Adaptive inexact semismooth Newton methods for the contact problem between two membranes

Jad Dabaghi, Vincent Martin, Martin Vohralík

Adaptive inexact semismooth Newton methods for the contact problem between two membranes*

Jad Dabaghi [†‡] Vincent Martin [§] Martin Vohralík[†‡]

October 19, 2018

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 variational inequalities. 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 the developed adaptive method.

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

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

$$\begin{aligned} \mathbb{E}\boldsymbol{X}_h &= \boldsymbol{F}, \\ \boldsymbol{K}(\boldsymbol{X}_h) &\geq \boldsymbol{0}, \quad \boldsymbol{G}(\boldsymbol{X}_h) \geq \boldsymbol{0}, \quad \boldsymbol{K}(\boldsymbol{X}_h) \perp \boldsymbol{G}(\boldsymbol{X}_h), \end{aligned} \tag{1}$$

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) typically represents the discretization of a linear partial differential equation (PDE) (the model example for this study is described further in (5)). The second line of (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), see for example the overview of Facchinei and Pang [38, 39], the book of Bonnans *et al.* [16], and the survey of Aganagić [1]. In particular, we mention

*This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647134 GATIPOR).

†Inria, 2 rue Simone Iff, 75589 Paris, France

‡Université Paris-Est, CERMICS (ENPC), 77455 Marne-la-Vallée 2, France

§Université technologie de Compiègne (UTC), 60200, France

the approach by interior point method of Wright [61], the active set strategy by Kanzow [44], and the primal-dual active set strategy by Hintermüler *et al.* [41]. 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 [30, 37, 38, 39]. The C-functions are not smooth in the classical sense (Fréchet-differentiable), but admit a weaker smoothness called the Clarke derivative, see [28]. Semismooth Newton algorithms like the Newton-min are often used in practice as they show local quadratic convergence properties, see [14, 15, 13, 38, 39].

The goal of the present study is to perform an a posteriori analysis of problem (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). This yields an equivalent formulation of (1) that requests to find a vector $\boldsymbol{X}_h \in \mathbb{R}^n$ such that

$$\mathcal{S}(\boldsymbol{X}_h) = \boldsymbol{0}, \tag{2}$$

where $\mathcal{S} : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear non-differentiable functional. Next, let any semismooth nonlinear solver be applied to system (2), yielding at a semismooth step $k \geq 1$ a linear system

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^k = \boldsymbol{B}^{k-1}, \tag{3}$$

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 (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 (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 (3). Similarly, $\boldsymbol{X}_h^{k,i}$ does not solve (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 books of Verfürth [60], Ainsworth [2] and Repin [55] for a general introduction). Concerning a posteriori error estimates for variational inequalities discretized as in (1) or (2), let us mention the pioneering work of Brezzi, Hager, and Raviart [21, 22], next Ainsworth, Oden, and Lee [3], Kornhuber [47], Repin [56] and Bürg and Schröder [24]. For the elliptic obstacle problem we can more precisely mention the papers of Chen and Nochetto [27], see also the references therein, Veeser [59], Bartels and Carstensen [6], and Braess [17]. Not to solve (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 [23, 31, 33, 40, 46] for the case of inexact Newton methods and in [36, 38, 39, 45, 50] 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 (3) and (2). We follow the approach by equilibrated flux reconstructions, solving auxiliary local problems (see Destuynder and Métivet [32] and Braess and Schöberl [19]). 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 [8], Coorevits, Hild, Pelle [29], Louf, Combe, Pelle [49], Jiránek *et al.* [43], Ern and Vohralík [34], Arioli, Georgoulis, Loghin [4], and Papež *et al.* [52], our estimate takes the form

$$e(\boldsymbol{X}_h^{k,i}) \leq \eta(\boldsymbol{X}_h^{k,i}) = \eta_{\text{disc}}^{k,i} + \eta_{\text{lin}}^{k,i} + \eta_{\text{alg}}^{k,i}, \tag{4}$$

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 $\eta_{\text{alg}}^{k,i}$ (respectively the linearization estimator $\eta_{\text{lin}}^{k,i}$) 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 (3) and (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

$$\left\{\begin{array}{llll} -\mu_1\Delta u_1 - \lambda = f_1 & \text{in} & \Omega, & \\ -\mu_2\Delta u_2 + \lambda = f_2 & \text{in} & \Omega, & \\ (u_1-u_2)\lambda = 0, \quad u_1 - u_2 \geq 0, \quad \lambda \geq 0 & \text{in} & \Omega, & \\ u_1 = g & \text{on} & \partial\Omega, & \\ u_2 = 0 & \text{on} & \partial\Omega, & \end{array}\right. \tag{5}$$

