_h^{\mathrm{int}},\mathcal{N}_h^{\mathrm{int}}}$ defined by

$$\left(\mathbb{J}_3(\boldsymbol{X}_h^{k-1})\right)_{l,l} := \frac{\lambda_h^{k-1}(\boldsymbol{a}_l)}{\sqrt{\left(u_{1h}^{k-1}(\boldsymbol{a}_l)-u_{2h}^{k-1}(\boldsymbol{a}_l)\right)^2+\left(\lambda_h^{k-1}(\boldsymbol{a}_l)\right)^2}}-1.$$

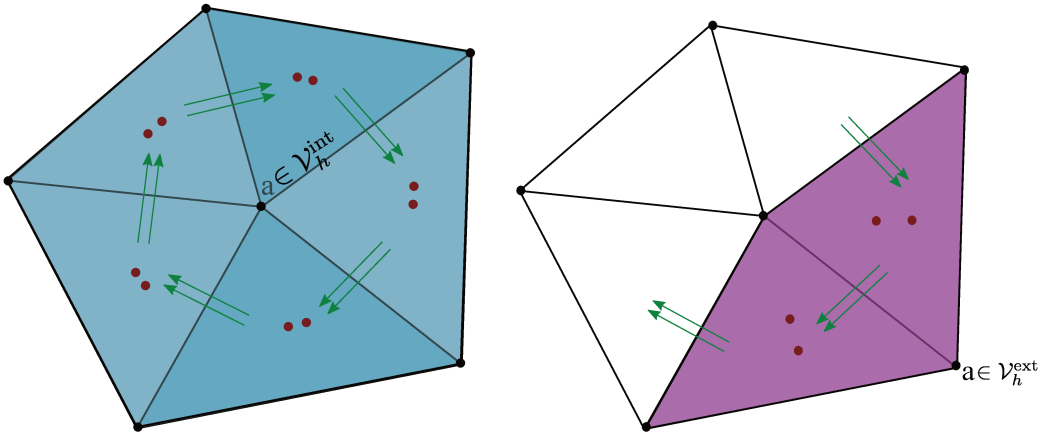

Figure 1.13: Degrees of freedom for the space $\mathbf{RT}_1$ in the patch $\omega_h^{\boldsymbol{a}}$. The bullets in red are internal degrees of freedom and the arrows in green represent edge degrees of freedom. Left: internal patch, right: external patch.

Construction of the discretization flux reconstruction when $p = 1$

Recall that we are interested in implementing the following mixed system (see also Section 1.4.1):

$$\begin{aligned}
\left(\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, \boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} - \left(\gamma_{\alpha h}^{k,i,\boldsymbol{a}}, \boldsymbol{\nabla}\cdot\boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} &= -\left(\mu_\alpha \psi_{h,\boldsymbol{a}} \boldsymbol{\nabla} u_{\alpha h}^{k,i}, \boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} && \forall \boldsymbol{\tau}_h \in \mathbf{V}_h^{\boldsymbol{a}},\\
\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} &= \left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} && \forall q_h \in Q_h^{\boldsymbol{a}},
\end{aligned} \tag{1.127}$$

where the right-hand sides are defined by

$$\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}} := \left(f_\alpha - (-1)^\alpha \lambda_h^{k,i}(\boldsymbol{a}) - r_{\alpha h}^{k,i}\right)\psi_{h,\boldsymbol{a}} - \mu_\alpha \boldsymbol{\nabla} u_{\alpha h}^{k,i}\cdot\boldsymbol{\nabla}\psi_{h,\boldsymbol{a}} \quad \forall \boldsymbol{a} \in \mathcal{V}_h, \tag{1.128}$$

and the spaces $\mathbf{V}_h^{\boldsymbol{a}}$, $Q_h^{\boldsymbol{a}}$ defined by (1.65) and (1.66). Recall that $k \geq 1$ stands for the semismooth Newton iteration and $i \geq 0$ indicates the linear iterative algebraic step. Recall that in general, for a triangle K, $\dim(\mathbf{RT}_1(K)) = 8$ (the degrees of freedom are interpreted as 2 flux per edge and 2 internal degrees of freedom). For the example of Figure 1.13, $\dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}})) = 20$ when $\boldsymbol{a} \in \mathcal{V}_h^{\rm int}$ and $\dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}})) = 10$ when $\boldsymbol{a} \in \mathcal{V}_h^{\rm ext}$. We treat the case of an external vertex $\boldsymbol{a} \in \mathcal{V}_h^{\rm ext}$.

We also denote by $\mathcal{V}_{\rm d}^1(\omega_h^{\boldsymbol{a}})$ the set of Lagrange nodes of discontinuous piecewise first-order polynomials in the patch $\omega_h^{\boldsymbol{a}}$ and by $\mathcal{V}_{\omega_h^{\boldsymbol{a}}}$ the set of vertices in the patch $\omega_h^{\boldsymbol{a}}$. For example, in Figure 1.13, $\mathrm{card}(\mathcal{V}_{\rm d}^1(\omega_h^{\boldsymbol{a}})) = 15$ and $\mathrm{card}(\mathcal{V}_{\omega_h^{\boldsymbol{a}}}) = 6$ when $\boldsymbol{a} \in \mathcal{V}_h^{\rm int}$ and $\mathrm{card}(\mathcal{V}_{\rm d}^1(\omega_h^{\boldsymbol{a}})) = 6$ and $\mathrm{card}(\mathcal{V}_{\omega_h^{\boldsymbol{a}}}) = 4$ when $\boldsymbol{a} \in \mathcal{V}_h^{\rm ext}$. The first order Raviart–Thomas basis functions in the patch are denoted by $(\boldsymbol{\Phi}_j)_{1\leq j\leq \dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))}$ and the Lagrange basis functions of continuous piecewise first-order polynomials are denoted by $(\psi_i)_{1\leq i\leq \mathcal{N}_h^{\rm int}}$. We have

$$\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}} = \sum_{1\leq i\leq \dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))} \boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}(i)\boldsymbol{\Phi}_i,$$

$$\gamma_{\alpha h}^{k,i,\boldsymbol{a}}|_K = \sum_{1\leq j\leq 3} \gamma_{\alpha h}^{k,i,\boldsymbol{a}}(j)\psi_j|_K,$$

$$u_{\alpha h}^{k,i} = \sum_{\boldsymbol{a}' \in \mathcal{V}_{\omega_h^{\boldsymbol{a}}}} u_{\alpha h}^{k,i}(\boldsymbol{a}') \psi_{\boldsymbol{a}'}.$$

Taking in (1.127) $\boldsymbol{\tau}_h = \boldsymbol{\Phi}_j, \forall 1 \leq j \leq \dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))$, and $q_h = (\psi_j)_{1 \leq j \leq \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))}$ we get the linear system

$$\mathbb{A} \boldsymbol{U}_h^{k,i} = \mathbb{F}$$

where the block matrix $\mathbb{A} \in \mathbb{R}^{\dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))+\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))} \times \mathbb{R}^{\dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))+\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))}$ is defined by

$$\mathbb{A} := \left(\begin{array}{c|c} \mathbb{A}_1 & \mathbb{A}_2 \\ \hline -\mathbb{A}_2 & 0 \end{array} \right)$$

with

$$\begin{aligned} \mathbb{A}_1 &:= \left(\boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j \right)_{1 \leq i,j \leq \dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))}, \\ \mathbb{A}_2 &:= -\left(\psi_i, \boldsymbol{\nabla} \cdot \boldsymbol{\Phi}_j \right)_{1 \leq i \leq \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}})), 1 \leq j \leq \dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))}. \end{aligned}$$

The vector of unknowns $\boldsymbol{U}_h^{k,i} \in \mathbb{R}^{\dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))+\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))}$ is defined by

$$\left(\boldsymbol{U}_h^{k,i} \right)^T := \left(\boldsymbol{\sigma}_{\alpha h, \mathrm{disc}}^{k,i,\boldsymbol{a}}(1), \cdots, \boldsymbol{\sigma}_{\alpha h, \mathrm{disc}}^{k,i,\boldsymbol{a}}(\dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}}))), \gamma_{\alpha h}^{k,i,\boldsymbol{a}}(1), \cdots, \gamma_{\alpha h}^{k,i,\boldsymbol{a}}(\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))) \right).$$

The right-hand side $\mathbb{F}$ is decomposed as a matrix vector product as follows:

$$\mathbb{F} := \widetilde{\mathbb{F}} \boldsymbol{Y}_h \tag{1.129}$$

where

$$\widetilde{\mathbb{F}} := \left(\begin{array}{c|c|c|c} \widetilde{\mathbb{F}}_1 & 0 & 0 & 0 \\ \hline \widetilde{\mathbb{F}}_2 & \widetilde{\boldsymbol{F}}_3 & \widetilde{\mathbb{F}}_4 & \widetilde{\boldsymbol{F}}_5 \end{array} \right)$$

with $\widetilde{\mathbb{F}}_1 \in \mathbb{R}^{\dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}})), \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))}$ defined by

$$\widetilde{\mathbb{F}}_1 := -\mu_\alpha \left(\psi_{\boldsymbol{a}} \boldsymbol{\nabla} \psi_i, \boldsymbol{\Phi}_j \right)_{1 \leq i \leq \dim(\mathbf{RT}_1(\omega_h^{\boldsymbol{a}})), 1 \leq j \leq \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))},$$

$\widetilde{\mathbb{F}}_2 \in \mathbb{R}^{\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}})), \dim(\mathcal{V}_{\omega_h^{\boldsymbol{a}}})}$ defined by

$$\widetilde{\mathbb{F}}_2 := -\mu_\alpha \left(\boldsymbol{\nabla} \psi_i \cdot \boldsymbol{\nabla} \psi_{\boldsymbol{a}}, \psi_j \right)_{1 \leq i \leq \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}})), 1 \leq j \leq \dim(\mathcal{V}_{\omega_h^{\boldsymbol{a}}})},$$

$\widetilde{\boldsymbol{F}}_3 \in \mathbb{R}^{\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))}$ defined by

$$\widetilde{\boldsymbol{F}}_3 := \left(\psi_{\boldsymbol{a}} f, \psi_j \right)_{1 \leq j \leq \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))},$$

$\widetilde{\mathbb{F}}_4 \in \mathbb{R}^{\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}})), \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))}$ defined by

$$\widetilde{\mathbb{F}}_4 := -\left(\psi_i \psi_{\boldsymbol{a}}, \psi_j \right)_{1 \leq i,j \leq \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))},$$

and $\widetilde{\boldsymbol{F}}_5 \in \mathbb{R}^{\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))}$ defined by

$$\widetilde{\boldsymbol{F}}_5 := \left(\psi_{\boldsymbol{a}}, \psi_j \right)_{1 \leq j \leq \mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))}.$$

The vector $\boldsymbol{Y}_h \in \mathbb{R}^{\dim(\mathcal{V}_{\omega_h^{\boldsymbol{a}}})+\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))+2}$ is defined by

$$\boldsymbol{Y}_h^T := \left(u_{\alpha h}^{k,i}(1), \cdots, u_{\alpha h}^{k,i}(\dim(\mathcal{V}_{\omega_h^{\boldsymbol{a}}})), 1, r_{\alpha h}^{k,i}(1), \cdots, r_{\alpha h}^{k,i}(\mathrm{card}(\mathcal{V}_{\mathrm{d}}^1(\omega_h^{\boldsymbol{a}}))), \lambda_h^{k,i}(\boldsymbol{a}) \right).$$

When, $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$, we follow the same methodology but we impose mean value zero for the test function q_h.

Construction of the contact estimator when $p = 1$

$$\begin{aligned}
\eta_{C,K}^{k,i,\mathrm{pos}} &:= 2\left(\lambda_h^{k,i,\mathrm{pos}}, u_{1h}^{k,i} - u_{2h}^{k,i}\right)_K, \\
&= 2\sum_{\boldsymbol{a}\in\mathcal{V}_K}\sum_{\boldsymbol{b}\in\mathcal{V}_K}\lambda_h^{k,i,\mathrm{pos}}(\boldsymbol{a})\left(u_{1h}^{k,i} - u_{2h}^{k,i}\right)(\boldsymbol{b})\left(\psi_{\boldsymbol{a}}, \psi_{\boldsymbol{b}}\right)_K, \\
&= 2\sum_{\boldsymbol{a}\in\mathcal{V}_K}\sum_{\boldsymbol{b}\in\mathcal{V}_K}\lambda_h^{k,i,\mathrm{pos}}(\boldsymbol{a})\left(u_{1h}^{k,i} - u_{2h}^{k,i}\right)(\boldsymbol{b})\left(\hat{\psi}_{\boldsymbol{a}}, \hat{\psi}_{\boldsymbol{b}}\right)_{\hat{K}}|\mathrm{dJ}_K|, \\
&= 2[\boldsymbol{X}_{1h}^{k,i} - \boldsymbol{X}_{2h}^{k,i}]|_K\hat{\mathbb{M}}[\boldsymbol{X}_{3h}^{k,i}]|_K^T|\mathrm{dJ}_K|,
\end{aligned} \tag{1.130}$$

where $\boldsymbol{X}_{1h}^{k,i}$ is the matricial representation of $u_{1h}^{k,i}$ in the Lagrange basis, $\boldsymbol{X}_{2h}^{k,i}$ is the matricial representation of $u_{2h}^{k,i}$ in the Lagrange basis, and $\boldsymbol{X}_{3h}^{k,i}$ is the matricial representation of $\lambda_h^{k,i}$ in the Lagrange basis. Furthermore, $\hat{\mathbb{M}}$ denotes the finite element mass matrix in the reference element, and $|\mathrm{dJ}_K|$ is the determinant of the Jacobian matrix associated to the mapping $\mathrm{J}_K : K \to \hat{K}$.

1.10 Conclusions

In this work, we have designed an adaptive inexact semismooth Newton method with adaptive stopping criteria for the problem of contact between two membranes. We proved an optimal a posteriori error estimate between the exact and approximate solution on each semismooth Newton step $k \geq 1$ and on each algebraic solver step $i \geq 1$. This estimate enables to distinguish the different error components. Our numerical experiments for $p = 1$ confirm that the adaptive inexact Newton-min method is much faster in comparison with the exact and inexact Newton-min ones. Moreover, in contrast to these standard methods, the adaptive inexact method presented here provides an accurate estimation of the error between the exact solution and its approximation. Implementation with high order polynomial degree is under investigation. In Chapter 2 we derive a posteriori estimates for parabolic variational inequalities.

Chapter 2

A posteriori error estimates and adaptive stopping criteria for a parabolic variational inequality

Abstract

We develop a posteriori error estimates for a parabolic variational inequality. This problem is discretized with conforming Lagrange finite elements of order $p \geq 1$ in space and with the backward Euler scheme in time. The nonlinearity is treated with any inexact semismooth Newton algorithm. In the case $p = 1$ and for exact solvers, we derive an a posteriori error estimate simulteanously in the energy error norm and in a norm approximating the time derivative error. Next, when $p \geq 1$, we provide an a posteriori error estimate in the energy norm which is valid at each step of the linearization and algebraic resolutions. Our estimate, based on equilibrated flux reconstructions, distinguishes the discretization, linearization, and algebraic error components. We build an adaptive inexact semismooth Newton algorithm based on stopping the iterations of both solvers when the estimators of the corresponding error components do not affect significantly the overall error estimate.

Keywords: parabolic variational inequality, complementarity condition, semismooth Newton method, algebraic solver, multigrid, a posteriori error estimate, adaptivity, stopping criterion.

2.1 Introduction

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain. Let $H^1(\Omega)$ be the space of L^2 functions on the domain Ω which admit a weak gradient in $[L^2(\Omega)]^2$ and $H^1_0(\Omega)$ its zero-trace subspace. Consider the affine space $H^1_g(\Omega) := \{v \in H^1(\Omega),\ v = g \text{ on } \partial\Omega\}$. The dual space of $H^1_0(\Omega)$ is $H^{-1}(\Omega)$. The duality pairing between any Sobolev space $\mathcal{H}$ and its corresponding dual space $\mathcal{H}^*$ is denoted by $\langle\cdot,\cdot\rangle_{\mathcal{H}^*,\mathcal{H}}$ and when $\mathcal{H} = [H^1_0(\Omega)]^2$, the indices are discarded. Let $\mathcal{A}$ be a linear continuous operator, $\mathcal{A} : [H^1(\Omega)]^2 \to [H^{-1}(\Omega)]^2$, coercive on $[H^1_0(\Omega)]^2$ associated to a bilinear form $a(\cdot,\cdot) : [H^1(\Omega)]^2 \times [H^1(\Omega)]^2 \to \mathbb{R}$ satisfying

$$\langle \mathcal{A}\boldsymbol{u}, \boldsymbol{v}\rangle := a(\boldsymbol{u}, \boldsymbol{v}).$$

Let $\boldsymbol{\mathcal{K}}_g$ be a nonempty closed convex subset of $H^1_g(\Omega) \times H^1_0(\Omega)$ and let $\boldsymbol{\mathcal{K}}^{\mathrm{t}}_g$ be its evolutive-in-time version:

$$\boldsymbol{\mathcal{K}}^{\mathrm{t}}_g := \left\{\boldsymbol{v} \in L^2(0,T;H^1_g(\Omega)) \times L^2(0,T;H^1_0(\Omega)),\ \boldsymbol{v}(t) \in \boldsymbol{\mathcal{K}}_g \text{ a.e. in }]0,T[\right\}. \quad (2.1)$$

We consider the following parabolic variational inequality: For the data $\boldsymbol{f} := (f_1, f_2) \in [L^2(0,T;L^2(\Omega))]^2$ and $\boldsymbol{u}^0 \in \boldsymbol{\mathcal{K}}_g$, we search $\boldsymbol{u} \in \boldsymbol{\mathcal{K}}^{\mathrm{t}}_g$ such that $\partial_t \boldsymbol{u} \in [L^2(0,T;H^{-1}(\Omega))]^2$, and $\forall \boldsymbol{v} \in \boldsymbol{\mathcal{K}}^{\mathrm{t}}_g$

$$\begin{aligned} &\int_0^T \langle \partial_t \boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}\rangle(t)\,\mathrm{d}t + \int_0^T \langle \mathcal{A}\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}\rangle(t)\,\mathrm{d}t \geq \int_0^T \langle \boldsymbol{f}, \boldsymbol{v} - \boldsymbol{u}\rangle(t)\,\mathrm{d}t, \\ &\boldsymbol{u}(0) = \boldsymbol{u}^0. \end{aligned} \quad (2.2)$$

Problem (2.2) belongs to the wide class of parabolic variational inequalities of the first kind, see Glowinski [93] and Lions [121] for a general introduction. Existence and uniqueness of a weak solution $\boldsymbol{u} \in \boldsymbol{\mathcal{K}}^{\mathrm{t}}_g$ for (2.2) is well understood today, see [94, 95, 121] and the references therein. When $\boldsymbol{\mathcal{K}}^{\mathrm{t}}_g$ is the scalar space $L^2(0,T;H^1_0(\Omega))$, and $a(u,v) = (\boldsymbol{\nabla} u, \boldsymbol{\nabla} v)_\Omega$, problem (2.2) turns into the parabolic heat equation: find $u \in L^2(0,T;H^1_0(\Omega))$ such that

$$\begin{aligned} &\int_0^T \langle \partial_t u, v\rangle(t)\,\mathrm{d}t + \int_0^T \langle \mathcal{A}u, v\rangle(t)\,\mathrm{d}t = \int_0^T \langle f, v\rangle(t)\,\mathrm{d}t \quad \forall v \in L^2(0,T;H^1_0(\Omega)) \\ &u(0) = u_0, \end{aligned} \quad (2.3)$$

and following the fundamental Lions Theorem [63, Chapter 18], it admits a unique weak solution $u \in L^2(0,T,H^1_0(\Omega)) \cap \mathcal{C}^0(0,T,L^2(\Omega))$. If $T = 0$ and $\partial_t u = 0$, the problem (2.2) is a stationnary variational inequality and it has been explored in detail in many surveys [17, 76, 88, 89, 105, 122, 124, 148].

Concerning the spatial discretization of variational inequalities, the $\mathbb{P}_1$ finite element method is commonly employed as it yields conforming solutions, see Chen and Nochetto [53], Veeser [153], Braess [31], or Ben Belgacem *et al.* [17]. Other approaches based on discontinuous Galerkin methods have been suggested recently, see Wang, Han, and Cheng [159]. Note that in Chapter 1 of this Thesis, we proposed a $\mathbb{P}_p$ finite element discretization for the stationnary contact between two membranes. The discretization in time often uses the backward Euler scheme.

Among the spectrum of resolution methods for the discrete counterpart of (2.2), let us mention the interior point method of Wright [160], the active set strategy by Kanzow [111], and the semismooth Newton methods (see [64,87–89]). In this work, we use a saddle-point Lagrangian formulation giving rise at each time step n to a nonlinear system of algebraic equations of the form

$$\mathcal{S}^n(\boldsymbol{X}_h^n) = 0, \tag{2.4}$$

where $\mathcal{S}$ is a nonlinear operator and $\boldsymbol{X}_h^n$ the unknown vector of degrees of freedom. We employ any semismooth linearization procedure starting from an initial guess $\boldsymbol{X}_h^{n,0}$ and giving at each step $k \geq 1$ the system of linear algebraic equations

$$\mathbb{A}^{n,k-1}\boldsymbol{X}_h^{n,k} = \boldsymbol{F}^{n,k-1}, \tag{2.5}$$

where the matrix $\mathbb{A}^{n,k-1}$ and the vector $\boldsymbol{F}^{n,k-1}$ are constructed from $\boldsymbol{X}_h^{n,k-1}$. Solving (2.5) with a direct method may be very expensive. In this work, we rather consider an inexact linearization procedure giving at each iterative linear algebraic step $i \geq 0$ and each linearization step $k \geq 1$ a residual vector $\boldsymbol{R}_h^{n,k,i}$ defined by

$$\boldsymbol{R}_h^{n,k,i} := \boldsymbol{F}^{n,k-1} - \mathbb{A}^{n,k-1}\boldsymbol{X}_h^{n,k,i}. \tag{2.6}$$

In the present work, we focus on answering the following questions: To which precision should (2.6) be solved? To which precision should (2.5) be resolved? Can we estimate the total error, as well as each error component of the overall discretization? Can we reduce the number of iterations? Our key tool to propose answers is the a posteriori error analysis.

A huge amount of work has been performed in the past on a posteriori error estimates for partial differential equations. We can mention the pioneering work of Prager and Synge [138], Ladevèze [118] and the books of Verfürth [155] and Ainsworth [5] for a general introduction. For elliptic variational inequalities, we can mention the contributions [6,43,116,144] and for the elliptic obstacle problem, we refer to the papers [31,53,55,99–101,153]. In contrast to the last references, in Dabaghi, Martin, and Vohralík [62] (or in Chapter 1), a $\mathbb{P}_p$ conforming finite element yielding a nonconforming approximation for $p \geq 2$ for a contact problem between two membranes is employed; not only the discretization error is estimated, but all the error components, namely the finite element discretization error, the semismooth linearization error, and the iterative algebraic error.

In the context of parabolic problems, the a posteriori analysis has received significant attention over the past decade. For parabolic equations, we mention Verfürth [154], Bernardi, Bergham, and Mghazli [26], Ern and Vohralík [80], and Ern, Smears, and Vohralík [78,79], where in particular in [78], local efficiency in space and in time for the estimators is proven. For parabolic variational inequalities, the edifice seems still under construction. We can mention Moon, Nochetto, Petersdorff, and Zhang [129], and Achdou, Hecht, and Pommier [3]. In the present work, we follow the methodology of [78] and Chapter 1 of this thesis (or [62]) to derive a posteriori error estimates for a parabolic variational inequality with estimation of each component of the error. In particular, it enables to define adaptive stopping

criteria for nonlinear semismooth and linear algebraic solvers, which is new to the best of our knowledge. Importantly, it enables to save many unnecessary iterations.

To exemplify our approach, we consider the extension of the model problem studied in Chapter 1 as an unsteady parabolic variational inequality. Several important difficulties arise for the a posteriori analysis of problem (2.2):

1) Denoting by $\boldsymbol{u}_{h\tau}^{k,i} := \left(u_{1h\tau}^{k,i}, u_{2h\tau}^{k,i}\right)$ the space-time numerical approximation, where here the indices k, i merely indicate the presence of inexact linearization and algebraic solvers and where $\boldsymbol{u}_{h\tau}^{k,i}$ is piecewise continuous in time and piecewise polynomial of degree p for each variable in space, $\boldsymbol{u}_{h\tau}^{k,i}$ is nonconforming in the sense that $\boldsymbol{u}_{h\tau}^{k,i} \notin \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$. Denoting by $\lambda_{h\tau}^{k,i}$ the discrete counterpart of the Lagrange multiplier λ, the same phenomenon occurs in the sense that $\lambda_{h\tau}^{k,i}$ is not also conforming. 2) We cannot easily provide, as for the classical parabolic heat equation, an a posteriori upper bound for the time derivative $\left\|\partial_t\left(\boldsymbol{u} - \boldsymbol{u}_{h\tau}^{k,i}\right)\right\|_{[L^2(0,T;H^{-1}(\Omega))]^2}$. It seems simply possible in the case where the membranes are separated. To tackle this difficulty we construct an element $\boldsymbol{z} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ such that $\|\boldsymbol{u} - \boldsymbol{z}\|_{\left[L^2(0,T;H_0^1(\Omega))\right]^2}$ is closely linked to $\left\|\partial_t\left(\boldsymbol{u} - \boldsymbol{u}_{h\tau}^{k,i}\right)\right\|_{[L^2(0,T;H^{-1}(\Omega))]^2}$ and such that the a posteriori error estimate holds as

$$\|\boldsymbol{u} - \boldsymbol{z}\|_{\left[L^2(0,T;H_0^1(\Omega))\right]^2} \leq 2\left\{\sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\eta_K^n(\boldsymbol{u}_{h\tau}^{k,i})\right)^2(t)\right\}^{\frac{1}{2}}.$$

In this case, we will be able to estimate the error measured as

$$\left\|\boldsymbol{u} - \boldsymbol{u}_{h\tau}^{k,i}\right\|_{\left[L^2(0,T;H_0^1(\Omega))\right]^2}^2 + \|\boldsymbol{u} - \boldsymbol{z}\|_{\left[L^2(0,T;H_0^1(\Omega))\right]^2}^2. \tag{2.7}$$

This presentation is structured as follows. We first present the model problem, its weak formulation, and its discretization with the backward Euler scheme in time and the conforming $\mathbb{P}_p$ ($p \geq 1$) finite element method in space. In particular, we show that our nonlinear system may be seen as a system of parabolic partial differential equations with complementarity constraints. In the spirit of Chapter 1, we give the expression of the discrete complementarity constraints in the Lagrange basis and its dual basis. Then, we present the concept of inexact semismooth Newton methods to solve our system of algebraic inequalities at each time step. Next, we provide the a posteriori analysis following the approach of the equilibrated flux reconstructions. In particular, we derive an a posteriori error estimate for linear finite elements ($p = 1$) at each time step n when the semismooth Newton solver as well as the algebraic iterative solver have converged. Then we can estimate the error as defined in (2.7). We next provide a second a posteriori error estimate, valid for $p \geq 1$ at each semismooth linearization iteration $k \geq 1$ and each iterative algebraic solver iteration $i \geq 0$. This estimate only bounds the first component of (2.7), but in particular distinguishes the different error components, namely the discretization error, the semismooth linearization error, and the algebraic error, leading to an adaptive inexact semismooth Newton algorithm for parabolic problems.

2.2 Model problem and setting

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain and $T > 0$ be the final simulation time. The model problem we consider here is to find u_1, u_2, and λ such that

$$\left\{\begin{array}{lll} \partial_t u_1 - \mu_1 \Delta u_1 - \lambda = f_1 & \text{in} & \Omega \times]0,T[\,, \\ \partial_t u_2 - \mu_2 \Delta u_2 + \lambda = f_2 & \text{in} & \Omega \times]0,T[\,, \\ (u_1 - u_2)\lambda = 0, \quad u_1 - u_2 \geq 0, \quad \lambda \geq 0 & \text{in} & \Omega \times]0,T[\,, \\ u_1 = g & \text{on} & \partial\Omega \times]0,T[\,, \\ u_2 = 0 & \text{on} & \partial\Omega \times]0,T[\,, \\ u_1(\boldsymbol{x},0) = u_1^0(\boldsymbol{x}),\ u_2(\boldsymbol{x},0) = u_2^0(\boldsymbol{x}),\ u_1^0(\boldsymbol{x}) - u_2^0(\boldsymbol{x}) \geq 0 & \text{in} & \Omega. \end{array}\right. \tag{2.8}$$

The two first equations of (2.8) are of parabolic type. Next, the third line of (2.8) are the linear complementarity conditions saying that either $u_1 - u_2 = 0$ and $\lambda > 0$, or $u_1 - u_2 > 0$ and $\lambda = 0$. The sources terms $(f_1, f_2) \in [L^2(0,T;L^2(\Omega))]^2$ are supposed constant on each time interval I_n defined below. The real coefficients μ_1 and μ_2 are positif and for the sake of simplicity, we assume that $g > 0$ is a constant. Thus, dirichlet boundary conditions are prescribed for the data u_1 and u_2. Observe that when $u_1 - u_2 > 0$ and $\lambda = 0$, problem (2.8) is equivalent to solve two separated heat equations, and when $T = 0$ and $\partial_t u_1 = \partial_t u_2 = 0$, (2.8) becomes the stationary contact problem between two membranes treated in Chapter 1.

We define the Lebesgue set

$$\Lambda := \left\{\chi \in L^2(\Omega),\ \chi \geq 0 \text{ a.e. in } \Omega\right\}$$

and the space-time Sobolev spaces

$$V_g := L^2(0,T;H^1_g(\Omega)), \quad V_0 := L^2(0,T;H^1_0(\Omega)), \quad V_0^* := L^2(0,T;H^{-1}(\Omega)),$$
$$\Psi = L^2(0,T;\Lambda).$$

The standard notations $\boldsymbol{\nabla}$ and $\boldsymbol{\nabla}\cdot$ are used respectively for the weak gradient and divergence operators. For a nonempty set $\mathcal{O}$ of $\mathbb{R}^2$, we denote its Lebesgue measure by $|\mathcal{O}|$ and the $L^2(\mathcal{O})$ scalar product by $(u,v)_{\mathcal{O}} := \int_{\mathcal{O}} uv \,\mathrm{dx}$ for $u, v \in L^2(\mathcal{O})$. We also use the following notations $\|v\|^2_{\mathcal{O}} := (v,v)_{\mathcal{O}}$ and $\|\boldsymbol{\nabla} v\|^2_{\mathcal{O}} := (\boldsymbol{\nabla} v, \boldsymbol{\nabla} v)_{\mathcal{O}}$. Then, we define the space energy norm:

$$\forall \boldsymbol{v} = (v_1, v_2) \in [H^1_0(\mathcal{O})]^2, \quad |||\boldsymbol{v}|||_{\mathcal{O}} := \left\{\sum_{\alpha=1}^{2} \mu_\alpha \|\boldsymbol{\nabla} v_\alpha\|^2_{\mathcal{O}}\right\}^{\frac{1}{2}}. \tag{2.9}$$

For a scalar-valued function $v \in V_0$, we define the space-time energy norm by

$$\|v\|_{V_0} := \left\{\int_0^T \|\boldsymbol{\nabla} v\|^2_\Omega(t)\,\mathrm{dt}\right\}^{\frac{1}{2}}.$$

Analogously to (2.9), for a vector-valued function $\boldsymbol{v} = (v_1, v_2) \in V_0 \times V_0$, we define the space-time energy norm by

$$|||\boldsymbol{v}|||_{V_0} := \left\{\int_0^T \left(\sum_{\alpha=1}^{2} \mu_\alpha \|\boldsymbol{\nabla} v_\alpha\|^2_\Omega\right)(t)\,\mathrm{dt}\right\}^{\frac{1}{2}}. \tag{2.10}$$

Besides, the Poincaré–Friedrichs and the Poincaré–Wirtinger inequalities, see [14, 137], state that if $\overline{v}_{\mathcal{O}}$ denotes the mean value of v and $h_{\mathcal{O}}$ the diameter of $\mathcal{O}$, then

$$\|v\|_{\mathcal{O}} \leq C_{\mathrm{PF}} h_{\mathcal{O}} \|\boldsymbol{\nabla} v\|_{\mathcal{O}} \qquad \forall v \in H_0^1(\mathcal{O}), \tag{2.11a}$$

$$\|v - \overline{v}_{\mathcal{O}}\|_{\mathcal{O}} \leq C_{\mathrm{PW}} h_{\mathcal{O}} \|\boldsymbol{\nabla} v\|_{\mathcal{O}} \qquad \forall v \in H^1(\mathcal{O}). \tag{2.11b}$$

2.3 Weak solution using a saddle point formulation

In this section we are interested in obtaining a variational formulation for Problem (2.8).

For the data $(f_1, f_2) \in [L^2(0,T; L^2(\Omega))]^2$ and the intial condition $(u_1^0, u_2^0) \in H_g^1(\Omega) \times H_0^1(\Omega)$, the weak formulation associated to (2.8) consists in: find $(u_1, u_2, \lambda) \in V_g \times V_0 \times \Psi$ such that $\partial_t u_\alpha \in V_0^*$, $\alpha = 1, 2$ and satisfying $\forall t \in]0,T[$ $\forall (v_1, v_2, \chi) \in H_0^1(\Omega) \times H_0^1(\Omega) \times \Lambda$

$$\begin{aligned} &\sum_{\alpha=1}^{2} \langle \partial_t u_\alpha(t), v_\alpha \rangle + \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_\alpha(t), \boldsymbol{\nabla} v_\alpha \right)_\Omega - \left(\lambda(t), v_1 - v_2 \right)_\Omega = \sum_{\alpha=1}^{2} \left(f_\alpha, v_\alpha \right)_\Omega , \\ &\left(\chi - \lambda(t), u_1(t) - u_2(t) \right)_\Omega \geq 0. \end{aligned} \tag{2.12}$$

Proposition 2.3.1. *Problems (2.8) and (2.12) are equivalent in the sense that any triple $(u_1, u_2, \lambda) \in V_g \times V_0 \times \Psi$ is a solution of (2.8) (in the sense of distributions) if and only if it is a solution of (2.12).*

Proof. The proof follows the argument of [17, Proposition 1]. Assume that $(u_1, u_2, \lambda) \in V_g \times V_0 \times \Psi$ is a solution of (2.8). Multiplying the first line of (2.8) for each $t \in]0,T[$ by a test function $v_1 \in H_0^1(\Omega)$ and the second line of (2.8) by a test function $v_2 \in H_0^1(\Omega)$, summing these two equations and employing the Green formula give the first line of (2.12). Furthermore, as $\lambda(t) \left(u_1 - u_2 \right)(t) = 0$ $\forall t \in]0,T[$, for any $\chi \in \Lambda$ we have

$$\begin{aligned} \left(\chi - \lambda(t), u_1(t) - u_2(t) \right)_\Omega &= \left(\chi, u_1(t) - u_2(t) \right)_\Omega - \left(\lambda(t), u_1(t) - u_2(t) \right)_\Omega \\ &= \left(\chi, u_1(t) - u_2(t) \right) \geq 0. \end{aligned}$$

Therefore, (u_1, u_2, λ) is a solution to (2.12).

Conversely, assume that (u_1, u_2, λ) is a solution to (2.12). Taking $v_1 \in \mathcal{D}(\Omega) \subset H_0^1(\Omega)$ and taking $v_2 = 0$ in (2.12), we get

$$\langle \partial_t u_1(t), v_1 \rangle + \mu_1 \left(\boldsymbol{\nabla} u_1(t), \boldsymbol{\nabla} v_1 \right)_\Omega - \left(\lambda(t), v_1 \right)_\Omega = \left(f_1, v_1 \right)_\Omega . \tag{2.13}$$

In the distributional sense (2.13) reads

$$\langle \partial_t u_1(t), v_1 \rangle_{\mathcal{D}'(\Omega), \mathcal{D}(\Omega)} - \mu_1 \langle \Delta u_1, v_1 \rangle_{\mathcal{D}'(\Omega), \mathcal{D}(\Omega)} = \langle \lambda(t) + f_1, v_1 \rangle_{\mathcal{D}'(\Omega), \mathcal{D}(\Omega)}.$$

Then,

$$\partial_t u_1(t) - \mu_1 \Delta u_1(t) - \lambda(t) = f_1 \quad \text{in} \quad \mathcal{D}'(\Omega).$$

Next, taking $v_2 \in \mathcal{D}(\Omega) \subset H_0^1(\Omega)$ and $v_1 = 0$ and employing the same methodology we get

$$\partial_t u_2(t) - \mu_2 \Delta u_2(t) + \lambda(t) = f_2 \quad \text{in} \quad \mathcal{D}'(\Omega).$$

Furthermore, if the second line of (2.12) is satisfied, taking $\chi = \lambda(t) + \mathbb{1}_{\mathcal{O}} \in \Lambda$ for $\mathcal{O}$ any measurable subset of Ω, and next $\chi = 0$, we obtain

$$u_1(t) - u_2(t) \geq 0, \quad \text{and} \quad (\lambda(t), u_1(t) - u_2(t))_\Omega \leq 0.$$

As $\lambda(t) \geq 0$ it yields $\lambda(t)\,(u_1 - u_2)\,(t) = 0$. Thus, (u_1, u_2, λ) is a solution to (2.8) and this concludes the proof. □

Remark 2.3.2. *Problem (2.12) can be interpreted as a system of parabolic variational equalities with the linear complementarity constraints: find $(u_1, u_2, \lambda) \in V_g \times V_0 \times \Psi$ such that $\partial_t u_\alpha \in V_0^*$ and satisfying $\forall t \in]0, T[$ and $\forall v_\alpha \in H_0^1(\Omega)$*

$$\begin{aligned} &\sum_{\alpha=1}^{2} \langle \partial_t u_\alpha(t), v_\alpha \rangle + \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_\alpha(t), \boldsymbol{\nabla} v_\alpha\right)_\Omega - \left(\lambda(t), v_1 - v_2\right)_\Omega = \sum_{\alpha=1}^{2} \left(f_\alpha, v_\alpha\right)_\Omega \\ &(u_1 - u_2)\,(t) \geq 0, \quad \lambda(t) \geq 0, \quad \lambda(t)\,(u_1 - u_2)\,(t) = 0. \end{aligned} \tag{2.14}$$

2.4 Weak formulation using a reduced problem

We showed in Section 2.3 that the problem (2.8) admits a weak formulation given by (2.12). Now we show that problem (2.8) can be seen as a parabolic variational inequality.

Let $\boldsymbol{\mathcal{K}}_g$ be the nonempty closed convex set defined by

$$\boldsymbol{\mathcal{K}}_g := \left\{ (v_1, v_2) \in H_g^1(\Omega) \times H_0^1(\Omega),\ v_1 - v_2 \geq 0 \text{ a.e. in } \Omega \right\} \tag{2.15}$$

and $\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ its evolutive-in-time version defined by (2.1). Note that since $(g, 0) \in V_g \times V_0$, $\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ is nonempty. Introducing the simplified notations

$$\begin{aligned} &a(\boldsymbol{u}, \boldsymbol{v}) := \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_\alpha, \boldsymbol{\nabla} v_\alpha\right)_\Omega, \quad b(\boldsymbol{v}, \chi) := \left(\chi, v_1 - v_2\right)_\Omega, \\ &l(\boldsymbol{v}) := \sum_{\alpha=1}^{2} \left(f_\alpha, v_\alpha\right), \quad \langle \partial_t \boldsymbol{u}, \boldsymbol{v} \rangle := \sum_{\alpha=1}^{2} \langle \partial_t u_\alpha, v_\alpha \rangle, \end{aligned} \tag{2.16}$$

the problem (2.12) reads: find $(u_1, u_2, \lambda) \in V_g \times V_0 \times \Psi$ such that $\partial_t u_\alpha \in V_0^*$, $\alpha = 1, 2$, and satisfying $\forall t \in]0, T[$ $\forall (v_1, v_2, \chi) \in H_0^1(\Omega) \times H_0^1(\Omega) \times \Lambda$

$$\begin{aligned} &\langle \partial_t \boldsymbol{u}, \boldsymbol{v} \rangle + a(\boldsymbol{u}, \boldsymbol{v}) - b(\boldsymbol{v}, \lambda) = l(\boldsymbol{v}), \\ &b(\boldsymbol{u}, \chi - \lambda) \geq 0. \end{aligned} \tag{2.17}$$

Proposition 2.4.1. *If the triple $(u_1, u_2, \lambda) \in V_g \times V_0 \times \Psi$ is a solution of (2.17) then it is also a solution of the following parabolic variational inequality: given $\boldsymbol{u}^0 \in \boldsymbol{\mathcal{K}}_g$, search $\boldsymbol{u} := (u_1, u_2) \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ such that $\partial_t u_\alpha \in V_0^*$, $\alpha = 1, 2$, and $\forall \boldsymbol{v} := (v_1, v_2) \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$, there holds*

$$\int_0^T \langle \partial_t \boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u} \rangle(t)\, \mathrm{d}t + \int_0^T a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u})(t)\, \mathrm{d}t \geq \int_0^T l(\boldsymbol{v} - \boldsymbol{u})(t)\, \mathrm{d}t. \tag{2.18}$$

Proof. Let $(u_1, u_2, \lambda) \in V_g \times V_0 \times \Psi$ be the solution of (2.17). Using Remark 2.3.2, the complementarity constraints $(u_1 - u_2)(t) \geq 0$ immediately provides $\boldsymbol{u} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$. Next, for any $\boldsymbol{v} = (v_1, v_2) \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$, $w_\alpha(t) := (v_\alpha - u_\alpha)(t) \in H_0^1(\Omega)$, $\forall \alpha = 1, 2$ and using w_α as test functions in (2.17) we get

$$\langle \partial_t \boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u} \rangle + a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) - (\lambda, v_1 - v_2 - (u_1 - u_2))_\Omega = l(\boldsymbol{v} - \boldsymbol{u}).$$

But, $(\lambda, u_1 - u_2) = 0$ and $-(\lambda, v_1 - v_2) \leq 0$. Thus, we obtain

$$\langle \partial_t \boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u} \rangle + a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq l(\boldsymbol{v} - \boldsymbol{u}). \tag{2.19}$$

Integrating in time (2.19) provides the reduced formulation (2.18) which concludes the proof. □

Remark 2.4.2. *It is possible to weaken (2.18) proceeding as follows. Using [83, Theorem 5.9.3] we have* $\forall \boldsymbol{v} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$

$$\int_0^T \langle \partial_t v_\alpha, v_\alpha - u_\alpha \rangle(t)\,\mathrm{d}t = \int_0^T \langle \partial_t (v_\alpha - u_\alpha), v_\alpha - u_\alpha \rangle(t)\,\mathrm{d}t + \int_0^T \langle \partial_t u_\alpha, v_\alpha - u_\alpha \rangle(t)\,\mathrm{d}t$$
$$= \frac{1}{2}\int_0^T \frac{\mathrm{d}}{\mathrm{d}t} \|(v_\alpha - u_\alpha)\|^2(t)\,\mathrm{d}t + \int_0^T \langle \partial_t u_\alpha, v_\alpha - u_\alpha \rangle(t)\,\mathrm{d}t.$$

Then, employing (2.18) we get

$$\begin{aligned}
&\sum_{\alpha=1}^{2} \int_0^T (\partial_t v_\alpha, v_\alpha - u_\alpha)_\Omega\,\mathrm{d}t + \sum_{\alpha=1}^{2} \int_0^T \mu_\alpha (\boldsymbol{\nabla} u_\alpha, \boldsymbol{\nabla}(v_\alpha - u_\alpha))_\Omega\,\mathrm{d}t \\
&\geq \int_0^T \sum_{\alpha=1}^{2} (f_\alpha, v_\alpha - u_\alpha)_\Omega\,\mathrm{d}t + \frac{1}{2}\sum_{\alpha=1}^{2} \left(\underbrace{\|v_\alpha(T) - u_\alpha(T)\|_\Omega^2}_{\geq 0} - \|v_\alpha(0) - u_\alpha(0)\|_\Omega^2 \right), \\
&\geq \int_0^T \sum_{\alpha=1}^{2} (f_\alpha, v_\alpha - u_\alpha)_\Omega\,\mathrm{d}t - \frac{1}{2}\sum_{\alpha=1}^{2} \|v_\alpha(0) - u_\alpha(0)\|_\Omega^2.
\end{aligned} \tag{2.20}$$

Thus, problem (2.18) implies the property $\forall \boldsymbol{v} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$:

$$\int_0^T \langle \partial_t \boldsymbol{v}, \boldsymbol{v} - \boldsymbol{u} \rangle\,\mathrm{d}t + \int_0^T a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u})\,\mathrm{d}t \geq \int_0^T l(\boldsymbol{v} - \boldsymbol{u})\,\mathrm{d}t - \sum_{\alpha=1}^{2} \frac{1}{2}\left\|v_\alpha(0) - u_\alpha^0\right\|_\Omega^2. \tag{2.21}$$

2.5 Discretization and semismooth Newton linearization

We now present the discretization and linearization of our model. Anagolously to Sections 2.4 and 2.3, we will provide a discretization of the parabolic variational inequality (2.18) and a discretization of the saddle-point formulation (2.12). The discretization relies on the backward Euler scheme in time and on the conforming finite element method of degree $p \geq 1$ in space.

2.5.1 Setting

For the time discretization, we introduce a division of the interval $[0, T]$ into subintervals $I_n := [t_{n-1}, t_n]$, $1 \leq n \leq N_\mathrm{t}$, such that $0 = t_0 < t_1 < \cdots < t_{N_\mathrm{t}} = T$. The time steps are denoted by $\Delta t_n = t_n - t_{n-1}$, $n = 1, \cdots, N_\mathrm{t}$.

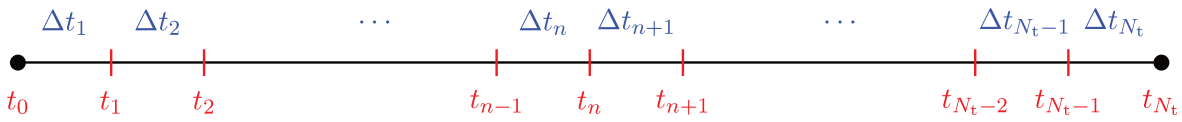

Figure 2.1: Time discretization of the model.

For the space discretization, we consider a conforming simplicial meshe $\mathcal{T}_h$ of the domain Ω, *i.e.* $\mathcal{T}_h$ is a set of triangles verifying

$$\bigcup_{K \in \mathcal{T}_h} \overline{K} = \overline{\Omega},$$

where the intersection of the closure of two elements of $\mathcal{T}_h$ is either an empty set, a vertex, or an edge. The set of vertices of $\mathcal{T}_h$ is denoted by $\mathcal{V}_h$ and is partitionned into the interior vertices $\mathcal{V}_h^{\mathrm{int}}$ and the boundary vertices $\mathcal{V}_h^{\mathrm{ext}}$. We denote by $\mathcal{N}_h^{\mathrm{int}}$ the number of interior vertices. The vertices of an element $K \in \mathcal{T}_h$ are collected in the set $\mathcal{V}_K$. Denote by h_K the diameter of a triangle K and $h := \max_{K \in \mathcal{T}_h} h_K$. Furthermore, for $\boldsymbol{a} \in \mathcal{V}_h$, let the patch $\omega_h^{\boldsymbol{a}} \subset \Omega$ be the domain made up of the elements of $\mathcal{T}_h$ that share $\boldsymbol{a}$. The vector $\mathbf{n}_{\omega_h^{\boldsymbol{a}}}$ stands for its outward unit normal. In the sequel, we use the discrete conforming space of piecewise polynomial functions $\forall 1 \leq n \leq N_\mathrm{t}$

$$X_h^p := \left\{ v_h \in \mathcal{C}^0(\overline{\Omega});\ v_h|_K \in \mathbb{P}_p(K)\ \ \forall K \in \mathcal{T}_h \right\} \subset H^1(\Omega),$$

where $\mathbb{P}_p(K)$ stands for the set of polynomials of total degree less than or equal to $p \geq 1$ on the element $K \in \mathcal{T}_h$. We also denote by $\mathcal{V}_\mathrm{d}^p$ the set of the Lagrange nodes degrees of freedom of the space X_h^p and by $\mathcal{N}_\mathrm{d}^p$ its cardinality. The internal degrees of freedom are collected in the set $\mathcal{V}_\mathrm{d}^{p,\mathrm{int}}$ and the boundary ones are collected in the set $\mathcal{V}_\mathrm{d}^{p,\mathrm{ext}}$. The Lagrange basis functions of X_h^p are denoted by $(\psi_{h,\boldsymbol{x}_l})_{1 \leq l \leq \mathcal{N}_\mathrm{d}^p}$ for $\boldsymbol{x}_l \in \mathcal{V}_\mathrm{d}^p$. We recall that $\psi_{h,\boldsymbol{x}_l}(\boldsymbol{x}_l) = 1$ for all $\boldsymbol{x}_l \in \mathcal{V}_\mathrm{d}^p$ and $\psi_{h,\boldsymbol{x}_l}(\boldsymbol{x}_{l'}) = 0$ for all $(\boldsymbol{x}_{l'})_{1 \leq l' \neq l \leq \mathcal{N}_\mathrm{d}^p} \in \mathcal{V}_\mathrm{d}^p$. In the particular case $p = 1$, the set $\mathcal{V}_\mathrm{d}^1$ coincides with $\mathcal{V}_h$ and the Lagrange basis functions are the "hat" basis functions that are denoted by $\psi_{h,\boldsymbol{a}}$, $\boldsymbol{a} \in \mathcal{V}_h$. Still in this case, we denote

$$M_{\boldsymbol{a}} := (\psi_{h,\boldsymbol{a}}, 1)_{\omega_h^{\boldsymbol{a}}} = \frac{|\omega_h^{\boldsymbol{a}}|}{3}.$$

We also introduce the boundary-aware set and space

$$X_{gh}^p := \{ v_h \in X_h^p,\ v_h = g \text{ on } \partial\Omega \} \subset H_g^1(\Omega), \quad X_{0h}^p := X_h^p \cap H_0^1(\Omega),$$

and the convex set

$$\boldsymbol{\mathcal{K}}_{gh}^p := \left\{ (v_{1h}, v_{2h}) \in X_{gh}^p \times X_{0h}^p,\ v_{1h}(\boldsymbol{x}_l) - v_{2h}(\boldsymbol{x}_l) \geq 0\ \ \forall (\boldsymbol{x}_l)_{1 \leq l \leq \mathcal{N}_\mathrm{d}^p} \in \mathcal{V}_\mathrm{d}^p \right\}. \tag{2.22}$$

Observe that $\boldsymbol{\mathcal{K}}_{gh}^p \not\subset \boldsymbol{\mathcal{K}}_g$ for $p \geq 2$ and that $\boldsymbol{\mathcal{K}}_{gh}^1 \subset \boldsymbol{\mathcal{K}}_g$ holds only in the case $p = 1$, see [17, 53, 62, 153].

2.5.2 Discrete reduced problem and discrete saddle-point problem

We begin this section by defining the continuous linear form

$$c_n(\boldsymbol{u}_h^n, \boldsymbol{v}_h) := \frac{1}{\Delta t_n} \sum_{\alpha=1}^{2} (u_{\alpha h}^n, v_{\alpha h})_\Omega \,, \quad 1 \leq n \leq N_\mathrm{t}. \tag{2.23}$$

Given $\boldsymbol{u}_h^0 \in \boldsymbol{\mathcal{K}}_g$, the discrete reduced problem corresponding to (2.18) consists in searching for all $1 \leq n \leq N_\mathrm{t}$, $\boldsymbol{u}_h^n \in \boldsymbol{\mathcal{K}}_{gh}^p$ such that $\forall \boldsymbol{v}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$

$$c_n\left(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h - \boldsymbol{u}_h^n\right) + a(\boldsymbol{u}_h^n, \boldsymbol{v}_h - \boldsymbol{u}_h^n) \geq l(\boldsymbol{v}_h - \boldsymbol{u}_h^n). \tag{2.24}$$

Proposition 2.5.1. *The discrete problem (2.24) admits a unique solution that is nonconforming in the sense that $\boldsymbol{u}_h^n \notin \boldsymbol{\mathcal{K}}_{gh}^p$ when $p \geq 2$.*

Proof. $(H^1(\Omega))^2$ is a Hilbert space and $\boldsymbol{\mathcal{K}}_{gh}^p$ is a nonempty closed convex set of $(H^1(\Omega))^2$. Furthermore, $a + c_n$ is a bilinear and continuous form on $(H^1(\Omega))^2 \times (H^1(\Omega))^2$ and coercive on $(H_0^1(\Omega))^2 \times (H_0^1(\Omega))^2$. Besides, $l - c_n$ defines a linear and continuous form on $(H^1(\Omega))^2$. Thus, as a result of the Lions–Stampacchia theorem (see [37]) problem (2.24) admits a unique solution $\boldsymbol{u}_h^n \notin \boldsymbol{\mathcal{K}}_{gh}^p$ when $p > 1$. □

Moreover, following the methodology of [17, 43, 62, 99] we define for $1 \leq n \leq N_\mathrm{t}$ the functions λ_{1h}^n and λ_{2h}^n in X_h^p by

$$\begin{aligned}
\langle \lambda_{1h}^n, z_{1h} \rangle_h &:= \frac{1}{\Delta t_n} \left(u_{1h}^n - u_{1h}^{n-1}, z_{1h}\right)_\Omega + \mu_1 \left(\boldsymbol{\nabla} u_{1h}^n, \boldsymbol{\nabla} z_{1h}\right)_\Omega - (f_1, z_{1h})_\Omega \quad \forall z_{1h} \in X_{0h}^p, \\
\langle \lambda_{2h}^n, z_{2h} \rangle_h &:= -\frac{1}{\Delta t_n} \left(u_{2h}^n - u_{2h}^{n-1}, z_{2h}\right)_\Omega - \mu_2 \left(\boldsymbol{\nabla} u_{2h}^n, \boldsymbol{\nabla} z_{2h}\right)_\Omega + (f_2, z_{2h})_\Omega \quad \forall z_{2h} \in X_{0h}^p, \\
\langle \lambda_{1h}^n, \psi_{h,\boldsymbol{x}_l} \rangle_h &:= 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_\mathrm{d}^{p,\mathrm{ext}}, \\
\langle \lambda_{2h}^n, \psi_{h,\boldsymbol{x}_l} \rangle_h &:= 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_\mathrm{d}^{p,\mathrm{ext}},
\end{aligned} \tag{2.25}$$

where for all $(w_h, v_h) \in X_h^p \times X_h^p$

$$\langle w_h, v_h \rangle_h := \begin{cases} \displaystyle\sum_{\boldsymbol{a} \in \mathcal{V}_h} w_h(\boldsymbol{a}) v_h(\boldsymbol{a}) M_{\boldsymbol{a}} & \text{if} \quad p = 1, \qquad (2.26) \\ (w_h, v_h)_\Omega & \text{if} \quad p \geq 2. \qquad (2.27) \end{cases}$$

Lemma 2.5.2. *Let $1 \leq n \leq N_\mathrm{t}$ be a time step and $(u_{1h}^n, u_{2h}^n) \in \boldsymbol{\mathcal{K}}_{gh}^p$ be the solution of the reduced discrete problem (2.24). Then, the functions λ_{1h}^n and λ_{2h}^n defined by (2.25) coincide and we set $\lambda_h^n := \lambda_{1h}^n = \lambda_{2h}^n \in X_h^p$.*

Proof. We substract the first two equations of (2.25) taking $z_{1h} = z_{2h} = \psi_{h,\boldsymbol{x}_l}$ with $\boldsymbol{x}_l$ any internal Lagrange node to get

$$\langle \lambda_{1h}^n - \lambda_{2h}^n, \psi_{h,\boldsymbol{x}_l} \rangle_h = \frac{1}{\Delta t_n} \left(u_{1h}^n - u_{1h}^{n-1} + u_{2h}^n - u_{2h}^{n-1}, \psi_{h,\boldsymbol{x}_l} \right)_\Omega$$
$$+ \left(\mu_1 \boldsymbol{\nabla} u_{1h}^n + \mu_2 \boldsymbol{\nabla} u_{2h}^n, \boldsymbol{\nabla} \psi_{h,\boldsymbol{x}_l} \right)_\Omega - \left(f_1 + f_2, \psi_{h,\boldsymbol{x}_l} \right)_\Omega \quad \forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}.$$

Taking $v_{1h} := u_{1h}^n + \psi_{h,\boldsymbol{x}_l}$ and $v_{2h} := u_{2h}^n + \psi_{h,\boldsymbol{x}_l}$ in (2.24) where $\psi_{h,\boldsymbol{x}_l}$ is the Lagrange basis function associated to $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$ and noting that $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$, we see

$$\langle \lambda_{1h}^n - \lambda_{2h}^n, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}. \tag{2.28}$$

In the same way, considering $v_{1h} := u_{1h}^n - \psi_{h,\boldsymbol{x}_l}$ and $v_{2h} := u_{2h}^n - \psi_{h,\boldsymbol{x}_l}$, we have $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ and we get

$$\langle \lambda_{1h}^n - \lambda_{2h}^n, \psi_{h,\boldsymbol{x}_l} \rangle_h \leq 0 \quad \forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}. \tag{2.29}$$

Finally, combining (2.28) and (2.29) with the last two equations of (2.25) provides

$$\langle \lambda_{1h}^n - \lambda_{2h}^n, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0 \quad \forall 1 \leq l \leq \mathcal{N}_{\mathrm{d}}^p.$$

For $p \geq 2$, $\lambda_{1h}^n - \lambda_{2h}^n \in X_h^p$ is $L^2(\Omega)$-orthogonal to all test functions in the space X_h^p, which implies $\lambda_{1h}^n = \lambda_{2h}^n$. For $p = 1$, $\lambda_{1h}^n = \lambda_{2h}^n$ holds true because $M_{\boldsymbol{a}} > 0$. □

Furthermore, $\lambda_h^n \in X_h^p$ satisfies the following property:

Lemma 2.5.3. *Let $(u_{1h}^n, u_{2h}^n) \in \boldsymbol{\mathcal{K}}_{gh}^p$ be the solution of the reduced problem (2.24) and let λ_h^n be defined by (2.25). Then, there holds*

$$\langle \lambda_h^n, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}.$$

Proof. Let $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$ be an internal node. First observe that $(v_{1h}, v_{2h}) := (u_{1h}^n + \psi_{h,\boldsymbol{x}_l}, u_{2h}^n) \in \boldsymbol{\mathcal{K}}_{gh}^p$. The conclusion follows from the reduced problem (2.24), the characterization (2.25) with $z_{1h} = \psi_{h,\boldsymbol{x}_l} \in X_{0h}^p$ and Lemma 2.5.2 giving $\forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$

$$\mu_1 \left(\boldsymbol{\nabla} u_{1h}^n, \boldsymbol{\nabla} \psi_{h,\boldsymbol{x}_l} \right)_\Omega - \left(f_1, \psi_{h,\boldsymbol{x}_l} \right)_\Omega + \frac{1}{\Delta t_n} \left(u_{1h}^n - u_{1h}^{n-1}, \psi_{h,\boldsymbol{x}_l} \right)_\Omega = \langle \lambda_{1h}, \psi_{h,\boldsymbol{x}_l} \rangle_h$$
$$= \langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0.$$

□

Following Lemma 2.5.3 and Chapter 1 we suggest to define the discrete convex set for λ_h^n by

$$\Lambda_h^p := \left\{ v_h \in X_h^p;\ \langle v_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0\ \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}},\ \langle v_h, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0\ \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{ext}} \right\}. \tag{2.30}$$

Observe that $\Lambda_h^p \not\subset \Lambda$ for $p \geq 2$ and in the case $p = 1$, Λ_h^p reduces to

$$\Lambda_h^1 = \left\{ v_h \in X_{0h}^1;\ v_h(\boldsymbol{a}) \geq 0 \quad \forall \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}} \right\} \subset \Lambda. \tag{2.31}$$

Note that for any $\chi_h^n \in \Lambda_h^p$ and any $\boldsymbol{v}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$,

$$\langle \chi_h^n, v_{1h}^n - v_{2h}^n \rangle_h = \sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}} (v_{1h}^n - v_{2h}^n)(\boldsymbol{x}_l) \langle \chi_h^n, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0. \tag{2.32}$$

Given $(u_{1h}^0, u_{2h}^0) \in \boldsymbol{\mathcal{K}}_{gh}^p$, a discrete formulation built by the Galerkin method, corresponding to problem (2.12) consists in searching $(u_{1h}^n, u_{2h}^n, \lambda_h^n) \in X_{gh}^p \times X_{0h}^p \times \Lambda_h^p$ such that for all $(z_{1h}, z_{2h}, \chi_h) \in X_{0h}^p \times X_{0h}^p \times \Lambda_h^p$ and for all $\chi_h \in \Lambda_h^p$

$$\begin{aligned} &\frac{1}{\Delta t_n} \sum_{\alpha=1}^{2} \left(u_{\alpha h}^n - u_{\alpha h}^{n-1}, z_{\alpha h}\right)_\Omega + \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_{\alpha h}^n, \boldsymbol{\nabla} z_{\alpha h}\right)_\Omega - \langle \lambda_h^n, z_{1h} - z_{2h} \rangle_h = \sum_{\alpha=1}^{2} \left(f_\alpha, z_{\alpha h}\right)_\Omega, \\ &\langle \chi_h - \lambda_h^n, u_{1h}^n - u_{2h}^n \rangle_h \geq 0. \end{aligned} \tag{2.33}$$

To prove the next Lemma, we construct the basis $(\Theta_{h,\boldsymbol{x}_l})_{1 \leq l \leq \mathcal{N}_{\mathrm{d}}^p}$ of X_h^p, dual to $(\psi_{h,\boldsymbol{x}_l})_{1 \leq l \leq \mathcal{N}_{\mathrm{d}}^p}$, satisfying

$$\begin{aligned} \langle \Theta_{h,\boldsymbol{x}_l}, \psi_{h,\boldsymbol{x}_l} \rangle_h &= 1 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p, \\ \langle \Theta_{h,\boldsymbol{x}_l}, \psi_{h,\boldsymbol{x}_l^*} \rangle_h &= 0 \quad \forall \boldsymbol{x}_{l^*} \in \mathcal{V}_{\mathrm{d}}^p,\ \boldsymbol{x}_{l^*} \neq \boldsymbol{x}_l. \end{aligned} \tag{2.34}$$

This procedure has been recently introduced in [62]. Note that each vector of the dual basis $\Theta_{h,\boldsymbol{x}_l}$ can be determined by inverting a diagonal (lumped mass) matrix for $p = 1$ and the finite element mass matrix for $p \geq 2$; importantly, all $\Theta_{h,\boldsymbol{x}_l}$, $1 \leq l \leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ belong to Λ_h^p. Note, however, that the support of $\Theta_{h,\boldsymbol{x}_l}$ is typically not local. Then we have

Lemma 2.5.4. *Let $1 \leq n \leq N_{\mathrm{t}}$ be a time step. For any solution $(u_{1h}^n, u_{2h}^n, \lambda_h^n)$ of problem (2.33), the pair (u_{1h}^n, u_{2h}^n) is a solution of problem (2.24). Conversely, for any solution (u_{1h}^n, u_{2h}^n) of problem (2.24), defining the function $\lambda_h^n = \lambda_{\alpha h}^n$, $\alpha = 1, 2$, by (2.25), the triple $(u_{1h}^n, u_{2h}^n, \lambda_h^n)$ is a solution to problem (2.33).*

Proof. For the case $p = 1$, the proof is a direct extension of [18, Lemma 13] and for $p \geq 2$ it employs the arguments of Chapter 1 Lemma 1.2.4. Let $p \geq 1$ and let $(u_{1h}^n, u_{2h}^n, \lambda_h^n)$ be the solution of problem (2.33). First we prove that $\boldsymbol{u}_h^n \in \boldsymbol{\mathcal{K}}_{gh}^p$. Decomposing $u_{1h}^n - u_{2h}^n$ in the Lagrange basis we obtain

$$u_{1h}^n - u_{2h}^n = \sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p} (u_{1h}^n - u_{2h}^n)(\boldsymbol{x}_l) \psi_{h,\boldsymbol{x}_l}.$$

Next, note that for $\chi_h, \lambda_h^n \in \Lambda_h^p$, $\chi_h + \lambda_h^n \in \Lambda_h^p$. Thereore, taking $\chi_h + \lambda_h^n$ as a test function in (2.33), we get

$$\langle \chi_h, u_{1h}^n - u_{2h}^n \rangle_h \geq 0 \quad \forall \chi_h \in \Lambda_h^p,$$

and then

$$\sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p} (u_{1h}^n - u_{2h}^n)(\boldsymbol{x}_l) \langle \chi_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall \chi_h \in \Lambda_h^p. \tag{2.35}$$

Finally, taking in (2.35) $\chi_h = \Theta_{h,\boldsymbol{x}_l^*}$, for all $\boldsymbol{x}_l^* \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$, we obtain

$$\sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p} (u_{1h}^n - u_{2h}^n)(\boldsymbol{x}_l) \left\langle \Theta_{h,\boldsymbol{x}_l^*}, \psi_{h,\boldsymbol{x}_l} \right\rangle_h = (u_{1h}^n - u_{2h}^n)(\boldsymbol{x}_l^*) \geq 0 \quad \forall \boldsymbol{x}_l^* \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}},$$

which proves that $\boldsymbol{u}_h^n \in \boldsymbol{\mathcal{K}}_{gh}^p$.

Now, we prove (2.24). Let $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$. Taking $z_{1h} := v_{1h} - u_{1h}^n \in X_{0h}^p$ and $z_{2h} := v_{2h} - u_{2h}^n \in X_{0h}^p$ as test functions in (2.33) provides

$$\begin{aligned} \langle \lambda_h^n, v_{1h} - v_{2h} \rangle_h - \langle \lambda_h^n, u_{1h}^n - u_{2h}^n \rangle_h &= a(\boldsymbol{u}_h^n, \boldsymbol{v}_h - \boldsymbol{u}_h^n) - l(\boldsymbol{v}_h - \boldsymbol{u}_h^n) \\ &\quad + c_n \left(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h - \boldsymbol{u}_h^n \right). \end{aligned} \tag{2.36}$$

Using (2.32) with $\lambda_h^n \in \Lambda_h^p$ and $\boldsymbol{v}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$ and taking $\chi_h = 0 \in \Lambda_h^p$ in (2.33) gives

$$\langle \lambda_h^n, v_{1h} - v_{2h} \rangle_h \geq 0, \quad \langle -\lambda_h^n, u_{1h}^n - u_{2h}^n \rangle_h \geq 0. \tag{2.37}$$

Combining (2.36) and (2.37) provides (2.24).

Conversely, let $(u_{1h}^n, u_{2h}^n) \in \boldsymbol{\mathcal{K}}_{gh}^p$ be the solution of the reduced problem (2.24) and let $(z_{1h}, z_{2h}) \in X_{0h}^p \times X_{0h}^p$ be arbitrary. The Lagrange multiplier λ_h^n defined by (2.25) combined with Lemma 2.5.2 and Lemma 2.5.3 yields $\lambda_h^n \in \Lambda_h^p$. Next, subtracting the two equations of (2.25) gives the first line of (2.33). It remains to prove the second line of (2.33). Let now $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$. The first line in (2.33) now implies (2.36) and the reduced problem (2.24) yields

$$-\langle \lambda_h^n, u_{1h}^n - u_{2h}^n \rangle_h + \langle \lambda_h^n, v_{1h} - v_{2h} \rangle_h \geq 0 \quad \forall (v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p. \tag{2.38}$$

For $v_{1h} := u_{1h}^n - \displaystyle\sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}} u_{1h}(\boldsymbol{x}_l)\psi_{h,\boldsymbol{x}_l} \in X_{gh}^p$ and $v_{2h} := 0 \in X_{0h}^p$, $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$, and using the definition of Λ_h^p, we have $\langle \lambda_h^n, v_{1h} - v_{2h} \rangle_h = 0$ and the inequality (2.38) yields

$$-\langle \lambda_h^n, u_{1h}^n - u_{2h}^n \rangle_h \geq 0.$$

To conclude the proof, we use (2.32) with $\boldsymbol{u}_h^n \in \boldsymbol{\mathcal{K}}_{gh}^p$ and for any $\chi_h \in \Lambda_h^p$. □

Remark 2.5.5. *Taking in* (2.33) $\chi_h = 0$ *and next* $\chi_h = 2\lambda_h^n \in \Lambda_h^p$ *gives*

$$\langle \lambda_h^n, u_{1h}^n - u_{2h}^n \rangle_h = 0.$$

As $\boldsymbol{u}_h^n \in \boldsymbol{\mathcal{K}}_{gh}^p$ *and* $\lambda_h^n \in \Lambda_h^p$, *we obtain the discrete complementarity condition valid* $\forall p \geq 1$:

$$\begin{aligned} &(u_{1h}^n - u_{2h}^n)(\boldsymbol{x}_l) \geq 0 && \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}, \\ &\langle \lambda_h^n, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0, && \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}, \\ &\langle \lambda_h^n, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0 && \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{ext}}, \\ &\langle \lambda_h^n, u_{1h}^n - u_{2h}^n \rangle_h = 0. \end{aligned} \tag{2.39}$$

2.5.3 Numerical resolution and discrete complementarity constraints

This section mimics Chapter 1, Section 1.2.3, then we only give the key ingredients. We are interested in expressing the discrete problem (2.33) under an algebraic form employing for the discrete Lagrange multiplier λ_h the Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^p}$, and the basis $(\Theta_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^p}$ dual to $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^p}$. Concerning the discrete solution $u_{1h}^n \in X_{gh}^p$ we use the lifting $u_{1h}^n = u_{1h}^{*,n} + g$ where $u_{1h}^{*,n} \in X_{0h}^p$ and $g > 0$ is the boundary value. The matricial representation of the lifting is denoted by $\boldsymbol{X}_{1h}^n \in \mathbb{R}^{\mathcal{N}_{\rm d}^{p,\rm int}}$ and is defined as (1.33). The matricial representation $\boldsymbol{X}_{2h}^n$ for the discrete functional u_{2h}^n is expressed in the Lagrange basis $\psi_{h,\boldsymbol{x}_l}$ as (1.34). We define the finite element mass matrix $\mathbb{M} \in \mathbb{R}^{\mathcal{N}_{\rm d}^p,\mathcal{N}_{\rm d}^p}$ by

$$\mathbb{M}_{l,m} := (\psi_{h,\boldsymbol{x}_l}, \psi_{h,\boldsymbol{x}_m})_\Omega \quad 1 \leq l,m \leq \mathcal{N}_{\rm d}^p. \tag{2.40}$$

For $\boldsymbol{x}_{l'}$ an internal node we extract from $\mathbb{M}$ the rectangular matrix $\widehat{\mathbb{M}} \in \mathbb{R}^{\mathcal{N}_{\rm d}^{p,\rm int},\mathcal{N}_{\rm d}^p}$ defined as

$$\left(\widehat{\mathbb{M}}_{l,l'}\right)_{\substack{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}\\ 1\leq l'\leq \mathcal{N}_{\rm d}^p}} := (\mathbb{M}_{j,j'})_{\substack{\boldsymbol{x}_j\in\mathcal{V}_{\rm d}^{p,\rm int}\\ \boldsymbol{x}_j\in\mathcal{V}_{\rm d}^p}}. \tag{2.41}$$

We also extract from $\mathbb{M}$ the square matrix denoted by $\overset{\circ}{\mathbb{M}} \in \mathbb{R}^{\mathcal{N}_{\rm d}^{p,\rm int},\mathcal{N}_{\rm d}^{p,\rm int}}$ corresponding to internal nodes defined as

$$\left(\overset{\circ}{\mathbb{M}}_{l,l'}\right)_{\substack{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}\\ 1\leq l'\leq \mathcal{N}_{\rm d}^{p,\rm int}}} := (\mathbb{M}_{j,j'})_{\substack{\boldsymbol{x}_j\in\mathcal{V}_{\rm d}^{p,\rm int}\\ \boldsymbol{x}_j\in\mathcal{V}_{\rm d}^{p,\rm int}}}. \tag{2.42}$$

Case of Lagrange basis for λ_h, $p = 1$

In this case, $\mathcal{N}_{\rm d}^{p,\rm int} = \mathcal{N}_h^{\rm int}$ and the discrete Lagrange multiplier λ_h^n admits a matricial representation $\boldsymbol{X}_{3h}^n$ in the Lagrange basis $(\psi_{h,\boldsymbol{a}_l})_{1\leq l\leq \mathcal{N}_h^{\rm int}}$ as (1.35). The first line of (2.33) reads

$$\mathbb{E}_1^n \boldsymbol{X}_h^n = \boldsymbol{F}^n, \tag{2.43}$$

where $\boldsymbol{X}_h^n := [\boldsymbol{X}_{1h}^n, \boldsymbol{X}_{2h}^n, \boldsymbol{X}_{3h}^n]^T$ is the unknown vector, $\mathbb{E}_1^n \in \mathbb{R}^{2\mathcal{N}_h^{\rm int},3\mathcal{N}_h^{\rm int}}$ is a rectangular matrix defined by

$$\mathbb{E}_1^n := \begin{bmatrix} \mu_1\mathbb{S} + \frac{1}{\Delta t_n}\overset{\circ}{\mathbb{M}} & \mathbf{0} & -\mathbb{D} \\ \mathbf{0} & \mu_2\mathbb{S} + \frac{1}{\Delta t_n}\overset{\circ}{\mathbb{M}} & +\mathbb{D} \end{bmatrix},$$

where the stiffness matrix $\mathbb{S} \in \mathbb{R}^{\mathcal{N}_h^{\rm int},\mathcal{N}_h^{\rm int}}$ is defined as (1.36), $\mathbb{D}$ is the diagonal mass lumped matrix defined as (1.36), and $\overset{\circ}{\mathbb{M}}$ is the finite element mass matrix defined above. The right-hand side vector $\boldsymbol{F}^n$ is defined by blocks ($[\boldsymbol{F}^n]^T := [\boldsymbol{F}_1^n, \boldsymbol{F}_2^n]^T$) as

$$\begin{aligned} (\boldsymbol{F}_1^n)_l &:= \left(f_1 + \frac{1}{\Delta t_n}u_{1h}^{n-1}, \psi_{h,\boldsymbol{a}_l}\right)_\Omega \quad 1 \leq l \leq \mathcal{N}_h^{\rm int}, \\ (\boldsymbol{F}_2^n)_l &:= \left(f_2 + \frac{1}{\Delta t_n}u_{2h}^{n-1}, \psi_{h,\boldsymbol{a}_l}\right)_\Omega \quad 1 \leq l \leq \mathcal{N}_h^{\rm int}. \end{aligned} \tag{2.44}$$

The linear complementarity constraints are easily computed as (1.38) and problem (2.33), when $p = 1$, is equivalent to search $\boldsymbol{X}_h^n \in R^{3\mathcal{N}_h^{\rm int}}$ such that

$$\begin{aligned} &\mathbb{E}_1^n \boldsymbol{X}_h^n = \boldsymbol{F}^n, \\ &\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n \geq 0, \quad \boldsymbol{X}_{3h}^n \geq 0, \quad (\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n) \cdot \boldsymbol{X}_{3h}^n = 0. \end{aligned} \tag{2.45}$$

Case of Lagrange basis for λ_h, $p \geq 2$

In this configuration, the discrete Lagrange multiplier $\lambda_h^n \in X_h^p$ admits a matricial representation $\widetilde{\boldsymbol{X}}_{3h}^n$ in the lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^p}$ defined as (1.42). Next, the first line of (2.33) reads

$$\widetilde{\mathbb{E}}_p^n \boldsymbol{X}_h^n = \boldsymbol{F}^n, \tag{2.46}$$

where $\boldsymbol{X}_h^n := \left[\boldsymbol{X}_{1h}^n, \boldsymbol{X}_{2h}^n, \widetilde{\boldsymbol{X}}_{3h}^n\right]^T \in \mathbb{R}^{2\mathcal{N}_{\rm d}^{p,\rm int}+\mathcal{N}_{\rm d}^p}$ is the unknown vector, $\widetilde{\mathbb{E}}_p^n \in \mathbb{R}^{2\mathcal{N}_{\rm d}^{p,\rm int},2\mathcal{N}_{\rm d}^{p,\rm int}+\mathcal{N}_{\rm d}^p}$ is a rectangular matrix defined by

$$\widetilde{\mathbb{E}}_p^n := \begin{bmatrix} \mu_1\mathbb{S} + \frac{1}{\Delta t_n}\overset{\circ}{\mathbb{M}} & \mathbf{0} & -\widehat{\mathbb{M}} \\ \mathbf{0} & \mu_2\mathbb{S} + \frac{1}{\Delta t_n}\overset{\circ}{\mathbb{M}} & +\widehat{\mathbb{M}} \end{bmatrix}, \tag{2.47}$$

where the stiffness matrix $\mathbb{S} \in \mathbb{R}^{\mathcal{N}_{\rm d}^{p,\rm int},\mathcal{N}_{\rm d}^{p,\rm int}}$ is defined as (1.43) and the finite element mass matrices $\overset{\circ}{\mathbb{M}}$ and $\widehat{\mathbb{M}}$ defined above. The right-hand side vector $\boldsymbol{F}^n$ is defined as (2.44) where the hat functions $(\psi_{h,\boldsymbol{a}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}}$ are replaced by the Lagrange basis functions $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}}$. The complementarity constraints are expressed in the Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}}$ as (1.44), (1.45), and (1.46). Finally, when $p \geq 2$, problem (2.33) is equivalent to search $\boldsymbol{X}_h^n \in \mathbb{R}^{2\mathcal{N}_{\rm d}^{p,\rm int}+\mathcal{N}_{\rm d}^p}$ such that

$$\begin{aligned} &\widetilde{\mathbb{E}}_p^n \boldsymbol{X}_h^n = \boldsymbol{F}^n, \\ &\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n \geq 0, \quad \widehat{\mathbb{M}}\widetilde{\boldsymbol{X}}_{3h}^n \geq 0, \quad (\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n) \cdot \widehat{\mathbb{M}}\widetilde{\boldsymbol{X}}_{3h}^n = 0. \end{aligned} \tag{2.48}$$

Case of dual basis for λ_h, $p \geq 2$

To finish, we provide a characterization of problem (2.33) in the basis $(\Theta_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}}$ dual to $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}}$ under an algebraic form when $p \geq 2$. The lifting $u_{1h}^* \in X_{0h}^p$ and $u_{2h}^n \in X_{0h}^p$ admit a matricial representation in the Lagrange basis $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}}$ given by (1.33) and (1.34). The discrete Lagrange multiplier λ_h^n is decomposed in the basis $(\Theta_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}}$ as (1.48). The first line of (2.33) reads

$$\mathbb{E}_p^n \boldsymbol{X}_h^n = \boldsymbol{F}^n, \tag{2.49}$$

where $\boldsymbol{X}_h^n := [\boldsymbol{X}_{1h}^n, \boldsymbol{X}_{2h}^n, \boldsymbol{X}_{3h}^n]^T \in \mathbb{R}^{3\mathcal{N}_{\rm d}^{p,\rm int}}$ is the unknown vector, and $\mathbb{E}_p^n \in \mathbb{R}^{2\mathcal{N}_{\rm d}^{p,\rm int},3\mathcal{N}_{\rm d}^{p,\rm int}}$ is a rectangular matrix defined by

$$\mathbb{E}_p^n := \begin{bmatrix} \mu_1\mathbb{S} + \frac{1}{\Delta t_n}\overset{\circ}{\mathbb{M}} & \mathbf{0} & -\mathbb{I}_{\rm d} \\ \mathbf{0} & \mu_2\mathbb{S} + \frac{1}{\Delta t_n}\overset{\circ}{\mathbb{M}} & +\mathbb{I}_{\rm d} \end{bmatrix}, \tag{2.50}$$

where $\mathbb{I}_{\mathrm{d}} \in R^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}},\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ is the identity matrix, $\mathbb{S} \in R^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}},\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ is the stiffness matrix, and $\overset{\circ}{\mathbb{M}}$ the finite element mass matrix defined above. The right-hand side vector $\boldsymbol{F}^n$ is also constructed as above. The complementarity constraints of (2.33) are expressed in the basis $(\Theta_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ as (1.49), (1.50), and (1.53). Thus, when $p \geq 2$, problem (2.33) is equivalent to search $\boldsymbol{X}_h^n \in R^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ such that

$$\begin{aligned} &\mathbb{E}_p^n \boldsymbol{X}_h^n = \boldsymbol{F}^n, \\ &\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n \geq 0, \quad \boldsymbol{X}_{3h}^n \geq 0, \quad (\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n) \cdot \boldsymbol{X}_{3h}^n = 0. \end{aligned} \tag{2.51}$$

Comments

As Chapter 1, Section 1.2.3, we obtained at each time step $1 \leq l \leq N_{\mathrm{t}}$ and for $p = 1$, discrete complementarity constraints in internal vertices of the mesh and the matrix $\mathbb{E}_1^n$ is similar to the one of Chapter 1 up to a scaled mass matrix. The system (2.45) is practical for implementation. For $p \geq 2$, we obtained a discrete complementarity problem (2.48) expressed in the Lagrange nodes. The construction of the rectangular matrix $\widetilde{\mathbb{E}}_p^n$ required to compute the mass matrices $\widehat{\mathbb{M}}$ and $\overset{\circ}{\mathbb{M}}$ which has some cost. The characterization of problem (2.33), when $p \geq 2$, in the basis $\Theta_{h,\boldsymbol{x}_l}$ shows that the construction of the rectangular matrix $\mathbb{E}_p^n$ depends on the identity matrix and the constraints are expressed in a very convenient way. Nevertheless, in this case, the unknown vector $\boldsymbol{X}_h^n$ has $\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ coordinates expressed in the basis $\Theta_{h,\boldsymbol{x}_l}$. So, to express it afterwards in the Lagrange basis, it requires to invert the finite element mass matrix which has some cost. In the sequel we consider the case $p \geq 1$ and the expression of (2.33) written in the basis $(\Theta_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$.

2.5.4 C-functions

We now express the complementarity constraints given by the second line of (2.51) via non-differentiable equations. Let us recall that a function $f : (\mathbb{R}^m)^2 \to \mathbb{R}^m$ $(m \geq 1)$ is a C-function or a complementarity function if

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in (\mathbb{R}^m)^2 \qquad f(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{0} \quad \Longleftrightarrow \quad \boldsymbol{x} \geq \mathbf{0}, \quad \boldsymbol{y} \geq \mathbf{0}, \quad \boldsymbol{x} \cdot \boldsymbol{y} = 0.$$

Examples of C-functions are the min function

$$(\min\{\boldsymbol{x}, \boldsymbol{y}\})_l := \min\{\boldsymbol{x}_l, \boldsymbol{y}_l\} \qquad l = 1, \ldots, m, \tag{2.52}$$

the Fischer–Burmeister function

$$(f_{\mathrm{FB}}(\boldsymbol{x}, \boldsymbol{y}))_l := \sqrt{\boldsymbol{x}_l^2 + \boldsymbol{y}_l^2} - (\boldsymbol{x}_l + \boldsymbol{y}_l) \quad l = 1, \ldots, m,$$

or the Mangasarian function

$$(f_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{y}))_l := \xi(|\boldsymbol{x}_l - \boldsymbol{y}_l|) - \xi(\boldsymbol{y}_l) - \xi(\boldsymbol{x}_l) \quad l = 1, \ldots, m,$$

where $\xi : \mathbb{R} \mapsto \mathbb{R}$ is an increasing function satisfying $\xi(\mathbf{0}) = \mathbf{0}$. The min function, the Fischer–Burmeister function, and the Mangasarian function are not Fréchet

differentiable. For more details on C-functions see [88,89]. Let $\tilde{\mathbf{C}}$ be any C-function, *i.e.*, satisfying (for $m = \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$) $\tilde{\mathbf{C}}(\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n, \boldsymbol{X}_{3h}^n) = 0 \iff \boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n \geq 0$, $\boldsymbol{X}_{3h}^n \geq 0$, and $[\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n] \cdot \boldsymbol{X}_{3h}^n = 0$. Then, introducing the function $\mathbf{C} : \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}} \to \mathbb{R}^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ defined as $\mathbf{C}(\boldsymbol{X}_h^n) = \tilde{\mathbf{C}}(\boldsymbol{X}_{1h}^n + g\mathbf{1} - \boldsymbol{X}_{2h}^n, \boldsymbol{X}_{3h}^n)$, the problem (2.51) can be equivalently rewritten as

$$\left\{ \begin{array}{ll} \mathbb{E}_p^n \boldsymbol{X}_h^n & = \boldsymbol{F}^n, \\ \mathbf{C}(\boldsymbol{X}_h^n) & = \mathbf{0}. \end{array} \right. \tag{2.53}$$

Now, we proceed as Chapter 1, Section 1.3. We establish at each time step $1 \leq n \leq N_{\mathrm{t}}$ an inexact semismooth Newton algorithm.

2.5.5 Linearization by semismooth Newton method

We provide in this section the linearization of system (2.53). Observe that the $2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ lines of (2.53) are linear and the nonlinearity occurs in the last $\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ lines of (2.53). Even if the function $\mathbf{C}$ is not Fréchet differentiable, it is locally Lipschitz and continuous. As a result of the Rademacher Theorem (see [58, 88, 89]), the function $\mathbf{C}$ is differentiable almost everywhere. More precisely, it belongs to the class of strong semismooth functions. The semismooth Newton linearization is defined as follows: Let an initial guess $\boldsymbol{X}_h^{n,0} \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ be given. Typically, $\boldsymbol{X}_h^{n,0} := \boldsymbol{X}_h^{n-1}$ where $\boldsymbol{X}_h^{n-1}$ is the last iterate from the previous time step (including possibly inexact solvers). At step $k \geq 1$ one looks for $\boldsymbol{X}_h^{n,k} \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ such that

$$\mathbb{A}^{n,k-1} \boldsymbol{X}_h^{n,k} = \boldsymbol{B}^{n,k-1}, \tag{2.54}$$

where $\mathbb{A}^{n,k-1} \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}, 3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ is a matrix, and $\boldsymbol{B}^{n,k-1} \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ the right-hand side vector. More precisely,

$$\mathbb{A}^{n,k-1} := \left[\begin{array}{c} \mathbb{E}_p^n \\ \mathbf{J}_{\mathbf{C}}(\boldsymbol{X}_h^{n,k-1}) \end{array} \right] \quad \boldsymbol{B}^{n,k-1} := \left[\begin{array}{c} \boldsymbol{F}^n \\ \mathbf{J}_{\mathbf{C}}(\boldsymbol{X}_h^{n,k-1}) \boldsymbol{X}_h^{n,k-1} - \mathbf{C}(\boldsymbol{X}_h^{n,k-1}) \end{array} \right]. \tag{2.55}$$

Here, the notation $\mathbf{J}_{\mathbf{C}}(\boldsymbol{X}_h^{n,k-1})$ stands for the Jacobian matrix in the sense of Clarke. For the construction of $\mathbf{J}_{\mathbf{C}}(\boldsymbol{X}_h^{n,k-1})$ when the C-function $\mathbf{C}$ is the min function refer to Chapter 1 Section 1.3.1.

For an "exact" resolution of (2.53), choose a tolerance $\varepsilon_{\mathrm{lin}}$ up to the machine precision and stop the linearization when the relative linearization residual $\boldsymbol{R}_{\mathrm{lin}}^{n,k}$ satisfies

$$\left\| \boldsymbol{R}_{\mathrm{lin}}^{n,k} \right\| := \frac{\left\| \begin{pmatrix} \boldsymbol{F}^n - \mathbb{E}_p^n \boldsymbol{X}_h^{n,k} \\ \mathbf{C}(\boldsymbol{X}_h^{n,k}) \end{pmatrix} \right\|}{\left\| \begin{pmatrix} \boldsymbol{F}^{n,0} - \mathbb{E}_p^n \boldsymbol{X}_h^{n,0} \\ \mathbf{C}(\boldsymbol{X}_h^{n,0}) \end{pmatrix} \right\|} \leq \varepsilon_{\mathrm{lin}}. \tag{2.56}$$

For an inexact resolution of (2.54), we choose an iterative algebraic solver indexed by $i \geq 0$. Given an initial guess $\boldsymbol{X}_h^{n,k,0} \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$, often taken as $\boldsymbol{X}_h^{n,k,0} := \boldsymbol{X}_h^{n,k-1}$, where $\boldsymbol{X}_h^{n,k-1}$ is the last available iterate from the previous semismooth Newton step (including possibly inexact algebraic solver), the residual is defined for $i \geq 0$ by

$$\boldsymbol{R}_h^{n,k,i} := \boldsymbol{B}^{n,k-1} - \mathbb{A}^{n,k-1} \boldsymbol{X}_h^{n,k,i}. \tag{2.57}$$

In fact, the residual $\boldsymbol{R}_h^{n,k,i} \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ is a block vector

$$\boldsymbol{R}_h^{n,k,i} := \left[\boldsymbol{R}_{1h}^{n,k,i}, \boldsymbol{R}_{2h}^{n,k,i}, \boldsymbol{R}_{3h}^{n,k,i}\right] \tag{2.58}$$

where $\boldsymbol{R}_{1h}^{n,k,i} \in \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ is the component of the whole residual vector associated to the first membrane, $\boldsymbol{R}_{2h}^{n,k,i} \in \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ is the part associated to the second membrane, and $\boldsymbol{R}_{3h}^{n,k,i} \in \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ is the part associated to the Lagrange multiplier. At the current step $k \geq 1$, we choose a tolerance $\varepsilon_{\mathrm{alg}}^k$ and stop the iterations when the relative algebraic residual $\boldsymbol{R}_{\mathrm{alg}}^{n,k,i}$ satisfies

$$\left\|\boldsymbol{R}_{\mathrm{alg}}^{n,k,i}\right\| := \frac{\left\|\boldsymbol{R}_h^{n,k,i}\right\|}{\left\|\boldsymbol{B}^{n,k-1} - \mathbb{A}^{n,k-1}\boldsymbol{X}_h^{n,k,0}\right\|} \leq \varepsilon_{\mathrm{alg}}^k, \tag{2.59}$$

where the term $\varepsilon_{\mathrm{alg}}^k$ is commonly called the "forcing term".

When the algebraic stopping criterion is satisfied ((2.59) or later (2.113)), update the solution:

$$\boldsymbol{X}_h^{n,k} := \boldsymbol{X}_h^{n,k,i}.$$

Concerning the stopping criterion of the nonlinear Newton solver, we request either (2.56) (with the index k replaced by k, i) or (2.114). When the stopping criterion is satisfied, update the solution:

$$\boldsymbol{X}_h^{n} := \boldsymbol{X}_h^{n,k,i}.$$

In this way, u_{1h}^{n-1}, u_{2h}^{n-1} are the functional representations of the vectors $\boldsymbol{X}_{1h}^{n-1,k,i}$ and $\boldsymbol{X}_{2h}^{n-1,k,i}$ when the criteria are met, *i.e.* when k, i are the final iterates.

2.6 A posteriori error analysis

2.6.1 Preamble

So far, at each time step $1 \leq n \leq N_{\mathrm{t}}$, we have presented the nonlinear system (2.53) giving in particular the degrees of freedom of the numerical solution $\boldsymbol{X}_h^{n,k,i} \in \mathbb{R}^{3\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ where $k \geq 1$ is the semismooth Newton step and $i \geq 0$ is the algebraic solver step. The functional representation of the vectors $\boldsymbol{X}_{1h}^{n,k,i}$ and $\boldsymbol{X}_{2h}^{n,k,i}$, denoted by $u_{1h}^{n,k,i}$ and $u_{2h}^{n,k,i}$ are given by (1.33) and (1.34) for $p \geq 1$, while the function of $\boldsymbol{X}_{3h}^{n,k,i}$ denoted by $\lambda_h^{n,k,i}$ is given by (1.35) if $p = 1$, or by (1.48) if $p \geq 2$. Obviously, $\left(u_{1h}^{n,k,i}, u_{2h}^{n,k,i}, \lambda_h^{n,k,i}\right) \in X_{gh}^p \times X_{0h}^p \times X_h^p \quad \forall 1 \leq n \leq N_{\mathrm{t}}$. Next, we associate to the functions in space $u_{1h}^{n,k,i} \in X_{gh}^p$, and $u_{2h}^{n,k,i} \in X_{0h}^p$, $1 \leq n \leq N_{\mathrm{t}}$, their space-time representation $u_{1h\tau}^{k,i}$, respectively $u_{2h\tau}^{k,i}$ as follows:

$$u_{1h\tau}^{k,i}|_{I_n} := \frac{u_{1h}^{n,k,i} - u_{1h}^{n-1}}{\Delta t_n}\left(t - t^n\right) + u_{1h}^{n,k,i} \quad \forall 1 \leq n \leq N_{\mathrm{t}}, \tag{2.60}$$

$$u_{2h\tau}^{k,i}|_{I_n} := \frac{u_{2h}^{n,k,i} - u_{2h}^{n-1}}{\Delta t_n}\left(t - t^n\right) + u_{2h}^{n,k,i} \quad \forall 1 \leq n \leq N_{\mathrm{t}}. \tag{2.61}$$

Concerning the discrete Lagrange multiplier $\lambda_h^{n,k,i} \in X_h^p$, its space-time representation is defined by $\lambda_{h\tau}^{k,i}$ as follows:

$$\lambda_{h\tau}^{k,i}|_{I_n} := \lambda_h^{n,k,i} \quad \textit{i.e.} \text{ piecewise constant in time.} \tag{2.62}$$

Note that this construction ensures that $u_{\alpha h\tau}^{k,i}$ are piecewise affine and continuous in time, so that $\partial_t u_{\alpha h\tau}^{k,i} \in V_0^*$. In the expressions of $u_{1h\tau}^{k,i}$, $u_{2h\tau}^{k,i}$ and $\lambda_{h\tau}^{k,i}$, the indices k, i are kept merely to indicate the presence of inexact solvers. Note that $u_{\alpha h}^{n-1}$ are equal to $u_{\alpha h}^{n-1,k,i}$ for the last iterates k and i (when the stopping criteria are met).