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 (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 ($u_1 = u_2$, $\lambda \geq 0$). This kind of variational inequalities, 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.* [42] and Rodrigues [58] for general concepts. For the quadratic finite element approximation of the Signorini problem refer to [9, 5]. The existence and uniqueness of a weak solution of (5) follows by Lions and Stampacchia [48], an a priori analysis for linear finite elements was performed in [10, 11], and an a posteriori analysis was undertaken in [12]. Therein, however, it was supposed that the discrete system (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 2, we give details on the model problem (5) and its finite element discretization for all polynomial degrees $p \geq 1$, which is new to the best of our knowledge when $p > 1$. This leads to algebraic systems of the form (1). In Section 3, we present the concept of the inexact semismooth Newton method giving rise to systems (2)–(3). The various flux reconstructions, in particular employing a multilevel mesh hierarchy for the algebraic error components following Papež *et al.* [52], are described in Section 4. Next, Section 5 is dedicated to the construction of the a posteriori error estimate of the form (4). In Section 6, we present the adaptive inexact semismooth algorithm and in Section 7, we prove the converse inequality to (4) up to a generic constant and up to a typically small contact term, assessing the quality of our estimates. Finally, Section 8 is devoted, for $p = 1$, to numerical experiments that confirm the theoretical results and Section 9 summarizes our conclusions.

2 Model problem and its finite element discretization

In this section, we set up the notation, describe in details the model problem (5), and introduce its finite element discretization for all polynomial degrees $p \geq 1$.

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 $L^2(\Omega)$. Moreover, we define the sets $H^1_g(\Omega) := \{v \in H^1(\Omega),\ v = g \text{ on } \partial\Omega\}$ and $\Lambda := \{\chi \in L^2(\Omega),\ \chi \geq 0 \text{ a.e. in } \Omega\}$. 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 [7, 53], state that if $\overline{v}_\mathcal{O}$ denotes the mean value of v and $h_\mathcal{O}$ the diameter of $\mathcal{O}$, then

$$\|v\|_\mathcal{O} \leq C_{\mathrm{PF}} h_\mathcal{O} \|\boldsymbol{\nabla} v\|_\mathcal{O} \qquad \forall v \in H^1_0(\mathcal{O}), \tag{6a}$$

$$\|v - \overline{v}_\mathcal{O}\|_\mathcal{O} \leq C_{\mathrm{PW}} h_\mathcal{O} \|\boldsymbol{\nabla} v\|_\mathcal{O} \qquad \forall v \in H^1(\mathcal{O}). \tag{6b}$$

The constants $C_{\rm PF}$ and $C_{\rm PW}$ can be precisely estimated in many cases. In particular, if $\mathcal{O}$ is convex, $C_{\rm PW}$ can be taken as $\frac{1}{\pi}$, see [7, 53] whereas $C_{\rm PF}$ is at most 1. Then, we define the energy norm:

$$\forall \boldsymbol{v} = (v_1, v_2) \in [H_0^1(\mathcal{O})]^2, \quad |||\boldsymbol{v}|||_{\mathcal{O}} := \left\{ \sum_{\alpha=1}^{2} \mu_\alpha \left\| \boldsymbol{\nabla} v_\alpha \right\|_{\mathcal{O}}^2 \right\}^{\frac{1}{2}}. \tag{7}$$

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:

$$\forall v \in H^{-1}(\mathcal{O}), \quad |||v|||_{H_*^{-1}(\mathcal{O})} := \sup_{\psi \in H_0^1(\mathcal{O}),\ \max(\mu_1^{\frac{1}{2}}, \mu_2^{\frac{1}{2}}) \|\boldsymbol{\nabla}\psi\|_{\mathcal{O}} = 1} \langle v, \psi \rangle. \tag{8}$$

2.1 Continuous and reduced problem

For $(f_1, f_2) \in [L^2(\Omega)]^2$ and g a positive constant, the weak formulation corresponding to (5) consists in: find $(u_1, u_2, \lambda) \in H_g^1(\Omega) \times H_0^1(\Omega) \times \Lambda$ such that

$$\begin{cases} \displaystyle\sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_\alpha, \boldsymbol{\nabla} v_\alpha \right)_\Omega - \left(\lambda, v_1 - v_2 \right)_\Omega = \sum_{\alpha=1}^{2} \left(f_\alpha, v_\alpha \right)_\Omega \quad \forall\, (v_1, v_2) \in [H_0^1(\Omega)]^2, \\ \left(\chi - \lambda, u_1 - u_2 \right)_\Omega \geq 0 \quad \forall \chi \in \Lambda. \end{cases} \tag{9}$$

Following [11, Proposition 1], problems (5) and (9) are equivalent and the latter admits a unique weak solution. Setting $\boldsymbol{u} := (u_1, u_2)$, $\boldsymbol{v} := (v_1, v_2)$, and defining

$$a(\boldsymbol{u}, \boldsymbol{v}) := \sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_\alpha, \boldsymbol{\nabla} v_\alpha \right)_\Omega, \ b(\boldsymbol{v}, \chi) := \left(\chi, v_1 - v_2 \right)_\Omega, \ l(\boldsymbol{v}) := \sum_{\alpha=1}^{2} \left(f_\alpha, v_\alpha \right)_\Omega, \tag{10}$$

problem (9) rewrites: find $\boldsymbol{u} \in H_g^1(\Omega) \times H_0^1(\Omega)$ and $\lambda \in \Lambda$ such that

$$\begin{cases} a(\boldsymbol{u}, \boldsymbol{v}) - b(\boldsymbol{v}, \lambda) &=& l(\boldsymbol{v}) & \forall \boldsymbol{v} \in [H_0^1(\Omega)]^2, \\ b\left(\boldsymbol{u}, \chi - \lambda\right) &\geq& 0 & \forall \chi \in \Lambda. \end{cases}$$