For each unknown we note

$$u_{1h\tau}^{n,k,i} := u_{1h\tau}^{k,i}|_{I_n}, \quad u_{2h\tau}^{n,k,i} := u_{2h\tau}^{k,i}|_{I_n}.$$

Note that consequently

$$\partial_t u_{1h\tau}^{n,k,i}|_{I_n} = \frac{1}{\Delta t_n}\left(u_{1h}^{n,k,i} - u_{1h}^{n-1}\right), \quad \partial_t u_{2h\tau}^{n,k,i}|_{I_n} = \frac{1}{\Delta t_n}\left(u_{2h}^{n,k,i} - u_{2h}^{n-1}\right).$$

Our a posteriori analysis relies on the equilibrated flux reconstructions following the concept of Destuynder and Métivet [66], Braess and Schöberl [33], Ern and Vohralík [81], Dabaghi, Martin, and Vohralík [62] (see also Chapter 1 of this thesis). We will construct a discretization flux reconstruction $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i} \in \mathbf{H}(\mathrm{div},\Omega)$ and an algebraic error flux reconstruction $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{n,k,i} \in \mathbf{H}(\mathrm{div},\Omega)$. More precisely, the discretization flux reconstruction is obtained by solving mixed finite element systems on the patches $\omega_h^{\boldsymbol{a}}$ around the mesh vertices $\boldsymbol{a} \in \mathcal{V}_h$ on the mesh $\mathcal{T}_h$ while the algebraic flux $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{n,k,i}$ is obtained via solving local problems on a hierarchy of nested grids.

We consider two cases. First, we establish an a posteriori error estimate when $p = 1$ and when both the algebraic and linearization solvers have converged. Next, we derive an a posteriori error estimate when $p \geq 1$ at any semismooth linearization step $k \geq 1$ and any step of the iterative algebraic solver $i \geq 0$.

We first start by giving a functional representation of (2.57). We associate respectively with $\boldsymbol{R}_{1h}^{n,k,i}$ and $\boldsymbol{R}_{2h}^{n,k,i}$ elementwise discontinuous polynomials $r_{1h}^{n,k,i}$ and $r_{2h}^{n,k,i}$ of degree $p \geq 1$ that vanish on the boundary of Ω. These can be easily computed solving on each element $K \in \mathcal{T}_h$ a small problem with an element mass matrix given as follows. For $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$, denote by $N_{h,\boldsymbol{x}_l}$ the number of mesh elements forming the support of the basis function $\psi_{h,\boldsymbol{x}_l}$. Then, $\forall K \in \mathcal{T}_h$, $\forall \alpha \in \{1,2\}$, define $r_{\alpha h}^{n,k,i}|_K \in \mathbb{P}_p(K)$ by

$$(r_{\alpha h}^{n,k,i}, \psi_{h,\boldsymbol{x}_l})_K := \frac{(\boldsymbol{R}_{\alpha h}^{n,k,i})_l}{N_{h,\boldsymbol{x}_l}} \quad \text{and} \quad r_{\alpha h}^{k,i}|_{\partial K \cap \partial\Omega} := 0$$

for all basis functions $\psi_{h,\boldsymbol{x}_l}$, $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$ nonzero on K. It is easily seen that the first $2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ lines of (2.57) then read

$$\begin{aligned}
\mu_1 \left(\boldsymbol{\nabla} u_{1h}^{n,k,i}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{x}_l}\right)_\Omega &= \left(f_1|_{I_n} + \tilde{\lambda}_{h,l}^{n,k,i} - r_{1h}^{n,k,i} - \partial_t u_{1h\tau}^{n,k,i}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega \quad \forall l = 1,\dots,\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}},\\
\mu_2 \left(\boldsymbol{\nabla} u_{2h}^{n,k,i}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{x}_l}\right)_\Omega &= \left(f_2|_{I_n} - \tilde{\lambda}_{h,l}^{n,k,i} - r_{2h}^{n,k,i} - \partial_t u_{2h\tau}^{n,k,i}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega \quad \forall l = 1,\dots,\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}},
\end{aligned} \tag{2.63}$$

where

$$\tilde{\lambda}_{h,l}^{n,k,i} = \begin{cases} \lambda_h^{n,k,i}(\boldsymbol{x}_l) \text{ if } p=1 \text{ (here } \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} = \mathcal{V}_h^{\mathrm{int}}), \\ \lambda_h^{n,k,i} \text{ (function } \lambda_h^{n,k,i}, \text{ the index } l \text{ is discarded) if } p \geq 2. \end{cases} \tag{2.64}$$

As in Chapter 1, we also use the shorthand notation

$$\tilde{\lambda}_{h,\boldsymbol{a}}^{n,k,i} = \begin{cases} \lambda_h^{n,k,i}(\boldsymbol{a}) & \text{if} \quad p=1, \\ \lambda_h^{n,k,i} & \text{if} \quad p \geq 2. \end{cases} \tag{2.65}$$

The functional representation of (2.57) given by (2.63) is essential for our a posteriori analysis as we will see in the sequel. Our fluxes $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{n,k,i}, \boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i}$ are reconstructed in the Raviart–Thomas subspaces of $\mathbf{H}(\mathrm{div},\Omega)$. We recall that $\mathbf{H}(\mathrm{div},\Omega)$ stands for the space of $[L^2(\Omega)]^2$ functions having a weak divergence in $L^2(\Omega)$. The Raviart–Thomas spaces of order $p \geq 1$ [38, 50, 142, 146] are defined by

$$\mathbf{RT}_p(\Omega) := \{\boldsymbol{\tau}_h \in \mathbf{H}(\mathrm{div},\Omega),\ \boldsymbol{\tau}_h|_K \in \mathbf{RT}_p(K)\ \ \forall K \in \mathcal{T}_h\},$$

where $\mathbf{RT}_p(K) := [\mathbb{P}_p(K)]^2 + \vec{\boldsymbol{x}}\mathbb{P}_p(K)$, with $\vec{\boldsymbol{x}} = [x_1, x_2]^T$. For $\boldsymbol{a} \in \mathcal{V}_h$, let

$$\mathbf{RT}_p(\omega_h^{\boldsymbol{a}}) := \{\boldsymbol{\tau}_h \in \mathbf{H}(\mathrm{div},\omega_h^{\boldsymbol{a}}),\ \boldsymbol{\tau}_h|_K \in \mathbf{RT}_p(K),\ \forall K \in \mathcal{T}_h \text{ such that } K \subset \omega_h^{\boldsymbol{a}}\},$$

and let $\mathbb{P}_p^{\mathrm{d}}(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}})$ stand for piecewise discontinuous polynomials of order p in the patch $\omega_h^{\boldsymbol{a}}$. Define consequently the spaces $\mathbf{V}_h^{\boldsymbol{a}}$ and $Q_h^{\boldsymbol{a}}$, when $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$ by

$$\begin{aligned} \mathbf{V}_h^{\boldsymbol{a}} &:= \left\{\boldsymbol{\tau}_h \in \mathbf{RT}_p(\omega_h^{\boldsymbol{a}}),\ \boldsymbol{\tau}_h \cdot \mathbf{n}_{\omega_h^{\boldsymbol{a}}} = 0 \text{ on } \partial\omega_h^{\boldsymbol{a}}\right\}, \\ Q_h^{\boldsymbol{a}} &:= \left\{q_h \in \mathbb{P}_p^{\mathrm{d}}(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}}),\ (q_h,1)_{\omega_h^{\boldsymbol{a}}} = 0\right\}, \end{aligned} \tag{2.66}$$

and when $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{ext}}$ by

$$\begin{aligned} \mathbf{V}_h^{\boldsymbol{a}} &:= \left\{\boldsymbol{\tau}_h \in \mathbf{RT}_p(\omega_h^{\boldsymbol{a}}),\ \boldsymbol{\tau}_h \cdot \mathbf{n}_{\omega_h^{\boldsymbol{a}}} = 0 \ \text{ on } \ \partial\omega_h^{\boldsymbol{a}} \backslash \partial\Omega\right\}, \\ Q_h^{\boldsymbol{a}} &:= \mathbb{P}_p(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}}). \end{aligned} \tag{2.67}$$

2.6.2 Discretization flux reconstructions

For all time steps $1 \leq n \leq N_{\mathrm{t}}$, and for all $\alpha \in \{1,2\}$, let $\left(u_{1h}^{n,k,i}, u_{2h}^{n,k,i}, \lambda_h^{n,k,i}\right)$ be the approximate solution given by (2.57), verifying in particular (2.63). For each vertex $\boldsymbol{a} \in \mathcal{V}_h$, define $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i,\boldsymbol{a}} \in \mathbf{V}_h^{\boldsymbol{a}}$ and $\gamma_{\alpha h}^{n,k,i,\boldsymbol{a}} \in Q_h^{\boldsymbol{a}}$, by solving:

$$\begin{aligned} \left(\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i,\boldsymbol{a}}, \boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} - \left(\gamma_{\alpha h}^{n,k,i,\boldsymbol{a}}, \boldsymbol{\nabla}\cdot\boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} &= -\left(\mu_\alpha \psi_{h,\boldsymbol{a}} \boldsymbol{\nabla} u_{\alpha h}^{n,k,i}, \boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} \quad \forall \boldsymbol{\tau}_h \in \mathbf{V}_h^{\boldsymbol{a}}, \\ \left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} &= \left(\tilde{g}_{\alpha h}^{n,k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} \quad \forall q_h \in Q_h^{\boldsymbol{a}}, \end{aligned} \tag{2.68}$$

where the spaces $\mathbf{V}_h^{\boldsymbol{a}}$ and $Q_h^{\boldsymbol{a}}$ are defined by (2.66)-(2.67). The right-hand sides are defined by

$$\tilde{g}_{\alpha h}^{n,k,i,\boldsymbol{a}} := \left(f_\alpha|_{I_n} - (-1)^\alpha \tilde{\lambda}_{h,\boldsymbol{a}}^{n,k,i} - r_{\alpha h}^{n,k,i} - \partial_t u_{\alpha h\tau}^{n,k,i}|_{\omega_h^{\boldsymbol{a}}}\right)\psi_{h,\boldsymbol{a}} - \mu_\alpha \boldsymbol{\nabla} u_{\alpha h}^{n,k,i} \cdot \boldsymbol{\nabla}\psi_{h,\boldsymbol{a}}. \tag{2.69}$$

Note that

$$\left(\tilde{g}_{\alpha h}^{n,k,i,\boldsymbol{a}}, 1\right)_{\omega_h^{\boldsymbol{a}}} = 0. \tag{2.70}$$

This implies the Neumann compatibility condition for (2.68). Indeed, this follows immediately from (2.63) for the hat test functions $\psi_{h,\boldsymbol{a}}$ (see also Remark 1.4.5). At each time step $1 \leq n \leq N_{\mathrm{t}}$ the discretization flux reconstruction is defined by

$$\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i} = \sum_{\boldsymbol{a}\in\mathcal{V}_h} \boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i,\boldsymbol{a}}. \tag{2.71}$$

Proposition 2.6.1. *The flux reconstruction* $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i} \in \mathbf{H}(\mathrm{div},\Omega)$ *and satisfies the equilibration property* $\forall K \in \mathcal{T}_h$

$$\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i}, q_h\right)_K = \left(f_\alpha|_{I_n} - (-1)^\alpha \lambda_h^{n,k,i} - r_{\alpha h}^{n,k,i} - \partial_t u_{\alpha h\tau}^{n,k,i}, q_h\right)_K \quad \forall q_h \in \mathbb{P}_p(K). \tag{2.72}$$

Proof. The proof is a direct extension of Chapter 1, Proposition 1.4.6, with merely replacing f_α by $f_\alpha|_{I_n} - \partial_t u_{\alpha h\tau}^{n,k,i}$.

□

Remark 2.6.2. *Equation (2.72) reads*

$$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i} = \boldsymbol{\Pi}_{\mathbb{P}_p}(f_\alpha|_{I_n}) - (-1)^\alpha \lambda_h^{n,k,i} - r_{\alpha h}^{n,k,i} - \partial_t u_{\alpha h\tau}^{n,k,i}.$$

2.6.3 Algebraic error flux reconstructions

The algebraic error flux reconstruction is defined as in Chapter 1 Section 1.4.2 giving:

Proposition 2.6.3. *For a time step* $1 \leq n \leq N_{\mathrm{t}}$*, a semismooth iteration* $k \geq 1$ *and an algebraic solver iteration* $i \geq 0$*, the algebraic flux* $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{n,k,i}$ *satisfies*

$$\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{n,k,i} \in \mathbf{H}(\mathrm{div},\Omega) \quad \textit{and} \quad \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{n,k,i} = r_{\alpha h}^{n,k,i}. \tag{2.73}$$

2.6.4 Total flux reconstructions

We define the total flux reconstructions as the sum of all the contributions of the component fluxes:

$$\boldsymbol{\sigma}_{\alpha h}^{n,k,i} := \boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i} + \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{n,k,i} \quad \alpha = 1,2. \tag{2.74}$$

In particular, the total flux reconstructions satisfy the following property:

Proposition 2.6.4. *For any time step* $1 \leq n \leq N_{\mathrm{t}}$*, any semismooth Newton iteration* $k \geq 1$*, and any iterative linear algebraic iteration* $i \geq 0$*, we have*

$$\begin{aligned}
&\boldsymbol{\sigma}_{\alpha h}^{n,k,i} \in \mathbf{H}(\mathrm{div},\Omega),\\
&\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{n,k,i}, q_h\right)_K = \left(f_\alpha|_{I_n} - (-1)^\alpha \lambda_h^{n,k,i} - \partial_t u_{\alpha h\tau}^{n,k,i}, q_h\right)_K \quad \forall q_h \in \mathbb{P}_p(K)\ \forall K \in \mathcal{T}_h.
\end{aligned}$$

To conclude this section we define the space-time function (piecewise constant in time) for each flux reconstruction as follows:

$$\begin{aligned}
&\left(\boldsymbol{\sigma}_{\alpha h\tau}^{k,i}, \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{disc}}^{k,i}, \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{alg}}^{k,i}\right) \in \left[L^2(0,T;\mathbf{H}(\mathrm{div},\Omega))\right]^3,\\
&\boldsymbol{\sigma}_{\alpha h\tau}^{k,i}|_{I_n} = \boldsymbol{\sigma}_{\alpha h}^{n,k,i},\ \ \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{disc}}^{k,i}|_{I_n} = \boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{n,k,i},\ \ \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{alg}}^{k,i}|_{I_n} = \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{n,k,i},\ \ \forall 1 \leq n \leq N_{\mathrm{t}}.
\end{aligned} \tag{2.75}$$

2.6.5 An a posteriori error estimate for $p = 1$ and exact solvers

In this section we establish an a posteriori error estimate between the exact solution $\boldsymbol{u} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ and the approximate numerical solution ($p = 1$) when the semismooth Newton solver and the iterative algebraic solver have converged. In this case, we discard the indices k and i and the approximate solution $\boldsymbol{u}_{h\tau}$ is conforming in the sense $\boldsymbol{u}_{h\tau} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$.

Definition 2.6.5. *Let $1 \leq n \leq N_{\mathrm{t}}$, $K \in \mathcal{T}_h$, and $\alpha = 1, 2$. We define the residual estimator $\eta^n_{\mathrm{R},K,\alpha}$, the flux estimator $\eta^n_{\mathrm{F},K,\alpha}$, and the constraints estimator $\eta^n_{\mathrm{C},K}$ by the temporal functions for all $t \in I_n$*

$$\eta^n_{\mathrm{R},K,\alpha}(t) := \frac{h_K}{\pi}\mu_\alpha^{-\frac{1}{2}} \left\| f_\alpha|_{I_n} - \partial_t u^n_{\alpha h\tau} - (-1)^\alpha \lambda^n_h - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}^n_{\alpha h\tau}\right\|_K, \tag{2.76}$$

$$\eta^n_{\mathrm{F},K,\alpha}(t) := \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} u^n_{\alpha h\tau} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}^n_{\alpha h\tau}\right\|_K, \tag{2.77}$$

$$\eta^n_{\mathrm{C},K}(t) := 2\left(\lambda^n_{h\tau}, u^n_{1h\tau} - u^n_{2h\tau}\right)_K. \tag{2.78}$$

The estimators (2.76)–(2.78) reflect various violations of physical properties of the approximate solution $(u^n_{1h\tau}, u^n_{2h\tau}, \lambda^n_{h\tau})$: $\eta^n_{\mathrm{R},K,\alpha}$ and $\eta^n_{\mathrm{F},K,\alpha}$ represent the nonconformity of the flux, *i.e.*, the fact that $-\mu_\alpha \boldsymbol{\nabla} u^n_{\alpha h\tau} \notin L^2(0,T;\mathbf{H}(\mathrm{div},\Omega))$; $\eta^n_{\mathrm{C},K}$ reflects inconsistencies in the complementarity conditions at the discrete level, *i.e.*, the fact that $(u^n_{1h\tau} - u^n_{2h\tau})\lambda^n_{h\tau} \neq 0$.

We now present our energy error estimate.

Theorem 2.6.6. *Let $\boldsymbol{u} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ be the exact solution satisfying (2.18). Let $\boldsymbol{u}_{h\tau} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ and $\lambda_{h\tau} \in \Psi$ be the approximate solution when $p = 1$ for exact solvers. Consider the equilibrated flux reconstructions $\boldsymbol{\sigma}_{\alpha h\tau} \in L^2(0,T;\mathbf{H}(\mathrm{div},\Omega))$ given by (2.74) and (2.75). For the error estimators defined by (2.76)–(2.78), there holds*

$$\begin{aligned}
&\left|\!\left|\!\left| \boldsymbol{u} - \boldsymbol{u}_{h\tau}\right|\!\right|\!\right|^2_{V_0} + \sum_{\alpha=1}^{2} \left\| \left(u_\alpha - u_{\alpha h\tau}\right)(\cdot,T)\right\|^2_\Omega \leq \eta^2 := \\
&\sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \left(\sum_{\alpha=1}^{2} \left(\eta^n_{\mathrm{R},K,\alpha} + \eta^n_{\mathrm{F},K,\alpha}\right)^2 + \eta^n_{\mathrm{C},K} \right)(t)\,\mathrm{dt} + \sum_{\alpha=1}^{2} \left\| \left(u_\alpha - u_{\alpha h\tau}\right)(\cdot,0)\right\|^2_\Omega .
\end{aligned} \tag{2.79}$$

To prove Theorem 2.6.6, we first introduce the following lemma.

Lemma 2.6.7. *Let a be the bilinear form and l be the continuous form defined in (2.16). Let $\boldsymbol{u} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ be the solution of the reduced problem (2.18) and let $\boldsymbol{y} := (y_1, y_2) \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ be arbitrary. Then, the vector $\boldsymbol{y}^* := (y_1^*, y_2^*) := (u_1 - y_1, u_2 - y_2) \in \left[L^2(0,T;H^1_0(\Omega))\right]^2$ satisfies*

$$\begin{aligned}
A &:= \int_0^T \left(l(\boldsymbol{y}^*) - \langle \partial_t \boldsymbol{u}_{h\tau}, \boldsymbol{y}^*\rangle - a(\boldsymbol{u}_{h\tau}, \boldsymbol{y}^*)\right)(t)\,\mathrm{dt} - \int_0^T \sum_{\alpha=1}^{2} \left((-1)^\alpha \lambda_{h\tau}, y^*_\alpha\right)_\Omega (t)\,\mathrm{dt} \\
&\leq \left\{ \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta^n_{\mathrm{R},K,\alpha} + \eta^n_{\mathrm{F},K,\alpha}\right)^2 (t)\,\mathrm{dt} \right\}^{\frac{1}{2}} \left|\!\left|\!\left| \boldsymbol{y}^*\right|\!\right|\!\right|_{V_0}.
\end{aligned} \tag{2.80}$$

Proof. Adding and subtracting $\boldsymbol{\sigma}_{\alpha h\tau}(t) \in \mathbf{H}(\mathrm{div},\Omega)$ and using the Green formula with $y_\alpha^*(t) \in H_0^1(\Omega)$, $\alpha = 1,2$, we have

$$A = \int_0^T \sum_{\alpha=1}^{2} \left(f_\alpha - \partial_t u_{\alpha h\tau} - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h\tau} - (-1)^\alpha \lambda_{h\tau}, y_\alpha^*\right)_\Omega (t)\,\mathrm{dt}$$
$$- \int_0^T \sum_{\alpha=1}^{2} \left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} u_{\alpha h\tau} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau}, \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} y_\alpha^*\right)_\Omega (t)\,\mathrm{dt}.$$

Denoting by $\overline{w}_K$ the mean value over K of $w \in L^2(\Omega)$ and using Proposition 2.6.4, one has, for all $t \in I_n$, $1 \le n \le N_\mathrm{t}$

$$\begin{aligned}&\left(f_\alpha|_{I_n} - \partial_t u_{\alpha h\tau}^n - (-1)^\alpha \lambda_{h\tau}^n - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h\tau}^n, y_\alpha^*\right)_K (t)\\ =&\left(\mu_\alpha^{-\frac{1}{2}} \left(f_\alpha|_{I_n} - \partial_t u_{\alpha h\tau}^n - (-1)^\alpha \lambda_{h\tau}^n - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h\tau}^n\right), \mu_\alpha^{\frac{1}{2}} \left(y_\alpha^* - (\overline{y_\alpha^*})_K\right)\right)_K (t).\end{aligned}$$

Using the Cauchy–Schwarz inequality and next the Poincaré–Wirtinger inequality (2.11b) with $C_\mathrm{PW} = \frac{1}{\pi}$ for convex set (see Section 1.2), we get, at any time $t \in I_n$ and for all $K \in \mathcal{T}_h$,

$$\begin{aligned}&\left(f_\alpha|_{I_n} - \partial_t u_{\alpha h\tau}^n - (-1)^\alpha \lambda_{h\tau}^n - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h\tau}^n, y_\alpha^*\right)_K (t)\\ \le& h_K C_\mathrm{PW} \mu_\alpha^{-\frac{1}{2}} \left\|f_\alpha - \partial_t u_{\alpha h\tau}^n - (-1)^\alpha \lambda_{h\tau}^n - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h\tau}^n\right\|_K \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} y_\alpha^*\right\|_K (t),\\ =& \eta_{\mathrm{R},K,\alpha}^n \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} y_\alpha^*\right\|_K (t).\end{aligned} \tag{2.81}$$

Next, as a result of the Cauchy–Schwarz inequality, we have, for $t \in I_n$ and $K \in \mathcal{T}_h$,

$$\begin{aligned}&\left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} u_{\alpha h\tau}^n + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau}^n, \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} y_\alpha^*\right)_K (t),\\ \le& \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} u_{\alpha h\tau}^n + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau}^n\right\|_K \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} y_\alpha^*\right\|_K (t),\\ =& \eta_{\mathrm{F},K,\alpha}^n \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} y_\alpha^*\right\|_K (t).\end{aligned} \tag{2.82}$$

Therefore, combining (2.81)–(2.82) and applying next the Cauchy–Schwarz inequality we get

$$A \le \left\{\sum_{n=1}^{N_\mathrm{t}} \int_{I_n} \sum_{\alpha=1}^{2} \sum_{K\in\mathcal{T}_h} \left(\eta_{\mathrm{R},K,\alpha}^n + \eta_{\mathrm{F},K,\alpha}^n\right)^2 (t)\,\mathrm{dt}\right\}^{\frac{1}{2}} |||\boldsymbol{y}^*|||_{V_0}$$

which concludes the proof. □

Proof of Theorem 2.6.6. Observe that [83, Theorem 5.9.3] gives

$$\frac{1}{2}\sum_{\alpha=1}^{2} \left\|\left(u_\alpha - u_{\alpha h\tau}\right)(\cdot,T)\right\|_\Omega^2 = \frac{1}{2}\sum_{\alpha=1}^{2} \left\|\left(u_\alpha - u_{\alpha h\tau}\right)(\cdot,0)\right\|_2^2$$
$$+ \int_0^T \sum_{\alpha=1}^{2} \langle \partial_t (u_\alpha - u_{\alpha h\tau}), u_\alpha - u_{\alpha h\tau}\rangle (t)\,\mathrm{dt}.$$

Then posing $B := |||\boldsymbol{u} - \boldsymbol{u}_{h\tau}|||_{V_0}^2 + \frac{1}{2}\sum_{\alpha=1}^{2} \|(u_\alpha - u_{\alpha h\tau})(\cdot, T)\|_\Omega^2$, we get

$$B = |||\boldsymbol{u} - \boldsymbol{u}_{h\tau}|||_{V_0}^2 + \int_0^T \sum_{\alpha=1}^{2} \langle \partial_t(u_\alpha - u_{\alpha h\tau}), u_\alpha - u_{\alpha h\tau}\rangle(t)\,\mathrm{dt} + \frac{1}{2}\sum_{\alpha=1}^{2} \|(u_\alpha - u_{\alpha h\tau})(\cdot, 0)\|_\Omega^2 .$$

Thus, using the definition (2.10), we get

$$\begin{aligned} B = &\int_0^T \sum_{\alpha=1}^{2} \left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}(u_\alpha - u_{\alpha h\tau}), \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}(u_\alpha - u_{\alpha h\tau})\right)_\Omega (t)\mathrm{dt} \\ &+ \int_0^T \sum_{\alpha=1}^{2} \langle \partial_t(u_\alpha - u_{\alpha h\tau}), u_\alpha - u_{\alpha h\tau}\rangle(t)\,\mathrm{dt} + \frac{1}{2}\sum_{\alpha=1}^{2} \|(u_\alpha - u_{\alpha h\tau})(\cdot, 0)\|_\Omega^2 . \end{aligned}$$

Then, using the weak formulation (2.18) with $\boldsymbol{v} = \boldsymbol{u}_{h\tau} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$, we get

$$\begin{aligned} B \leq &\int_0^T \sum_{\alpha=1}^{2} (f_\alpha - \partial_t u_{\alpha h\tau}, u_\alpha - u_{\alpha h\tau})_\Omega (t)\,\mathrm{dt} \\ &- \int_0^T \sum_{\alpha=1}^{2} \mu_\alpha (\boldsymbol{\nabla} u_{\alpha h\tau}, \boldsymbol{\nabla}(u_\alpha - u_{\alpha h\tau}))_\Omega (t)\,\mathrm{dt} + \frac{1}{2}\sum_{\alpha=1}^{2} \|(u_\alpha - u_{\alpha h\tau})(\cdot, 0)\|_\Omega^2 . \end{aligned}$$

Adding and subtracting $\int_0^T \sum_{\alpha=1}^{2} ((-1)^\alpha \lambda_{h\tau}, u_\alpha - u_{\alpha h\tau})_\Omega (t)\,\mathrm{dt}$ and noting that $\int_0^T (-\lambda_{h\tau}, u_1 - u_2)_\Omega (t)\,\mathrm{dt} \leq 0$ because $\lambda_{h\tau} \in \Psi$ we obtain

$$\begin{aligned} B \leq &\int_0^T \sum_{\alpha=1}^{2} (f_\alpha - \partial_t u_{\alpha h\tau} - (-1)^\alpha \lambda_{h\tau}, u_\alpha - u_{\alpha h\tau})_\Omega (t)\,\mathrm{dt} \\ &- \int_0^T \sum_{\alpha=1}^{2} \mu_\alpha (\boldsymbol{\nabla} u_{\alpha h\tau}, \boldsymbol{\nabla}(u_\alpha - u_{\alpha h\tau}))_\Omega (t)\,\mathrm{dt} + \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{\eta_{\mathrm{C},K}^n}{2}(t)\,\mathrm{dt} \\ &+ \frac{1}{2}\sum_{\alpha=1}^{2} \|(u_\alpha - u_{\alpha h\tau})(\cdot, 0)\|_\Omega^2 . \end{aligned}$$

Employing Lemma 2.6.7 with $\boldsymbol{y} = \boldsymbol{u}_{h\tau} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ and using the Young inequality $A_1 A_2 \leq \frac{1}{2}(A_1^2 + A_2^2)$, $\forall A_1, A_2 \geq 0$, we get the desired result. □

An attempt to control the temporal derivative

So far, we have established an a posteriori error estimate between the exact solution $\boldsymbol{u} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ and its approximate solution $\boldsymbol{u}_{h\tau} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ in the energy norm on the space V_0. As we mentionned in the introduction, we cannot easily estimate the norm

$\|\partial_t(\boldsymbol{u}-\boldsymbol{u}_{h\tau})\|_{[V_0^*]^2}$. We now give our replacement result. Given $\boldsymbol{u}\in\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ and for the approximate solution $\boldsymbol{u}_{h\tau}\in\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$, let $\boldsymbol{z}\in\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ be such that, for all $\boldsymbol{v}\in\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$,

$$\int_0^T a(\boldsymbol{z}-\boldsymbol{u},\boldsymbol{v}-\boldsymbol{z})(t)\,\mathrm{dt}\geq-\int_0^T\sum_{\alpha=1}^{2}\langle\partial_t(u_\alpha-u_{\alpha h\tau})-(-1)^\alpha\lambda_{h\tau},v_\alpha-z_\alpha\rangle(t)\,\mathrm{dt}. \tag{2.83}$$

As a result of the Lions–Stampacchia theorem, problem (2.83) is well posed.

Now, we give an a posteriori error estimate on the error $|||\boldsymbol{u}-\boldsymbol{z}|||_{V_0}$.

Theorem 2.6.8. *Let $\boldsymbol{u}\in\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ be the solution of the weak formulation given by (2.18) and let $\boldsymbol{z}\in\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ be the solution of (2.83). Assume that the hypotheses of Theorem 2.6.6 hold. Let the total estimator η be defined by (2.79). Then,*

$$|||\boldsymbol{u}-\boldsymbol{z}|||_{V_0}\leq 2\eta. \tag{2.84}$$

Proof. Setting $\boldsymbol{w}^*:=\boldsymbol{u}-\boldsymbol{z}$, we have

$$|||\boldsymbol{w}^*|||_{V_0}^2=\int_0^T\sum_{\alpha=1}^{2}\mu_\alpha\,\|\boldsymbol{\nabla}(u_\alpha-z_\alpha)\|_\Omega^2(t)\,\mathrm{dt}.$$

But, for $\boldsymbol{v}=\boldsymbol{u}\in\boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$, we get from (2.83)

$$\begin{aligned}\int_0^T\sum_{\alpha=1}^{2}\mu_\alpha\left(\boldsymbol{\nabla}(z_\alpha-u_\alpha),\boldsymbol{\nabla}w_\alpha^*\right)_\Omega(t)\,\mathrm{dt}\geq&-\int_0^T\sum_{\alpha=1}^{2}\langle\partial_t(u_\alpha-u_{\alpha h\tau}),w_\alpha^*\rangle(t)\,\mathrm{dt}\\&-\int_0^T\sum_{\alpha=1}^{2}\left(-(-1)^\alpha\lambda_{h\tau},w_\alpha^*\right)_\Omega(t)\,\mathrm{dt}.\end{aligned}$$

Therefore,

$$|||\boldsymbol{w}^*|||_{V_0}^2\leq\int_0^T\sum_{\alpha=1}^{2}\langle\partial_t(u_\alpha-u_{\alpha h\tau})-(-1)^\alpha\lambda_{h\tau},w_\alpha^*\rangle(t)\,\mathrm{dt} \tag{2.85}$$

and also

$$\begin{aligned}|||\boldsymbol{w}^*|||_{V_0}^2\leq&\int_0^T\sum_{\alpha=1}^{2}\langle\partial_t(u_\alpha-u_{\alpha h\tau})-(-1)^\alpha\lambda_{h\tau},w_\alpha^*\rangle(t)\,\mathrm{dt}\\&+\int_0^T\sum_{\alpha=1}^{2}\mu_\alpha\left(\boldsymbol{\nabla}(u_\alpha-u_{\alpha h\tau}),\boldsymbol{\nabla}w_\alpha^*\right)_\Omega(t)\,\mathrm{dt}\\&-\int_0^T\sum_{\alpha=1}^{2}\mu_\alpha\left(\boldsymbol{\nabla}(u_\alpha-u_{\alpha h\tau}),\boldsymbol{\nabla}w_\alpha^*\right)_\Omega(t)\,\mathrm{dt}.\end{aligned}$$

Employing the weak formulation (2.18) with $\boldsymbol{v} = \boldsymbol{z} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ we get

$$\begin{aligned}
|||\boldsymbol{w}^*|||_{V_0}^2 \leq & \int_0^T \sum_{\alpha=1}^{2} \left(f_\alpha - \partial_t u_{\alpha h\tau} - (-1)^\alpha \lambda_{h\tau}, w_\alpha^*\right)_\Omega (t)\,\mathrm{d}t \\
& - \int_0^T \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_{\alpha h\tau}, \boldsymbol{\nabla} w_\alpha^*\right)_\Omega (t)\,\mathrm{d}t \\
& - \int_0^T \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h\tau}\right), \boldsymbol{\nabla} w_\alpha^*\right)_\Omega (t)\,\mathrm{d}t.
\end{aligned} \tag{2.86}$$

To bound the two first terms of (2.86), we employ Lemma 2.6.7 with $\boldsymbol{y} = \boldsymbol{z} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ to get

$$\begin{aligned}
|||\boldsymbol{w}^*|||_{V_0}^2 \leq & \left\{ \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{R},K,\alpha}^n + \eta_{\mathrm{F},K,\alpha}^n\right)^2 (t)\,\mathrm{d}t \right\}^{\frac{1}{2}} |||\boldsymbol{w}^*|||_{V_0} \\
& - \int_0^T \sum_{\alpha=1}^{2} \left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h\tau}\right), \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} w_\alpha^*\right)_\Omega (t)\,\mathrm{d}t.
\end{aligned}$$

The Cauchy–Schwarz inequality and the Young inequality give

$$\begin{aligned}
-\int_0^T \sum_{\alpha=1}^{2} \left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h\tau}\right), \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} w_\alpha^*\right)_\Omega (t)\,\mathrm{d}t &\leq |||\boldsymbol{u} - \boldsymbol{u}_{h\tau}|||_{V_0} |||\boldsymbol{w}^*|||_{V_0} \\
&\leq |||\boldsymbol{u} - \boldsymbol{u}_{h\tau}|||_{V_0}^2 + \frac{1}{4} |||\boldsymbol{w}^*|||_{V_0}^2 .
\end{aligned} \tag{2.87}$$

Employing again the Young inequality, we have

$$\begin{aligned}
\left\{ \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{R},K,\alpha}^n + \eta_{\mathrm{F},K,\alpha}^n\right)^2 \right\}^{\frac{1}{2}} |||\boldsymbol{w}^*|||_{V_0} \leq & \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{R},K,\alpha}^n + \eta_{\mathrm{F},K,\alpha}^n\right)^2 \\
& + \frac{1}{4} |||\boldsymbol{w}^*|||_{V_0}^2 .
\end{aligned} \tag{2.88}$$

Finally, combining (2.87)–(2.88) with (2.79) we get

$$|||\boldsymbol{w}^*|||_{V_0}^2 \leq 4\eta^2$$

and thus the desired result. □

We can bound the energy norm of $\boldsymbol{u} - \boldsymbol{z}$ by the H^{-1} norms of the errors in $\lambda_{h\tau}^{k,i}$ and in $\partial_t u_{\alpha h\tau}^{k,i}$.

Lemma 2.6.9. *Assuming the hypotheses of Theorem 2.6.8, we have*

$$\begin{aligned}
|||\boldsymbol{u} - \boldsymbol{z}|||_{V_0} \leq & \left(\int_0^T \sum_{\alpha=1}^{2} \left\| \mu_\alpha^{-\frac{1}{2}} \partial_t \left(u_\alpha - u_{\alpha h\tau}\right) \right\|_{H^{-1}(\Omega)}^2 (t)\,\mathrm{d}t \right)^{\frac{1}{2}} \\
& + \left(\mu_1^{-\frac{1}{2}} + \mu_2^{-\frac{1}{2}}\right) \left(\int_0^T \|\lambda_{h\tau} - \lambda\|_{H^{-1}(\Omega)}^2 (t)\,\mathrm{d}t \right)^{\frac{1}{2}} .
\end{aligned} \tag{2.89}$$

Proof. From (2.85), denoting $\boldsymbol{w}^* := \boldsymbol{u} - \boldsymbol{z}$, we have

$$|||\boldsymbol{w}^*|||_{V_0}^2 \leq \int_0^T \sum_{\alpha=1}^{2} \langle \partial_t \left(u_\alpha - u_{\alpha h\tau}\right), w_\alpha^* \rangle(t)\,\mathrm{dt} + \int_0^T \left(\lambda_{h\tau}, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt}. \quad (2.90)$$

Next,

$$\int_0^T \left(\lambda_{h\tau}, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt} = \int_0^T \left(\lambda_{h\tau} - \lambda, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt} + \int_0^T \left(\lambda, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt}.$$

Observe that

$$\int_0^T \left(\lambda, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt} = \int_0^T \left(\lambda, u_1 - u_2\right)_\Omega (t)\,\mathrm{dt} - \int_0^T \left(\lambda, z_1 - z_2\right)_\Omega (t)\,\mathrm{dt}.$$

From (2.8) $\lambda(u_1 - u_2) = 0$, and as $\lambda \in \Psi$ and $\boldsymbol{z} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$, we have $\int_0^T \left(\lambda_{h\tau}, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt} \leq 0$ and thus

$$\int_0^T \left(\lambda_{h\tau}, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt} \leq \int_0^T \left(\lambda_{h\tau} - \lambda, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt}.$$

Finally,

$$|||\boldsymbol{w}^*|||_{V_0}^2 \leq \int_0^T \sum_{\alpha=1}^{2} \langle \partial_t \left(u_\alpha - u_{\alpha h\tau}\right), w_\alpha^* \rangle(t)\,\mathrm{dt} + \int_0^T \left(\lambda_{h\tau} - \lambda, w_1^* - w_2^*\right)_\Omega (t)\,\mathrm{dt}. \quad (2.91)$$

Furthermore, denoting by A_1 the first term in the right-hand side of (2.91) we have,

$$\begin{aligned}
A_1 &= \int_0^T \sum_{\alpha=1}^{2} \left\langle \mu_\alpha^{-\frac{1}{2}} \partial_t \left(u_\alpha - u_{\alpha h\tau}\right), \mu_\alpha^{\frac{1}{2}} w_\alpha^* \right\rangle (t)\,\mathrm{dt}, \\
&= \int_0^T \sum_{\alpha=1}^{2} \frac{\left\langle \mu_\alpha^{-\frac{1}{2}} \partial_t \left(u_\alpha - u_{\alpha h\tau}\right), \mu_\alpha^{\frac{1}{2}} w_\alpha^* \right\rangle}{\left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} w_\alpha^* \right\|_\Omega} \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} w_\alpha^* \right\|_\Omega (t)\,\mathrm{dt}, \\
&\leq \int_0^T \sum_{\alpha=1}^{2} \sup_{\Phi_\alpha \in H_0^1(\Omega)} \frac{\left\langle \mu_\alpha^{-\frac{1}{2}} \partial_t \left(u_\alpha - u_{\alpha h\tau}\right), \mu_\alpha^{\frac{1}{2}} \Phi_\alpha \right\rangle}{\left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} \Phi_\alpha \right\|_\Omega} \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} w_\alpha^* \right\|_\Omega (t)\,\mathrm{dt}, \\
&= \int_0^T \sum_{\alpha=1}^{2} \left\| \mu_\alpha^{-\frac{1}{2}} \partial_t \left(u_\alpha - u_{\alpha h\tau}\right) \right\|_{H^{-1}(\Omega)} \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} w_\alpha^* \right\|_\Omega (t)\,\mathrm{dt}.
\end{aligned}$$

The Cauchy–Schwarz inequality gives

$$A_1 \leq \left(\int_0^T \sum_{\alpha=1}^{2} \left\| \mu_\alpha^{-\frac{1}{2}} \partial_t \left(u_\alpha - u_{\alpha h\tau}\right) \right\|_{H^{-1}(\Omega)}^2 (t)\,\mathrm{dt} \right)^{\frac{1}{2}} |||\boldsymbol{w}^*|||_{V_0}. \quad (2.92)$$

To bound the second term A_2 of (2.91) we proceed as follows

$$A_2 = \int_0^T \left(\mu_1^{-\frac{1}{2}} \left(\lambda_{h\tau} - \lambda\right), \mu_1^{\frac{1}{2}} w_1^* \right)_\Omega (t)\,\mathrm{dt} - \int_0^T \left(\mu_2^{-\frac{1}{2}} \left(\lambda_{h\tau} - \lambda\right), \mu_2^{\frac{1}{2}} w_2^* \right)_\Omega (t)\,\mathrm{dt}.$$

Next, we have as result of the Cauchy–Schwarz inequality

$$\begin{aligned}
A_2 &\le \int_0^T \sum_{\alpha=1}^{2} \left\| \mu_\alpha^{-\frac{1}{2}} (\lambda_{h\tau} - \lambda) \right\|_{H^{-1}(\Omega)} \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} w_\alpha^* \right\|_\Omega (t) \, \mathrm{dt} \\
&\le \left(\int_0^T \sum_{\alpha=1}^{2} \left\| \mu_\alpha^{-\frac{1}{2}} (\lambda_{h\tau} - \lambda) \right\|_{H^{-1}(\Omega)}^2 (t) \, \mathrm{dt} \right)^{\frac{1}{2}} \left(\int_0^T \sum_{\alpha=1}^{2} \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} w_\alpha^* \right\|_\Omega^2 (t) \, \mathrm{dt} \right)^{\frac{1}{2}} \\
&\le \left(\mu_1^{-\frac{1}{2}} + \mu_2^{-\frac{1}{2}} \right) \left(\int_0^T \| \lambda_{h\tau} - \lambda \|_{H^{-1}(\Omega)}^2 (t) \, \mathrm{dt} \right)^{\frac{1}{2}} |||\boldsymbol{w}^*|||_{V_0} .
\end{aligned} \tag{2.93}$$

Combining (2.91), (2.92), and (2.93) we obtain the desired result. □

2.6.6 An a posteriori error estimate for $p \ge 1$ and each step $k \ge 1$, $i \ge 0$

In this section we devise an a posteriori error estimate which is valid at any time step $1 \le n \le N_\mathrm{t}$, at any semismooth Newton step $k \ge 1$, and at any algebraic step $i \ge 0$. Several difficulties arise. First, when $p = 1$, the complementarity constraints given by the second line of (2.45) expressed in the internal vertices of the mesh are not valid in general. At convergence, the constraints $u_{1h}^n - u_{2h}^n \ge 0$ and $\lambda_h^n \ge 0$ are satisfied but $(u_{1h}^n - u_{2h}^n) \cdot \lambda_h^n \ne 0$ in general. When $p \ge 2$, the complementarity constraints expressed in the Lagrange basis (see the second line of (2.48)) or in its dual basis (see the second line of (2.51)) can be violated. Even at convergence, we have $u_{1h}^n - u_{2h}^n \not\ge 0$, $\lambda_h^n \not\ge 0$, and $(u_{1h}^n - u_{2h}^n) \cdot \lambda_h^n \ne 0$. Consequently, we have to work with a nonconforming space-time solution $\boldsymbol{u}_{h\tau}^{k,i} \notin \boldsymbol{\mathcal{K}}_g^\mathrm{t}$. Following the concept of Chapter 1, Section 1.5, we employ the decomposition

$$\lambda_h^{n,k,i} = \lambda_h^{n,k,i,\mathrm{pos}} + \lambda_h^{n,k,i,\mathrm{neg}}$$

where

$$\lambda_h^{n,k,i,\mathrm{pos}} = \max\left\{ \lambda_h^{n,k,i}, 0 \right\}, \quad \text{and} \quad \lambda_h^{n,k,i,\mathrm{neg}} = \min\left\{ \lambda_h^{n,k,i}, 0 \right\}.$$

We introduce the potential $\boldsymbol{s}_{h\tau}^{k,i} := \left(s_{1h\tau}^{k,i}, s_{2h\tau}^{k,i} \right) \in \boldsymbol{\mathcal{K}}_g^\mathrm{t}$ as a piecewise affine and continuous function in time over the whole time interval $]0,T[$, verifying $s_{1h\tau}^{k,i}(t) - s_{2h\tau}^{k,i}(t) \ge 0 \ \forall t \in]0,T[$. When $p = 1$, for all $1 \le n \le N_\mathrm{t}$, for all $\boldsymbol{a} \in \mathcal{V}_h^\mathrm{int}$, a possibility is to construct $\boldsymbol{s}_h^{n,k,i} := \left(s_{1h}^{n,k,i}, s_{2h}^{n,k,i} \right) \in \boldsymbol{\mathcal{K}}_{gh}^1$ by

$$\boldsymbol{s}_h^{n,k,i}(\boldsymbol{a}) := \begin{cases} \boldsymbol{u}_h^{k,i}(\boldsymbol{a}) = \left(u_{1h}^{n,k,i}(\boldsymbol{a}), u_{2h}^{n,k,i}(\boldsymbol{a}) \right) & \text{if } u_{1h}^{n,k,i}(\boldsymbol{a}) \ge u_{2h}^{n,k,i}(\boldsymbol{a}), \\ \left(\dfrac{u_{1h}^{n,k,i}(\boldsymbol{a}) + u_{2h}^{n,k,i}(\boldsymbol{a})}{2}, \dfrac{u_{1h}^{n,k,i}(\boldsymbol{a}) + u_{2h}^{n,k,i}(\boldsymbol{a})}{2} \right) & \text{if } u_{1h}^{n,k,i}(\boldsymbol{a}) < u_{2h}^{n,k,i}(\boldsymbol{a}). \end{cases} \tag{2.94}$$

Definition 2.6.10. *We define* $\forall 1 \le n \le N_{\mathrm{t}}$ *the following error estimators*

$$
\begin{aligned}
\eta_{\mathrm{R,K},\alpha}^{n,k,i}(t) &:= h_\Omega C_{\mathrm{PF}} \mu_\alpha^{-\frac{1}{2}} \left\| f_\alpha|_{I_n} - \partial_t s_{\alpha h\tau}^{n,k,i} - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h\tau}^{n,k,i} - (-1)^\alpha \lambda_h^{n,k,i} \right\|_K (t),\\
\eta_{\mathrm{F,K},\alpha}^{n,k,i}(t) &:= \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} s_{\alpha h\tau}^{n,k,i} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau}^{n,k,i} \right\|_K (t),\\
\eta_{\mathrm{C},K}^{n,k,i,\mathrm{pos}}(t) &:= 2 \left(\lambda_{h\tau}^{n,k,i,\mathrm{pos}}, u_{1h\tau}^{n,k,i} - u_{2h\tau}^{n,k,i} \right)_K (t),\\
\eta_{\mathrm{nonc},1,K}^{n,k,i}(t) &:= h_\Omega C_{\mathrm{PF}} \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} \right)^{\frac{1}{2}} \left\| \lambda_{h\tau}^{n,k,i,\mathrm{neg}} \right\|_K (t),\\
\eta_{\mathrm{nonc},2,K,\alpha}^{n,k,i}(t) &:= \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} \left(s_{\alpha h\tau}^{n,k,i} - u_{\alpha h\tau}^{n,k,i} \right) \right\|_K (t),\\
\eta_{\mathrm{nonc},3,K}^{n,k,i}(t) &:= 2 \left(\lambda_{h\tau}^{n,k,i,\mathrm{pos}}, \left(s_{1h\tau}^{n,k,i} - u_{1h\tau}^{n,k,i} \right) - \left(s_{2h\tau}^{n,k,i} - u_{2h\tau}^{n,k,i} \right) \right)_K (t).
\end{aligned}
$$

We observe that the estimators given by Definition 2.6.10 are slightly different from the ones provided in Definition 2.6.5. Indeed, in the estimators $\eta_{\mathrm{R,K},\alpha}^{n,k,i}$ and $\eta_{\mathrm{F,K},\alpha}^{n,k,i}$, there appears a $s_{\alpha h\tau}^{n,k,i}$ in place of $u_{\alpha h\tau}^{n,k,i}$, and h_Ω instead of h_K for instance. The constraints estimator $\eta_{\mathrm{C},K}^{n,k,i,\mathrm{pos}}$ is as in Definition 2.6.5 (remember that $\lambda_{h\tau}^n \ge 0$ at convergence for $p=1$) and expresses that $\lambda_{h\tau}^{n,k,i} \left(u_{1h\tau}^{n,k,i} - u_{2h\tau}^{n,k,i} \right) = 0$ is not valid. Next, $\eta_{\mathrm{nonc},1,K}^{n,k,i}$, $\eta_{\mathrm{nonc},2,K,\alpha}^{n,k,i}$, and $\eta_{\mathrm{nonc},3,K}^{n,k,i}$ are nonconformity estimators expressing the possible negativity of the discrete Lagrange multiplier and measuring how far the potential reconstruction $\boldsymbol{s}_{h\tau}^{n,k,i}$ is from the displacements $\boldsymbol{u}_{h\tau}^{n,k,i}$. Note that for $p=1$, the estimators $\eta_{\mathrm{nonc},1,K}^{n,k,i}$, $\eta_{\mathrm{nonc},2,K,\alpha}^{n,k,i}$, and $\eta_{\mathrm{nonc},3,K}^{n,k,i}$ turn into semismooth linearization estimators such that $\eta_{\mathrm{lin},1,K}^{n,k,i} := \eta_{\mathrm{nonc},1,K}^{n,k,i}$, $\eta_{\mathrm{lin},2,K,\alpha}^{n,k,i} := \eta_{\mathrm{nonc},2,K,\alpha}^{n,k,i}$, and $\eta_{\mathrm{lin},3,K}^{n,k,i} := \eta_{\mathrm{nonc},3,K}^{n,k,i}$. At convergence, for $p=1$, $\lambda_{h\tau}^{n,k,i,\mathrm{pos}} = \lambda_{h\tau}^n$, $\lambda_{h\tau}^{n,k,i,\mathrm{neg}} = 0$, $\boldsymbol{s}_{h\tau}^{n,k,i} = \boldsymbol{u}_{h\tau}^{n,k,i}$, and then $\eta_{\mathrm{lin},1,K}^{n,k,i} = \eta_{\mathrm{lin},2,K,\alpha}^{n,k,i} = \eta_{\mathrm{lin},3,K}^{n,k,i} = 0$.

Theorem 2.6.11. *Let* $(u_1,u_2,\lambda) \in V_g \times V_0 \times \Psi$ *be the exact solution and let* $\boldsymbol{u}_{h\tau}^{k,i} \notin \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ *with* $\lambda_{h\tau}^{k,i}$ *be the approximate solution issued from inexact linearization and algebraic solvers at each time step* $1 \le n \le N_{\mathrm{t}}$. *Consider the total equilibrated flux reconstruction* $\boldsymbol{\sigma}_{\alpha h\tau}^{k,i} \in L^2(0,T,\mathbf{H}(\mathrm{div},\Omega))$ *given by* (2.74) *and* (2.75). *Let* $\boldsymbol{s}_{h\tau}^{k,i} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ *and consider the estimators of Definition 2.6.10. Then, for*

$$
\begin{aligned}
\left(\eta^{k,i}\right)^2 := \Bigg(& \left\{ \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{R,K},\alpha}^{n,k,i} \right)^2 (t)\,\mathrm{dt} \right\}^{\frac{1}{2}} + \left\{ \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{F,K},\alpha}^{n,k,i} \right)^2 (t)\,\mathrm{dt} \right\}^{\frac{1}{2}} \\
& + \left\{ \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \left(\eta_{\mathrm{nonc},1,K}^{n,k,i} \right)^2 (t)\,\mathrm{dt} \right\}^{\frac{1}{2}} \Bigg)^2 + \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{C},K}^{n,k,i,\mathrm{pos}}(t)\,\mathrm{dt} \\
& + \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{nonc},3,K}^{n,k,i}(t)\,\mathrm{dt} + \sum_{\alpha=1}^{2} \left\| \left(u_\alpha - s_{\alpha h\tau}^{n,k,i} \right)(\cdot,0) \right\|_\Omega^2,
\end{aligned}
\tag{2.95}
$$

we have the a posteriori error estimate

$$\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{u}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0} \leq \eta^{k,i} + \left\{\sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{\alpha=1}^{2}\sum_{K\in\mathcal{T}_h}\left(\eta_{\mathrm{nonc},2,K,\alpha}^{n,k,i}\right)^2(t)\,\mathrm{dt}\right\}^{\frac{1}{2}}. \tag{2.96}$$

Proof. The structure of the proof differs from that of Chapter 1, Theorem 1.5.1. We start by the triangle inequality, leading to

$$\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{u}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0} \leq \left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0} + \left|\!\left|\!\left|\boldsymbol{s}_{h\tau}^{k,i}-\boldsymbol{u}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0}. \tag{2.97}$$

The second term of (2.97) immediately equals to

$$\left|\!\left|\!\left|\boldsymbol{s}_{h\tau}^{k,i}-\boldsymbol{u}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0}^2 = \sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{\alpha=1}^{2}\sum_{K\in\mathcal{T}_h}\left(\eta_{\mathrm{nonc},2,K,\alpha}^{n,k,i}\right)^2(t)\,\mathrm{dt}. \tag{2.98}$$

Next, observe that

$$\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0}^2 \leq \left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0}^2 + \frac{1}{2}\sum_{\alpha=1}^{2}\left\|\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)(\cdot,T)\right\|_\Omega^2.$$

Employing the fact that

$$\begin{aligned}\frac{1}{2}\sum_{\alpha=1}^{2}\left\|\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)(\cdot,T)\right\|_\Omega^2 =& \frac{1}{2}\sum_{\alpha=1}^{2}\left\|\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)(\cdot,0)\right\|_\Omega^2 \\ &+ \sum_{\alpha=1}^{2}\int_0^T\left\langle\partial_t\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right), u_\alpha - s_{\alpha h\tau}^{k,i}\right\rangle(t)\,\mathrm{dt},\end{aligned}$$

we have

$$\begin{aligned}\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0}^2 \leq& \sum_{\alpha=1}^{2}\int_0^T\mu_\alpha\left(\boldsymbol{\nabla}\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right), \boldsymbol{\nabla}\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)\right)_\Omega(t)\,\mathrm{dt} \\ &+ \sum_{\alpha=1}^{2}\int_0^T\left\langle\partial_t\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right), u_\alpha - s_{\alpha h\tau}^{k,i}\right\rangle(t)\,\mathrm{dt} + \frac{1}{2}\sum_{\alpha=1}^{2}\left\|\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)(\cdot,0)\right\|_\Omega^2.\end{aligned}$$

We now use the weak formulation (2.18) with $\boldsymbol{v} = \boldsymbol{s}_{h\tau}^{k,i} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ to get

$$\begin{aligned}\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0}^2 \leq& \sum_{\alpha=1}^{2}\int_0^T\left(f_\alpha - \partial_t s_{\alpha h\tau}^{k,i}, u_\alpha - s_{\alpha h\tau}^{k,i}\right)_\Omega(t) \\ &- \sum_{\alpha=1}^{2}\int_0^T\mu_\alpha\left(\boldsymbol{\nabla}s_{\alpha h\tau}^{k,i}, \boldsymbol{\nabla}\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)\right)_\Omega(t)\,\mathrm{dt} + \frac{1}{2}\sum_{\alpha=1}^{2}\left\|\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)(\cdot,0)\right\|_\Omega^2.\end{aligned}$$

Adding and subtracting $\boldsymbol{\sigma}_{\alpha h\tau}^{k,i} \in L^2(0,T;\mathbf{H}(\mathrm{div},\Omega))$ and using the Green formula with $\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)(t) \in H_0^1(\Omega)\ \forall t \in]0,T[$, we obtain

$$\begin{aligned}\left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{s}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0}^2 \leq & \sum_{\alpha=1}^{2}\int_0^T \left(f_\alpha - \partial_t s_{\alpha h\tau}^{k,i} - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h\tau}^{k,i}, u_\alpha - s_{\alpha h\tau}^{k,i}\right)_\Omega (t)\,\mathrm{dt} \\ & - \sum_{\alpha=1}^{2}\int_0^T \left(\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla} s_{\alpha h\tau}^{k,i} + \mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}_{\alpha h\tau}^{k,i}, \mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)\right)_\Omega (t)\,\mathrm{dt} \\ & + \frac{1}{2}\sum_{\alpha=1}^{2}\left\|\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)(\cdot,0)\right\|_\Omega^2.\end{aligned}$$

Furthermore, adding and subtracting $\displaystyle\sum_{\alpha=1}^{2}\int_0^T \left((-1)^\alpha \lambda_{h\tau}^{k,i}, u_\alpha - s_{\alpha h\tau}^{k,i}\right)_\Omega (t)\,\mathrm{dt}$, we get

$$\left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{s}_{h\tau}^{k,i}\right|\!\right|\!\right|_{V_0}^2 \leq A + B + C + \frac{1}{2}\sum_{\alpha=1}^{2}\left\|\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)(\cdot,0)\right\|_\Omega^2 \tag{2.99}$$

with

$$\begin{aligned}A &:= \sum_{\alpha=1}^{2}\int_0^T \left(f_\alpha - \partial_t s_{\alpha h\tau}^{k,i} - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h\tau}^{k,i} - (-1)^\alpha\lambda_{h\tau}^{k,i}, u_\alpha - s_{\alpha h\tau}^{k,i}\right)_\Omega (t)\,\mathrm{dt}, \\ B &:= -\sum_{\alpha=1}^{2}\int_0^T \left(\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla} s_{\alpha h\tau}^{k,i} + \mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}_{\alpha h\tau}^{k,i}, \mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}\left(u_\alpha - s_{\alpha h\tau}^{k,i}\right)\right)_\Omega (t)\,\mathrm{dt}, \\ C &:= \sum_{\alpha=1}^{2}\int_0^T \left((-1)^\alpha\lambda_{h\tau}^{k,i}, u_\alpha - s_{\alpha h\tau}^{k,i}\right)_\Omega (t)\,\mathrm{dt}.\end{aligned} \tag{2.100}$$

To bound A and B we proceed as follows. We apply the Cauchy–Schwarz inequality and next the Poincaré–Friedrichs inequality (2.11a) to get

$$A \leq \left(\sum_{n=1}^{N_\mathrm{t}}\int_{I_n}\sum_{\alpha=1}^{2}\sum_{K\in\mathcal{T}_h}\left(\eta_{\mathrm{R,K},\alpha}^{n,k,i}\right)^2(t)\,\mathrm{dt}\right)^{\frac{1}{2}}\left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{s}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0}, \tag{2.101}$$

and directly

$$B \leq \left(\sum_{n=1}^{N_\mathrm{t}}\int_{I_n}\sum_{\alpha=1}^{2}\sum_{K\in\mathcal{T}_h}\left(\eta_{\mathrm{F,K},\alpha}^{n,k,i}\right)^2(t)\,\mathrm{dt}\right)^{\frac{1}{2}}\left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{s}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0}. \tag{2.102}$$

It remains to bound the term C. Observe that

$$\begin{aligned}C = & \int_0^T \left(-\lambda_{h\tau}^{k,i,\mathrm{neg}}, u_1 - s_{1h\tau}^{k,i}\right)(t)\,\mathrm{dt} + \int_0^T \left(\lambda_{h\tau}^{k,i,\mathrm{neg}}, u_2 - s_{2h\tau}^{k,i}\right)_\Omega (t)\,\mathrm{dt} \\ & + \int_0^T \left(-\lambda_{h\tau}^{k,i,\mathrm{pos}}, u_1 - s_{1h\tau}^{k,i}\right)(t)\,\mathrm{dt} + \int_0^T \left(\lambda_{h\tau}^{k,i,\mathrm{pos}}, u_2 - s_{2h\tau}^{k,i}\right)_\Omega (t)\,\mathrm{dt}.\end{aligned}$$

There holds

$$\begin{aligned}
&\int_0^T \left(-\lambda_{h\tau}^{k,i,\mathrm{pos}}, u_1 - s_{1h\tau}^{k,i} - (u_2 - s_{2h\tau}^{k,i})\right)(t)\,\mathrm{dt} \\
&= \int_0^T \underbrace{-b(\boldsymbol{u}, \lambda_{h\tau}^{k,i,\mathrm{pos}})}_{\leq 0}(t)\,\mathrm{dt} + \int_0^T b(\boldsymbol{s}_{h\tau}^{k,i} - \boldsymbol{u}_{h\tau}^{k,i}, \lambda_{h\tau}^{k,i,\mathrm{pos}})(t)\,\mathrm{dt} \\
&\quad + \int_0^T b(\boldsymbol{u}_{h\tau}^{k,i}, \lambda_{h\tau}^{k,i,\mathrm{pos}})(t)\,\mathrm{dt}.
\end{aligned}$$

But, $\boldsymbol{u} \in \boldsymbol{\mathcal{K}}_g^{\mathrm{t}}$ and $\lambda_{h\tau}^{k,i,\mathrm{pos}}(t) \geq 0\ \forall t \in]0,T[$, so that $-b(\boldsymbol{u}, \lambda_{h\tau}^{k,i,\mathrm{pos}}) \leq 0$. Thus,

$$C \leq C_1 + C_2 + C_3$$

with

$$\begin{aligned}
C_1 &:= \int_0^T \left(-\lambda_{h\tau}^{k,i,\mathrm{neg}}, u_1 - s_{1h\tau}^{k,i} - \left(u_2 - s_{2h\tau}^{k,i}\right)\right)(t)\,\mathrm{dt}, \\
C_2 &:= \int_0^T b(\boldsymbol{s}_{h\tau}^{k,i} - \boldsymbol{u}_{h\tau}^{k,i}, \lambda_{h\tau}^{k,i,\mathrm{pos}})(t)\,\mathrm{dt}, \\
C_3 &= \int_0^T b(\boldsymbol{u}_{h\tau}^{k,i}, \lambda_{h\tau}^{k,i,\mathrm{pos}})(t)\,\mathrm{dt}.
\end{aligned}$$

The Cauchy–Schwarz inequality and the Poincaré–Friedrichs inequality (2.11a) yield

$$\begin{aligned}
C_1 &\leq h_\Omega C_{\mathrm{PF}} \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)^{\frac{1}{2}} \left(\sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \left\|\lambda_{h\tau}^{n,k,i,\mathrm{neg}}\right\|_K^2 (t)\,\mathrm{dt}\right)^{\frac{1}{2}} \left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{s}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0} \\
&= \left(\sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\eta_{\mathrm{nonc},1,K}^{n,k,i}\right)^2 (t)\,\mathrm{dt}\right)^{\frac{1}{2}} \left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{s}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0}.
\end{aligned} \tag{2.103}$$

Next, we have

$$C_2 = \frac{1}{2} \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \eta_{\mathrm{nonc},3,K}^{n,k,i}(t)\,\mathrm{dt}. \tag{2.104}$$

Furthermore, we have

$$\begin{aligned}
C_3 &= \frac{1}{2} \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} 2\left(\lambda_{h\tau}^{n,k,i,\mathrm{pos}}, u_{1h\tau}^{n,k,i} - u_{2h\tau}^{n,k,i}\right)(t)\,\mathrm{dt} \\
&= \frac{1}{2} \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \eta_{\mathrm{C},K}^{n,k,i,\mathrm{pos}}(t).
\end{aligned} \tag{2.105}$$

Finally, combining (2.99)–(2.105) we get

$$
\begin{aligned}
\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0}^2 \leq & \left(\sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{\alpha=1}^{2}\sum_{K\in\mathcal{T}_h}\left(\eta_{\mathrm{R,K},\alpha}^{n,k,i}\right)^2(t)\,\mathrm{dt}\right)^{\frac{1}{2}}\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0} \\
& + \left(\sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{\alpha=1}^{2}\sum_{K\in\mathcal{T}_h}\left(\eta_{\mathrm{F,K},\alpha}^{n,k,i}\right)^2(t)\,\mathrm{dt}\right)^{\frac{1}{2}}\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0} \\
& + \left(\sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{K\in\mathcal{T}_h}\left(\eta_{\mathrm{nonc},1,K}^{n,k,i}\right)^2(t)\,\mathrm{dt}\right)^{\frac{1}{2}}\left|\!\left|\!\left|\boldsymbol{u}-\boldsymbol{s}_{h\tau}^{n,k,i}\right|\!\right|\!\right|_{V_0} \\
& + \sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{K\in\mathcal{T}_h}\frac{1}{2}\left(\eta_{\mathrm{nonc},3,K}^{n,k,i}+\eta_{\mathrm{C},K}^{n,k,i,\mathrm{pos}}\right)(t)\,\mathrm{dt} \\
& + \frac{1}{2}\sum_{\alpha=1}^{2}\left\|\left(u_\alpha-s_{\alpha h\tau}^{k,i}\right)(\cdot,0)\right\|_{\Omega}^2.
\end{aligned}
\tag{2.106}
$$

Employing the Young inequality ($ab \leq \frac{1}{2}(a^2+b^2)\ \ \forall\,(a,b)\geq 0$), and using (2.98) provide the desired result.

□

2.7 Adaptivity

In Section 2.6.6, we derived an a posteriori error estimate between the exact solution and approximate solution at each time step $1 \leq n \leq N_{\mathrm{t}}$, at each semismooth Newton step $k \geq 1$, and each algebraic iterative solver step $i \geq 0$. In the spirit of Chapter 1 Section 1.5 and Section 1.6, we provide an a posteriori error estimate distinguishing the different error components when $p = 1$. This new estimate provides adaptive stopping criteria based on the distinction of each error component.

Definition 2.7.1. *We define the total discretization error estimator $\eta_{\mathrm{disc}}^{n,k,i}$, the total semismooth linearization error estimator $\eta_{\mathrm{lin}}^{n,k,i}$, and the total algebraic error estimator $\eta_{\mathrm{alg}}^{n,k,i}$ respectively by*

$$
\begin{aligned}
\eta_{\mathrm{disc}}^{k,i} := & \left\{3\left(\left\{\sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{K\in\mathcal{T}_h}\sum_{\alpha=1}^{2}\left(\eta_{\mathrm{R,K},\alpha}^{n,k,i}\right)^2\right\}^{\frac{1}{2}}\right. \\
& \left.+\left\{\sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{K\in\mathcal{T}_h}\sum_{\alpha=1}^{2}\left\|\mu_\alpha^{\frac{1}{2}}\boldsymbol{\nabla}s_{\alpha h\tau}^{n,k,i}+\mu_\alpha^{-\frac{1}{2}}\boldsymbol{\sigma}_{\alpha h\tau,\mathrm{disc}}^{n,k,i}\right\|_{\Omega}^2\right\}^{\frac{1}{2}}\right)^2+\left|\sum_{n=1}^{N_{\mathrm{t}}}\int_{I_n}\sum_{K\in\mathcal{T}_h}\eta_{\mathrm{C},K}^{n,k,i,\mathrm{pos}}\right|\Bigg\}^{\frac{1}{2}},
\end{aligned}
\tag{2.107}
$$

$$\eta_{\mathrm{lin}}^{k,i} := \left\{ 3 \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\eta_{\mathrm{nonc},1,K}^{n,k,i} \right)^2 + \left| \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \eta_{\mathrm{nonc},3,K}^{n,k,i} \right| \right\}^{\frac{1}{2}} + \left\{ \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{nonc},2,K,\alpha}^{n,k,i} \right)^2 \right\}^{\frac{1}{2}}, \tag{2.108}$$

$$\eta_{\mathrm{alg}}^{k,i} := 3^{\frac{1}{2}} \left\{ \sum_{n=1}^{N_{\mathrm{t}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \sum_{\alpha=1}^{2} \left\| \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{alg}}^{n,k,i} \right\|_K^2 (t) \, \mathrm{dt} \right\}^{\frac{1}{2}}, \tag{2.109}$$

$$\eta_{\mathrm{init}} := \left\{ \sum_{\alpha=1}^{2} \left\| \left(u_\alpha - s_{\alpha h\tau}^{n,k,i} \right) (\cdot, 0) \right\|_\Omega^2 \right\}^{\frac{1}{2}}. \tag{2.110}$$

Corollary 2.7.2. *For $p = 1$, we have the following a posteriori error estimate distinguishing separately the error components:*

$$\left| \left| \left| \boldsymbol{u} - \boldsymbol{u}_{h\tau}^{k,i} \right| \right|_{V_0} \leq \eta_{\mathrm{disc}}^{k,i} + \eta_{\mathrm{lin}}^{k,i} + \eta_{\mathrm{alg}}^{k,i} + \eta_{\mathrm{init}}. \tag{2.111}$$

Proof. The triangle inequality gives

$$\begin{aligned} \eta_{\mathrm{F,K},\alpha}^{n,k,i} &= \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} s_{\alpha h\tau}^{n,k,i} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau}^{n,k,i} \right\|_K, \\ &\leq \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} s_{\alpha h\tau}^{n,k,i} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{disc}}^{n,k,i} \right\|_K + \left\| \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{alg}}^{n,k,i} \right\|_K. \end{aligned}$$

Next, using the Minkowski inequality to separate the algebraic contribution from the discretization one and employing after the result $(A_1 + A_2 + A_3)^2 \leq 3\,(A_1^2 + A_2^2 + A_3^2)$ for $A_1, A_2, A_3 \geq 0$ to gather the discretization terms, we obtain the desired result. □

Adaptive inexact semismooth Newton algorithm

We finally present our adaptive inexact semismooth Newton algorithm. Following the concept of Chapter 1, Section 1.6, it is designed to only perform the linearization and algebraic resolution with minimal necessary precision, and thus to avoid unnecessary iterations. Let γ_{lin} and γ_{alg} be two positive parameters typically of order 0.1, representing the desired relative size of the algebraic and linearization errors. Note that as the estimators of Definition 2.7.1 are global, we consider their restrictions $\eta_{\mathrm{disc}}^{n,k,i}$, $\eta_{\mathrm{lin}}^{n,k,i}$, and $\eta_{\mathrm{alg}}^{n,k,i}$ to the time interval I_n defined by

$$\eta_{\mathrm{disc}}^{n,k,i} := \left(\int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\sum_{\alpha=1}^{2} 6 \left(\left(\eta_{\mathrm{R,K},\alpha}^{n,k,i} \right)^2 + \left\| \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} s_{\alpha h\tau}^{n,k,i} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{disc}}^{n,k,i} \right\|_K^2 \right) + |\eta_{\mathrm{C},K}^{n,k,i,\mathrm{pos}}| \right) (t) \, \mathrm{dt} \right)^{\frac{1}{2}},$$

$$\eta_{\mathrm{lin}}^{n,k,i} := \left(\int_{I_n} 2 \left(\sum_{K \in \mathcal{T}_h} \left(3 \left(\eta_{\mathrm{nonc},1,K}^{n,k,i} \right)^2 + |\eta_{\mathrm{nonc},3,K}^{n,k,i}| \right) + \left| \left| \left| \boldsymbol{s}_{h\tau}^{n,k,i} - \boldsymbol{u}_{h\tau}^{n,k,i} \right| \right|_\Omega^2 \right) (t) \, \mathrm{dt} \right)^{\frac{1}{2}}$$

$$\eta_{\mathrm{alg}}^{n,k,i} := \left(3|I_n| \sum_{K \in \mathcal{T}_h} \sum_{\alpha=1}^{2} \left\| \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h\tau,\mathrm{alg}}^{n,k,i} \right\|_K^2 \right)^{\frac{1}{2}}.$$

Supposing that η_{init} is small enough, we propose:

Algorithm 2 Adaptive inexact semismooth Newton algorithm on each time step n

0. Choose an initial vector $\boldsymbol{X}_h^{n,0} \in \mathbb{R}^{3\mathcal{N}_{\text{d}}^{p,\text{int}}}$ and set $k = 1$.

1. From $\boldsymbol{X}_h^{n,k-1}$ define $\mathbb{A}^{n,k-1} \in \mathbb{R}^{3\mathcal{N}_{\text{d}}^{p,\text{int}},3\mathcal{N}_{\text{d}}^{p,\text{int}}}$ and $\boldsymbol{B}^{n,k-1} \in \mathbb{R}^{3\mathcal{N}_{\text{d}}^{p,\text{int}}}$ by (2.55).

2. Consider the linear system

$$\mathbb{A}^{n,k-1}\boldsymbol{X}_h^{n,k} = \boldsymbol{B}^{n,k-1}. \tag{2.112}$$

3. Set $\boldsymbol{X}_h^{n,k,0} = \boldsymbol{X}_h^{n,k-1}$ as initial guess for the iterative linear solver, set $i = 0$.

4*a*. Perform $\nu \geq 1$ steps of a chosen linear solver for (2.112), starting from $\boldsymbol{X}_h^{n,k,i}$. Set $i = i + \nu$.

This yields on step i an approximation $\boldsymbol{X}_h^{n,k,i}$ to $\boldsymbol{X}_h^{n,k}$ satisfying

$$\mathbb{A}^{n,k-1}\boldsymbol{X}_h^{n,k,i} = \boldsymbol{B}^{n,k-1} - \boldsymbol{R}^{n,k,i}.$$

4*b*. Compute the estimators of Definition 2.7.1 and check the stopping criterion for the linear solver in the form:

$$\eta_{\text{alg}}^{n,k,i} \leq \gamma_{\text{alg}} \max\left\{\eta_{\text{disc}}^{n,k,i}, \eta_{\text{lin}}^{n,k,i}\right\}. \tag{2.113}$$

If satisfied, set $\boldsymbol{X}_h^{n,k} = \boldsymbol{X}_h^{n,k,i}$. If not go back to **4*a***.

5. Check the stopping criterion for the nonlinear solver in the form

$$\eta_{\text{lin}}^{n,k,i} \leq \gamma_{\text{lin}}\eta_{\text{disc}}^{n,k,i}. \tag{2.114}$$

If satisfied, return $\boldsymbol{X}_h^n = \boldsymbol{X}_h^{n,k}$. If not, set $k = k + 1$ and go back to **1**.

2.8 Conclusion

In this work, we focused on an unsteady parabolic variational inequality. We employed the $\mathbb{P}_p$ finite element method for the discretization in space and we used the backward Euler scheme for the discretization in time. We designed a posteriori error estimates when $p = 1$ and at convergence of the semismooth Newton solver and the iterative algebraic solver. In this case, we estimate both energy and time derivative errors. Next, we extended the study to all polynomial degrees at each semismooth Newton step $k \geq 1$ and each iterative linear algebraic solver step $i \geq 0$, for the energy error only. We finally proposed an adaptive inexact semismooth Newton algorithm based on the a posteriori error estimators we derived. Numerical experiments with the Newton-min algorithm in combination wih the iterative algebraic GMRES solver are under investigation.

Chapter 3

A posteriori error estimates and adaptive stopping criteria for a compositional two-phase flow with nonlinear complementarity constraints

Abstract

In this work, we develop an a-posteriori-steered algorithm for a compositional two-phase flow with exchange of components between the phases in porous media. As a model problem, we choose the two-phase liquid–gas flow with appearance and disappearance of the gas phase formulated as a system of nonlinear evolutive partial differential equations with nonlinear complementarity constraints. The discretization of our model is based on the backward Euler scheme in time and the finite volume scheme in space. The resulting nonlinear system is solved via an inexact semismooth Newton method. The key ingredient for the a posteriori analysis are the discretization, linearization, and algebraic flux reconstructions allowing to devise estimators for each error component. These enable to formulate criteria for stopping the iterative algebraic solver and the iterative linearization solver whenever the corresponding error components do not affect significantly the overall error. Numerical experiments are performed using the Newton-min algorithm as well as the Newton–Fischer–Burmeister algorithm in combination with the GMRES iterative linear solver to show the efficiency of our adaptive method.

Keywords: compositional multiphase flow, phase transition, complementarity condition, semismooth Newton method, a posteriori error estimate, adaptivity, stopping criterion

In Chapter 1 we studied a stationnary variational and in Chapter 2 we studied a parabolic variational inequality. In this chapter, we present our work on a posteriori error estimates for a two-phase flow with phase transition in porous media. The results of this chapter are taken from the article [20] submitted for publication.

3.1 Introduction

The storage of radioactive waste in deep geological layers generates broad interest among researchers and engineers concerned with the ecosystem preservation and protection. This storage induces, on a long time-scale, a gas (hydrogen) emission affecting heavily the environment and its sustainable and renewable ressources. The mathematical models describing these complex phenomena are part of the large category of strongly nonlinear evolutive multiphase multi-compositional equations where numerical simulation appears to be the only viable approach to finding a solution. A key point investigated today is the reduction of the computational cost of the numerical resolution employing an adaptive strategy based on a posteriori error estimates [62, 70, 71, 81, 158].

In this work, we consider a simpler situation described by a compositional two-phase flow in an isotropic porous medium in two space dimensions. The two miscible fluids involved are liquid and gas, and exchange components. To be coherent with the physical aspects of the problem, at the beginning of the simulation, the medium is monophasic liquid, *i.e.*, completely filled with the water component (the amount of hydrogen is negligible and completely dissolved in the liquid). Afterwards, the quantity of hydrogen increases, and it will be partially gaseous. At this stage, the flow is two-phase liquid–gas. In a usual scenario, at the end of the simulation, the production of gas hydrogen stops and the medium comes back to monophasic liquid.

The mathematical model expressing the behavior of two fluids with or without components in a porous medium relies on a strongly nonlinear system of partial differential equations where the unknowns are the pressure and saturation of the phases, see the book of Chen [52]. In Chavent and Jaffré [49] a reduction of these two-phase (without components) equations to a system of a single parabolic saturation equation coupled with an elliptic pressure equation is introduced, replacing the two pressure unknowns (one per phase) by only one pressure unknown, called the global pressure. A formulation for the compositional compressible two-phase flow liquid–gas by the global pressure has been recently proposed in Amaziane *et al.* [7]. Another formulation providing interesting results is the method of negative saturations, see Panfilov and Rasoulzadeh [135] and Panfilov and Panvilova [134].

Concerning the numerical methods employed for the discretization of the compositional multiphase models, we mention the finite differences, finite volumes, finite elements, mixed finite elements, and discontinuous Galerkin methods, see the books [13, 49, 51, 52, 98, 145, 151] and the references therein for a general introduction. The finite volume method is a popular approach and is commonly used in practice as it satisfies by construction local mass balance and is easy to implement, see [54, 85, 106].

One difficulty encountered by engineers is in handling the appearance/disappearance of the phases. From a mathematical standpoint, we can mention the pioneering works of [59] and [90] that are relevant for compositional multiphase flows. Nevertheless, it often leads to irregular convergence behaviour if the phase states are quickly changing. More recently, the approach consists in formulating the phase transitions as a set of local inequality constraints, which are then directly integrated into the nonlinear solver using nonlinear complementarity conditions. For

a two-phase industrial application, we can mention the work of Bourgeat, Jurak, and Smaï [30], Lauser *et al.* [119], Jaffré and Sboui [107] where in the last reference the appearance and disappearance of the gas phase is treated by Henry's law giving rise to a system of nonlinear equations coupled with nonlinear complementarity conditions. Next, in Ben Gharbia *et al.* [24], the same approach is introduced with as main novely the application of an exact semismooth Newton solver to treat the nonlinearities on the complementarity constraints. Usually, the nonlinear system is not solved exactly, leading to the concept of an inexact semismooth Newton method which is a popular approach to speed the convergence. Such approaches can be found in [65, 72, 114] for the case of inexact Newton methods and in [86, 92, 112, 127] for inexact semismooth Newton methods. For convergence results of semismooth Newton algorithms refer to [21–23, 88, 89].

In this work, we use the mathematical model of [24] and we are interested in deriving a posteriori error estimates, in order to formulate adaptive stopping criteria for our inexact semismooth solvers to save computational time. The a posteriori error estimates give a fully computable upper bound on the overall error between the exact solution and the approximate solution and localize the error at each simulation time and each element of the simulation domain. There is a well-developed litterature on a posteriori error estimates for partial differential equations. Related to our formulation, we first mention the fundamental work of Prager and Synge [138], the books of Ainsworth and Oden [5] and Repin [143], and the work of Ladevèze [118], where upper bounds for the error inspired from Prager and Synge's identity are derived. More recently, one approach consists in obtaining the so-called potential and equilibrated flux reconstructions solving auxiliary local problems (see Destuynder and Métivet [66], Braess and Schöberl [33], Ern and Vohralík [82], see also the references therein). Concerning a posteriori error estimate for variational inequalities, one can point out the pioneering work of Brezzi, Hager, and Raviart [39, 40], Kornhuber [116], Chen and Nochetto [53], Veeser [153], Repin [144] and Ben Belgacem *et al.* [19]. In particular in [19], a posteriori error estimates are given for exact solvers and recently, in [62] (see Chapter 1 and Chapter 2 of this Thesis), a posteriori error estimates are derived for inexact semismooth solvers and provide adaptive stopping criteria. The concept of adaptive stopping criteria relies on stopping the nonlinear and linear iterations whenever the associated estimators do not affect significantly the overall error, see [9, 62, 81, 108, 128]. For multiphase flows, devising a posteriori error estimates between the exact solution and approximate solution seems very ambitious and is still an open problem. Indeed, the existence of a weak solution relies on several strong assumptions and to construct upper bounds for energy norm errors seems somewhat inaccessible. In [44] an estimation between the exact solution and the approximate solution for the L^2 norm in time and H^{-1} in space has been derived in the case of a two-phase flow with only one component per phase. In general, for multiphase compositional flows, the alternative is to construct estimators as upper bounds for some dual norm of a residual, see [70], [71], and [158]. Constructing a posteriori error estimates and devise adaptive stopping criteria for inexact semismooth Newton solvers when the phase transition occurs has never been presented to the best of our knowledges. Therefore we will try to fill this gap.

We organize our paper as follows. In Sections 3.2 and 3.3 we introduce the

model problem, its finite difference discretization in time, and finite volume discretization in space. Next, in Section 3.4, we show that any inexact semismooth Newton method can be employed to solve the nonlinear system stemming from the discretization. Section 3.5 is devoted to the description of the various potential and flux reconstruction enabling to obtain a posteriori error estimators distinguishing all error components, namely the discretization error, the semismooth linearization error, and the algebraic error. In Section 3.6 we show numerical experiments when the semismooth min and Fischer–Burmeister solvers are employed in one dimensional space, and Section 3.7 summarizes our findings.

3.2 Setting

The methodology is presented for the sake of clarity in 2 space dimensions but can be extended to 3 or 1 without difficulties. We assume that the porous medium domain Ω is an open bounded connected polygon. We are interested in solving the model of appearance/disappearance of the gas phase thanks to nonlinear complementarity conditions over the time interval $(0, t_{\mathrm{F}})$, $t_{\mathrm{F}} > 0$, and devise a posteriori error estimates.

3.2.1 Functional spaces

First, we recall the definition of some Sobolev spaces. Let $H^1(\Omega)$ be the space of L^2 functions on the domain Ω which admit a weak gradient in $[L^2(\Omega)]^2$ and $H^1_0(\Omega)$ its zero-trace subspace. Similarly, $\mathbf{H}(\mathrm{div}, \Omega)$ stands for the space of $[L^2(\Omega)]^2$ functions having a weak divergence in $L^2(\Omega)$. The standard notation $\boldsymbol{\nabla}$ and $\boldsymbol{\nabla}\cdot$ are used respectively for the weak gradient and divergence. For a nonempty bounded set $\mathcal{O}$ of $\mathbb{R}^2$, we denote its Lebesgue measure by $|\mathcal{O}|$ and the $L^2(\mathcal{O})$ scalar product by $(u, v)_{\mathcal{O}} = \int_{\mathcal{O}} uv \,\mathrm{dx}$ for $u, v \in L^2(\mathcal{O})$. We also use the following notations: $\|v\|^2_{\mathcal{O}} := (v, v)_{\mathcal{O}}$, and $\|\boldsymbol{\nabla} v\|^2_{\mathcal{O}} := (\boldsymbol{\nabla} v, \boldsymbol{\nabla} v)_{\mathcal{O}}$. Besides, the Poincaré–Friedrichs and the Poincaré–Wirtinger inequalities, see [14, 137], state that if $\overline{v}_{\mathcal{O}}$ denotes the mean value of v on $\mathcal{O}$ and $h_{\mathcal{O}}$ the diameter of $\mathcal{O}$, then

$$\begin{aligned} \|v\|_{\mathcal{O}} &\le C_{\mathrm{PF}} h_{\mathcal{O}} \|\boldsymbol{\nabla} v\|_{\mathcal{O}} \quad \forall v \in H^1_0(\mathcal{O}), \\ \|v - \overline{v}_{\mathcal{O}}\|_{\mathcal{O}} &\le C_{\mathrm{PW}} h_{\mathcal{O}} \|\boldsymbol{\nabla} v\|_{\mathcal{O}} \quad \forall v \in H^1(\mathcal{O}). \end{aligned}$$

The constants C_{PF} and C_{PW} can be precisely estimated in many cases. In particular, if $\mathcal{O}$ is convex, C_{PW} can be taken as $\frac{1}{\pi}$, see [14, 137] whereas $C_{\mathrm{PF}} = 1$ is always possible.

3.2.2 The compositional two-phase model

We consider a compositional thermal biphasic flow in the porous medium Ω. The porous medium is characterized by its porosity ϕ and its absolute permeability $\underline{\mathbf{K}}$, both of which are assumed constant in space and time for the sake of simplicity. When the porous medium Ω is anisotropic, the positive constant $\underline{\mathbf{K}}$ is replaced by a symmetric positive definite matrix. Usually, the porous medium is caracterized

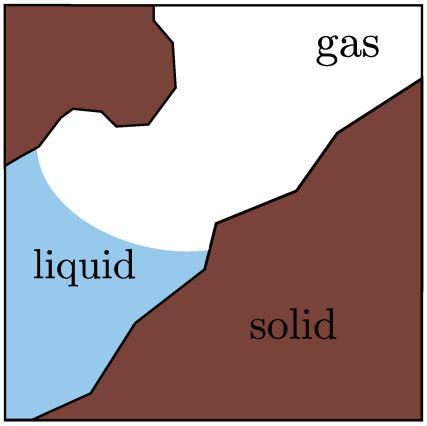

Figure 3.1: Porosity of interstices and two phases: liquid and gas.

by rocks and contains empty spaces called pores that are permeable to flows (see Figure 3.1). The porosity ϕ is defined by:

$$\phi := \frac{\text{Volume of the pores}}{\text{Total volume}}. \tag{3.1}$$

The phases are collected in the set $\mathcal{P} = \{\mathrm{l}, \mathrm{g}\}$ where "l" stands for the liquid phase and "g" for the gas one. Each of the considered fluids can be composed of two components: water (denoted by "w") and hydrogen (denoted by "h"). The set of components is defined by $\mathcal{C} = \{\mathrm{w}, \mathrm{h}\}$ and we denote by $\mathcal{C}^p$ the set of components present in the phase p and $\mathcal{P}_c$ the set of phases containing the component c. For a given phase $p \in \mathcal{P}$, S^p denotes its saturation, P^p its pressure and for each component $c \in \mathcal{C}^p$, χ_c^p is the molar fraction of the component c in phase p. Because of the interactions of forces between the fluids and the solid matrix and the curvature of the surface contact between the two fluids, we have an additional pressure called the capillary pressure depending on the saturation S^{l} with higher wettability, see [131], defined as

$$P_{\mathrm{cp}}(S^{\mathrm{l}}) = P^{\mathrm{g}} - P^{\mathrm{l}}. \tag{3.2}$$

Here, P_{cp} is a given function of the liquid saturation S^{l} and in the litterature, the suggestions of Brooks and Corey or Van Genuchten are commonly used, see [103].

The unknowns of the model below will be S^{l} (saturation of the liquid phase), P^{l} (pressure of the liquid phase), and $\chi_{\mathrm{h}}^{\mathrm{l}}$ (molar fraction of hydrogen in the liquid phase).

For a phase $p \in \mathcal{P}$ and for a given component $c \in \mathcal{C}^p$, $\rho_c^p(P^p, \chi_c^p)$ represents its molar density, $C_c^p(P^p, \chi_c^p)$ its molar concentration, $\mathbf{J}_c^p(P^p, S^p, \chi_c^p)$ its Fick flux, and D_c^p its molecular diffusion coefficient supposed constant. Furthermore, for a given phase $p \in \mathcal{P}$, $\mu^p(P^p, \chi_c^p)$ stands for its dynamic viscosity and $k_{\mathrm{r}}^p(S^p)$ represents its relative permeability. The relative permeability is typically an increasing function of S^p satisfying $k_{\mathrm{r}}^p(0) = 0$. Then, M_c represents the molar mass of the component c and $g = 9.81\mathrm{m.s}^{-2}$ is the gravity acceleration constant. We provide briefly an illustration of the relative permeabilities and the capillary pressure for the Van Genuchten model.

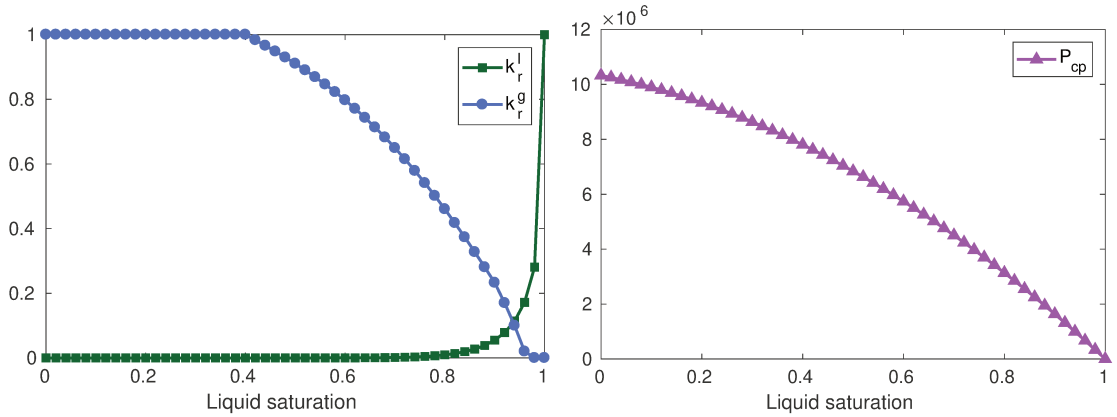

Figure 3.2: Van Genuchten model. Relative permeability of the liquid and gas phases (left), capillary pressure (right).

From Figure 3.2 we observe that $k_{\mathrm{r}}^{\mathrm{l}}$ vanishes in the interval $[0, S_{\mathrm{m}}]$ with $S_{\mathrm{m}} \approx 0.75$. It means that the liquid, when $S^{\mathrm{l}} \leq S_{\mathrm{m}}$ cannot be displaced by the gas. Similarly, $k_{\mathrm{r}}^{\mathrm{g}}$ vanishes on the interval $[S_{\mathrm{M}}, 1]$ with S_{M} close to 1. It means that the gas, when $S^{\mathrm{g}} \leq 1 - S_{\mathrm{M}}$ cannot be displaced by the liquid. In the literature, S_{m} (respectively $1 - S_{\mathrm{M}}$) is called the residual saturation of liquid (respectively of the gas). From Figure 3.2, right, we observe the typical behavior of the capillary pressure. It is a decreasing function of the liquid saturation.

We recall some elementary properties. The molar density of phase $p \in \mathcal{P}$ is defined as the sum of the molar densities of the components present in the phase:

$$\rho^p := \rho_{\mathrm{w}}^p + \rho_{\mathrm{h}}^p.$$

The molar concentration of phase $p \in \mathcal{P}$ is defined as the sum of the molar concentrations of the components present in the phase:

$$C^p := C_{\mathrm{w}}^p + C_{\mathrm{h}}^p := \frac{\rho_{\mathrm{w}}^p}{M_{\mathrm{w}}} + \frac{\rho_{\mathrm{h}}^p}{M_{\mathrm{h}}}. \tag{3.3}$$

Furthermore, the molar fraction of component $c \in \mathcal{C}^p$ is defined by

$$\chi_c^p := \frac{C_c^p}{C^p}, \quad \text{so that} \quad \chi_{\mathrm{w}}^p + \chi_{\mathrm{h}}^p = 1. \tag{3.4}$$

The Fick's law for any component $c \in \mathcal{C}^p$ gives

$$\mathbf{J}_c^p := -\phi M_c S^p C^p D_c^p \boldsymbol{\nabla} \chi_c^p.$$

The molecular diffusion in a phase $p \in \mathcal{P}$ is supposed negligible compared to the global displacement of this phase which implies

$$\mathbf{J}_{\mathrm{h}}^p + \mathbf{J}_{\mathrm{w}}^p = 0. \tag{3.5}$$

Next, as the pores are completely occupied by the fluids, we have the closure equation

$$S^{\mathrm{l}} + S^{\mathrm{g}} = 1. \tag{3.6}$$

To finish, the Darcy velocity $\mathbf{q}^p$ for any phase $p \in \mathcal{P}$ is defined by:

$$\mathbf{q}^p = -\underline{\mathbf{K}} \frac{k_{\mathrm{r}}^p(S^p)}{\mu^p} \left[\boldsymbol{\nabla} P^p - \rho^p g \boldsymbol{\nabla} z \right].$$

where z stands for the vertical coordinate.

We make the following assumptions:

Assumption 3.2.1. *We assume that the fluid is at thermodynamic equilibrium and that the water is incompressible and only present in the liquid phase:*

$\rho_{\mathrm{w}}^{\mathrm{l}}$ *is a constant,* $\rho_{\mathrm{w}}^{\mathrm{g}} = 0$, $\rho^{\mathrm{g}} = \rho_{\mathrm{h}}^{\mathrm{g}}$, $\chi_{\mathrm{h}}^{\mathrm{g}} = 1$, $\chi_{\mathrm{w}}^{\mathrm{g}} = 0$, $\mathcal{C}^{\mathrm{l}} := \{\mathrm{w}, \mathrm{h}\}$, $\mathcal{C}^{\mathrm{g}} := \{\mathrm{h}\}$, $\mathcal{P}_{\mathrm{w}} := \{\mathrm{l}\}$, $\mathcal{P}_{\mathrm{h}} := \{\mathrm{l}, \mathrm{g}\}$.

Next, we suppose that the liquid solution is an ideal diluted solution and the gas is slightly compressible:

$$C_{\mathrm{h}}^{\mathrm{l}} \ll C_{\mathrm{w}}^{\mathrm{l}} \quad \textit{and} \quad \rho^{\mathrm{g}} = \beta^{\mathrm{g}} P^{\mathrm{g}},$$

where β^{g} *is a compressibility constant.*

From Assumption 3.2.1 and (3.5) we obtain $\mathbf{J}_{\mathrm{h}}^{\mathrm{g}} = \mathbf{J}_{\mathrm{w}}^{\mathrm{g}} = 0$. Next, equation (3.4) combined with Assumption 3.2.1 gives

$$\chi_{\mathrm{w}}^{\mathrm{l}} \approx 1 \quad \text{and} \quad \chi_{\mathrm{h}}^{\mathrm{l}} \approx \frac{C_{\mathrm{h}}^{\mathrm{l}}}{C_{\mathrm{w}}^{\mathrm{l}}}. \tag{3.7}$$

Finally, equation (3.3) and (3.7) yield

$$\rho_{\mathrm{h}}^{\mathrm{l}} \approx \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}} \quad \text{with} \quad \beta^{\mathrm{l}} = \rho_{\mathrm{w}}^{\mathrm{l}} \frac{M_{\mathrm{h}}}{M_{\mathrm{w}}}. \tag{3.8}$$

Under Assumption 3.2.1, Fick's law for the hydrogen component in the liquid phase reads

$$\mathbf{J}_{\mathrm{h}}^{\mathrm{l}} = -\phi M_{\mathrm{h}} S^{\mathrm{l}} \left(\frac{\rho_{\mathrm{w}}^{\mathrm{l}}}{M_{\mathrm{w}}} + \frac{\beta^{\mathrm{l}}}{M_{\mathrm{h}}} \chi_{\mathrm{h}}^{\mathrm{l}} \right) D_{\mathrm{h}}^{\mathrm{l}} \boldsymbol{\nabla} \chi_{\mathrm{h}}^{\mathrm{l}}, \tag{3.9}$$

and the Darcy velocities reads

$$\begin{aligned} \mathbf{q}^{\mathrm{l}} &= -\underline{\mathbf{K}} \frac{k_{\mathrm{r}}^{\mathrm{l}}(S^{\mathrm{l}})}{\mu^{\mathrm{l}}} \left[\boldsymbol{\nabla} P^{\mathrm{l}} - \left[\rho_{\mathrm{w}}^{\mathrm{l}} + \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}} \right] g \boldsymbol{\nabla} z \right], \\ \mathbf{q}^{\mathrm{g}} &= -\underline{\mathbf{K}} \frac{k_{\mathrm{r}}^{\mathrm{g}}(1 - S^{\mathrm{l}})}{\mu^{\mathrm{g}}} \left[\boldsymbol{\nabla} \left[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}}) \right] - \beta^{\mathrm{g}} \left[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}}) \right] g \boldsymbol{\nabla} z \right]. \end{aligned} \tag{3.10}$$

In the sequel, all approximate equations will be considered to be exact equations.

3.2.3 Governing partial differential equations and nonlinear complementarity constraints

The system of partial differential equation representing the mass conservation for the two components, water and hydrogen, has the following form:

$$\begin{aligned} \partial_t(\phi \rho_{\mathrm{w}}^{\mathrm{l}} S^{\mathrm{l}} + \phi \rho_{\mathrm{w}}^{\mathrm{g}} S^{\mathrm{g}}) + \boldsymbol{\nabla} \cdot (\rho_{\mathrm{w}}^{\mathrm{l}} \mathbf{q}^{\mathrm{l}} + \rho_{\mathrm{w}}^{\mathrm{g}} \mathbf{q}^{\mathrm{g}} + \mathbf{J}_{\mathrm{w}}^{\mathrm{l}} + \mathbf{J}_{\mathrm{w}}^{\mathrm{g}}) = Q_{\mathrm{w}}, \\ \partial_t(\phi \rho_{\mathrm{h}}^{\mathrm{l}} S^{\mathrm{l}} + \phi \rho_{\mathrm{h}}^{\mathrm{g}} S^{\mathrm{g}}) + \boldsymbol{\nabla} \cdot (\rho_{\mathrm{h}}^{\mathrm{l}} \mathbf{q}^{\mathrm{l}} + \rho_{\mathrm{h}}^{\mathrm{g}} \mathbf{q}^{\mathrm{g}} + \mathbf{J}_{\mathrm{h}}^{\mathrm{l}} + \mathbf{J}_{\mathrm{h}}^{\mathrm{g}}) = Q_{\mathrm{h}}, \end{aligned} \tag{3.11}$$

where Q_c is a source term representing the outflow of the component $c \in \mathcal{C}$. To model the appearance of the gas phase we employ Henry's law, see [24, 107], giving

$$HP^{\mathrm{g}} = \rho_{\mathrm{h}}^{\mathrm{l}},$$

with $H = \tilde{H} M_{\mathrm{h}}$ where $\tilde{H}$ is Henry's constant. Next, using (3.6), (3.2), and (3.8) yields

$$1 - S^{\mathrm{l}} > 0 \quad \text{and} \quad H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}} = 0. \tag{3.12}$$

If the gas phase does not exist, using (3.6), (3.2), and [107, Section 3.2] we get

$$1 - S^{\mathrm{l}} = 0 \quad \text{and} \quad HP^{\mathrm{l}} - \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}} > 0. \tag{3.13}$$

Thus, using (3.12) and (3.13) we get nonlinear complementarity constraints:

$$1 - S^{\mathrm{l}} \geq 0, \quad H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}} \geq 0, \quad [1 - S^{\mathrm{l}}] \cdot \left[H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}}\right] = 0. \tag{3.14}$$

Finally, using Assumption 3.2.1, (3.11), and (3.14) our two-phase flow model with exchange between phases is governed by the following system: find S^{l}, P^{l}, $\chi_{\mathrm{h}}^{\mathrm{l}}$ such that

$$\begin{aligned}
&\partial_t l_{\mathrm{w}} + \boldsymbol{\nabla} \cdot \boldsymbol{\Phi}_{\mathrm{w}} = Q_{\mathrm{w}}, \\
&\partial_t l_{\mathrm{h}} + \boldsymbol{\nabla} \cdot \boldsymbol{\Phi}_{\mathrm{h}} = Q_{\mathrm{h}}, \\
&1 - S^{\mathrm{l}} \geq 0,\ H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}} \geq 0, [1 - S^{\mathrm{l}}] \cdot \left[H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}}\right] = 0.
\end{aligned} \tag{3.15}$$

Here, the component fluxes $\boldsymbol{\Phi}_c, c \in \mathcal{C}$, are defined by

$$\boldsymbol{\Phi}_{\mathrm{w}} = -\rho_{\mathrm{w}}^{\mathrm{l}} \mathbf{q}^{\mathrm{l}} - \mathbf{J}_{\mathrm{h}}^{\mathrm{l}}, \tag{3.16}$$

$$\boldsymbol{\Phi}_{\mathrm{h}} = -\beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}} \mathbf{q}^{\mathrm{l}} - \beta^{\mathrm{g}} \left[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})\right] \mathbf{q}^{\mathrm{g}} + \mathbf{J}_{\mathrm{h}}^{\mathrm{l}}, \tag{3.17}$$

where $\mathbf{q}^{\mathrm{l}}$, $\mathbf{q}^{\mathrm{g}}$, and $\mathbf{J}_{\mathrm{h}}^{\mathrm{l}}$ are defined in (3.10) and (3.9) and the amounts of components w and h per unit volume are defined by

$$\begin{aligned}
l_{\mathrm{w}} &= \phi \rho_{\mathrm{w}}^{\mathrm{l}} S^{\mathrm{l}}, \\
l_{\mathrm{h}} &= \phi \beta^{\mathrm{l}} \chi_{\mathrm{h}}^{\mathrm{l}} S^{\mathrm{l}} + \phi \beta^{\mathrm{g}} \left[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})\right] \left[1 - S^{\mathrm{l}}\right].
\end{aligned} \tag{3.18}$$

For the sake of simplicity, we assume that no-flow boundary conditions are prescribed for all the component fluxes,

$$\boldsymbol{\Phi}_c \cdot \boldsymbol{n}_\Omega = 0 \quad \text{on} \quad \partial\Omega \times (0, t_{\mathrm{F}}) \quad c \in \{\mathrm{w}, \mathrm{h}\} \tag{3.19}$$

with $\boldsymbol{n}_\Omega$ the outward unit normal vector to Ω. At $t = 0$ we prescribe the initial amount of each component

$$l_c(\cdot, 0) = l_c^0 \quad \forall c \in \{\mathrm{w}, \mathrm{h}\}. \tag{3.20}$$

3.3 Discretization and numerical approximation

We present in this section the discretization of our model. We use the backward Euler scheme in time and the cell-centered lowest order finite volume scheme in space.

3.3.1 Space-time meshes

For the time discretization, we consider an increasing sequence of points $\{t_n\}_{0\leq n\leq N_\mathrm{t}}$ such that $t_0 = 0$, $t_{N_\mathrm{t}} = t_\mathrm{F}$, and we introduce the interval $I_n = (t_{n-1}, t_n)$ and the time step $\tau_n = t_n - t_{n-1}$, $\forall 1 \leq n \leq N_\mathrm{t}$. For the space discretization, we consider $\mathcal{T}_h$ a family of conforming triangular meshes of the space domain Ω. We assume that $\mathcal{T}_h$ is formed by a set of triangles verifying $\bigcup_{K\in\mathcal{T}_h} \overline{K} = \overline{\Omega}$ where the intersection of two elements of $\mathcal{T}_h$ is either an empty set, a vertex, or an edge. We also define $H^1(\mathcal{T}_h)$ as the broken Sobolev space of L^2 functions on the domain Ω such that their restriction to any element K are H^1 in the element K. We denote by $\mathbb{P}^\mathrm{c}_m(\mathcal{T}_h)$ the space of continuous piecewise polynomials of degree $\leq m$ and by $\mathbb{P}^\mathrm{d}_m(\mathcal{T}_h)$ the broken polynomial space of discontinuous piecewise polynomials of degree $\leq m$. In the sequel, we will employ $m = 0$ and $m = 2$. We denote by D_m the set of Lagrange degrees of freedom associated to $\mathbb{P}^\mathrm{c}_m(\mathcal{T}_h)$. The set of vertices of $\mathcal{T}_h$ is denoted by $\mathcal{V}_h$ and is decomposed into interior vertices $\mathcal{V}_h^\mathrm{int}$ and boundary vertices $\mathcal{V}_h^\mathrm{ext}$. The vertices of an element $K \in \mathcal{T}_h$ are collected in the set $\mathcal{V}_K$. We denote by $\mathcal{E}_h$ the set of mesh edges. Boundary edges are collected in the set $\mathcal{E}_h^\mathrm{ext} = \{\sigma \in \mathcal{E}_h; \sigma \subset \partial\Omega\}$ and internal edges are collected in the set $\mathcal{E}_h^\mathrm{int} = \mathcal{E}_h \backslash \mathcal{E}_h^\mathrm{ext}$. Likewise, the edges of an element $K \in \mathcal{T}_h$ are collected in the set $\mathcal{E}_K$ and the later is decomposed into interior edges $\mathcal{E}_K^\mathrm{int}$ and boundary edges $\mathcal{E}_K^\mathrm{ext}$. We denote by N_sp the number of elements in the mesh $\mathcal{T}_h$. Furthermore, the notation $\boldsymbol{n}_{K,\sigma}$ stands for the outward unit normal vector to the element K on σ. We also assume that the family $\mathcal{T}_h$ is superadmissible in the sense that for all cells $K \in \mathcal{T}_h$ there exists a point $\boldsymbol{x}_K \in K$ (the cell center) and for all edges $\sigma \in \mathcal{E}_h$ there exists a point $\boldsymbol{x}_\sigma \in \sigma$ (the edge center) such that, for all edges $\sigma \in \mathcal{E}_K$, the line segment joining $\boldsymbol{x}_K$ with $\boldsymbol{x}_\sigma$ is orthogonal to σ, see [84]. For an interior edge $\sigma \in \mathcal{E}_h^\mathrm{int}$ shared by two elements K and L (denoted in the sequel by $\overline{\sigma} = \overline{K} \cap \overline{L}$) we define the distance between these elements $d_{KL} := \mathrm{dist}(\boldsymbol{x}_K, \boldsymbol{x}_L)$. Next, the vertical coordinate of any point $\boldsymbol{x}_K$ in the mesh $\mathcal{T}_h$ is denoted by z_K. For $\boldsymbol{a} \in D_m$, we call $\mathcal{T}_{\boldsymbol{a}}$ the patch around $\boldsymbol{a}$, *i.e.* the set of elements of $\mathcal{T}_h$ that share $\boldsymbol{a}$, and $\omega_h^{\boldsymbol{a}} \subset \Omega$ is the corresponding polygonal subdomain with $\mathbf{n}_{\omega_h^{\boldsymbol{a}}}$ its outward unit normal. The number of elements in $\mathcal{T}_{\boldsymbol{a}}$ is denoted by $|\mathcal{T}_{\boldsymbol{a}}|$. Note for instance that, in 2D, the patch for an interior edge degree of freedom contains exactly 2 elements and the patch for a vertex degree of freedom can contain a variable number of elements.

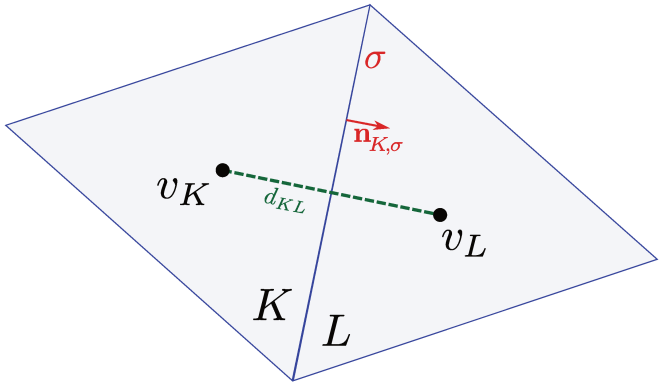

Figure 3.3: Illustration of the discretization of a gradient.

3.3.2 Finite volume discretization

Using the cell-centered finite volume method, the unknowns of the model are discretized using one value per cell: $\forall 1 \leq n \leq N_{\mathrm{t}}$ we let

$$\boldsymbol{U}^n := (\boldsymbol{U}_K^n)_{K \in \mathcal{T}_h} \in \mathbb{R}^{3N_{\mathrm{sp}}}, \qquad \boldsymbol{U}_K^n := \begin{pmatrix} S_K^n \\ P_K^n \\ \chi_K^n \end{pmatrix} \in \mathbb{R}^3,$$

where S_K^n, respectively P_K^n, respectively χ_K^n are the discrete elementwise unknowns approximating the values of S^{l}, respectively P^{l}, respectively $\chi_{\mathrm{h}}^{\mathrm{l}}$ in the element $K \in \mathcal{T}_h$. In the same way, $l_{c,K}^n$ approximates the value of l_c in the element $K \in \mathcal{T}_h$.

For a function of time v with sufficient regularity, we denote $v^n := v(t^n)$, $0 \leq n \leq N_{\mathrm{t}}$, and, for $1 \leq n \leq N_{\mathrm{t}}$, we define the backward differencing operator

$$\partial_t^n v := \frac{1}{\tau_n} \left(v^n - v^{n-1} \right). \tag{3.21}$$

To approximate the space gradient we use

$$(\boldsymbol{\nabla} v \cdot \boldsymbol{n}_{K,\sigma}, 1)_\sigma \approx |\sigma| \frac{v_L - v_K}{d_{KL}} \quad \text{if} \quad \sigma \in \mathcal{E}_K^{\mathrm{int}}, \quad \overline{\sigma} = \overline{K} \cap \overline{L}.$$

First, we discretize the water conservation equation. Let $K \in \mathcal{T}_h$. By integration over the element K we obtain

$$(\partial_t l_{\mathrm{w}} + \boldsymbol{\nabla} \cdot \boldsymbol{\Phi}_{\mathrm{w}}, 1)_K = (Q_{\mathrm{w}}, 1)_K .$$

The Green formula gives the approximation for $n = 1, \ldots, N_{\mathrm{t}}$

$$|K| \partial_t^n l_{\mathrm{w},K} + \sum_{\sigma \in \mathcal{E}_K} F_{\mathrm{w},K,\sigma}(\boldsymbol{U}^n) = |K| Q_{\mathrm{w,K}}^n, \tag{3.22}$$

where the discrete elementwise water source term and the discrete elementwise amount of water are given by

$$Q_{\mathrm{w,K}}^n := \int_{I_n} \frac{(Q_{\mathrm{w}}, 1)_K}{|K| \tau_n}(t) \, \mathrm{dt}, \quad \text{and} \quad l_{\mathrm{w},K}^n := \phi \rho_{\mathrm{w}}^{\mathrm{l}} S_K^n \quad \text{and,} \quad \partial_t^n l_{\mathrm{w},K} = \frac{1}{\tau_n} \left(l_{\mathrm{w},K}^n - l_{\mathrm{w},K}^{n-1} \right).$$

Let $\sigma \in \mathcal{E}_K^{\rm int}$, $\overline{\sigma} = \overline{K} \cap \overline{L}$. Then, the total flux across σ of the water component is given by

$$F_{{\rm w},K,\sigma}(\boldsymbol{U}^n) := \rho_{\rm w}^{\rm l}(\mathfrak{M}^{\rm l})_\sigma^n(\psi^{\rm l})_\sigma^n - (\mathrm{j}_{\rm h}^{\rm l})_\sigma^n, \tag{3.23}$$

with the discrete Fick term given by

$$(\mathrm{j}_{\rm h}^{\rm l})_\sigma^n := -|\sigma|\phi M_{\rm h} S_\sigma^n \left[\frac{\rho_{\rm w}^{\rm l}}{M_{\rm w}} + \frac{\beta^{\rm l}}{M_{\rm h}}\chi_\sigma^n\right] D_{\rm h}^{\rm l} \frac{\chi_L^n - \chi_K^n}{d_{KL}}, \tag{3.24}$$

the discrete liquid Darcy term given by

$$(\psi^{\rm l})_\sigma^n := -|\sigma|\frac{\mathbf{K}}{d_{KL}}\left[P_L^n - P_K^n - \left[\rho_{\rm w}^{\rm l} + \beta^{\rm l}\chi_\sigma^n\right] g[z_L - z_K]\right], \tag{3.25}$$

and the mobility of the liquid phase using an upwind approximation

$$(\mathfrak{M}^{\rm l})_\sigma^n := \frac{k_{\rm r}^{\rm l}(S_K^n)}{\mu^{\rm l}} \quad \text{if} \quad (\psi^{\rm l})_\sigma^n \geq 0, \quad (\mathfrak{M}^{\rm l})_\sigma^n := \frac{k_{\rm r}^{\rm l}(S_L^n)}{\mu^{\rm l}} \quad \text{if} \quad (\psi^{\rm l})_\sigma^n < 0, \tag{3.26}$$

where

$$S_\sigma^n := \frac{S_K^n + S_L^n}{2}, \quad \text{and} \quad \chi_\sigma^n := \frac{\chi_K^n + \chi_L^n}{2}. \tag{3.27}$$

Now, we discretize the hydrogen conservation equation. Let $K \in \mathcal{T}_h$. By integration over the element K we obtain

$$\left(\partial_t l_{\rm h} + \boldsymbol{\nabla}\cdot\boldsymbol{\Phi}_{\rm h}, 1\right)_K = \left(Q_{\rm h}, 1\right)_K.$$

The Green formula gives the approximation for $n = 1, \ldots, N_{\rm t}$

$$|K|\partial_t^n l_{{\rm h},K} + \sum_{\sigma\in\mathcal{E}_K} F_{{\rm h},K,\sigma}(\boldsymbol{U}^n) = |K|Q_{\rm h,K}^n, \tag{3.28}$$

where the discrete elementwise hydrogen source term and the discrete elementwise amount of hydrogen are given by

$$Q_{\rm h,K}^n := \int_{I_n} \frac{(Q_{\rm h},1)_K}{|K|\tau_n}(t)\,\mathrm{d}t, \quad l_{{\rm h},K}^n := \phi\beta^{\rm l}\chi_K^n S_K^n + \phi\beta^{\rm g}\left[P_K^n + P_{\rm cp}(S_K^n)\right]\left[1 - S_K^n\right].$$

Let $\sigma \in \mathcal{E}_K^{\rm int}$, $\sigma = \overline{K}\cap\overline{L}$. The total discrete flux across σ of the hydrogen component is given by

$$F_{{\rm h},K,\sigma}(\boldsymbol{U}^n) := \beta^{\rm l}\chi_\sigma^n(\mathfrak{M}^{\rm l})_\sigma^n(\psi^{\rm l})_\sigma^n + (\mathfrak{M}^{\rm g})_\sigma^n(\psi^{\rm g})_\sigma^n(\rho^{{\rm g},\star})_\sigma^n + (\mathrm{j}_{\rm h}^{\rm l})_\sigma^n, \tag{3.29}$$

where the discrete Fick term is given by (3.24), the discrete Darcy liquid term is given by (3.25), the mobility of the liquid phase is given by (3.26), and χ_σ^n is given by (3.27). Furthermore, the discrete Darcy gas term is given by

$$(\psi^{\rm g})_\sigma^n := -|\sigma|\frac{\mathbf{K}}{d_{KL}}\left[P_L^n + P_{\rm cp}(S_L^n) - P_K^n - P_{\rm cp}(S_K^n) - (\rho^{{\rm g},\star})_\sigma^n g[z_L - z_K]\right],$$

with

$$(\rho^{{\rm g},\star})_\sigma^n := \frac{(\rho^{\rm g})_K^n + (\rho^{\rm g})_L^n}{2}, \quad \text{and} \quad (\rho^{\rm g})_K^n := \beta^{\rm g}\left[P_K^n + P_{\rm cp}(S_K^n)\right].$$

Next, the mobility of the gas phase is

$$(\mathfrak{M}^{\mathrm{g}})_\sigma^n := \frac{k_{\mathrm{r}}^{\mathrm{g}}\,(1-S_K^n)}{\mu^{\mathrm{g}}} \quad \text{if} \quad (\psi^{\mathrm{g}})_\sigma^n \geq 0, \quad (\mathfrak{M}^{\mathrm{g}})_\sigma^n := \frac{k_{\mathrm{r}}^{\mathrm{g}}\,(1-S_L^n)}{\mu^{\mathrm{g}}} \quad \text{if} \quad (\psi^{\mathrm{g}})_\sigma^n < 0.$$

If $\sigma \in \mathcal{E}_K^{\mathrm{ext}} \subset \partial\Omega$, the homogeneous Neumann boundary condition yield

$$F_{\mathrm{w},K,\sigma}(\boldsymbol{U}^n) = F_{\mathrm{h},K,\sigma}(\boldsymbol{U}^n) = 0.$$

Thus, (3.22) and (3.28) define $\forall K \in \mathcal{T}_h$, $\forall c \in \{\mathrm{w},\mathrm{h}\}$, $\forall 1 \leq n \leq N_{\mathrm{t}}$ the nonlinear function $H_{c,K}^n : \mathbb{R}^{3N_{\mathrm{sp}}} \to \mathbb{R}$ defined by

$$H_{c,K}^n(\boldsymbol{U}^n) := |K|\partial_t^n l_{c,K} + \sum_{\sigma\in\mathcal{E}_K^{\mathrm{int}}} F_{c,K,\sigma}(\boldsymbol{U}^n) - |K|Q_{c,K}^n. \tag{3.30}$$

At each time step n, (3.30) gives a system of $2N_{\mathrm{sp}}$ nonlinear equations. As we have $3N_{\mathrm{sp}}$ unknowns, to close the system, we use the nonlinear complementarity conditions as follows.
Let F_K be the function discretizing elementwise $1 - S^{\mathrm{l}}$ and let G_K be the function discretizing elementwise $H[P^{\mathrm{l}} + P_{\mathrm{cp}}(S^{\mathrm{l}})] - \beta^{\mathrm{l}}\chi_{\mathrm{h}}^{\mathrm{l}}$ defined by:

$$\begin{aligned} F_K :\mathbb{R}^3 &\to \mathbb{R} \\ \boldsymbol{U}_K^n &\longmapsto 1 - S_K^n, \end{aligned}$$

$$\begin{aligned} G_K :\mathbb{R}^3 &\to \mathbb{R} \\ \boldsymbol{U}_K^n &\longmapsto H\left[P_K^n + P_{\mathrm{cp}}(S_K^n)\right] - \beta^{\mathrm{l}}\chi_K^n. \end{aligned}$$

Then, the finite volume scheme correponding to (3.15) reads: for all $1 \leq n \leq N_{\mathrm{t}}$ and all $K \in \mathcal{T}_h$

$$\begin{aligned} H_{c,K}^n(\boldsymbol{U}^n) &= 0 \quad \forall c \in \mathcal{C}, \\ F_K(\boldsymbol{U}_K^n) &\geq 0,\ G_K(\boldsymbol{U}_K^n) \geq 0,\ F_K(\boldsymbol{U}_K^n)\cdot G_K(\boldsymbol{U}_K^n) = 0. \end{aligned} \tag{3.31}$$

Observe that system (3.31) is written elementwise. We define the global version of the first $2N_{\mathrm{sp}}$ lines of system (3.31) by

$$\mathcal{H}^n(\boldsymbol{U}^n) = 0 \quad \text{where} \quad \mathcal{H}^n : \mathbb{R}^{3N_{\mathrm{sp}}} \to \mathbb{R}^{2N_{\mathrm{sp}}} \tag{3.32}$$

is define over K by the first line of (3.31). Inexact semismooth Newton methods will be employed to solve (3.31), as we detail in the next section.

3.4 Inexact semismooth Newton method

We detail in this section a semismooth Newton linearization associated to (3.31). We proceed in several steps. First, we briefly present the class of C-functions and the concept of semismoothness. Next, we give the linearization of (3.32) at each semismooth step.

3.4.1 C-functions

Definition 3.4.1. *A function* $f : \mathbb{R}^{N_{\text{sp}}} \times \mathbb{R}^{N_{\text{sp}}} \to \mathbb{R}^{N_{\text{sp}}}$ *is a complementarity function or (C-function) if* $\forall (\boldsymbol{a}, \boldsymbol{b}) \in \mathbb{R}^{N_{\text{sp}}} \times \mathbb{R}^{N_{\text{sp}}}$

$$f(\boldsymbol{a}, \boldsymbol{b}) = 0 \quad \iff \quad \boldsymbol{a} \geq 0, \quad \boldsymbol{b} \geq 0, \quad \boldsymbol{a} \cdot \boldsymbol{b} = 0.$$

Examples of C-functions are the minimum (min) function

$$(\min\{\boldsymbol{a}, \boldsymbol{b}\})_l = \min\{\boldsymbol{a}_l, \boldsymbol{b}_l\} \quad l = 1, \ldots, N_{\text{sp}}, \tag{3.33}$$

and the Fischer–Burmeister function

$$(f_{\text{FB}}(\boldsymbol{a}, \boldsymbol{b}))_l = \sqrt{\boldsymbol{a}_l^2 + \boldsymbol{b}_l^2} - (\boldsymbol{a}_l + \boldsymbol{b}_l) \quad l = 1, \ldots, N_{\text{sp}}. \tag{3.34}$$

For a direct application of the min function see [24, 62] and for more general details on C-functions see [88, 89]. For $1 \leq n \leq N_{\text{t}}$ let C^n be any C-function satisfying

$$\begin{aligned} & C^n\left((F_K(\boldsymbol{U}_K^n))_{K \in \mathcal{T}_h}, (G_K(\boldsymbol{U}_K^n))_{K \in \mathcal{T}_h}\right) = 0 \\ & \iff F_K(\boldsymbol{U}_K^n) \geq 0, \; G_K(\boldsymbol{U}_K^n) \geq 0, \; F_K(\boldsymbol{U}_K^n) \cdot G_K(\boldsymbol{U}_K^n) = 0 \\ & \forall K \in \mathcal{T}_h \end{aligned}$$

Introducing the function $\mathcal{C}^n : \mathbb{R}^{3N_{\text{sp}}} \to \mathbb{R}^{N_{\text{sp}}}$ defined as

$$\mathcal{C}^n(\boldsymbol{U}^n) = C^n\left((F_K(\boldsymbol{U}_K^n))_{K \in \mathcal{T}_h}, (G_K(\boldsymbol{U}_K^n))_{K \in \mathcal{T}_h}\right), \tag{3.35}$$

problem (3.31) then reads, for all $1 \leq n \leq N_{\text{t}}$

$$\begin{aligned} \mathcal{H}^n(\boldsymbol{U}^n) &= 0, \\ \mathcal{C}^n(\boldsymbol{U}^n) &= 0. \end{aligned} \tag{3.36}$$

The disadvantage of introducing the C-function is that the problem would no longer be $\mathcal{C}^1$, thus causing problems for the local quadratic convergence of the Newton algorithm. Nevertheless, the C-functions that are commonly used are locally Lipschitz and continuous, thus differentiable almost everywhere as a result of the Rademacher Theorem (see [58, 88]). More precisely they belong to the class of strong semismooth functions. Then it is possible (see [24, 29, 88, 89]) to build a semismooth Newton scheme.

3.4.2 Inexact semismooth Newton method

For $1 \leq n \leq N_{\text{t}}$ and $\boldsymbol{U}^{n,0} \in \mathbb{R}^{3N_{\text{sp}}}$ fixed (typically $\boldsymbol{U}^{n,0} = \boldsymbol{U}^{n-1}$), the semismooth Newton algorithm generates a sequence $(\boldsymbol{U}^{n,k})_{k \geq 1}$, with $\boldsymbol{U}^{n,k} \in \mathbb{R}^{3N_{\text{sp}}}$ given by the system of linear algebraic equations:

$$\mathbb{A}^{n,k-1} \boldsymbol{U}^{n,k} = \boldsymbol{B}^{n,k-1}, \tag{3.37}$$

where the Jacobian matrix $\mathbb{A}^{n,k-1} \in \mathbb{R}^{3N_{\text{sp}}, 3N_{\text{sp}}}$ and the right hand side vector $\boldsymbol{B}^{n,k-1} \in \mathbb{R}^{3N_{\text{sp}}}$ are defined by

$$\mathbb{A}^{n,k-1} := \begin{bmatrix} \mathbf{J}_{\mathcal{H}^n}(\boldsymbol{U}^{n,k-1}) \\ \mathbf{J}_{\mathcal{C}^n}(\boldsymbol{U}^{n,k-1}) \end{bmatrix}, \tag{3.38}$$

$$\boldsymbol{B}^{n,k-1} := \begin{bmatrix} \mathbf{J}_{\mathcal{H}^n}(\boldsymbol{U}^{n,k-1})\boldsymbol{U}^{n,k-1} - \mathcal{H}^n(\boldsymbol{U}^{n,k-1}) \\ \mathbf{J}_{\mathcal{C}^n}(\boldsymbol{U}^{n,k-1})\boldsymbol{U}^{n,k-1} - \mathcal{C}^n(\boldsymbol{U}^{n,k-1}) \end{bmatrix}. \tag{3.39}$$

Note that here the $3N_{\rm sp}$ lines of (3.30) are nonlinear and the semismooth linearity occurs in the last $N_{\rm sp}$ lines.
Here $\mathbf{J}_{\mathcal{H}^n}(\boldsymbol{U}^{n,k-1})$ is the Jacobian matrix of the function $\mathcal{H}^n$ at point $\boldsymbol{U}^{n,k-1}$ obtained by a Newton linearization; and $\mathbf{J}_{\mathcal{C}^n}(\boldsymbol{U}^{n,k-1})$ is the Jacobian matrix of the semismooth function "in the sense of Clarke", see [24, 29, 58, 88, 89]. For example, if we consider the semismooth function min of (3.33) and if we denote by $\boldsymbol{Y}$ the vector whose each component is defined by $\boldsymbol{Y}_l := HP'_{\rm cp}(S^{n,k-1}_{K_l})$ for $1 \le l \le N_{\rm sp}$ and if we define by $\mathbb{K}$ and $\mathbb{L}$ the matrices by

$$\begin{aligned} \mathbb{K} &:= \left[-\mathbf{Id}_{N_{\rm sp}\times N_{\rm sp}}, \mathbf{0}_{N_{\rm sp}\times N_{\rm sp}}, \mathbf{0}_{N_{\rm sp}\times N_{\rm sp}}\right], \\ \mathbb{L} &:= \left[\mathrm{diag}\boldsymbol{Y}_{N_{\rm sp}\times N_{\rm sp}}, H \times \mathbf{Id}_{N_{\rm sp}\times N_{\rm sp}}, -\beta^{\rm l} \times \mathbf{Id}_{N_{\rm sp}\times N_{\rm sp}}\right], \end{aligned}$$

then, the $l^{\rm th}$ row of the matrix $\mathbf{J}_{\mathcal{C}^n}(\boldsymbol{U}^{n,k-1})$ is either given by the $l^{\rm th}$ row of $\mathbb{K}$ if

$$1 - S^{n,k-1}_{K_l} \le H\left[P^{n,k-1}_{K_l} + P_{\rm cp}(S^{n,k-1}_{K_l})\right] - \beta^{\rm l}\chi^{n,k-1}_{K_l},$$

or by the $l^{\rm th}$ line of $\mathbb{L}$ if

$$H\left[P^{n,k-1}_{K_l} + P_{\rm cp}(S^{n,k-1}_{K_l})\right] - \beta^{\rm l}\chi^{n,k-1}_{K_l} < 1 - S^{n,k-1}_{K_l}.$$

Next, the approximate solution to (3.37) is obtained using an iterative algebraic solver. For $1 \le n \le N_{\rm t}$, a fixed semismooth Newton step $k \ge 1$, and an initial guess $\boldsymbol{U}^{n,k,0}$ (usually, $\boldsymbol{U}^{n,k,0} = \boldsymbol{U}^{n,k-1}$) the iterative algebraic solver generates a sequence $(\boldsymbol{U}^{n,k,i})_{i\ge 0}$ satisfying

$$\mathbb{A}^{n,k-1}\boldsymbol{U}^{n,k,i} = \boldsymbol{B}^{n,k-1} - \boldsymbol{R}^{n,k,i} \tag{3.40}$$

where $\boldsymbol{R}^{n,k,i} \in \mathbb{R}^{3N_{\rm sp}}$ is the algebraic residual vector. Below, it will be convenient to use the detailed form of the first two equations of (3.40):

$$\frac{|K|}{\tau_n}\left[l_{c,K}\left(\boldsymbol{U}^{n,k-1}\right) - l^{n-1}_{c,K} + \mathcal{L}^{n,k,i}_{c,K}\right] + \sum_{\sigma\in\mathcal{E}^{\rm int}_K} \mathcal{F}^{n,k,i}_{c,K,\sigma} - |K|Q^n_{c,K} + \boldsymbol{R}^{n,k,i}_{c,K} = 0, \quad \forall K \in \mathcal{T}_h \tag{3.41}$$

with the linear perturbation in the accumulation defined by

$$\mathcal{L}^{n,k,i}_{c,K} := \sum_{K'\in\mathcal{T}_h} \frac{\partial l^n_{c,K}}{\partial \boldsymbol{U}_{K'}}(\boldsymbol{U}^{n,k-1})\left[\boldsymbol{U}^{n,k,i}_{K'} - \boldsymbol{U}^{n,k-1}_{K'}\right],$$

and the linearized component flux by

$$\mathcal{F}^{n,k,i}_{c,K,\sigma} := \sum_{K'\in\mathcal{T}_h} \frac{\partial F_{c,K,\sigma}}{\partial \boldsymbol{U}_{K'}}\left(\boldsymbol{U}^{n,k-1}\right)\left[\boldsymbol{U}^{n,k,i}_{K'} - \boldsymbol{U}^{n,k-1}_{K'}\right] + F_{c,K,\sigma}\left(\boldsymbol{U}^{n,k-1}\right). \tag{3.42}$$

3.5 A posteriori error estimates

3.5.1 Preamble

In this section we establish an a posteriori error estimate between the exact solution and its approximate numerical solution at each semismooth Newton step $k \geq 1$ and each linear algebraic step $i \geq 0$. We start by giving some additional generic notations. Concerning the discrete unknowns, as we employed the cell-centered finite volume method, for each time step $0 \leq n \leq N_{\mathrm{t}}$ and for each $k \geq 1$ and $i \geq 0$, the discrete liquid pressure as well as the discrete liquid saturation and the discrete molar fraction of liquid hydrogen are piecewise constant in space. To carry out properly the a posteriori analysis, the discrete pressures and the discrete molar fraction of liquid hydrogen should belong to $H^1(\Omega)$ which is not the case as they are discontinuous at the cell interfaces. Therefore we assume that, from the constant finite volume unknowns, we have constructed discontinuous piecewise quadratic-in-space functions $P_h^{n,k,i} \in \mathbb{P}_2^{\mathrm{d}}(\mathcal{T}_h)$ (liquid pressure), $\chi_h^{n,k,i} \in \mathbb{P}_2^{\mathrm{d}}(\mathcal{T}_h)$ (molar fraction). We will also employ continuous piecewise quadratic functions $\tilde{P}_h^{n,k,i} \in \mathbb{P}_2^{\mathrm{c}}(\mathcal{T}_h)$, and $\tilde{\chi}_h^{n,k,i} \in \mathbb{P}_2^{\mathrm{c}}(\mathcal{T}_h)$. As an intermediate of computation we will also need to construct a discontinuous piecewise quadratic-in-space gas pressure function $P_h^{\mathrm{g},n,k,i}$, see Section 3.5.3. The saturation and thus the amount of water and hydrogen are defined in $\mathbb{P}_0^{\mathrm{d}}(\mathcal{T}_h)$ by

$$S_h^{n,k,i}|_K(\boldsymbol{x}) = S_K^{n,k,i}, \quad l_{\mathrm{w},h}^{n,k,i}|_K(\boldsymbol{x}) = l_{\mathrm{w},K}^{n,k,i}, \quad l_{\mathrm{h},h}^{n,k,i}|_K = l_{\mathrm{h},K}^{n,k,i}, \quad \forall K \in \mathcal{T}_h. \tag{3.43}$$

From the above space functions, we define the space-time functions as continuous and piecewise affine in time (*i.e.* in $\mathbb{P}_1^{\mathrm{c}}(0,t_{\mathrm{F}})$) by

$$\begin{aligned} &P_{h\tau}^{n,k,i}(t^n) = P_h^{n,k,i},\ \tilde{P}_{h\tau}^{n,k,i}(t^n) = \tilde{P}_h^{n,k,i},\ l_{c,h\tau}^{n,k,i}(t^n) = l_{c,h}^{n,k,i}, \\ &S_{h\tau}^{n,k,i}(t^n) = S_h^{n,k,i},\ \chi_{h\tau}^{n,k,i}(t^n) = \chi_h^{n,k,i},\ \tilde{\chi}_{h\tau}^{n,k,i}(t^n) = \tilde{\chi}_h^{n,k,i}. \end{aligned} \tag{3.44}$$

Concerning the source terms, we define the space-time function $Q_{c,h\tau}$ such that $(Q_{c,h\tau})|_{K\times I_n} = Q_{c,K}^n$, thus piecewise constant in time and in space. To finish we assume that the initial condition (3.20) holds. For the a posteriori analysis, the goal would be to find an upper bound of the form:

$$\left\|P^{\mathrm{l}} - P_{h\tau}^{n,k,i}\right\|_\sharp + \left\|S^{\mathrm{l}} - S_{h\tau}^{n,k,i}\right\|_\sharp + \left\|\chi_{\mathrm{h}}^{\mathrm{l}} - \chi_{h\tau}^{n,k,i}\right\|_\sharp \leq \eta,$$

with $\|\cdot\|_\sharp$ some norm and η only depending on the approximate solution. This kind of estimate has to our knowledge not been established for compositional multiphase flow. In the litterature, such an a posteriori error estimate has been derived for a two-phase flow with one component per phase, see [44]. We thus follow the methodology proposed in [70, 158] by considering some dual norm of the residual. We first start by defining appropriate spaces for the unknowns. Let X, Y, $\widehat{Y}$ be the spaces and let Z be the set defined by:

$$\begin{aligned} X &:= L^2((0,t_{\mathrm{F}}); H^1(\Omega)), \\ Y &:= H^1((0,t_{\mathrm{F}}); L^2(\Omega)), \\ \widehat{Y} &:= H^1((0,t_{\mathrm{F}}); L^\infty(\Omega)), \\ Z &:= \left\{v \in L^2((0,t_{\mathrm{F}}); L^\infty(\Omega)),\ v \geq 0 \text{ on } \Omega \times (0,t_{\mathrm{F}})\right\}. \end{aligned}$$

We denote by X_n the restriction of the energy space X to the time interval I_n, $X_n := L^2(I_n; H^1(\Omega))$.We equip the spaces X and X_n with the norms

$$\|\varphi\|_X := \left\{ \sum_{n=1}^{N_t} \|\varphi\|_{X_n}^2 \, \mathrm{dt} \right\}^{\frac{1}{2}}, \quad \|\varphi\|_{X_n} := \left\{ \int_{I_n} \sum_{K \in \mathcal{T}_h} \|\varphi\|_{X,K}^2 \, \mathrm{dt} \right\}^{\frac{1}{2}}, \tag{3.45}$$

with $\|\varphi\|_{X,K}^2 := \varepsilon h_K^{-2} \|\varphi\|_K^2 + \|\boldsymbol{\nabla}\varphi\|_K^2$.

Note that $\varepsilon = 0$ is to be chosen when homogeneous Dirichlet conditions are prescribed on the boundary $\partial\Omega$, whereas $\varepsilon > 0$ enables to take into account Neumann boundary conditions. Then h_K^{-2} is a scaling term.

3.5.2 Weak solution

Let $Q_c \in L^2((0,t_F); L^2(\Omega)) \ \forall c \in \mathcal{C}$. We assume that there exists a unique weak solution satisfying:

Assumption 3.5.1.

$$S^l \in \widehat{Y}, \quad 1 - S^l \in Z, \quad l_w \in Y, \quad l_h \in Y, \tag{3.46}$$

$$P^l \in X, \quad \chi_h^l \in X, \tag{3.47}$$

$$\boldsymbol{\Phi}_c \in [L^2((0,t_F); \mathbf{H}(\mathrm{div}, \Omega))]^2 \quad \forall c \in \mathcal{C}, \tag{3.48}$$

$$\int_0^{t_F} (\partial_t l_c, \varphi)_\Omega (t) \, \mathrm{dt} - \int_0^{t_F} (\boldsymbol{\Phi}_c, \boldsymbol{\nabla}\varphi)_\Omega (t) \, \mathrm{dt} = \int_0^{t_F} (Q_c, \varphi)_\Omega (t) \, \mathrm{dt} \quad \forall \varphi \in X \quad \forall c \in \mathcal{C} \tag{3.49}$$

$$\int_0^{t_F} \left(\lambda - \left(1 - S^l\right), H[P^l + P_{cp}(S^l)] - \beta^l \chi_h^l \right)_\Omega (t) \, \mathrm{dt} \geq 0 \quad \forall \lambda \in Z, \tag{3.50}$$

the initial condition (3.20) *holds where* l_c *and* $\boldsymbol{\Phi}_c$ *are defined by* (3.18), (3.16), *and* (3.17). (3.51)

Proposition 3.5.2. *Under assumptions* (3.46), (3.47), *and* (3.48), *the two first conservation equations given by the strong formulation* (3.15) *are equivalent to the weak formulation given by* (3.49). *Furthermore, the nonlinear complementarity conditions given by the third line of* (3.15) *are equivalent to* (3.50).

Proof. Suppose first that the strong formulation (3.15) holds. Let c be any component from the set $\mathcal{C}$ and let φ be a test function belonging to the space X. As $l_c \in Y$ and $\boldsymbol{\Phi}_c \in [L^2((0,t_F); \mathbf{H}(\mathrm{div}, \Omega))]^2$ the Green formula combined with the homogeneous Neumann condition (3.19) gives

$$\int_0^{t_F} (\partial_t l_c, \varphi)_\Omega (t) \, \mathrm{dt} - \int_0^{t_F} (\boldsymbol{\Phi}_c, \boldsymbol{\nabla}\varphi)_\Omega (t) \, \mathrm{dt} = \int_0^{t_F} (Q_c, \varphi)_\Omega (t) \, \mathrm{dt}.$$

Conversely, if (3.49) is satisfied, we have by doing an integration by parts

$$\int_0^{t_F} (\partial_t l_c + \boldsymbol{\nabla} \cdot \boldsymbol{\Phi}_c - Q_c, \varphi)_\Omega (t) \, \mathrm{dt} = 0 \quad \forall \varphi \in X.$$

Then, as $\varphi \in X$ and $\partial_t l_c + \boldsymbol{\nabla}\cdot\boldsymbol{\Phi}_c - Q_c \in L^2((0,t_{\rm F}),L^2(\Omega))$ it yields (see [37, Lemma 8.1])

$$\partial_t l_c + \boldsymbol{\nabla}\cdot\boldsymbol{\Phi}_c - Q_c = 0.$$

Suppose that the third line of the strong formulation (3.15) holds. Let $\lambda \in Z$. We have $1 - S^{\rm l} \in Z$ and

$$\begin{aligned}&\int_0^{t_{\rm F}} \left(\lambda - \left(1-S^{\rm l}\right), H[P^{\rm l} + P_{\rm cp}(S^{\rm l})] - \beta^{\rm l}\chi^{\rm l}_{\rm h}\right)_\Omega (t)\,{\rm dt}\\ &= \int_0^{t_{\rm F}} \left(\lambda, H\left[P^{\rm l} + P_{\rm cp}(S^{\rm l})\right] - \beta^{\rm l}\chi^{\rm l}_{\rm h}\right)_\Omega (t)\,{\rm dt} \geq 0.\end{aligned}$$

Conversely, suppose that the assumption (3.50) is satisfied. For $\lambda = 0 \in Z$ we have

$$\int_0^{t_{\rm F}} \left(1 - S^{\rm l}, H[P^{\rm l} + P_{\rm cp}(S^{\rm l})] - \beta^{\rm l}\chi^{\rm l}_{\rm h}\right)_\Omega (t)\,{\rm dt} \leq 0. \tag{3.52}$$

Next, for $\lambda(x,t) = 1 - S^{\rm l}(x,t) + \mathbb{1}_{\mathcal{O}\times[t-\zeta,t+\zeta]}$ where $\zeta > 0$ and $\mathcal{O}$ is any measurable subset of Ω we have $\lambda \in Z$ as $1 - S^{\rm l} \in Z$; thus

$$H[P^{\rm l}(x,t) + P_{\rm cp}(S^{\rm l}(x,t))] - \beta^{\rm l}\chi^{\rm l}_{\rm h}(x,t) \geq 0. \tag{3.53}$$

Therefore, combining (3.52), (3.53) and the assumption $1 - S^{\rm l} \geq 0$, we get

$$\left[1 - S^{\rm l}\right] \cdot \left[H\left[P^{\rm l} + P_{\rm cp}(S^{\rm l})\right] - \beta^{\rm l}\chi^{\rm l}_{\rm h}\right] = 0.$$

□

3.5.3 Error measure

As discussed in Preamble 3.5.1, the natural choice is to consider an error measure constructed from the dual norm of a residual supplemented by the nonconformity of the liquid pressure and the molar fraction of liquid hydrogen following [70] and the references therein. As we treat the phase transitions, we also have to add a term checking the complementarity constraints.

Definition 3.5.3. *For the discrete approximations $P^{n,k,i}_{h\tau}$ and $\chi^{n,k,i}_{h\tau}$ belonging to $L^2(I_n;H^1(\mathcal{T}_h))$ to be defined later in Section 3.5.5 and $S^{n,k,i}_{h\tau}$ given by (3.43)–(3.44), the residual associated to assumption (3.49) is defined for any $\varphi \in X_n$ by*

$$\langle \mathcal{R}_c(S^{n,k,i}_{h\tau}, P^{n,k,i}_{h\tau}, \chi^{n,k,i}_{h\tau}), \varphi\rangle_{X_n',X_n} := \int_{I_n} \left\{\left(Q_c - \partial_t l^{n,k,i}_{c,h\tau}, \varphi\right)_\Omega + \left(\boldsymbol{\Phi}^{n,k,i}_{c,h\tau}, \boldsymbol{\nabla}\varphi\right)_\Omega\right\}(t)\,{\rm dt}, \tag{3.54}$$

and its dual norm is defined by

$$\left\|\mathcal{R}_c(S^{n,k,i}_{h\tau}, P^{n,k,i}_{h\tau}, \chi^{n,k,i}_{h\tau})\right\|_{X_n'} := \sup_{\varphi\in X_n, \|\varphi\|_{X_n}=1} \langle \mathcal{R}_c(S^{n,k,i}_{h\tau}, P^{n,k,i}_{h\tau}, \chi^{n,k,i}_{h\tau}), \varphi\rangle_{X_n',X_n},$$

where $\boldsymbol{\Phi}^{n,k,i}_{c,h\tau}$, $c \in \mathcal{C}$, are the discrete fluxes corresponding to (3.16) and (3.17) defined by

$$\begin{aligned}\boldsymbol{\Phi}^{n,k,i}_{{\rm w},h\tau} &:= \rho^{\rm l}_{\rm w}\mathbf{q}^{n,k,i}_{h\tau} - \mathbf{J}^{n,k,i}_{{\rm h},h\tau},\\ \boldsymbol{\Phi}^{n,k,i}_{{\rm h},h\tau} &:= \beta^{\rm l}\chi^{n,k,i}_{h\tau}\mathbf{q}^{n,k,i}_{h\tau} + \beta^{\rm g}\left[P_{\rm cp}(S^{n,k,i}_{h\tau}) + P^{n,k,i}_{h\tau}\right]\mathbf{q}^{{\rm g},n,k,i}_{h\tau} + \mathbf{J}^{n,k,i}_{{\rm h},h\tau},\end{aligned}$$

where the discrete Darcy space-time vectorial functions $\mathbf{q}_{h\tau}^{n,k,i}$ *and* $\mathbf{q}_{h\tau}^{\mathrm{g},n,k,i}$ *and the discrete liquid Fick space-time vectorial function* $\mathbf{J}_{\mathrm{h},h\tau}^{n,k,i}$ *are defined by*

$$\begin{aligned}
\mathbf{q}_{h\tau}^{n,k,i} &:= -\underline{\mathbf{K}}\frac{k_{\mathrm{r}}^{\mathrm{l}}(S_{h\tau}^{n,k,i})}{\mu^{\mathrm{l}}}\left[\boldsymbol{\nabla} P_{h\tau}^{n,k,i} - \left[\rho_{\mathrm{w}}^{\mathrm{l}} + \beta^{\mathrm{l}}\chi_{h\tau}^{n,k,i}\right] g\boldsymbol{\nabla} z\right],\\
\mathbf{q}_{h\tau}^{\mathrm{g},n,k,i} &:= -\underline{\mathbf{K}}\frac{k_{\mathrm{r}}^{\mathrm{g}}(1-S_{h\tau}^{n,k,i})}{\mu^{\mathrm{g}}}\left[\boldsymbol{\nabla} P_{h\tau}^{\mathrm{g},n,k,i} - \beta^{\mathrm{g}} P_{h\tau}^{\mathrm{g},n,k,i} g\boldsymbol{\nabla} z\right],\\
\mathbf{J}_{\mathrm{h},h\tau}^{n,k,i} &:= -\phi M_{\mathrm{h}} S_{h\tau}^{n,k,i}\left[\frac{\rho_{\mathrm{w}}^{\mathrm{l}}}{M_{\mathrm{w}}} + \frac{\beta^{\mathrm{l}}}{M_{\mathrm{h}}}\chi_{h\tau}^{n,k,i}\right] D_{\mathrm{h}}^{\mathrm{l}}\boldsymbol{\nabla}\chi_{h\tau}^{n,k,i},
\end{aligned}$$

where the space-time function $P_{h\tau}^{\mathrm{g},n,k,i}$ *is built from* $P_{h\tau}^{n,k,i}$ *and* $S_{h\tau}^{n,k,i}$ *in Section 3.5.5 below. Furthermore, we define the residual equation associated to assumption (3.50) as*

$$\mathcal{R}_{\mathrm{e}}(S_{h\tau}^{n,k,i}, P_{h\tau}^{n,k,i}, \chi_{h\tau}^{n,k,i}) := \frac{1}{\alpha}\left[\int_{I_n}\left(1 - S_{h\tau}^{n,k,i}, H\left[P_{h\tau}^{n,k,i} + P_{\mathrm{cp}}(S_{h\tau}^{n,k,i})\right] - \beta^{\mathrm{l}}\chi_{h\tau}^{n,k,i}\right)_{\Omega}(t)\,\mathrm{dt}\right],$$

with $\alpha > 0$ *a rescaling constant.*

We define our error measure by

$$\begin{aligned}
\mathcal{N}^{n,k,i} := &\left\{\sum_{c\in\mathcal{C}}\left\|\mathcal{R}_c(S_{h\tau}^{n,k,i}, P_{h\tau}^{n,k,i}, \chi_{h\tau}^{n,k,i})\right\|_{X_n'}^2\right\}^{\frac{1}{2}} + \left\{\left[\mathcal{N}_P^{n,k,i}(P_{h\tau}^{n,k,i})\right]^2 + \left[\mathcal{N}_\chi^{n,k,i}(\chi_{h\tau}^{n,k,i})\right]^2\right\}^{\frac{1}{2}}\\
&+ \mathcal{R}_{\mathrm{e}}(S_{h\tau}^{n,k,i}, P_{h\tau}^{n,k,i}, \chi_{h\tau}^{n,k,i}),
\end{aligned} \tag{3.55}$$

with

$$\mathcal{N}_P^{n,k,i}(P_{h\tau}^{n,k,i}) := \inf_{\delta_{\mathrm{l}}\in X_n}\left\{\sum_{c\in\mathcal{C}^{\mathrm{l}}}\int_{I_n}\left\|\boldsymbol{\Upsilon}_{\mathrm{l},c}(P_{h\tau}^{n,k,i})(t) - \boldsymbol{\Upsilon}_{\mathrm{l},c}(\delta_{\mathrm{l}})(t)\right\|^2\,\mathrm{dt}\right\}^{\frac{1}{2}}, \tag{3.56}$$

$$\mathcal{N}_\chi^{n,k,i}(\chi_{h\tau}^{n,k,i}) := \inf_{\theta\in X_n}\left\{\int_{I_n}\left\|\boldsymbol{\Psi}(\chi_{h\tau}^{n,k,i})(t) - \boldsymbol{\Psi}(\theta)(t)\right\|^2\,\mathrm{dt}\right\}^{\frac{1}{2}}, \tag{3.57}$$

where the function $\boldsymbol{\Upsilon}_{\mathrm{l},c}$ is defined by

$$\begin{aligned}
\boldsymbol{\Upsilon}_{\mathrm{l,w}}(\varphi) &:= -\underline{\mathbf{K}}\frac{k_{\mathrm{r}}^{\mathrm{l}}(S_{h\tau}^{n,k,i})}{\mu^{\mathrm{l}}}\rho_{\mathrm{w}}^{\mathrm{l}}\boldsymbol{\nabla}\varphi \quad \forall\varphi\in L^2(I_n, H^1(\mathcal{T}_h))\\
\boldsymbol{\Upsilon}_{\mathrm{l,h}}(\varphi) &:= -\underline{\mathbf{K}}\frac{k_{\mathrm{r}}^{\mathrm{l}}(S_{h\tau}^{n,k,i})}{\mu^{\mathrm{l}}}\beta^{\mathrm{l}}\chi_{\mathrm{h}}^{\mathrm{l}}\boldsymbol{\nabla}\varphi \quad \forall\varphi\in L^2(I_n; H^1(\mathcal{T}_h)),
\end{aligned}$$

and the function $\boldsymbol{\Psi}$ defined by

$$\boldsymbol{\Psi}(\varphi) := -\phi M_{\mathrm{h}} S_{h\tau}^{n,k,i}\left[\frac{\rho_{\mathrm{w}}^{\mathrm{l}}}{M_{\mathrm{w}}} + \frac{\beta^{\mathrm{l}}}{M_{\mathrm{h}}}\chi_{h\tau}^{n,k,i}\right] D_{\mathrm{h}}^{\mathrm{l}}\boldsymbol{\nabla}\varphi \quad \forall\varphi\in L^2(I_n; H^1(\mathcal{T}_h)).$$

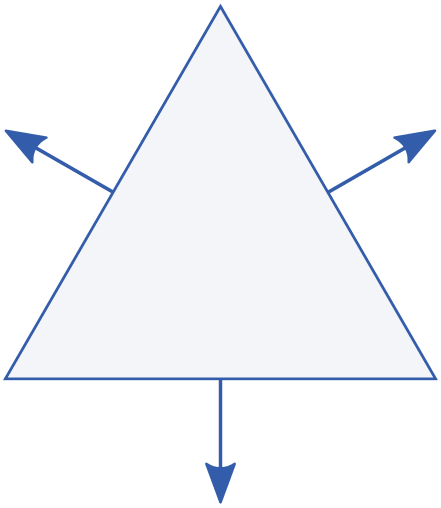

Figure 3.4: Degrees of freedom for the space $\mathbf{RT}_0$.

3.5.4 Equilibrated component flux reconstructions

Let $1 \leq n \leq N_\mathrm{t}$, a semismooth Newton linearization iteration $k \geq 1$, and an algebraic solver iteration $i \geq 0$ be fixed. We are interested in finding an upper bound for the error measure $\mathcal{N}^{n,k,i}$ defined in (3.55). To do so, we employ the methodology of the equilibrated flux reconstruction in the context of the cell-centered finite volume method [70,71,81]. The subspace of $\mathbf{H}(\mathrm{div},\Omega)$ we use in the sequel is the lowest-order Raviart–Thomas space, see Raviart and Thomas [142], or Roberts and Thomas [146], or Brezzi and Fortin [38] and is defined by

$$\mathbf{RT}_0(\Omega) := \{\boldsymbol{w}_h \in \mathbf{H}(\mathrm{div},\Omega), \boldsymbol{w}_h|_K \in \mathbf{RT}_0(K)\ \forall K \in \mathcal{T}_h\},$$
$$\mathbf{RT}_0(K) := [\mathbb{P}_0(K)]^2 + \boldsymbol{x}\cdot\mathbb{P}_0(K) \text{ with } \boldsymbol{x} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

For a function $\boldsymbol{v} \in \mathbf{RT}_0(K)$, we recall that its 3 degrees of freedom are given by $(\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma}, 1)_\sigma,\ \ \sigma \in \mathcal{E}_K$.

For all component $c \in \mathcal{C}$, for all $K \in \mathcal{T}_h$, and for all $\sigma \in \mathcal{E}_K^{\mathrm{int}}$ we can define from (3.22), (3.28), (3.41) the different component flux reconstructions in $\mathbf{RT}_0(\mathcal{T}_h)$, namely the discretization flux reconstruction $\boldsymbol{\Theta}_{c,h,\mathrm{disc}}^{n,k,i}$, the linearization flux reconstruction $\boldsymbol{\Theta}_{c,h,\mathrm{lin}}^{n,k,i}$, and the algebraic flux reconstruction $\boldsymbol{\Theta}_{c,h,\mathrm{alg}}^{n,k,i}$ as follows

$$\left(\boldsymbol{\Theta}_{c,h,\mathrm{disc}}^{n,k,i}\cdot\boldsymbol{n}_{K,\sigma}, 1\right)_\sigma := F_{c,K,\sigma}\left(\boldsymbol{U}^{n,k,i}\right), \tag{3.58}$$
$$\left(\boldsymbol{\Theta}_{c,h,\mathrm{lin}}^{n,k,i}\cdot\boldsymbol{n}_{K,\sigma}, 1\right)_\sigma := \mathcal{F}_{c,K,\sigma}^{n,k,i} - F_{c,K,\sigma}\left(\boldsymbol{U}^{n,k,i}\right), \tag{3.59}$$
$$\boldsymbol{\Theta}_{c,h,\mathrm{alg}}^{n,k,i,\nu} := \boldsymbol{\Theta}_{c,h,\mathrm{disc}}^{n,k,i+\nu} + \boldsymbol{\Theta}_{c,h,\mathrm{lin}}^{n,k,i+\nu} - \left(\boldsymbol{\Theta}_{c,h,\mathrm{disc}}^{n,k,i} + \boldsymbol{\Theta}_{c,h,\mathrm{lin}}^{n,k,i}\right) \tag{3.60}$$

with a fixed number $\nu > 0$ of additional algebraic iterations.

Here $F_{c,K,\sigma}$ is defined by (3.23) or (3.29), and $\mathcal{F}_{c,K,\sigma}^{n,k,i}$ is defined by (3.42). For the boundary conditions we set $\boldsymbol{\Theta}_{c,h,\mathrm{disc}}^{n,k,i}\cdot\boldsymbol{n}_{K,\sigma} = \boldsymbol{\Theta}_{c,h,\mathrm{lin}}^{n,k,i}\cdot\boldsymbol{n}_{K,\sigma} = \boldsymbol{\Theta}_{c,h,\mathrm{alg}}^{n,k,i,\nu}\cdot\boldsymbol{n}_{K,\sigma} = 0$ for

$\sigma \in \mathcal{E}_h^{\rm ext}$. Therefrom, we define $\forall c \in \mathcal{C}$, the total flux reconstruction $\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}$ by

$$\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu} = \boldsymbol{\Theta}_{c,h,\rm disc}^{n,k,i} + \boldsymbol{\Theta}_{c,h,\rm lin}^{n,k,i} + \boldsymbol{\Theta}_{c,h,\rm alg}^{n,k,i,\nu}. \tag{3.61}$$

Lemma 3.5.4. *The component fluxes* $\boldsymbol{\Theta}_{c,h,\rm disc}^{n,k,i}$, $\boldsymbol{\Theta}_{c,h,\rm lin}^{n,k,i}$, *and* $\boldsymbol{\Theta}_{c,h,\rm alg}^{n,k,i,\nu}$ *belong to* $\mathbf{H}(\mathrm{div},\Omega)$.

Proof. We prove the result only for the discretization flux $\boldsymbol{\Theta}_{c,h,\rm disc}^{n,k,i}$ as for the others it follows the same methodology. First, it is clear that $\boldsymbol{\Theta}_{c,h,\rm disc}^{n,k,i}|_K \in \mathbf{H}(\mathrm{div},K)\ \forall K \in \mathcal{T}_h$. Let $\sigma \in \mathcal{E}_h^{\rm int}$, $\overline{\sigma} = \overline{K} \cap \overline{L}$. Defining the jump of any vector $\boldsymbol{v}$ by

$$[\![\boldsymbol{v}]\!]_\sigma := (\boldsymbol{v}|_K)\,|_\sigma - (\boldsymbol{v}|_L)\,|_\sigma$$

we then have,

$$\begin{aligned}
[\![\boldsymbol{\Theta}_{c,h,\rm disc}^{n,k,i}]\!] \cdot \boldsymbol{n}_{K,\sigma} &= \left(\boldsymbol{\Theta}_{c,h,\rm disc}^{n,k,i}|_K\right)|_\sigma \cdot \boldsymbol{n}_{K,\sigma} + \left(\boldsymbol{\Theta}_{c,h,\rm disc}^{n,k,i}|_L\right)|_\sigma \cdot \boldsymbol{n}_{L,\sigma}, \\
&= \frac{1}{|\sigma|} F_{c,K,\sigma}\left(\boldsymbol{U}^{n,k,i}\right) + \frac{1}{|\sigma|} F_{c,L,\sigma}\left(\boldsymbol{U}^{n,k,i}\right), \\
&= 0 \text{ as the finite volume method is conservative.}
\end{aligned}$$

Thus, $\boldsymbol{\Theta}_{c,h,\rm disc}^{n,k,i} \in \mathbf{H}(\mathrm{div},\Omega)$. □

Note that it is possible in practice to change the definition (3.60), see [136] for a reconstruction based on a multigrid structure. We have,

Proposition 3.5.5. *Let* $1 \le n \le N_{\rm t}$, *a semismooth Newton iteration* $k \ge 1$, *and an algebraic solver iteration* $i \ge 0$ *be fixed and* $\nu > 0$. *For all* $c \in \mathcal{C}$ *and for all* $K \in \mathcal{T}_h$ *there holds,*

$$\left(Q_{c,K}^n - \frac{l_{c,K}(\boldsymbol{U}^{n,k-1}) - l_{c,K}^{n-1} + \mathcal{L}_{c,K}^{n,k,i+\nu}}{\tau_n} - \boldsymbol{\nabla}\cdot\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}, 1\right)_K = \boldsymbol{R}_{c,K}^{n,k,i+\nu}. \tag{3.62}$$

Proof. Employing the definition of the total fluxes (3.61), the definition of the component fluxes (3.58)–(3.60), and the Green formula we get

$$\left(-\boldsymbol{\nabla}\cdot\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}, 1\right)_K = -\sum_{\sigma\in\mathcal{E}_K^{\rm int}} \mathcal{F}_{c,K,\sigma}^{n,k,i+\nu}.$$

Thus, equation (3.41) at iterate $i+\nu$ yields the desired result. □

3.5.5 Phase pressure and molar fraction reconstructions

We present in this section the construction from the finite volume unknowns of the discontinuous quadratic liquid pressure and molar fraction of liquid hydrogen and next their continuous quadratic interpolant so as to preserve the physical properties imposed by the problem.

Let $1 \leq n \leq N_{\rm t}$, we define $\left(\boldsymbol{\xi}_h^{n,k,i}, \boldsymbol{\xi}_h^{{\rm g},n,k,i}\right) \in \mathbf{RT}_0(\mathcal{T}_h) \times \mathbf{RT}_0(\mathcal{T}_h)$ such that $\forall K \in \mathcal{T}_h$ and $\forall \sigma \in \mathcal{E}_K^{\rm int}$ such that $\overline{\sigma} = \overline{K} \cap \overline{L}$

$$\left(\boldsymbol{\xi}_h^{n,k,i} \cdot \boldsymbol{n}_{K,\sigma}, 1\right)_\sigma := -|\sigma| \frac{P_L^{n,k,i} - P_K^{n,k,i}}{d_{KL}},$$

$$\left(\boldsymbol{\xi}_h^{{\rm g},n,k,i} \cdot \boldsymbol{n}_K, 1\right)_\sigma := -|\sigma| \frac{P_L^{{\rm g},n,k,i} - P_K^{{\rm g},n,k,i}}{d_{KL}},$$

with

$$P_K^{{\rm g},n,k,i} = P_K^{n,k,i} + P_{\rm cp}(S_K^{n,k,i}).$$

The discontinuous piecewise quadratic liquid phase pressure $P_h^{n,k,i} \in \mathbb{P}_2^{\rm d}(\mathcal{T}_h)$ is such that $\forall K \in \mathcal{T}_h$

$$\left(-\boldsymbol{\nabla} P_h^{n,k,i}\right)|_K := \left(\boldsymbol{\xi}_h^{n,k,i}\right)_K \quad \text{and} \quad \frac{\left(P_h^{n,k,i}, 1\right)_K}{|K|} := P_K^{n,k,i}.$$

while the discontinuous quadratic gas phase pressure $P_h^{{\rm g},n,k,i} \in \mathbb{P}_2^{\rm d}(\mathcal{T}_h)$ satisfies $\forall K \in \mathcal{T}_h$

$$\left(-\boldsymbol{\nabla} P_h^{{\rm g},n,k,i}\right)|_K := \left(\boldsymbol{\xi}_h^{{\rm g},n,k,i}\right)_K, \text{ and } \frac{\left(P_h^{{\rm g},n,k,i}, 1\right)_K}{|K|} := P_K^{{\rm g},n,k,i}.$$

Untill now, we have transformed a constant in each cells onto a discontinuous $\mathbb{P}_2$ polynomial. This transformation unfortunately does not give the global continuity in space so that $P_h^{n,k,i}$ and $P_h^{{\rm g},n,k,i}$ do not belong to $H^1(\Omega)$. To do so, we use the Oswald interpolation operator, see [113, 156] that associates to the discontinuous piecewise polynomial $P_h^{n,k,i}$ its conforming interpolant.
Then, from $P_h^{n,k,i} \in \mathbb{P}_2^{\rm d}(\mathcal{T}_h)$, using the notations introduced at the beginning of Section 3.3.1, we define $\tilde{P}_h^{n,k,i} \in \mathbb{P}_2^{\rm c}(\mathcal{T}_h)$ by

$$\tilde{P}_h^{n,k,i}(\boldsymbol{a}) := \frac{1}{|\mathcal{T}_{\boldsymbol{a}}|} \sum_{K \in \mathcal{T}_{\boldsymbol{a}}} \left(P_h^{n,k,i}\right)|_K(\boldsymbol{a}) \quad \text{for} \quad \boldsymbol{a} \in D_2. \tag{3.63}$$

In the same way, we reconstruct a continuous $\mathbb{P}_2^{\rm c}(\mathcal{T}_h)$ molar fraction as follows. Let $1 \leq n \leq N_{\rm t}$, we define $\boldsymbol{\nu}_h^{n,k,i} \in \mathbf{RT}_0(\mathcal{T}_h)$ such that $\forall K \in \mathcal{T}_h$ and $\forall \sigma \in \mathcal{E}_K^{\rm int}$ such that $\overline{\sigma} = \overline{K} \cap \overline{L}$,

$$\left(\boldsymbol{\nu}_h^{n,k,i} \cdot \boldsymbol{n}_K, 1\right)_\sigma := -|\sigma| \frac{\chi_L^{n,k,i} - \chi_K^{n,k,i}}{d_{KL}}.$$

The discontinuous quadratic molar fraction $\chi_h^{n,k,i}$ is such that $\forall K \in \mathcal{T}_h$,

$$\left(-\boldsymbol{\nabla} \chi_h^{n,k,i}\right)|_K := \left(\boldsymbol{\nu}_h^{n,k,i}\right)_K \quad \text{and} \quad \frac{\left(\chi_h^{n,k,i}, 1\right)_K}{|K|} := \chi_K^{n,k,i}.$$

From the discontinuous polynomial $\chi_h^{n,k,i} \in \mathbb{P}_2^{\rm d}(\mathcal{T}_h)$, we construct its conforming interpolant, using the Oswald interpolation operator as follows

$$\tilde{\chi}_h^{n,k,i}(\boldsymbol{a}) := \frac{1}{|\mathcal{T}_{\boldsymbol{a}}|} \sum_{K \in \mathcal{T}_{\boldsymbol{a}}} \left(\chi_h^{n,k,i}\right)|_K(\boldsymbol{a}) \quad \boldsymbol{a} \in D_2. \tag{3.64}$$

Remark 3.5.6. *The constructions* (3.63) *and* (3.64) *give* $\tilde{P}_h^{n,k,i}$ *and* $\tilde{\chi}_h^{n,k,i} \in H^1(\Omega)$.

3.5.6 A posteriori error estimates

In this section we provide an upper bound for the error measure defined in (3.55) at each semismooth step $k \geq 1$ and each algebraic iteration $i \geq 0$. An important difficulty is that during the iterations in i and k, the approximation is no more conforming in the sense that the conditions

$$1 - S_{h\tau}^{n,k,i} \geq 0, \quad H\left[P_{h\tau}^{n,k,i} + P_{\rm cp}(S_{h\tau}^{n,k,i})\right] - \beta^{\rm l}\chi_{h\tau}^{n,k,i} \geq 0,$$
$$\left[1 - S_{h\tau}^{n,k,i}\right] \cdot \left[H\left[P_{h\tau}^{n,k,i} + P_{\rm cp}(S_{h\tau}^{n,k,i})\right] - \beta^{\rm l}\chi_{h\tau}^{n,k,i}\right] = 0$$

do not necessarily hold. We define for all $c \in \mathcal{C}$ the estimators linked to the finite volume discretization

$$\eta_{{\rm R},K,c}^{n,k,i,\nu} := \min\left\{C_{\rm PW}, \varepsilon^{-\frac{1}{2}}\right\} h_K \times$$
$$\left\| Q_{c,K}^n - \frac{1}{\tau_n}\left[l_{c,K}(\boldsymbol{U}^{n,k-1}) - l_{c,K}^{n-1} + \mathcal{L}_{c,K}^{n,k,i+\nu}\right] - \frac{\boldsymbol{R}_{c,K}^{n,k,i+\nu}}{|K|} - \boldsymbol{\nabla}\cdot\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}\right\|_K, \tag{3.65}$$

$$\eta_{{\rm F},K,c}^{n,k,i,\nu}(t) := \left\|\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu} - \boldsymbol{\Phi}_{c,h\tau}^{n,k,i}(t)\right\|_K \quad t \in I_n, \tag{3.66}$$

the estimators linked to the nonconformity of the liquid pressure and the molar fraction of liquid hydrogen

$$\eta_{{\rm NC},K,{\rm l},c}^{n,k,i}(t) := \left\|\boldsymbol{\Upsilon}_{{\rm l},c}(P_{h\tau}^{n,k,i})(t) - \boldsymbol{\Upsilon}_{{\rm l},c}(\tilde{P}_{h\tau}^{n,k,i})(t)\right\|_K \quad t \in I_n,\ c \in \mathcal{C}, \tag{3.67}$$

$$\eta_{{\rm NC},K,\chi}^{n,k,i}(t) := \left\|\boldsymbol{\Psi}(\chi_{h\tau}^{n,k,i})(t) - \boldsymbol{\Psi}(\tilde{\chi}_{h\tau}^{n,k,i})(t)\right\|_K \quad t \in I_n, \tag{3.68}$$

and the estimators linked respectively to the semismooth linearization and linear algebra

$$\eta_{{\rm NA},K,c}^{n,k,i,\nu} := \varepsilon^{-\frac{1}{2}}\frac{h_K}{\tau_n}\left\|l_{c,K}(\boldsymbol{U}^{n,k,i}) - l_{c,K}(\boldsymbol{U}^{n,k-1}) - \mathcal{L}_{c,K}^{n,k,i+\nu}\right\|_K, \tag{3.69}$$

$$\eta_{{\rm rem},K,c}^{n,k,i,\nu} := h_K|K|^{-1}\varepsilon^{-\frac{1}{2}}\left\|\boldsymbol{R}_{c,K}^{n,k,i+\nu}\right\|_K. \tag{3.70}$$

The estimators defined previously reflect various violations of physical properties of the approximate numerical solution $\boldsymbol{U}^{n,k,i}$: the residual estimator $\eta_{{\rm R},K,c}^{n,k,i,\nu}$ illustrates the fact that the discrete flux reconstruction $\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}$ does not necessarily satisfy exactly the first two lines of (3.15). Note that, when the source term $Q_{c,K}^n$ is constant in time and space ($Q_{c,K}^n = Q_c$), (3.62) leads to $\eta_{{\rm R},K,c}^{n,k,i,\nu} = 0$. The flux estimator $\eta_{{\rm F},K,c}^{n,k,i,\nu}$ given by (3.66) indicates how far is the flux at the discrete level from the equilibrated flux reconstruction. It is related to the temporal discretization, linearization, and algebraic errors. Next, the nonconformity estimators (3.67)-(3.68) show how far are the discrete discontinuous quadratic liquid pressure and molar fraction of liquid hydrogen from their interpolants in the energy space X. Finally, the estimator (3.69) is the nonlinear accumulation estimator and (3.70) is the algebraic remainder estimator. Observe that at convergence of the semismooth solver and the iterative algebraic solver ($k \to \infty$, $i \to \infty$), the estimators (3.69) and (3.70) vanish.

The following result provides an upper bound for the error measure (3.55) at each semismooth Newton step $k \geq 1$ and each algebraic solver step $i \geq 0$ of each time step $1 \leq n \leq N_{\mathrm{t}}$.

Theorem 3.5.7. *Consider a time step $1 \leq n \leq N_{\mathrm{t}}$, a semismooth Newton step $k \geq 1$, an algebraic solver steps $i \geq 0$, and $\nu > 0$ an additional algebraic iteration. Let $\left(S_h^{n,k,i}, P_h^{n,k,i}, \chi_h^{n,k,i}\right)$ be the approximate solution and let $\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}$, $\tilde{P}_{h\tau}^{n,k,i}$, and $\tilde{\chi}_{h\tau}^{n,k,i}$ be respectively the equilibrated flux reconstructions defined by (3.61), the liquid phase pressure reconstruction, and the molar fraction reconstruction defined in Section 3.5.5 with the convention (3.44). We have the following a posteriori error estimate*

$$\begin{aligned}
\mathcal{N}^{n,k,i} \leq{}& \left\{\sum_{c\in\mathcal{C}}\left\{\left\{\int_{I_n}\sum_{K\in\mathcal{T}_h}\left(\eta_{\mathrm{R},K,c}^{n,k,i,\nu}+\eta_{\mathrm{F},K,c}^{n,k,i,\nu}(t)+\eta_{\mathrm{NA},K,c}^{n,k,i,\nu}+\eta_{\mathrm{rem},K,c}^{n,k,i,\nu}\right)^2\mathrm{d}t\right\}^{\frac12}+\|Q_c-Q_{c,h\tau}\|_{X_n'}\right\}^2\right\}^{\frac12}\\
&+\mathcal{R}_{\mathrm{e}}(S_{h\tau}^{n,k,i},P_{h\tau}^{n,k,i},\chi_{h\tau}^{n,k,i})+\left\{\int_{I_n}\sum_{K\in\mathcal{T}_h}\left\{\sum_{c\in\mathcal{C}^{\mathrm{l}}}\left(\eta_{\mathrm{NC},K,\mathrm{l},c}^{n,k,i}(t)\right)^2+\left(\eta_{\mathrm{NC},K,\chi}^{n,k,i}(t)\right)^2\right\}\mathrm{d}t\right\}^{\frac12}.
\end{aligned}\tag{3.71}$$

Proof. The proof follows the one presented in [70, Corollary 4.4] with the difference in the treatment of the algebraic remainder and the presence of the residual associated to the constraints. Let $\varphi \in X_n$ such that $\|\varphi\|_{X_n} = 1$. The residual (3.54) is given by

$$\langle \mathcal{R}_c(S_{h\tau}^{n,k,i},P_{h\tau}^{n,k,i},\chi_{h\tau}^{n,k,i}),\varphi\rangle_{X_n',X_n}=\int_{I_n}\sum_{K\in\mathcal{T}_h}A_K(\varphi)(t)\,\mathrm{d}t,$$

where

$$A_K(\varphi):=\left(Q_c-\partial_t l_{c,h\tau}^{n,k,i},\varphi\right)_K+\left(\boldsymbol{\Phi}_{c,h\tau}^{n,k,i},\boldsymbol{\nabla}\varphi\right)_K.$$

Using (3.62) and noting that $\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}\in\mathbf{H}(\mathrm{div},\Omega)$, we have

$$\begin{aligned}
A_K(\varphi)={}&\left(Q_{c,K}^n-\frac{l_{c,K}(\boldsymbol{U}^{n,k-1})-l_{c,K}^{n-1}+\mathcal{L}_{c,K}^{n,k,i+\nu}}{\tau_n}-\boldsymbol{\nabla}\cdot\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}-\frac{\boldsymbol{R}_{c,K}^{n,k,i+\nu}}{|K|},\varphi\right)_K\\
&-\left(\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu}-\boldsymbol{\Phi}_{c,h\tau}^{n,k,i},\boldsymbol{\nabla}\varphi\right)_K-\left(\partial_t^n l_{c,h\tau}^{n,k,i}-\frac{l_{c,K}(\boldsymbol{U}^{n,k-1})-l_{c,K}^{n-1}+\mathcal{L}_{c,K}^{n,k,i+\nu}}{\tau_n},\varphi\right)_K\\
&-\left(\frac{1}{|K|}\boldsymbol{R}_{c,K}^{n,k,i+\nu},\varphi\right)_K+\left(Q_c-Q_{c,h\tau},\varphi\right)_K.
\end{aligned}\tag{3.72}$$

We bound separately each of the five terms in (3.72) denoted by $A_{j,K}(\varphi), j = 1,\ldots,5$. Observe from the equilibration property (3.62) that the first term $A_{1,K}(\varphi)$ in (3.72) is equal to

$$A_{1,K}(\varphi):=\left(Q_{c,K}^n-\frac{l_{c,K}(\boldsymbol{U}^{n,k-1})-l_{c,K}^{n-1}+\mathcal{L}_{c,K}^{n,k,i+\nu}}{\tau_n}-\frac{\boldsymbol{R}_{c,K}^{n,k,i+\nu}}{|K|}-\boldsymbol{\nabla}\cdot\boldsymbol{\Theta}_{c,h}^{n,k,i,\nu},\varphi-\overline{\varphi}_K\right)_K,\tag{3.73}$$

where $\overline{\varphi}_K$ is the mean value of φ on $K \in \mathcal{T}_h$. Next, we have as a result of the Poincaré–Wirtinger inequality

$$\|\varphi - \overline{\varphi}_K\|_K(t) \leq h_K C_{\mathrm{PW}} \|\boldsymbol{\nabla}\varphi\|_K(t) \leq h_K C_{\mathrm{PW}} \|\varphi\|_{X,K}(t).$$

As $\overline{\varphi}_K$ is the $L^2(K)$ orthogonal projection of $\varphi|_K$ on the space $\mathbb{P}_0(K)$ we have

$$\|\varphi - \overline{\varphi}_K\|_K \leq \|\varphi - c\|_K \quad \forall c \in \mathbb{P}_0(K).$$

Finally, we have

$$\|\varphi - \overline{\varphi}_K\|_K(t) \leq \|\varphi\|_K(t) = \frac{\varepsilon^{\frac{1}{2}} \|\varphi\|_K(t) h_K^{-1}}{\varepsilon^{\frac{1}{2}} h_K^{-1}} \leq \frac{\|\varphi\|_{X,K}(t)}{\varepsilon^{\frac{1}{2}} h_K^{-1}}. \tag{3.74}$$

Combining (3.73)–(3.74) provides the following upper bound:

$$A_{1,K}(\varphi) \leq \eta_{\mathrm{R},K,c}^{n,k,i,\nu} \|\varphi\|_{X,K}(t). \tag{3.75}$$

Using the Cauchy–Schwarz inequality, the second term $A_{2,K}(\varphi)$ of (3.72) is obviously bounded as

$$A_{2,K}(\varphi) \leq \eta_{\mathrm{F},K,c}^{n,k,i,\nu}(t) \|\varphi\|_{X,K}(t). \tag{3.76}$$

Concerning the third term $A_{3,K}(\varphi)$ of (3.72), observe first of all, employing (3.21), that it is equal to

$$A_{3,K}(\varphi) := \left(\frac{l_{c,K}(\boldsymbol{U}^{n,k,i}) - l_{c,K}(\boldsymbol{U}^{n,k-1}) - \mathcal{L}_{c,K}^{n,k,i+\nu}}{\tau_n}, \varphi \right)_K. \tag{3.77}$$

To bound (3.77) we use the Cauchy–Schwarz inequality giving

$$A_{3,K}(\varphi) \leq \eta_{\mathrm{NA},K,c}^{n,k,i,\nu} \varepsilon^{\frac{1}{2}} h_K^{-1} \|\varphi\|_K \leq \eta_{\mathrm{NA},K,c}^{n,k,i,\nu} \|\varphi\|_{X,K}. \tag{3.78}$$

To bound the space integral $A_{4,K}(\varphi)$ containing the algebraic remainder, we employ the Cauchy–Schwarz inequality and next the definition of the error measure (3.45) to get

$$A_{4,K}(\varphi) \leq \frac{1}{|K|} \left\| \boldsymbol{R}_{c,K}^{n,k,i+\nu} \right\|_K \varepsilon^{-\frac{1}{2}} h_K \|\varphi\|_{X,K}(t) = \eta_{\mathrm{rem},K,c}^{n,k,i,\nu} \|\varphi\|_{X,K}(t). \tag{3.79}$$

Finally, concerning the last bound $A_{5,K}(\varphi)$ we use

$$\int_{I_n} (Q_c - Q_{c,h\tau}, \varphi)_\Omega(t)\,\mathrm{dt} \leq \|Q_c - Q_{c,h\tau}\|_{X_n'} \|\varphi\|_{X_n}. \tag{3.80}$$

Thus, as $\|\varphi\|_{X_n} = 1$, combining (3.72), (3.75), (3.76), (3.78), (3.79), (3.80) and using the Cauchy–Schwarz inequality we get

$$\begin{aligned} \left\| \mathcal{R}_c(S_{h\tau}^{n,k,i}, P_{h\tau}^{n,k,i}, \chi_{h\tau}^{n,k,i}) \right\|_{X_n'} \leq & \left\{ \int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\eta_{\mathrm{R},K,c}^{n,k,i,\nu} + \eta_{\mathrm{F},K,c}^{n,k,i,\nu}(t) + \eta_{\mathrm{NA},K,c}^{n,k,i,\nu} + \eta_{\mathrm{rem},K,c}^{n,k,i,\nu} \right)^2 \mathrm{dt} \right\}^{\frac{1}{2}} \\ & + \|Q_c - Q_{c,h\tau}\|_{X_n'}. \end{aligned} \tag{3.81}$$

Next, as $\tilde{P}_{h\tau}^{n,k,i} \in X_n$ and $\tilde{\chi}_{h\tau}^{n,k,i} \in X_n$ we deduce from (3.56) and (3.57) that

$$\mathcal{N}_P^{n,k,i}(P_{h\tau}^{n,k,i}) \leq \left\{ \sum_{c \in \mathcal{C}^{\mathrm{l}}} \int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\eta_{\mathrm{NC},K,\mathrm{l},c}^{n,k,i}(t) \right)^2 \mathrm{dt} \right\}^{\frac{1}{2}} \tag{3.82}$$

and

$$\mathcal{N}_\chi^{n,k,i}(\chi_{h\tau}^{n,k,i}) \leq \left\{ \int_{I_n} \sum_{K \in \mathcal{T}_h} \left(\eta_{\mathrm{NC},K,\chi}^{n,k,i}(t) \right)^2 \mathrm{dt} \right\}^{\frac{1}{2}}. \tag{3.83}$$

Thus, combining (3.81)–(3.83) we get the desired result. □

So far, we have established an posteriori estimate between the exact and the approximate solution. We now provide an estimate distinguishing the different error components. For this purpose, we additionally define the positive and negative parts of each constraint as follows. For A, any real number, we define

$$A = A^+ + A^-, \quad \text{with} \quad A^+ := \max(0, A) \geq 0, \quad \text{and} \quad A^- := \min(0, A) \leq 0. \tag{3.84}$$

Definition 3.5.8. *Let $1 \leq n \leq N_{\mathrm{t}}$ be a time step, $k \geq 1$ be a semismooth Newton iteration, and $i \geq 0$ be an algebraic iteration. For any $c \in \mathcal{C}$, we define the discretization estimator, the linearization estimator, and the algebraic estimator by*

$$\begin{aligned} \eta_{\mathrm{disc}}^{n,k,i,\nu} := 2^{\frac{1}{2}} \Bigg\{ \sum_{K \in \mathcal{T}_h} \int_{I_n} \Bigg\{ \sum_{c \in \mathcal{C}} \Big(\eta_{\mathrm{R},K,c}^{n,k,i,\nu} + \left\| \boldsymbol{\Theta}_{c,h,\mathrm{disc}}^{n,k,i} - \boldsymbol{\Phi}_{c,h\tau}^{n,k,i}(t) \right\|_K + \eta_{\mathrm{NC},K,\mathrm{l},c}^{n,k,i}(t) \Big) \\ + \eta_{\mathrm{NC},K,\chi}^{n,k,i}(t) \Bigg\}^2 \mathrm{dt} \Bigg\}^{\frac{1}{2}} + \frac{1}{\alpha} \sum_{K \in \mathcal{T}_h} \int_{I_n} \eta_{\mathrm{P},K,\mathrm{pos}}^{n,k,i}(t) \, \mathrm{dt} \end{aligned} \tag{3.85}$$

$$\eta_{\mathrm{lin}}^{n,k,i} := \left\{ \sum_{c \in \mathcal{C}} \tau_n \sum_{K \in \mathcal{T}_h} \left(\left\| \boldsymbol{\Theta}_{c,h,\mathrm{lin}}^{n,k,i} \right\|_K + \eta_{\mathrm{NA},K,c}^{n,k,i,\nu} \right)^2 \right\}^{\frac{1}{2}} + \frac{1}{\alpha} \sum_{K \in \mathcal{T}_h} \int_{I_n} \eta_{\mathrm{P},K,\mathrm{neg}}^{n,k,i}(t) \, \mathrm{dt}, \tag{3.86}$$

$$\eta_{\mathrm{alg}}^{n,k,i,\nu} := \left\{ \sum_{c \in \mathcal{C}} \tau_n \sum_{K \in \mathcal{T}_h} \left(\left\| \boldsymbol{\Theta}_{c,h,\mathrm{alg}}^{n,k,i,\nu} \right\|_K + \eta_{\mathrm{rem},K,c}^{n,k,i,\nu} \right)^2 \right\}^{\frac{1}{2}}, \tag{3.87}$$

with

$$\eta_{\mathrm{P},K,\mathrm{pos}}^{n,k,i}(t) := \left(\left\{ 1 - S_{h\tau}^{n,k,i}(t) \right\}^+, \left\{ H\left[P_{h\tau}^{n,k,i}(t) + P_{\mathrm{cp}}\left(S_{h\tau}^{n,k,i}(t) \right) \right] - \beta^{\mathrm{l}} \chi_{h\tau}^{n,k,i}(t) \right\}^+ \right)_K, \tag{3.88}$$

$$\eta_{\mathrm{P},K,\mathrm{neg}}^{n,k,i}(t) := \left(\left\{ 1 - S_{h\tau}^{n,k,i}(t) \right\}^-, \left\{ H\left[P_{h\tau}^{n,k,i}(t) + P_{\mathrm{cp}}\left(S_{h\tau}^{n,k,i}(t) \right) \right] - \beta^{\mathrm{l}} \chi_{h\tau}^{n,k,i}(t) \right\}^- \right)_K. \tag{3.89}$$

Remark 3.5.9. *In Definition 3.5.8 we proposed three components of the error constructed from the various estimators defined in Section 3.5.6. Note that it is possible to bound the residual $\mathcal{R}_{\mathrm{e}}$ following the decomposition (3.84) and employing the property*

$$A_1 A_2 = \left[A_1^+ + A_1^-\right]\left[A_2^+ + A_2^-\right] \leq A_1^+ A_2^+ + A_1^- A_2^-. \tag{3.90}$$

The phase transition estimators $\eta^{n,k,i}_{\mathrm{P},K,\mathrm{pos}}(t)$ and $\eta^{n,k,i}_{\mathrm{P},K,\mathrm{neg}}(t)$ given by (3.88) and (3.89) are new to the best of our knowledge and give a control on the violation of the constraints: they evaluate the error due to the physical phase change between the liquid and the liquid–gas phase. At convergence of the semismooth and linear algebraic solver ($k \to \infty$, $i \to \infty$) $\eta^{n,\infty,\infty}_{\mathrm{P},K,\mathrm{neg}}(t) = 0$. Observe that when the gas phase appears in the triangle K, the estimator $\eta^{n,\infty,\infty}_{\mathrm{P},K,\mathrm{pos}}(t)$ is positive on the time interval corresponding to the state change. Otherwise it is vanishing. Therefore, our approach is heuristic in the sense that at convergence of the iterative algebraic solver and the semismooth solver $\eta^{n,k,i}_{\mathrm{lin}} \to 0$ and $\eta^{n,k,i,\nu}_{\mathrm{alg}} \to 0$. Note that the algebraic remainder estimator is always positive then when added to $\left\|\boldsymbol{\Theta}^{n,k,i,\nu}_{c,h,\mathrm{alg}}\right\|_K$ provides a non vanishing global algebraic estimator at the begining of the iterations.

Corollary 3.5.10. *For a given time step $1 \leq n \leq N_{\mathrm{t}}$, a semismooth Newton iteration $k \geq 1$, an algebraic iteration $i \geq 0$, and $\nu > 0$ additional algebraic sover steps, consider the estimators defined by (3.85)–(3.87). Assume moreover that the source term Q_c is piecewise constant in space and time. Then, we have*

$$\mathcal{N}^{n,k,i} \leq \eta^{n,k,i,\nu}_{\mathrm{disc}} + \eta^{n,k,i}_{\mathrm{lin}} + \eta^{n,k,i,\nu}_{\mathrm{alg}}.$$

Proof. The triangle inequality applied on the flux estimator gives

$$\eta^{n,k,i,\nu}_{\mathrm{F},K,c}(t) \leq \left\|\boldsymbol{\Theta}^{n,k,i}_{c,h,\mathrm{disc}} - \boldsymbol{\Phi}^{n,k,i}_{c,h\tau}(t)\right\|_K + \left\|\boldsymbol{\Theta}^{n,k,i}_{c,h,\mathrm{lin}}\right\|_K + \left\|\boldsymbol{\Theta}^{n,k,i,\nu}_{c,h,\mathrm{alg}}\right\|_K. \tag{3.91}$$

Plugging (3.91) in (3.71), and using after the Minkowski inequality to separate each component fluxes and each nonconform estimators provides the following bound for (3.71)

$$\begin{aligned}
\mathcal{N}^{n,k,i} \leq & \left\{\sum_{c\in\mathcal{C}}\int_{I_n}\sum_{K\in\mathcal{T}_h}\left(\eta^{n,k,i,\nu}_{\mathrm{R},K,c} + \left\|\boldsymbol{\Theta}^{n,k,i}_{c,h,\mathrm{disc}} - \boldsymbol{\Phi}^{n,k,i}_{c,h\tau}(t)\right\|_K\right)^2 \mathrm{dt}\right\}^{\frac{1}{2}} \\
& + \mathcal{R}_{\mathrm{e}}(S^{n,k,i}_{h\tau}, P^{n,k,i}_{h\tau}, \chi^{n,k,i}_{h\tau}) \\
& + \left\{\int_{I_n}\sum_{K\in\mathcal{T}_h}\left\{\sum_{c\in\mathcal{C}^{\mathrm{l}}}\left(\eta^{n,k,i}_{\mathrm{NC},K,\mathrm{l},c}(t)\right)^2 + \left(\eta^{n,k,i}_{\mathrm{NC},K,\chi}(t)\right)^2\right\}\mathrm{dt}\right\}^{\frac{1}{2}} \\
& + \left\{\sum_{c\in\mathcal{C}}\tau_n\sum_{K\in\mathcal{T}_h}(\left\|\boldsymbol{\Theta}^{n,k,i}_{c,h,\mathrm{lin}}\right\|_K + \eta^{n,k,i,\nu}_{\mathrm{NA},K,c})^2\right\}^{\frac{1}{2}} + \eta^{n,k,i,\nu}_{\mathrm{alg}}.
\end{aligned} \tag{3.92}$$

To bound $\mathcal{R}_{\mathrm{e}}(S^{n,k,i}_{h\tau}, P^{n,k,i}_{h\tau}, \chi^{n,k,i}_{h\tau})$ we employ (3.90) to get

$$\mathcal{R}_{\mathrm{e}}(S^{n,k,i}_{h\tau}, P^{n,k,i}_{h\tau}, \chi^{n,k,i}_{h\tau}) \leq \frac{1}{\alpha}\int_{I_n}\sum_{K\in\mathcal{T}_h}\left(\eta^{n,k,i}_{\mathrm{P},K,\mathrm{pos}} + \eta^{n,k,i}_{\mathrm{P},K,\mathrm{neg}}\right)\mathrm{dt}.$$

To conclude, it remains to bound the sum of the first and third term of (3.92). To do so, we employ the inequality $\left(\sum_{q=1}^{r} X_q^2\right)^{\frac{1}{2}} + \left(\sum_{q=1}^{r} Y_q^2\right)^{\frac{1}{2}} \leq \left(2\sum_{q=1}^{r}\left(X_q^2 + Y_q^2\right)\right)^{\frac{1}{2}}$ for all $X_q, Y_q \geq 0$ and next the identity $A^2 + B^2 \leq (A+B)^2$ for all $A, B \geq 0$ to obtain the desired result. □

3.5.7 Adaptive inexact semismooth Newton method using adaptive stopping criteria

In this section we develop an adaptive inexact semismooth Newton method. In the spirit of [9, 62, 81, 128], it is designed to perform the linearization and algebraic resolution with minimal necessary precision and thus to avoid unnecessary iterations. We rely on Corollary 3.5.10 that estimates the different error components. We define γ_{lin} and γ_{alg} as two positive parameters representing the desired relative size of the algebraic and linearization errors. We propose the following stopping criteria, balancing globally the algebraic, linearization, and discretization error components for our adaptive algorithm (see Algorithm 3)

$$\text{(a) } \eta_{\text{alg}}^{n,k,i,\nu} \leq \gamma_{\text{alg}} \max\left\{\eta_{\text{disc}}^{n,k,i,\nu}, \eta_{\text{lin}}^{n,k,i}\right\}, \quad \text{(b) } \eta_{\text{lin}}^{n,k,i} \leq \gamma_{\text{lin}} \eta_{\text{disc}}^{n,k,i,\nu}. \tag{3.93}$$

We propose the following adaptive inexact semismooth algorithm:

Algorithm 3 Adaptive inexact semismooth Newton algorithm

0. Choose an initial vector $\boldsymbol{U}^{n,0} \in \mathbb{R}^{3N_{\text{sp}}}$ and set $k = 1$.

1. From $\boldsymbol{U}^{n,k-1}$ define $\mathbb{A}^{n,k-1} \in \mathbb{R}^{3N_{\text{sp}},3N_{\text{sp}}}$ and $\boldsymbol{B}^{n,k-1} \in \mathbb{R}^{3N_{\text{sp}}}$ by (3.38) and (3.39).

2. Consider the linear system

$$\mathbb{A}^{n,k-1}\boldsymbol{U}^{n,k} = \boldsymbol{B}^{n,k-1}. \tag{3.94}$$

3. Set $\boldsymbol{U}^{n,k,0} = \boldsymbol{U}^{n,k-1}$ as initial guess for the iterative linear solver, set $i = 0$.

4*a*. Perform $\nu \geq 1$ steps of a chosen linear solver for (3.94), starting from $\boldsymbol{U}^{n,k,i}$.

This yields on step $i + \nu$ an approximation $\boldsymbol{U}^{n,k,i+\nu}$ to $\boldsymbol{U}^{n,k}$ satisfying

$$\mathbb{A}^{n,k-1}\boldsymbol{U}^{n,k,i+\nu} = \boldsymbol{B}^{n,k-1} - \boldsymbol{R}^{n,k,i+\nu}.$$

4*b*. Compute the estimators of Definition 3.5.8 and check the stopping criterion for the linear solver in the form (3.93)(a). Set $i = i + \nu$. If satisfied, set $\boldsymbol{U}^{n,k} = \boldsymbol{U}^{n,k,i}$. If not go back to **4*a***.

5. Check the stopping criterion for the nonlinear solver in the form (3.93)(b). If satisfied, return $\boldsymbol{U}^{n} = \boldsymbol{U}^{n,k}$. If not, set $k = k + 1$ and go back to **1**.

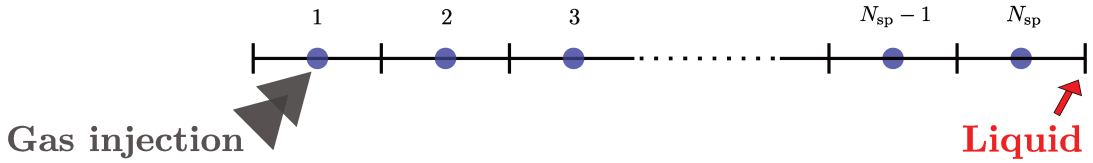

Figure 3.5: Definition of the geometry of the test case. The center of the finite volumes cells are represented in blue.

3.6 Numerical experiments

3.6.1 Settings

This section illustrates numerically our theoretical developments. We use the Couplex-gas benchmark proposed by Andra (French National Inventory of Radioactive Materials and Waste) [1] and the research group MoMaS (Mathematical Modeling and Numerical Simulation for Nuclear Waste Management Problems) [2]. We consider a homogeneous porous medium Ω in one dimension supposed to be horizontal with length $L = 200$m. Its constant porosity is fixed to $\phi = 0.15$ and its constant absolute permeability is equal to $\underline{\mathbf{K}} = 5 \times 10^{-20}\text{m}^2$. The porous medium is initially satured with liquid ($S^{\text{l}} = 1$) and contains no hydrogen ($\chi^{\text{l}}_{\text{h}} = 0$). We assume that gaseous hydrogen is injected constantly in time in the first cell K_1 ($Q^n_{\text{h},K_1} = 5.57 \times 10^{-6}\text{kg/m}^2\text{/year}$) and the water flow rate is zero. We have homogeneous Neumann boundary conditions on the left of the domain. For boundary conditions on the right, we assume that the gas injected will never reach the end of the domain, thus Dirichlet conditions are prescribed ($S^{\text{l}} = 1$, $P^{\text{l}} = 10^6\text{Pa}$, $\chi^{\text{l}}_{\text{h}} = 0$). As we consider a horizontal 1D case, gravitational effects are not taken into account in the numerical tests. The dynamic liquid phase viscosity $\mu^{\text{l}} = 10^{-9}\text{Pa.s}$, the dynamic gas phase viscosity $\mu^{\text{g}} = 9 \times 10^{-9}\text{Pa.s}$, the molar mass of water $M_{\text{w}} = 10^{-2}\text{kg.mol}^{-1}$, the molar mass of hydrogen $M_{\text{h}} = 2 \times 10^{-3}\text{kg.mol}^{-1}$, the molar density of water $\rho^{\text{l}}_{\text{w}} = 10^3\text{kg.m}^{-3}$, the molecular diffusion coefficient $D^{\text{l}}_{\text{h}} = 3 \times 10^{-9}\text{m}^2\text{.s}^{-1}$, and Henry's constant $\tilde{H} = 7.65 \times 10^{-6}\text{mol.Pa}^{-1}\text{.m}^{-3}$. We consider for the capillary pressure P_{cp} and the permeability of the liquid phase k^{l}_{r} and gas phase k^{g}_{r} the Van Genuchten–Mualem model:

$$
\begin{aligned}
P_{\text{cp}}(S^{\text{l}}) &= P_{\text{r}}\left(S_{\text{le}}^{-\frac{1}{m}} - 1\right)^{\frac{1}{n^*}}, \\
k^{\text{l}}_{\text{r}}(S^{\text{l}}) &= \sqrt{S_{\text{le}}}\left(1 - (1 - S_{\text{le}}^{\frac{1}{m}})^m\right)^2, \\
k^{\text{g}}_{\text{r}}(S^{\text{l}}) &= \sqrt{1 - S_{\text{le}}}\left(1 - S_{\text{le}}^{\frac{1}{m}}\right)^{2m},
\end{aligned}
$$

with

$$
S_{\text{le}} = \frac{S^{\text{l}} - S^{\text{l}}_{\text{res}}}{1 - S^{\text{l}}_{\text{res}} - S^{\text{g}}_{\text{res}}} \quad \text{and} \quad m = 1 - \frac{1}{n^*}.
$$

Here $P_{\text{r}} = 2 \times 10^6\text{Pa}$ is the reference pressure, $n^* = 1.49$ is a parameter depending on the porous medium, and $S^{\text{l}}_{\text{res}} = 0.4$, $S^{\text{g}}_{\text{res}} = 0$ are respectively the residual liquid saturation and residual gas saturation (see for more details [49]). We consider a uniform spatial mesh ($N_{\text{sp}} = 1000$ elements) and we use a constant time step $\tau_n = 5000$

years $\forall 1 \leq n \leq N_{\mathrm{t}}$. The final time of simulation is $t_{\mathrm{F}} = 5 \times 10^5$ years and the rescaling constant $\alpha = 2500$ years.
We consider two different semismooth Newton solvers. We first employ the Newton-min algorithm combined with the GMRES linear iterative algebraic solver for the system (3.37). Next, we employ the Newton–Fischer–Burmeister algorithm in combination with the GMRES solver. In both cases, an ILU preconditionner is used to speed up the GMRES solver. Other possibilities for preconditionners can be found in [117] and the references therein. For the computation of the algebraic flux reconstruction $\boldsymbol{\Theta}_{c,h,\mathrm{alg}}^{n,k,i,\nu}$, we use (3.60) with $\nu = 1$. We also define the algebraic and linearization residuals by

$$\boldsymbol{R}_{\mathrm{alg}}^{n,k,i} := \boldsymbol{B}^{n,k-1} - \mathbb{A}^{n,k-1}\boldsymbol{U}^{n,k,i}, \tag{3.95}$$

$$\boldsymbol{R}_{\mathrm{lin}}^{n,k,i} := \begin{bmatrix} \mathcal{H}^n(\boldsymbol{U}^{n,k,i}) \\ \mathcal{C}^n(\boldsymbol{U}^{n,k,i}) \end{bmatrix}, \tag{3.96}$$

where the nonlinear operators $\mathcal{H}^n$ and $\mathcal{C}^n$ are defined in (3.32) and (3.35).
Three different approaches are tested:
1) The *exact* semismooth Newton method. Here, both the linear and nonlinear solvers are iterated to "almost" convergence. More precisely, we take $\varepsilon_{\mathrm{alg}} = 10^{-12}$ and $\varepsilon_{\mathrm{lin}} = 10^{-7}$ and replace respectively the stopping criteria (3.93) of Algorithm 3 by criteria on the relative residuals,

$$\text{(a)}\ \frac{\left\|\boldsymbol{R}_{\mathrm{alg}}^{n,k,i}\right\|}{\|\boldsymbol{B}^{n,k-1}\|} \leq \varepsilon_{\mathrm{alg}}, \qquad \text{(b)}\ \frac{\left\|\boldsymbol{R}_{\mathrm{lin}}^{n,k,i}\right\|}{\|\mathcal{F}(\boldsymbol{U}^{n,0})\|} \leq \varepsilon_{\mathrm{lin}}. \tag{3.97}$$

2) The *inexact* semismooth Newton method. Here, (3.37) is solved only approximately. We use the following stopping criterion replacing (3.97)(a) for the iterative algebraic solver:

$$\text{(a)}\ \frac{\left\|\boldsymbol{R}_{\mathrm{alg}}^{n,k,i}\right\|}{\|\boldsymbol{B}^{n,k-1}\|} \leq \eta_k. \tag{3.98}$$

In the litterature, η_k is called the "forcing term" and under the assumption that the sequence $(\eta_k)_{k\geq 1}$ is uniformely less than 1, inexact Newton methods are locally convergent, see [65, 114]. We choose $\eta_k = \dfrac{1}{2^k}\dfrac{\boldsymbol{R}_{\mathrm{lin}}^{n,k,i}}{\|\mathcal{F}(\boldsymbol{U}^{n,0})\|}$ but other choices are possible, see [42, 73]. Concerning the stopping criterion for the semismooth Newton solver, we keep (3.97) (b).

3) The *adaptive inexact* semismooth Newton method (see Algorithm 3) that relies on the stopping criteria (3.93)(a) and (3.93)(b) with $\gamma_{\mathrm{alg}} = 10^{-3}$ and $\gamma_{\mathrm{lin}} = 10^{-3}$.

For the three methods, the criteria are computed every $\nu = 1$ linear iteration. In the sequel, when the stopping criterion of the nonlinear solver is satisfied , the index k will be denoted by $\overline{k}$, and similarly the index i at the various stopping criteria will be denoted by $\overline{i}$.

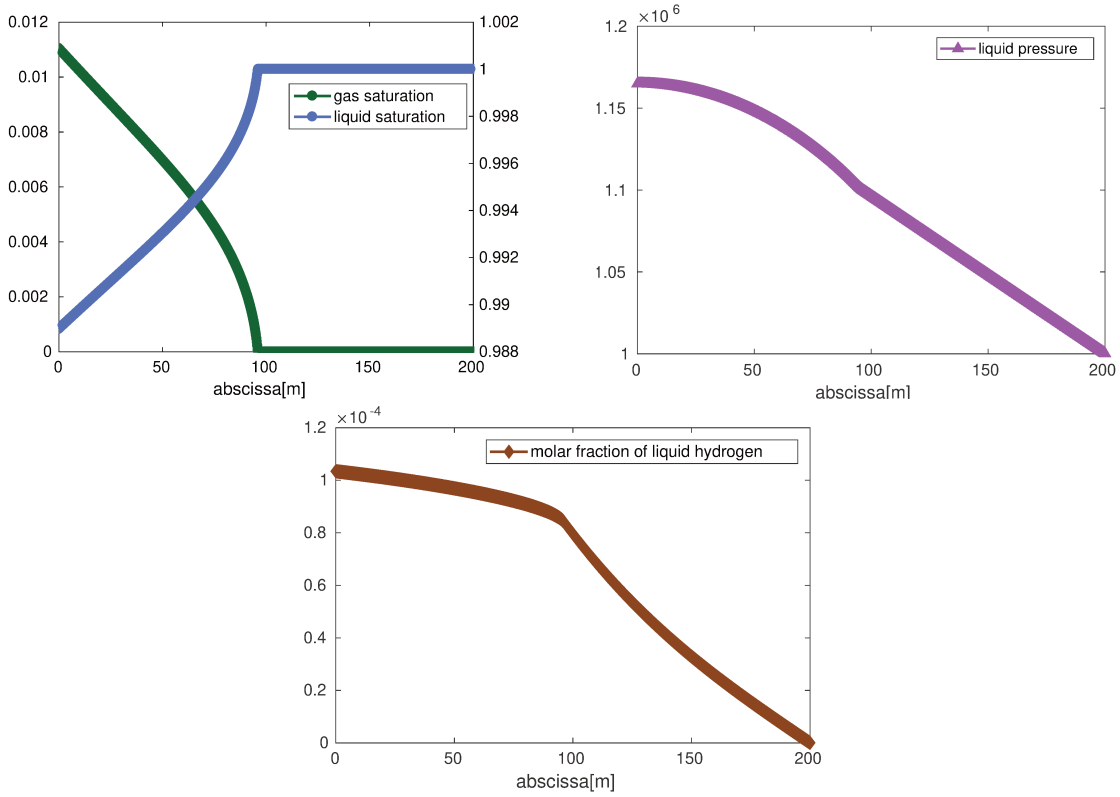

Figure 3.6: Solution at convergence ($k = \overline{k}$, $i = \overline{i}$) for $N_{\text{sp}} = 1000$ elements at $t = 1.05 \times 10^5$ years. Left: saturation of the phases, middle: pressure of the liquid phase, right: molar fraction of liquid hydrogen.

3.6.2 Newton-min

We consider the $2 \times N_{\text{sp}}$ equations given by the cell-centered finite volume discretization (3.30), where we recall that N_{sp} equations correspond to each component $c \in \mathcal{C}$. The nonlinear complementarity constraints are reformulated thanks to the semismooth min function as follows: $\forall K \in \mathcal{T}_h,\ \forall 1 \leq n \leq N_{\text{t}}$

$$
\begin{aligned}
&1 - S_K^n \geq 0, \quad H\left[P_K^n + P_{\text{cp}}(S_K^n)\right] - \beta^{\text{l}} \chi_K^n \geq 0, \\
&\left[1 - S_K^n\right] \cdot \left[H\left[P_K^n + P_{\text{cp}}(S_K^n)\right] - \beta^{\text{l}} \chi_K^n\right] = 0, \\
&\iff \min\left(1 - S_K^n, H\left[P_K^n + P_{\text{cp}}(S_K^n)\right] - \beta^{\text{l}} \chi_K^n\right) = 0.
\end{aligned}
$$

We then employ the Newton-min solver to treat the nonlinearities. Figure 3.6 displays the behavior of the solution at time $t = 1.05 \times 10^5$ years (corresponding to a two-phase regime) when the Newton-min and the GMRES solvers have converged. We observe from the three figures that the liquid pressure and the molar fraction of liquid hydrogen have increased almost everywhere and that the gas has spread in several cells of the domain. It is characteristic of a two-phase flow after appearance of the gas phase.

Figure 3.7 shows the possible violations of the nonlinear complementarity constraints during the iterations at the time step $t = 5 \times 10^4$ years, see the begining of Section 3.5.6. We have represented on the left figure the negative part of the

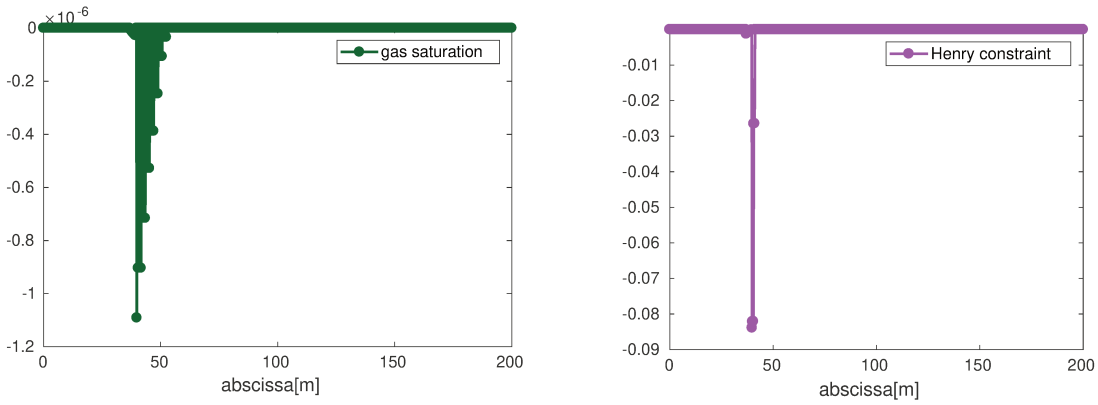

Figure 3.7: Complementarity constraints ($k = 4$, $i = 2$) at time $t = 5 \times 10^4$ years. Left: negative part of the saturation constraint, right: negative part of Henry's constraint.

saturation constraint $\left\{1 - S_{h\tau}^{n,k,i}\right\}^-$ and we observe its negativity in several cells. The same phenomenon occurs for the constraint given by Henry's law, see the right figure.

In Figure 3.8, we have displayed the behavior of the phase transition estimator $\eta_{\mathrm{P},K,\mathrm{pos}}^{n,k,i}(t)$ as a function of the abscissa at convergence ($k = \overline{k}$, $i = \overline{i}$). We recall that $\eta_{\mathrm{P},K,\mathrm{pos}}^{n,k,i}(t^n) = 0$ for all endpoints of all intervals I_n, $1 \leq n \leq N_{\mathrm{t}}$, see (3.88)–(3.89), so the estimator is shown in the middle of the time interval I_n, denoted by $t_n^\star$. In the left figure, we have chosen I_1 ($t_1^\star = 2500$ years), during which there is only one liquid phase and one observes that $\eta_{\mathrm{P,K}}^{n,k,i}(t) = 0$ over all $t \in I_1$. On the middle figure, the estimator is shown at $t_3^\star = 1.25 \times 10^4$ years (corresponding time interval I_3). It corresponds to the time interval when the gas phase starts to appear in the leftmost cell, which can be observed on the estimator. Then, in several cells close to the left boundary, we observe a peak corresponding to the activation of the two constraints $1 - S_{h\tau}^{n,k,i}(\cdot, t_3^\star) > 0$ and $H\left[P_{h\tau}^{n,k,i}(\cdot, t_3^\star) + P_{\mathrm{cp}}(S_{h\tau}^{n,k,i}(\cdot, t_3^\star))\right] - \beta^{\mathrm{l}} \chi_{h\tau}^{n,k,i}(\cdot, t_3^\star) > 0$, then the nonnegativity of the estimator. On the right figure, the estimator $\eta_{\mathrm{P,K}}^{n,k,i}(t_9^\star)$ is shown at $t_9^\star = 4.25 \times 10^4$ years, when the flow is two-phase liquid–gas. We see the localisation (near 45m) of the gas phase appearance on the domain Ω by a peak. Thus, the front between the one-phase and the two-phase regimes can be clearly noted thanks to the estimator.

Remark 3.6.1. *From this example, one can see that this estimator detects the error caused by the appearance of the gas phase whenever the gas spreads throughout the domain. It gives important tools for adaptive mesh refinement strategy that will be considered in a future work.*

Figure 3.9 represents at the fixed time value $t = 1.05 \times 10^5$ years the evolution of the various estimators and the behavior of the non relative residuals $\left\|\boldsymbol{R}_{\mathrm{lin}}^{n,k,i}\right\|$ and $\left\|\boldsymbol{R}_{\mathrm{alg}}^{n,k,i}\right\|$ given by (3.95) and (3.96) as a function of the Newton-min iterations when the stopping criteria (3.97)(a)-(3.97)(b), respectively (3.98)(a)-(3.97)(b), respectively (3.93)(a)-(3.93)(b) have been satisfied (1000 elements, k varies, $i = \overline{i}$).

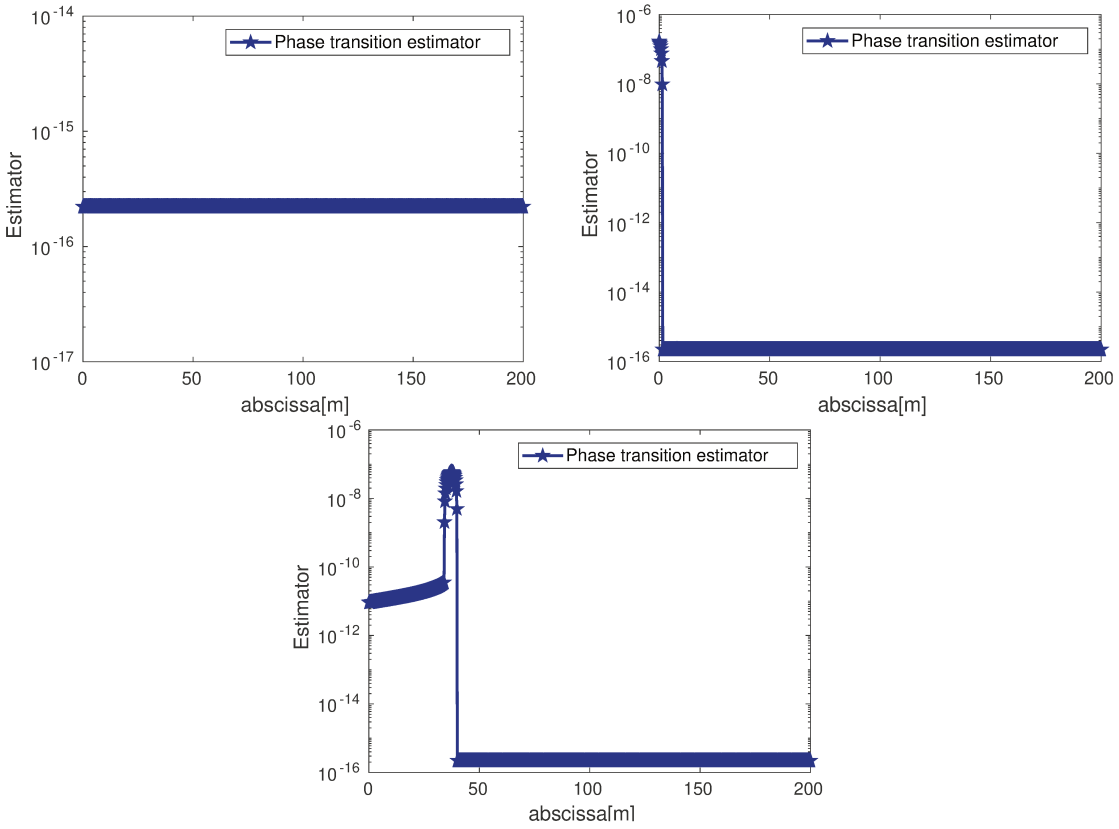

Figure 3.8: Phase transition estimator $\eta^{n,k,i}_{\mathrm{P},K,\mathrm{pos}}$ at convergence ($k = \overline{k}$, $i = \overline{i}$). Left: one-phase liquid. middle: appearance of gas phase, right: two-phase liquid–gas.

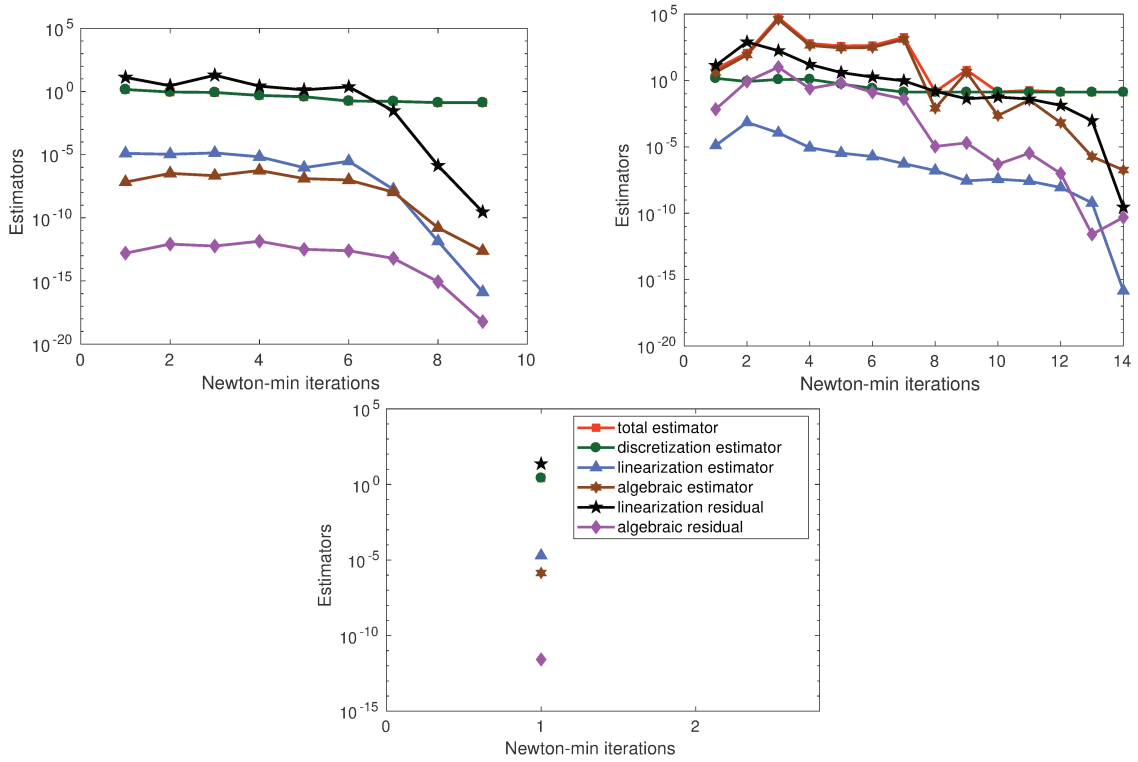

Figure 3.9: Estimators as a function of the Newton-min iterates k, $(i = \bar{i})$ at $t = 1.05 \times 10^5$. Exact (left), inexact (middle), and adaptive inexact (right) Newton-min methods.

In the exact resolution case, the discretization estimator globally dominates and coincides with the total estimator (the red and green curves are superimposed). The linearization estimator is small and decreases rapidly after $k = 6$. The algebraic estimator is small and takes values between 10^{-6} and 10^{-12}. Observe that the behavior of the linearization estimator (respectively algebraic estimator) mimics the one of the linearization residual (respectively algebraic residual) up to an important roughly constant shift. Note that the stopping criteria for exact and inexact Newton-min are based on the relative linearization and algebraic residuals see (3.97)(a)–(3.97)(b) which do not correspond to the curves of $\left\|\boldsymbol{R}_{\text{lin}}^{n,k,i}\right\|$ and $\left\|\boldsymbol{R}_{\text{alg}}^{n,k,i}\right\|$ that are non relative residuals. From the first Newton-min iteration, the discretization estimator is more or less constant, which means that the other components of the error do not influence the behavior of the total error estimator. Therefore, the semismooth linearization iterations can be stopped at the first Newton-min step. This is precisely the situation described the by our adaptive inexact Newton-min (figure on the right). We have displayed in the figure in the middle the number of Newton-min iterations required to satisfy the inexact stopping criterion (3.97)(b). We observe that the inexact method requires more semismooth Newton-min iterations to converge (14 iterations) than the exact one.

Figure 3.10 shows the evolution of the various estimators and the behavior of $\left\|\boldsymbol{R}_{\text{lin}}^{n,k,i}\right\|$ and $\left\|\boldsymbol{R}_{\text{alg}}^{n,k,i}\right\|$ given by (3.95) and (3.96) during the algebraic iterations of the first Newton-min step (1000 elements, $k = 1$, i varies). In the three methods,

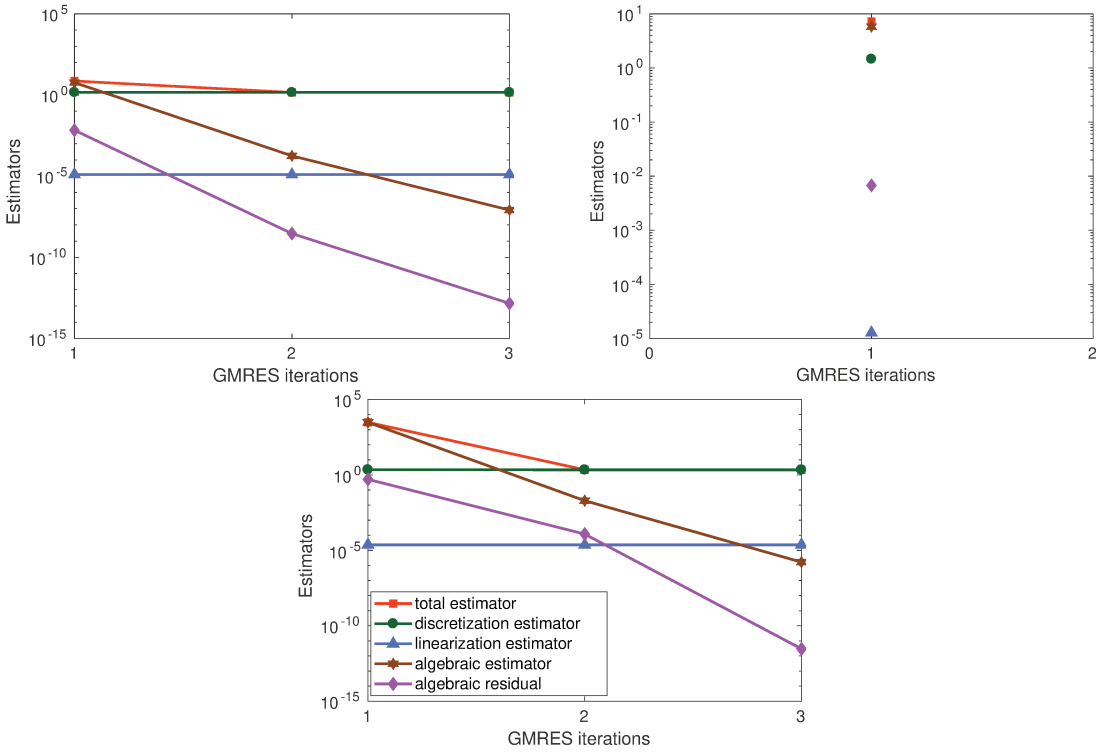

Figure 3.10: Estimators as a function of the algebraic iterations i for $k = 1$ at $t = 1.05 \times 10^5$. Exact (left), inexact (middle), and adaptive inexact (right) semismooth Newton-min methods.

the algebraic estimator is dominant and dominates the total estimator whereas the discretization and linearization estimators roughly stagnate. We observe that 3 GMRES iterations are needed to achieve the stopping criterion (3.97)(a) whereas in the inexact and adaptive inexact cases, 1 iteration, respectively 3 iterations, are required to satisfy the stopping criteria (3.98)(a), respectively (3.93)(a). For the three methods, the estimators are computed every $\nu = 1$ iteration.

In Figure 3.11 are displayed the number of Newton-min iterations and the total number of GMRES iterations required to satisfy the various stopping criteria at each time step of the simulation. In particular, the first graph shows that the inexact Newton-min method requires many more semismooth iterations to converge in comparison with the other methods. The second graph of Figure 3.11 shows that the exact Newton-min method is globally the most expensive method in terms of linear algebraic iterations and adaptive inexact Newton-min method is the cheapest one.

Figure 3.12 illustrates the overall performance of the three approaches. In the first graph, the cumulated number of Newton-min iterations for the three methods is displayed as a function of the time steps. The inexact Newton-min method requires approximately 1000 Newton-min iterations in total whereas exact Newton-min, respectively adaptive inexact Newton-min, require 550 iterations, respectively 100 iterations. The right part of Figure 3.12 focuses on the cumulated number of GMRES iterations for each method as a function of the time step. The adaptive inexact Newton-min method is the less expensive since it requires approximately 500,

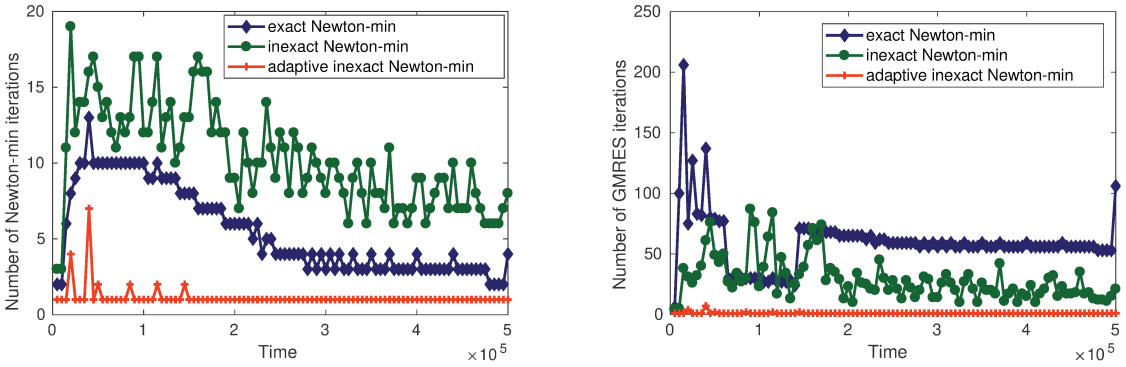

Figure 3.11: Number of Newton-min iterations at each time step (left), number of GMRES iterations at each time step (right).

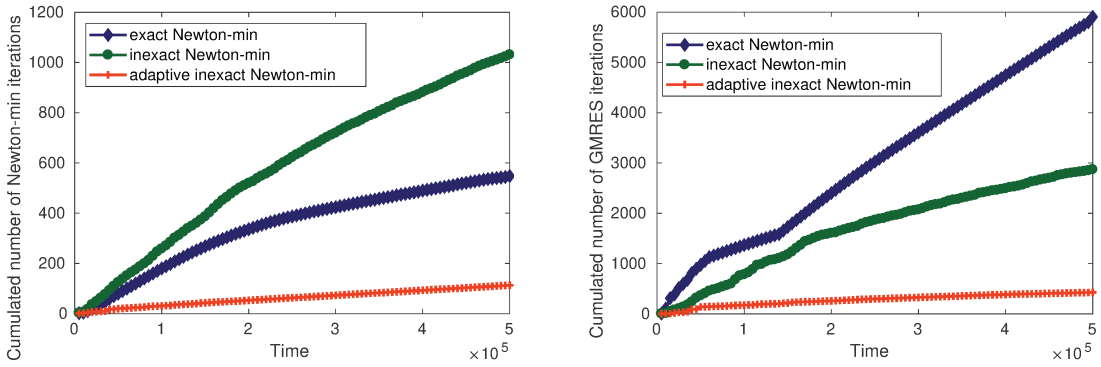

Figure 3.12: Cumulated number of Newton-min iterations as a function of time (left), and cumulated number of GMRES iterations as a function of time (right).

iterations whereas inexact Newton-min, and respectively adaptive inexact Newton-min, require 3000 iterations, respectively 6000 iterations, to finish the simulation. Thus, globally our approach yields an economy by a factor of roughly 6 with respect to inexact Newton-min and roughly 12 with respect to exact Newton-min in terms of total algebraic solver iterations.

On the three first graphs of Figure 3.13 (top left, top right, and middle left) is displayed the behavior of the solution at convergence ($k = \overline{k}$, $i = \overline{i}$) at the selected time $t = 1.05 \times 10^5$ years for the exact Newton-min resolution and adaptive inexact Newton-min resolution with the weights $\gamma_{\text{alg}} = \gamma_{\text{lin}} = 10^{-3}$. We observe a non consistency zone for the three graphs explained by the nonlinear stopping criterion in adaptive inexact resolution that stops earlier the semismooth iterations. The next three graphs of Figure 3.13 (middle right, bottom left, bottom right) show that at a time close to the final simulation time ($t = 3.5 \times 10^5$ years), the curves of the solutions given by exact Newton-min and adaptive inexact Newton-min almost coincide. Thus, our adaptive inexact semimooth Newton algorithm saves many Newton-min and GMRES iterations and generates a solution whose precision does not differ from the exact one more than by a fraction of the discretization error.

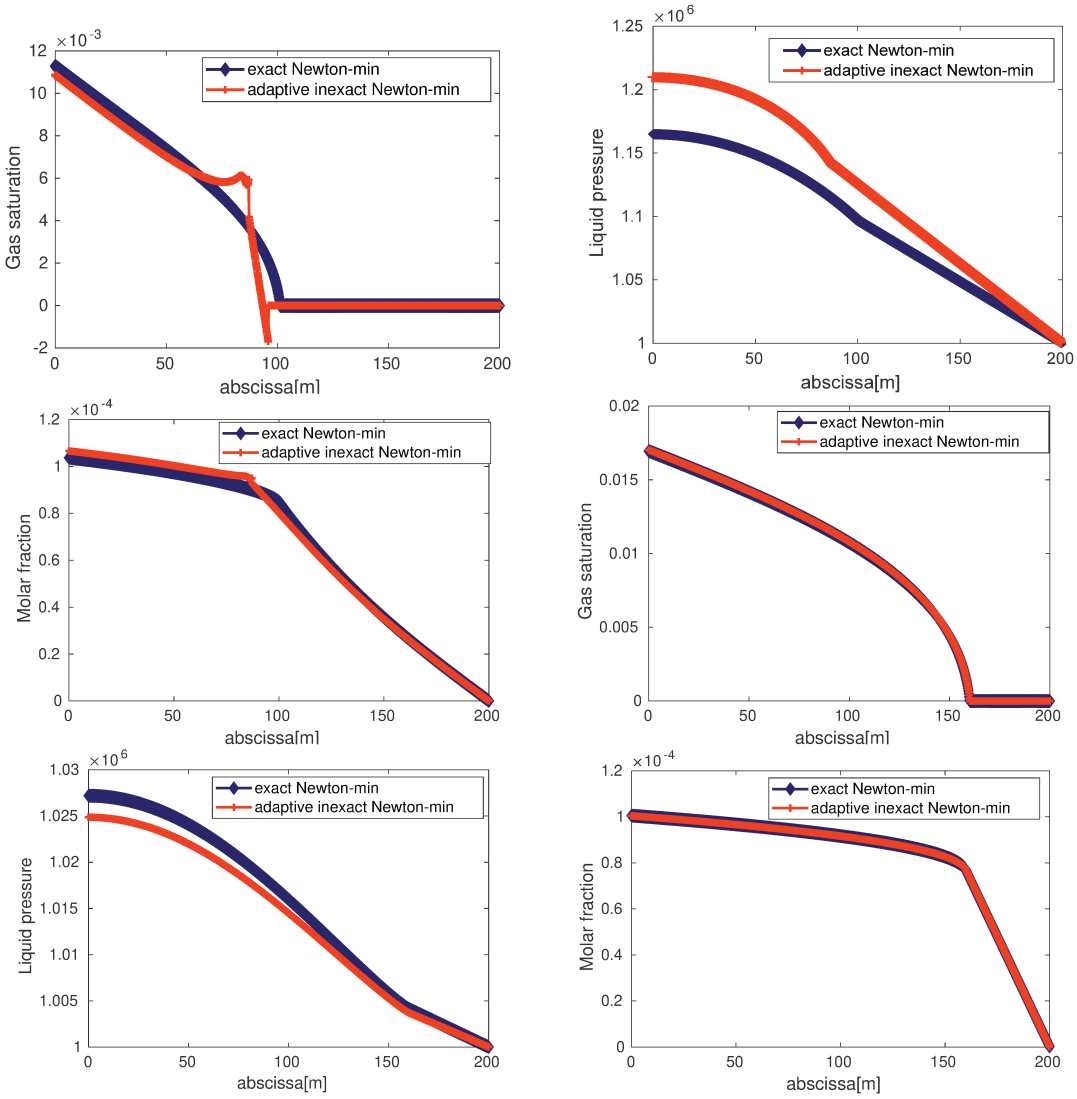

Figure 3.13: Gas saturation (top left), liquid pressure (top right), and molar fraction of liquid hydrogen (middle left) for exact Newton-min and adaptive inexact Newton-min at convergence at $t = 1.05 \times 10^5$ years. Gas saturation (middle right), liquid pressure (bottom left), and molar fraction of liquid hydrogen (bottom right) for exact Newton-min and adaptive inexact Newton-min at convergence at $t = 3.5 \times 10^5$ years.

Table 3.1: Total number of linear and nonlinear iterations for adaptive inexact Newton-min method for several parameters γ_{alg} and γ_{lin}.

$(\gamma_{\text{alg}}, \gamma_{\text{lin}})$	Cumulated Newton-min iterations	Cumulated GMRES iterations
$(10^{-1}, 10^{-1})$	100	366
$(10^{-3}, 10^{-3})$	113	427
$(10^{-6}, 10^{-3})$	108	967
$(10^{-3}, 10^{-6})$	351	1682
$(10^{-6}, 10^{-6})$	308	2019

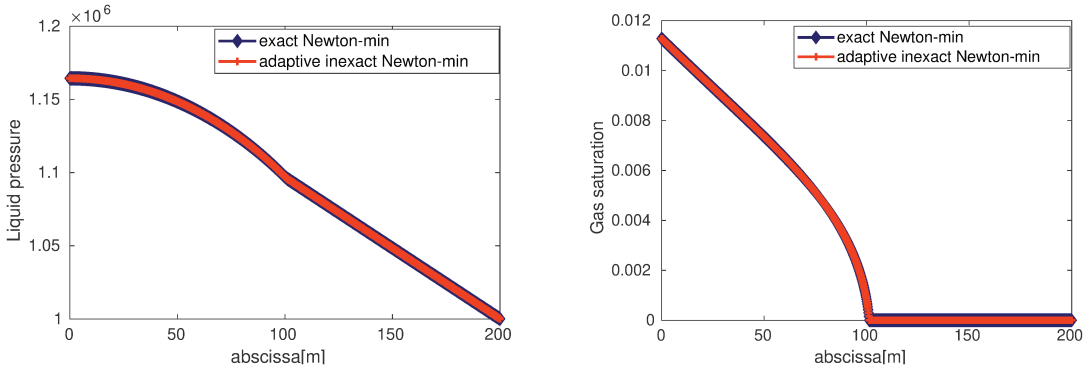

Figure 3.14: Liquid pressure (left) and gas saturation (right) for exact Newton-min and adaptive inexact Newton-min at convergence at time $t = 1.05 \times 10^5$ years with $\gamma_{\text{alg}} = 10^{-3}$ and $\gamma_{\text{lin}} = 10^{-6}$. The two curves superimpose.

3.6.3 Complements

In this section we carry out numerical simulations supplementing the previous results. We test the influence of the weights γ_{lin} and γ_{alg} on our adaptive inexact semismooth methodology. We also briefly provide results for the adaptive inexact Newton–Fischer–Burmesiter algorithm.

In Table 3.1 we give the cumulated number of Newton-min iterations and GMRES iterations to reach the end of the simulation for different weights γ_{alg} and γ_{lin}. We observe that decreasing the values of the weights will increase the number of required iterations and increasing the values of the weights, for example $(\gamma_{\text{alg}}, \gamma_{\text{lin}}) = (10^{-1}, 10^{-1})$, will decrease the required number of iterations.

In Figure 3.14 we test the influence of the weight γ_{lin} on the behavior of the solution. We take $\gamma_{\text{lin}} = 10^{-6}$, $\gamma_{\text{alg}} = 10^{-3}$, and the time value close to the begining of the simulation $t = 1.05 \times 10^5$ years when the semismooth Newton solver and the GMRES solver have converged $(k = \overline{k},\ i = \overline{i})$. Recall that in Figure 3.13, we considered the same time instant but with $\gamma_{\text{lin}} = 10^{-3}$. We thus see that the solution given by exact Newton-min and adaptive inexact Newton-min are almost identical with $\gamma_{\text{lin}} = 10^{-6}$. From this example, we deduce that for a time step close to the begining of the simulation, it is possible to increase the precision in the adaptive inexact resolution by decreasing the value of the weight γ_{lin}. Besides, even taking the smallest values for the weights γ_{lin} and γ_{alg}, $(\gamma_{\text{lin}} = \gamma_{\text{alg}} = 10^{-6})$ will

Table 3.2: Total number of nonlinear and linear iterations for the adaptive inexact Newton–Fischer–Burmeister method for several parameters γ_{alg} and γ_{lin} and for the exact Newton–Fischer–Burmeister method.

$(\gamma_{\text{alg}}, \gamma_{\text{lin}})$	Cumulated number of Newton–Fischer–Burmeister iterations	Cumulated number of GMRES iterations
$(10^{-1}, 10^{-1})$	100	428
$(10^{-3}, 10^{-3})$	119	751
$(10^{-3}, 10^{-6})$	482	2074
$(10^{-6}, 10^{-3})$	117	1694
Exact resolution	757	10089

obviously increase the cumulated number of GMRES iterations (2019 iterations see Table 3.1) and increase the accuracy but, the adaptive strategy is still economic in comparison to exact Newton-min resolution that requires 6000 iterations (see Figure 3.12).

To conclude this section, we present some results obtained by the Newton–Fischer–Burmeister algorithm. In this case, the nonlinear complementarity constraints can be reformulated thanks to the semismooth Fischer–Burmeister function, see (3.34), as follows: $\forall K \in \mathcal{T}_h, \quad \forall 1 \leq n \leq N_{\text{t}}$,

$$
\begin{aligned}
&1 - S_K^n \geq 0, \quad H\left[P_K^n + P_{\text{cp}}(S_K^n)\right] - \beta^{\text{l}} \chi_K^n \geq 0, \\
&\left[1 - S_K^n\right] \cdot \left[H\left[P_K^n + P_{\text{cp}}(S_K^n)\right] - \beta^{\text{l}} \chi_K^n\right] = 0, \\
&\iff f_{\text{FB}}(1 - S_K^n, H\left[P_K^n + P_{\text{cp}}(S_K^n)\right] - \beta^{\text{l}} \chi_K^n) = 0.
\end{aligned}
$$

Table 3.2 provides the behavior of the exact Newton–Fischer–Burmeister algorithm and of the adaptive inexact Newton–Fischer–Burmeister algorithm for several weights γ_{alg} and γ_{lin}. The adaptive strategy gives suitable results as it roughly saves 90% of the iterations from the exact resolution. Furthermore, we can observe that exact and adaptive inexact Newton-min provides better results in terms of computational cost than exact and adaptive inexact Newton–Fischer–Burmeister. This observation is in agreement with the fast convergence rate of the Newton-min algorithm [24, 88, 89].

Acknowledgements: We thank S. Yousef (IFPEN) for discussions on a posteriori error estimates and implementation.

3.7 Conclusion

We have studied a compositional two-phase liquid–gas flow with appearance/disappearance of the gas phase. We have employed the semismooth theory to treat the nonlinearities in the complementarity constraints. We have devised a posteriori error estimates between the exact and approximate solution, in particular

when the phase transition occurs and we have distinguished the different error components. In the numerical experiments, we have tested the quality of our adaptive strategy. In particular the results confirmed the strength of this approach.

Conclusion et perspectives

Dans cette thèse, nous avons étudié des systèmes d'équations aux dérivées partielles contenant des contraintes de complémentarité. Nous avons abordé trois modèles : un problème stationnaire de contact entre deux membranes, une inégalité variationnelle parabolique comme extension du modèle étudié dans le premier chapitre, et pour finir, un écoulement diphasique avec changement de phase en milieu poreux. Pour chacun de ces problèmes, nous avons utilisé des algorithmes de Newton semi-lisse inexacts pour traiter les non linéarités non différentiables. De même, nous avons établi des estimations d'erreur a posteriori donnant une borne supérieure garantie sur l'erreur, et permettant de distinguer et d'évaluer les différentes composantes de l'erreur. L'ensemble de nos essais numériques montrent que notre stratégie adaptative basée sur l'arrêt anticipé des itérations de nos solveurs (non linéaire et linéaire) n'affecte ni la précision, ni la qualité de la solution numérique obtenue.

Dans les Chapitres 1 et 2, la discrétisation des modèles repose sur la méthode des éléments finis de degré $p \geq 1$, alors que dans le Chapitre 3, elle repose sur la méthode des volumes finis centrés sur les mailles. A chaque fois, nous avons établi des estimations d'erreur a posteriori en utilisant la méthodologie de reconstruction des flux équilibrés. Pour les Chapitres 1 et 2, les flux de discrétisation et d'algèbre linéaire sont reconstruits en résolvant des systèmes mixtes sur des patchs, tandis qu'au Chapitre 3, les flux de discrétisation, de linéarisation et d'algèbre linéaire sont plus aisément calculés grâce aux flux normaux sur les faces issus de la méthode des volumes finis. Dans le Chapitre 1, nous avons également prouvé l'efficacité locale de nos estimateurs à l'exception de l'estimateur de contact qui est négligeable numériquement.

A l'issue de ce travail de thèse, plusieurs questions se profilent en toile de fond avec des perspectives sur quelques pistes qu'il serait utile d'approfondir par la suite. En effet, dans le cadre des Chapitres 1 et 2, il serait intéressant d'étendre notre discrétisation éléments finis aux discrétisations volumes finis ou de type Galerkine discontinue afin de comparer l'efficacité des temps de calcul obtenus. Ensuite, il serait utile d'étudier la convergence des algorithmes de résolution de type Newton semi-lisse ; en effet, nous sommes face à un problème de combinatoire car le Jacobien de Clarke est un ensemble de cardinal égal à $2^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ pour le problème stationnaire et $2^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ à chaque pas de temps pour le problème instationnaire où $\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ est le nombre de degrés de liberté. Ainsi, pour s'assurer de la convergence de l'agorithme, il faudra montrer que les $2^{\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}}$ matrices Jacobiennes potentielles sont inversibles. Par ailleurs, dans le Chapitre 1, la reconstruction des flux équilibrés qui exige la solution de problèmes locaux étant un peu coûteuse numériquement dans l'implémentation actuelle, il serait bénéfique d'optimiser le code afin de réduire le temps de calcul.

Cette optimisation serait également utile pour l'implémentation des estimateurs du Chapitre 2. Il est également prévu d'étendre cette implémentation à des polynômes de plus haut degrés. Cela permettra en outre de tester numériquement les propriétés de la base duale. De plus, dans le Chapitre 2, notre solution numérique dépend du temps et nous avons estimé l'erreur de discrétisation, de linéarisation et d'algèbre linéaire. Il serait opportun d'estimer aussi l'erreur de discrétisation temporelle afin d'obtenir un algorithme semi-lisse inexact adaptatif en temps. Sur un autre plan, malgré la richesse de la littérature sur les estimations d'erreur a posteriori pour les inégalités variationnelles, peu de travaux ont été réalisés sur les problèmes de contact entre plusieurs corps. Il serait intéressant d'appliquer notre méthodologie à ce type de problèmatiques. Il serait également pertinent, plus tard, d'étendre notre étude à un problème de contact entre deux membranes vibrantes.

Dans le Chapitre 3, nous avons étudié un problème diphasique compositionnel liquide-gaz avec échange d'hydrogène entre les deux phases en milieu poreux. Nous avons établi des estimations d'erreur a posteriori en tenant compte du changement de phase. En particulier, l'estimateur de changement de phase introduit permet d'identifier les cellules dans lesquelles le gaz apparaît. Une première piste intéressante pour la suite serait de raffiner adaptativement le maillage aux endroits où le gaz apparaît pour améliorer la précision de la solution numérique.

Une difficulté particulière des formulations diphasiques est de prouver l'existence et l'uncité de la solution faible. Cela est possible dans le contexte d'un écoulement diphasique avec un composant par phase. Pour les écoulements diphasiques compositionnels cela n'a en revanche jamais été effectué car c'est une question très compliquée. De plus, aucun travail n'a à ma connaissance été effectué sur les estimations d'erreur a posteriori pour des écoulements multiphasiques avec multiples transitions de phase. Il me semble opportun d'étudier plus en profondeur le travail effectué par Helmig, Wohlmuth, Hager et Lauser [119] avant d'essayer de formuler des estimateurs d'erreur dans le cadre de leur formulation. Il semble également intéressant d'étendre nos travaux aux écoulements diphasiques en milieux poreux et fracturés.

Enfin, je tiens à souligner une différence peu visible entre nos trois chapitres. Dans les Chapitres 1 et 2, les équations de diffusions du modèle continu font appel à un multiplicateur de Lagrange. D'ailleurs, lors de la construction des estimateurs d'erreur, l'estimateur de résidu utilise ce multiplicateur de Lagrange. Pour la formulation diphasique, les équations de conservation étant l'analogue des équations de diffusion présentes dans les Chapitres 1 et 2, elles n'utilisent pas de multiplicateur de Lagrange. En particulier, l'estimateur de résidu est construit sans un tel multiplicateur. Je m'interroge sur la possibilité de définir un modèle diphasique utilisant un multiplicateur de Lagrange et d'étendre la théorie de nos Chapitres 1 et 2 à une telle formulation.

Appendix A

Semi-lissité

Nous présentons pour aider le lecteur, les concepts de base de la notion de semi-lissité. Une étude plus complète sur le sujet est présentée dans [29, 88, 89].

A.1 Motivation

Considérons le système d'équations suivant :

$$F(\boldsymbol{x}) = 0 \tag{A.1}$$

où $F : \mathcal{E} \to \mathcal{F}$ est une fonction non lisse (non différentiable au sens de Fréchet) et $\mathcal{E}$ et $\mathcal{F}$ deux espaces vectoriels normés de même dimension finie notée p. On supposera que F est lipschitzienne dans un voisinage du zéro $\boldsymbol{x}_*$ recherché. Le caractère lipschitzien de F assure que cette dernière est Fréchet-différentiable presque partout d'après le Théorème de Rademacher [58]. La notion de fonction semi-lisse peut être motivée par le souhait de faire converger localement l'algorithme de Newton pour trouver un zéro $\boldsymbol{x}_*$ de (A.1). La généralisation de l'algorithme de Newton que l'on souhaite voir converger s'écrit localement

$$\boldsymbol{x}^k := \boldsymbol{x}^{k-1} - \left[\mathbb{A}^{k-1}\right]^{-1} F(\boldsymbol{x}^{k-1}) \quad \forall k \geq 1 \tag{A.2}$$

où $\left[\mathbb{A}^{k-1}\right]^{-1}$ est un élément inversible du différentiel de Clarke $\partial_{\mathrm{C}} F(\boldsymbol{x}^{k-1})$. Nous expliquerons plus bas ce qu'est un tel différentiel.

A.2 Différentiel de Clarke

Soient Ω un ouvert de $\mathcal{E}$ et $F : \Omega \to \mathcal{F}$ une fonction localement lipschitzienne sur Ω. Notons $\mathcal{D}_F := \{\boldsymbol{x} \in \Omega : \mathrm{F} \text{ est Fréchet-differentiable en } \boldsymbol{x}\}$. Par le Théorème de Rademacher, $\Omega \backslash \mathcal{D}_F$ est de mesure nulle.

Definition A.2.1 (B-différentiel et C-différentiel)**.** *Le B-différentiel (B honore Bouligant [147]) de F en $\boldsymbol{x}$ est l'ensemble noté et défini par*

$$\partial_{\mathrm{B}} F(\boldsymbol{x}) := \left\{\mathbb{J} \in \mathbb{R}^p \times \mathbb{R}^p : \exists \boldsymbol{x}^k \in \mathcal{D}_F \text{ tel que } \boldsymbol{x}^k \to \boldsymbol{x} \text{ et } F'(\boldsymbol{x}^k) \to \mathbb{J}\right\}. \tag{A.3}$$

Le C-différentiel (C pour Clarke) de F en $\boldsymbol{x}$ est l'enveloppe convexe du B-différentiel, à savoir

$$\partial_{\mathrm{C}} F(\boldsymbol{x}) := \mathrm{co}\left(\partial_{\mathrm{B}} F(\boldsymbol{x})\right). \tag{A.4}$$

On rappelle que l'enveloppe convexe d'une partie $\mathcal{A}$ de $\mathcal{E}$ est l'intersection de toutes les parties convexes de $\mathcal{E}$ qui contiennent $\mathcal{A}$.

Exemple : Si $F : \mathcal{E} \to \mathcal{F}$ est dérivable dans un voisinage de $\boldsymbol{x} \in \mathcal{E}$ et si F' est continue en $\boldsymbol{x}$ on a clairement que $\partial_{\mathrm{B}} F(\boldsymbol{x}) = \partial_{\mathrm{C}} F(\boldsymbol{x}) = \{F'(\boldsymbol{x})\}$.

A.3 Semi-lissité

On dit que $F : \Omega \to \mathcal{F}$ est semi-lisse en $\boldsymbol{x} \in \Omega$ si

1. F est lipschitzienne dans un voisinage de $\boldsymbol{x}$
2. F admet des dérivées directionnelles en $\boldsymbol{x}$ dans toutes les directions
3. lorsque $\boldsymbol{h} \to 0$ dans $\mathcal{E}$ on a

$$\sup_{\mathbb{J} \in \partial_{\mathrm{C}} F(\boldsymbol{x}+\boldsymbol{h})} \|F(\boldsymbol{x}+\boldsymbol{h}) - F(\boldsymbol{x}) - \mathbb{J}\boldsymbol{h}\| = o(\|\boldsymbol{h}\|). \tag{A.5}$$

On donne quelques propriétés des fonctions semi-lisses.

Proposition A.3.1. *Si $F : \Omega \to \mathcal{F}$ est $\mathcal{C}^1$ dans un voisinage de $\boldsymbol{x} \in \Omega$, alors F est semi-lisse en $\boldsymbol{x}$.*

Proposition A.3.2. *Si $F : \Omega \to \mathbb{R}$ est convexe sur un voisinage convexe de $\boldsymbol{x} \in \Omega$, alors F est semi-lisse en $\boldsymbol{x}$.*

Proposition A.3.3. *Les fonctions minimum, maximum et la fonction de Fischer–Burmeister sont semi-lisses.*

Bibliography

[1] http://www.andra.fr/international/.

[2] http://www.gdrmomas.org/ex_qualifications.html/.

[3] Y. Achdou, F. Hecht, and D. Pommier, *A posteriori error estimates for parabolic variational inequalities*, J. Sci. Comput., 37 (2008), pp. 336–366, https://doi.org/10.1007/s10915-008-9215-7.

[4] M. Aganagić, *Newton's method for linear complementarity problems*, Math. Programming, 28 (1984), pp. 349–362.

[5] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York, 2000, https://doi.org/10.1002/9781118032824.

[6] M. Ainsworth, J. T. Oden, and C.-Y. Lee, *Local a posteriori error estimators for variational inequalities*, Numer. Methods Partial Differential Equations, 9 (1993), pp. 23–33, https://doi.org/10.1002/num.1690090104.

[7] B. Amaziane, M. Jurak, and A. Žgaljić Keko, *Modeling compositional compressible two-phase flow in porous media by the concept of the global pressure*, Comput. Geosci., 18 (2014), pp. 297–309, https://doi.org/10.1007/s10596-013-9362-2.

[8] P. F. Antonietti, L. Beirão da Veiga, C. Lovadina, and M. Verani, *Hierarchical a posteriori error estimators for the mimetic discretization of elliptic problems*, SIAM J. Numer. Anal., 51 (2013), pp. 654–675, https://doi.org/10.1137/120873157.

[9] M. Arioli, E. H. Georgoulis, and D. Loghin, *Stopping criteria for adaptive finite element solvers*, SIAM J. Sci. Comput., 35 (2013), pp. A1537–A1559.

[10] S. Auliac, Z. Belhachmi, F. Ben Belgacem, and F. Hecht, *Quadratic finite elements with non-matching grids for the unilateral boundary contact*, ESAIM Math. Model. Numer. Anal., 47 (2013), pp. 1185–1203, https://doi.org/10.1051/m2an/2012064.

[11] I. Babuška and W. C. Rheinboldt, *Reliable error estimation and mesh adaptation for the finite element method*, in Computational methods in nonlinear mechanics (Proc. Second Internat. Conf., Univ. Texas, Austin, Tex., 1979), North-Holland, Amsterdam-New York, 1980, pp. 67–108.

[12] S. Bartels and C. Carstensen, *Averaging techniques yield reliable a posteriori finite element error control for obstacle problems*, Numer. Math., 99 (2004), pp. 225–249, https://doi.org/10.1007/s00211-004-0553-6.

[13] P. Bastian, *Numerical computation of multiphase flow in porous media*. Habilitationsschrift, 1999.

[14] M. Bebendorf, *A note on the Poincaré inequality for convex domains*, Z. Anal. Anwendungen, 22 (2003), pp. 751–756, https://doi.org/10.4171/ZAA/1170.

[15] R. Becker, C. Johnson, and R. Rannacher, *Adaptive error control for multigrid finite element methods*, Computing, 55 (1995), pp. 271–288, https://doi.org/10.1007/BF02238483.

[16] Z. Belhachmi and F. B. Belgacem, *Quadratic finite element approximation of the Signorini problem*, Math. Comp., 72 (2003), pp. 83–104, https://doi.org/10.1090/S0025-5718-01-01413-2.

[17] F. Ben Belgacem, C. Bernardi, A. Blouza, and M. Vohralík, *A finite element discretization of the contact between two membranes*, M2AN Math. Model. Numer. Anal., 43 (2008), pp. 33–52, https://doi.org/10.1051/m2an/2008041.

[18] F. Ben Belgacem, C. Bernardi, A. Blouza, and M. Vohralík, *On the unilateral contact between membranes. Part 1: Finite element discretization and mixed reformulation*, Math. Model. Nat. Phenom., 4 (2009), pp. 21–43, https://doi.org/10.1051/mmnp/20094102.

[19] F. Ben Belgacem, C. Bernardi, A. Blouza, and M. Vohralík, *On the unilateral contact between membranes. Part 2: a posteriori analysis and numerical experiments*, IMA J. Numer. Anal., 32 (2012), pp. 1147–1172, https://doi.org/10.1093/imanum/drr003.

[20] I. Ben Gharbia, J. Dabaghi, V. Martin, and M. Vohralík, *A posteriori error estimates and adaptive stopping criteria for a compositional two-phase flow with nonlinear complementarity constraints*. HAL Preprint 01919067, submitted for publication, 2018, https://hal.inria.fr/hal-01919067.

[21] I. Ben Gharbia and J. C. Gilbert, *Nonconvergence of the plain Newton-min algorithm for linear complementarity problems with a P-matrix*, Math. Program., 134 (2012), pp. 349–364, https://doi.org/10.1007/s10107-010-0439-6.

[22] I. Ben Gharbia and J. C. Gilbert, *An algorithmic characterization of* **P***-matricity*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 904–916, https://doi.org/10.1137/120883025.

[23] I. Ben Gharbia and J. C. Gilbert, *An algorithmic characterization of P-matricity II: adjustments, refinements, and validation.* HAL Preprint 01672197, submitted for publication, 2018, https://hal.inria.fr/hal-01672197v3.

[24] I. Ben Gharbia and J. Jaffré, *Gas phase appearance and disappearance as a problem with complementarity constraints*, Math. Comput. Simulation, 99 (2014), pp. 28–36, https://doi.org/10.1016/j.matcom.2013.04.021.

[25] A. Bensoussan and J.-L. Lions, *Applications of variational inequalities in stochastic control*, vol. 12 of Studies in Mathematics and its Applications, North-Holland Publishing Co., Amsterdam-New York, 1982. Translated from the French.

[26] A. Bergam, C. Bernardi, and Z. Mghazli, *A posteriori analysis of the finite element discretization of some parabolic equations*, Math. Comp., 74 (2005), pp. 1117–1138, https://doi.org/10.1090/S0025-5718-04-01697-7.

[27] C. Bernardi, Y. Maday, and F. Rapetti, *Discrétisations variationnelles de problèmes aux limites elliptiques*, vol. 45 of Mathématiques & Applications (Berlin) [Mathematics & Applications], Springer-Verlag, Berlin, 2004.

[28] D. Boffi, F. Brezzi, and M. Fortin, *Mixed finite element methods and applications*, vol. 44 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2013, https://doi.org/10.1007/978-3-642-36519-5.

[29] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization*, Universitext, Springer-Verlag, Berlin, second ed., 2006. Theoretical and practical aspects.

[30] A. Bourgeat, J. Mladen, and F. Smaï, *Two-phase, partially miscible flow and transport modeling in porous media; application to gas migration in a nuclear waste repository*, Comput. Geosci., 13 (2009), pp. 29–42.

[31] D. Braess, *A posteriori error estimators for obstacle problems—another look*, Numer. Math., 101 (2005), pp. 415–421, https://doi.org/10.1007/s00211-005-0634-1.

[32] D. Braess, V. Pillwein, and J. Schöberl, *Equilibrated residual error estimates are p-robust*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 1189–1197, https://doi.org/10.1016/j.cma.2008.12.010.

[33] D. Braess and J. Schöberl, *Equilibrated residual error estimator for edge elements*, Math. Comp., 77 (2008), pp. 651–672, https://doi.org/10.1090/S0025-5718-07-02080-7.

[34] S. C. Brenner, *Two-level additive Schwarz preconditioners for nonconforming finite element methods*, Math. Comp., 65 (1996), pp. 897–921, https://doi.org/10.1090/S0025-5718-96-00746-6.

[35] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer-Verlag, New York, 1994, https://doi.org/10.1007/978-1-4757-4338-8.

[36] H. Brezis, *Inéquations variationnelles paraboliques*, Séminaire Jean Leray, 7 (1971), pp. 1–10.

[37] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext, Springer, New York, 2011.

[38] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991, https://doi.org/10.1007/978-1-4612-3172-1.

[39] F. Brezzi, W. W. Hager, and P.-A. Raviart, *Error estimates for the finite element solution of variational inequalities*, Numer. Math., 28 (1977), pp. 431–443, https://doi.org/10.1007/BF01404345.

[40] F. Brezzi, W. W. Hager, and P.-A. Raviart, *Error estimates for the finite element solution of variational inequalities. II. Mixed methods*, Numer. Math., 31 (1978/79), pp. 1–16, https://doi.org/10.1007/BF01396010.

[41] W. L. Briggs, *A multigrid tutorial*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1987.

[42] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 450–481, https://doi.org/10.1137/0911026.

[43] M. Bürg and A. Schröder, *A posteriori error control of hp-finite elements for variational inequalities of the first and second kind*, Comput. Math. Appl., 70 (2015), pp. 2783–2802, https://doi.org/10.1016/j.camwa.2015.08.031.

[44] C. Cancès, I. S. Pop, and M. Vohralík, *An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow*, Math. Comp., 83 (2014), pp. 153–188, https://doi.org/10.1090/S0025-5718-2013-02723-8.

[45] C. Carstensen and S. A. Funken, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput., 21 (1999/00), pp. 1465–1484, https://doi.org/10.1137/S1064827597327486.

[46] C. Carstensen, D. Gallistl, and Y. Huang, *Saturation and reliable hierarchical a posteriori Morley finite element error control*, J. Comput. Math., 36 (2018), pp. 833–844.

[47] C. Carstensen, R. Lazarov, and S. Tomov, *Explicit and averaging a posteriori error estimates for adaptive finite volume methods*, SIAM J. Numer. Anal., 42 (2005), pp. 2496–2521, https://doi.org/10.1137/S0036142903425422.

[48] J. Céa, *Approximation variationnelle des problèmes aux limites*, Ann. Inst. Fourier (Grenoble), 14 (1964), pp. 345–444.

[49] G. Chavent and J. Jaffré, *Mathematical models and finite elements for reservoir simulation*, North Holland, 1986, https://doi.org/10.1016/j.matcom.2013.04.021.

[50] Z. Chen, *Finite element methods and their applications*, Scientific Computation, Springer-Verlag, Berlin, 2005.

[51] Z. Chen, *Reservoir simulation*, vol. 77 of CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007, https://doi.org/10.1137/1.9780898717075. Mathematical techniques in oil recovery.

[52] Z. Chen, G. Huan, and Y. Ma, *Computational methods for multiphase flows in porous media*, vol. 2 of Computational Science & Engineering, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[53] Z. Chen and R. H. Nochetto, *Residual type a posteriori error estimates for elliptic obstacle problems*, Numer. Math., 84 (2000), pp. 527–548, https://doi.org/10.1007/s002110050009.

[54] S. Chippada, C. N. Dawson, M. L. Martinez, and M. F. Wheeler, *A Godunov-type finite volume method for the system of shallow water equations*, Comput. Methods Appl. Mech. Engrg., 151 (1998), pp. 105–129, https://doi.org/10.1016/S0045-7825(97)00108-4. Symposium on Advances in Computational Mechanics, Vol. 3 (Austin, TX, 1997).

[55] F. Chouly, M. Fabre, P. Hild, J. Pousin, and Y. Renard, *Residual-based a posteriori error estimation for contact problems approximated by Nitsche's method*, IMA J. Numer. Anal., 38 (2018), pp. 921–954, https://doi.org/10.1093/imanum/drx024.

[56] F. Chouly and P. Hild, *A Nitsche-based method for unilateral contact problems: numerical analysis*, SIAM J. Numer. Anal., 51 (2013), pp. 1295–1307, https://doi.org/10.1137/12088344X.

[57] P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[58] F. H. Clarke, *Optimization and nonsmooth analysis*, vol. 5 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 1990, https://doi.org/10.1137/1.9781611971309.

[59] H. Class, R. Helmig, and P. Bastian, *Numerical simulation of non-isothermal multiphase multi-component processes in porous media. 1. An efficient solution technique.*, Adv. Water. Resour., 25 (2002), pp. 533–550.

[60] P. Coorevits, P. Hild, and J.-P. Pelle, *A posteriori error estimation for unilateral contact with matching and non-matching meshes*, Comput. Methods Appl. Mech. Engrg., 186 (2000), pp. 65–83, https://doi.org/10.1016/S0045-7825(99)00105-X.

[61] E. Creusé and S. Nicaise, *A posteriori error estimator based on gradient recovery by averaging for discontinuous Galerkin methods*, J. Comput. Appl. Math., 234 (2010), pp. 2903–2915, https://doi.org/10.1016/j.cam.2010.03.027.

[62] J. Dabaghi, V. Martin, and M. Vohralík, *Adaptive inexact semismooth Newton methods for the contact problem between two membranes*. HAL Preprint 01666845, submitted for publication, 2018, https://hal.inria.fr/hal-01666845.

[63] R. Dautray and J.-L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques. Vol. 8*, INSTN: Collection Enseignement. [INSTN: Teaching Collection], Masson, Paris, 1988. Évolution: semi-groupe, variationnel. [Evolution: semigroups, variational methods], Reprint of the 1985 edition.

[64] T. De Luca, F. Facchinei, and C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Math. Programming, 75 (1996), pp. 407–439.

[65] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408, https://doi.org/10.1137/0719025.

[66] P. Destuynder and B. Métivet, *Explicit error bounds in a conforming finite element method*, Math. Comp., 68 (1999), pp. 1379–1396, https://doi.org/10.1090/S0025-5718-99-01093-5.

[67] P. Deuflhard, *Newton methods for nonlinear problems*, vol. 35 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2011, https://doi.org/10.1007/978-3-642-23899-4. Affine invariance and adaptive algorithms, First softcover printing of the 2006 corrected printing.

[68] D. A. Di Pietro and A. Ern, *Mathematical aspects of discontinuous Galerkin methods*, vol. 69 of Mathématiques & Applications (Berlin) [Mathematics & Applications], Springer, Heidelberg, 2012, https://doi.org/10.1007/978-3-642-22980-0.

[69] D. A. Di Pietro, E. Flauraud, M. Vohralík, and S. Yousef, *A posteriori error estimates, stopping criteria, and adaptivity for multiphase compositional Darcy flows in porous media*, J. Comput. Phys., 276 (2014), pp. 163–187, https://doi.org/10.1016/j.jcp.2014.06.061.

[70] D. A. Di Pietro, E. Flauraud, M. Vohralík, and S. Yousef, *A posteriori error estimates, stopping criteria, and adaptivity for multiphase compositional Darcy flows in porous media*, J. Comput. Phys., 276 (2014), pp. 163–187.

[71] D. A. Di Pietro, M. Vohralík, and S. Yousef, *An a posteriori-based, fully adaptive algorithm with adaptive stopping criteria and mesh refinement for thermal multiphase compositional flows in porous media*, Comput. Math. Appl., 68 (2014), pp. 2331–2347, https://doi.org/10.1016/j.camwa.2014.08.008.

[72] S. C. Eisenstat and H. F. Walker, *Globally convergent inexact Newton methods*, SIAM J. Optim., 4 (1994), pp. 393–422, https://doi.org/10.1137/0804022.

[73] S. C. Eisenstat and H. F. Walker, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32, https://doi.org/10.1137/0917003. Special issue on iterative methods in numerical linear algebra (Breckenridge, CO, 1994).

[74] L. El Alaoui and A. Ern, *Residual and hierarchical a posteriori error estimates for nonconforming mixed finite element methods*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 903–929, https://doi.org/10.1051/m2an:2004044.

[75] Y. Epshteyn and B. Rivière, *Analysis of hp discontinuous Galerkin methods for incompressible two-phase flow*, J. Comput. Appl. Math., 225 (2009), pp. 487–509, https://doi.org/10.1016/j.cam.2008.08.026.

[76] B. Erdmann, M. Frei, R. H. W. Hoppe, R. Kornhuber, and U. Wiest, *Adaptive finite element methods for variational inequalities*, East-West J. Numer. Math., 1 (1993), pp. 165–197.

[77] A. Ern and J.-L. Guermond, *Theory and practice of finite elements*, vol. 159 of Applied Mathematical Sciences, Springer-Verlag, New York, 2004, https://doi.org/10.1007/978-1-4757-4355-5.

[78] A. Ern, I. Smears, and M. Vohralík, *Guaranteed, locally space-time efficient, and polynomial-degree robust a posteriori error estimates for high-order discretizations of parabolic problems*, SIAM J. Numer. Anal., 55 (2017), pp. 2811–2834, https://doi.org/10.1137/16M1097626.

[79] A. Ern, I. Smears, and M. Vohralík, *Equilibrated flux a posteriori error estimates in $L^2(H^1)$-norms for high-order discretizations of parabolic problems*, IMA Journal of Numerical Analysis, (2018), https://doi.org/10.1093/imanum.

[80] A. Ern and M. Vohralík, *A posteriori error estimation based on potential and flux reconstruction for the heat equation*, SIAM J. Numer. Anal., 48 (2010), pp. 198–223, https://doi.org/10.1137/090759008.

[81] A. Ern and M. Vohralík, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*, SIAM J. Sci. Comput., 35 (2013), pp. A1761–A1791, https://doi.org/10.1137/120896918.

[82] A. Ern and M. Vohralík, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM J. Numer. Anal., 53 (2015), pp. 1058–1081, https://doi.org/10.1137/130950100.

[83] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, 1997.

[84] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of numerical analysis, Vol. VII, Handb. Numer. Anal., VII, North-Holland, Amsterdam, 2000, pp. 713–1020.

[85] R. Eymard, R. Herbin, and A. Michel, *Mathematical study of a petroleum-engineering scheme*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 937–972, https://doi.org/10.1051/m2an:2003062.

[86] F. Facchinei and C. Kanzow, *A nonsmooth inexact Newton method for the solution of large-scale nonlinear complementarity problems*, Math. Programming, 76 (1997), pp. 493–512.

[87] F. Facchinei, C. Kanzow, and S. Sagratella, *Solving quasi-variational inequalities via their KKT conditions*, Math. Program., 144 (2014), pp. 369–412.

[88] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems. Vol. I*, Springer Series in Operations Research, Springer-Verlag, New York, 2003.

[89] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems. Vol. II*, Springer Series in Operations Research, Springer-Verlag, New York, 2003.

[90] R. W. Falta, K. Pruess, I. Javandel, and P. Witherspoon, *Numerical modeling of steam injection for the removal of nonaqeous phase liquids from the subsurface.*, Water. Resour. Res, 28 (1992), pp. 433–449.

[91] F. Fierro and A. Veeser, *A posteriori error estimators, gradient recovery by averaging, and superconvergence*, Numer. Math., 103 (2006), pp. 267–298, https://doi.org/10.1007/s00211-005-0671-9.

[92] Z. Ge, Q. Ni, and X. Zhang, *A smoothing inexact Newton method for variational inequalities with nonlinear constraints*, J. Inequal. Appl., (2017).

[93] R. Glowinski, *Numerical methods for nonlinear variational problems*, Scientific Computation, Springer-Verlag, Berlin, 2008. Reprint of the 1984 original.

[94] R. Glowinski, J.-L. Lions, and R. Trémolières, *Analyse numérique des inéquations variationnelles. Tome 1*, Dunod, Paris, 1976. Théorie générale premiéres applications, Méthodes Mathématiques de l'Informatique, 5.

[95] R. Glowinski, J.-L. Lions, and R. Trémolières, *Analyse numérique des inéquations variationnelles. Tome 2*, Dunod, Paris, 1976. Applications aux phénomènes stationnaires et d'évolution, Méthodes Mathématiques de l'Informatique, 5.

[96] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical analysis of variational inequalities*, vol. 8 of Studies in Mathematics and its Applications, North-Holland Publishing Co., Amsterdam-New York, 1981. Translated from the French.

[97] E. Godlewski and P.-A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws*, vol. 118 of Applied Mathematical Sciences, Springer-Verlag, New York, 1996, https://doi.org/10.1007/978-1-4612-0713-9.

[98] S. Gross and A. Reusken, *Numerical methods for two-phase incompressible flows*, vol. 40 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2011, https://doi.org/10.1007/978-3-642-19686-7.

[99] T. Gudi and K. Porwal, *A posteriori error control of discontinuous Galerkin methods for elliptic obstacle problems*, Math. Comp., 83 (2014), pp. 579–602, https://doi.org/10.1090/S0025-5718-2013-02728-7.

[100] T. Gudi and K. Porwal, *A remark on the a posteriori error analysis of discontinuous Galerkin methods for the obstacle problem*, Comput. Methods Appl. Math., 14 (2014), pp. 71–87, https://doi.org/10.1515/cmam-2013-0015.

[101] T. Gudi and K. Porwal, *A posteriori error estimates of discontinuous Galerkin methods for the Signorini problem*, J. Comput. Appl. Math., 292 (2016), pp. 257–278, https://doi.org/10.1016/j.cam.2015.07.008.

[102] W. Hackbusch, *Multigrid methods and applications*, vol. 4 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1985, https://doi.org/10.1007/978-3-662-02427-0.

[103] R. Helmig, *Multiphase flow and transport processes in the subsurface–A contribution to the modeling of hydrosystems*, Springer-Verlag, Berlin, Heidelberg, 1997.

[104] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888 (2003), https://doi.org/10.1137/S1052623401383558.

[105] I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek, *Solution of variational inequalities in mechanics*, vol. 66 of Applied Mathematical Sciences, Springer-Verlag, New York, 1988, https://doi.org/10.1007/978-1-4612-1048-1. Translated from the Slovak by J. Jarník.

[106] R. Huber and R. Helmig, *Node-centered finite volume discretizations for the numerical simulation of multiphase flow in heterogeneous porous media*, Comput. Geosci., 4 (2000), pp. 141–164, https://doi.org/10.1023/A:1011559916309.

[107] J. Jaffré and A. Sboui, *Henry' law and gas phase disappearance*, Transport in porous media, 82 (2010), pp. 521–526.

[108] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590, https://doi.org/10.1137/08073706X.

[109] C. Johnson, *Numerical solution of partial differential equations by the finite element method*, Cambridge University Press, Cambridge, 1987.

[110] L. V. Kantorovich, *Functional analysis and applied mathematics*, NBS Rep. 1509, U. S. Department of Commerce, National Bureau of Standards, Los Angeles, Calif., 1952. Translated by C. D. Benster.

[111] C. Kanzow, *An active set-type Newton method for constrained nonlinear systems*, in Complementarity: applications, algorithms and extensions (Madison, WI, 1999), vol. 50 of Appl. Optim., Kluwer Acad. Publ., Dordrecht, 2001, pp. 179–200, https://doi.org/10.1007/978-1-4757-3279-5_9.

[112] C. Kanzow, *Inexact semismooth Newton methods for large-scale complementarity problems*, Optim. Methods Softw., 19 (2004), pp. 309–325. The First International Conference on Optimization Methods and Software. Part II.

[113] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399, https://doi.org/10.1137/S0036142902405217.

[114] C. T. Kelley, *Iterative methods for linear and nonlinear equations*, vol. 16 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995. With separately available software.

[115] C. T. Kelley, *Solving nonlinear equations with Newton's method*, vol. 1 of Fundamentals of Algorithms, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003, https://doi.org/10.1137/1.9780898718898.

[116] R. Kornhuber, *A posteriori error estimates for elliptic variational inequalities*, Comput. Math. Appl., 31 (1996), pp. 49–60.

[117] S. Lacroix, Y. Vassilevski, J. Wheeler, and M. Wheeler, *Iterative solution methods for modeling multiphase flow in porous media fully implicitly*, SIAM J. Sci. Comput., 25 (2003), pp. 905–926, https://doi.org/10.1137/S106482750240443X.

[118] P. Ladevèze, *Comparaison de modèles de mécanique des milieux continus*, Thèse d'état, Université Paris VI, Paris, (1975).

[119] A. Lauser, C. Hager, R. Helmig, and B. Wohlmuth, *A new approach for phase transitions in miscible multi-phase flow in porous media*, Advances in Water Resources, 68 (2011), pp. 957–966.

[120] J. Liesen and Z. Strakoš, *Krylov subspace methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013. Principles and analysis.

[121] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod; Gauthier-Villars, Paris, 1969.

[122] J.-L. Lions and G. Stampacchia, *Variational inequalities*, Comm. Pure Appl. Math., 20 (1967), pp. 493–519.

[123] F. Louf, J.-P. Combe, and J.-P. Pelle, *Constitutive error estimator for the control of contact problems involving friction*, Comput. & Structures, 81 (2003), pp. 1759–1772, https://doi.org/10.1016/S0045-7949(03)00200-1.

[124] J. Lovíšek, *Optimal control of a variational inequality with possibly nonsymmetric linear operator. Application to the obstacle problems in mathematical physics*, Acta Math. Univ. Comenian. (N.S.), 63 (1994), pp. 1–23.

[125] A. Lozinski, M. Picasso, and V. Prachittham, *An anisotropic error estimator for the Crank-Nicolson method: application to a parabolic problem*, SIAM J. Sci. Comput., 31 (2009), pp. 2757–2783, https://doi.org/10.1137/080715135.

[126] R. Luce and B. I. Wohlmuth, *A local a posteriori error estimator based on equilibrated fluxes*, SIAM J. Numer. Anal., 42 (2004), pp. 1394–1414, https://doi.org/10.1137/S0036142903433790.

[127] J. M. Martínez and L. Q. Qi, *Inexact Newton methods for solving nonsmooth equations*, J. Comput. Appl. Math., 60 (1995), pp. 127–145. Linear/nonlinear iterative methods and verification of solution (Matsuyama, 1993).

[128] D. Meidner, R. Rannacher, and J. Vihharev, *Goal-oriented error control of the iterative solution of finite element equations*, J. Numer. Math., 17 (2009), pp. 143–172.

[129] K.-S. Moon, R. H. Nochetto, T. V. Petersdorff, and C.-S. Zhang, *A posteriori error analysis for parabolic variational inequalities*, ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique, 41 (2007), pp. 485–511, https://doi.org/10.1051/m2an:2007029.

[130] S. Nicaise and N. Soualem, *A posteriori error estimates for a nonconforming finite element discretization of the heat equation*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 319–348, https://doi.org/10.1051/m2an:2005009.

[131] J. Niessner and R. Helmig, *Multi-scale modeling of three-phase-three-component processes in heterogeneous porous media*, Advances in Water Resources, 30 (2007), pp. 2309–2325.

[132] M. A. Olshanskii and E. E. Tyrtyshnikov, *Iterative methods for linear systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014, https://doi.org/10.1137/1.9781611973464. Theory and applications.

[133] J. M. Ortega, *The Newton-Kantorovich theorem*, Amer. Math. Monthly, 75 (1968), pp. 658–660, https://doi.org/10.2307/2313800.

[134] M. Panfilov and I. Panfilova, *Method of negative saturations for flow with variable number of phases in porous media: extension to three-phase multicomponent case*, Comput. Geosci., 18 (2014), pp. 385–399.

[135] M. Panfilov and M. Rasoulzadeh, *Interfaces of phase transition and disappearance and method of negative saturation for compositional flow with diffusion and capillarity in porous media*, Transp. Porous. Med, 83 (2010), pp. 73–98.

[136] J. Papež, U. Rüde, M. Vohralík, and B. Wohlmuth, *Sharp algebraic and total a posteriori error bounds for h and p finite elements via a multilevel approach*. HAL Preprint 01662944, submitted for publication, 2017, https://hal.inria.fr/hal-01662944/.

[137] L. E. Payne and H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Rational Mech. Anal., 5 (1960), pp. 286–292 (1960), https://doi.org/10.1007/BF00252910.

[138] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math., 5 (1947), pp. 241–269, https://doi.org/10.1090/qam/25902.

[139] J.-P. Puel, *Inéquations variationnelles d'évolution paraboliques du 2ème ordre*, Séminaire Equations aux dérivées partielles (Polytechnique), 8 (1974), pp. 1–12.

[140] A. Quarteroni and A. Valli, *Numerical approximation of partial differential equations*, vol. 23 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1994.

[141] R. Rankin and B. Rivière, *A high order method for solving the black-oil problem in porous media*, Advances in Water Resources, 78 (2015), pp. 126–144.

[142] P.-A. Raviart and J.-M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975, Springer, Berlin, 1977, pp. 292–315. Lecture Notes in Math., Vol. 606.

[143] S. Repin, *A posteriori estimates for partial differential equations*, vol. 4 of Radon Series on Computational and Applied Mathematics, Walter de Gruyter GmbH & Co. KG, Berlin, 2008, https://doi.org/10.1515/9783110203042.

[144] S. I. Repin, *Functional a posteriori estimates for elliptic variational inequalities*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 348 (2007), pp. 147–164, 305, https://doi.org/10.1007/s10958-008-9093-4.

[145] B. Rivière, *Discontinuous Galerkin methods for solving elliptic and parabolic equations*, vol. 35 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008, https://doi.org/10.1137/1.9780898717440. Theory and implementation.

[146] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 523–639.

[147] S. M. Robinson, *Corrigenda to: "Local structure of feasible sets in nonlinear programming. III. Stability and sensitivity" [Math. Programming Stud. No. 30 (1987), 45–66; MR0874131 (88j:90198)]*, Math. Programming, 49 (1990/91), p. 143, https://doi.org/10.1007/BF01588784.

[148] J.-F. Rodrigues, *Obstacle problems in mathematical physics*, vol. 134 of North-Holland Mathematics Studies, North-Holland Publishing Co., Amsterdam, 1987. Notas de Matemática [Mathematical Notes], 114.

[149] Y. Saad, *Iterative methods for sparse linear systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, second ed., 2003, https://doi.org/10.1137/1.9780898718003.

[150] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869, https://doi.org/10.1137/0907058.

[151] W. T. Sha, *Novel porous media formulation for multiphase flow conservation equations*, Cambridge University Press, Cambridge, 2011, https://doi.org/10.1017/CBO9781139003407. With forewords by Alan Schriesheim, Wm. Howard Arnold and Charles Kelber.

[152] G. Strang, *The finite element method and approximation theory*, in Numerical Solution of Partial Differential Equations, II (SYNSPADE 1970) (Proc. Sympos., Univ. of Maryland, College Park, Md., 1970), Academic Press, New York, 1971, pp. 547–583.

[153] A. Veeser, *Efficient and reliable a posteriori error estimators for elliptic obstacle problems*, SIAM J. Numer. Anal., 39 (2001), pp. 146–167, https://doi.org/10.1137/S0036142900370812.

[154] R. Verfürth, *A posteriori error estimates for nonlinear problems: $L^r(0,T;W^{1,\rho}(\Omega))$-error estimates for finite element discretizations of parabolic equations*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 487–518, https://doi.org/10.1002/(SICI)1098-2426(199807)14:4<487::AID-NUM4>3.0.CO;2-G.

[155] R. Verfürth, *A posteriori error estimation techniques for finite element methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013, https://doi.org/10.1093/acprof:oso/9780199679423.001.0001.

[156] M. Vohralík, *Residual flux-based a posteriori error estimates for finite volume and related locally conservative methods*, Numer. Math., 111 (2008), pp. 121–158.

[157] M. Vohralík, *Guaranteed and fully robust a posteriori error estimates for conforming discretizations of diffusion problems with discontinuous coefficients*, J. Sci. Comput., 46 (2011), pp. 397–438, https://doi.org/10.1007/s10915-010-9410-1.

[158] M. Vohralík and M. F. Wheeler, *A posteriori error estimates, stopping criteria, and adaptivity for two-phase flows*, Comput. Geosci., 17 (2013), pp. 789–812, https://doi.org/10.1007/s10596-013-9356-0.

[159] F. Wang, W. Han, and X.-L. Cheng, *Discontinuous Galerkin methods for solving elliptic variational inequalities*, SIAM J. Numer. Anal., 48 (2010), pp. 708–733, https://doi.org/10.1137/09075891X.

[160] S. J. Wright, *Primal-dual interior-point methods*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997, https://doi.org/10.1137/1.9781611971453.

[161] I. Yotov, *A multilevel Newton-Krylov interface solver for multiphysics couplings of flow in porous media*, Numer. Linear Algebra Appl., 8 (2001), pp. 551–570, https://doi.org/10.1002/nla.263. Solution methods for large-scale non-linear problems (Pleasanton, CA, 2000).