Next, we define the convex set $\boldsymbol{\mathcal{K}}_g$ by

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

The reduced variational problem reads: find $\boldsymbol{u} = (u_1, u_2) \in \boldsymbol{\mathcal{K}}_g$ such that

$$a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq l(\boldsymbol{v} - \boldsymbol{u}) \qquad \forall \boldsymbol{v} = (v_1, v_2) \in \ \boldsymbol{\mathcal{K}}_g. \tag{11}$$

In [11], problems (11) and (9) are shown to be equivalent, in the sense that both of them admit the same unique solution $\boldsymbol{u}$.

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 $\mathcal{V}_h^{\rm int}$ and the boundary vertices $\mathcal{V}_h^{\rm ext}$. We denote by $N_h^{\rm int}$ 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

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

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 $\mathcal{V}_{\rm d}^p$ the set of the Lagrange degrees of freedom (values in the points $\boldsymbol{x}_l$) of the space X_h^p and by $\mathcal{N}_{\rm d}^p$ its cardinality. The internal degrees of freedom are collected in the set $\mathcal{V}_{\rm d}^{p,\rm int}$ (with $\mathcal{N}_{\rm d}^{p,\rm int}$ its cardinality) and the external ones are collected in the set $\mathcal{V}_{\rm d}^{p,\rm ext}$. The Lagrange basis functions of X_h^p are denoted by $(\psi_{h,\boldsymbol{x}_l})_{1\leq l\leq \mathcal{N}_{\rm d}^p}$ 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 $(\boldsymbol{x}_{l'})_{1\leq l'\neq l\leq \mathcal{N}_{\rm d}^p} \in \mathcal{V}_{\rm d}^p$. 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

$$\boldsymbol{\mathcal{K}}_{gh}^p := \left\{(v_{1h}, v_{2h}) \in X_{gh}^p \times X_{0h}^p,\ v_{1h}(\boldsymbol{x}_l) - v_{2h}(\boldsymbol{x}_l) \geq 0\ \ \forall\, (\boldsymbol{x}_l)_{1\leq l\leq \mathcal{N}_{\rm d}^{p,\rm int}} \in \mathcal{V}_{\rm d}^{p,\rm int}\right\}. \tag{12}$$

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 [10, 27, 59]. The discrete counterpart to (11) consists in: find $\boldsymbol{u}_h = (u_{1h}, u_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ such that

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) \geq l(\boldsymbol{v}_h - \boldsymbol{u}_h) \qquad \forall \boldsymbol{v}_h = (v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p. \tag{13}$$

As a result of the Lions–Stampacchia theorem, problem (13) admits a unique solution that is nonconforming ($\boldsymbol{u}_h \notin \boldsymbol{\mathcal{K}}_g$) when $p > 1$. Moreover, following the methodology of [11, equation (4.5)] or [24] we define the functions λ_{1h} and λ_{2h} in X_h^p by

$$\left\{\begin{array}{rcll} \langle\lambda_{1h}, z_{1h}\rangle_h & := & \mu_1\,(\boldsymbol{\nabla} u_{1h}, \boldsymbol{\nabla} z_{1h})_\Omega - (f_1, z_{1h})_\Omega & \forall z_{1h} \in X_{0h}^p, \\ \langle\lambda_{2h}, z_{2h}\rangle_h & := & -\mu_2\,(\boldsymbol{\nabla} u_{2h}, \boldsymbol{\nabla} z_{2h})_\Omega + (f_2, z_{2h})_\Omega & \forall z_{2h} \in X_{0h}^p, \\ \langle\lambda_{1h}, \psi_{h,\boldsymbol{x}_l}\rangle_h & := & 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm ext}, & \\ \langle\lambda_{2h}, \psi_{h,\boldsymbol{x}_l}\rangle_h & := & 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm ext}, & \end{array}\right. \tag{14}$$

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

$$\langle w_h, v_h\rangle_h := \left\{\begin{array}{ll} \displaystyle\sum_{\boldsymbol{a}\in\mathcal{V}_h} w_h(\boldsymbol{a}) v_h(\boldsymbol{a}) M_{\boldsymbol{a}} & \text{if} \quad p = 1, \qquad (15) \\ (w_h, v_h)_\Omega & \text{if} \quad p \geq 2. \qquad (16) \end{array}\right.$$

Note that (15) corresponds to the use of a mass lumping and will lead to particular properties in the case $p = 1$. Extending [10, Proposition 12] to the case $p > 1$ we have:

Lemma 2.1. *Let $(u_{1h}, u_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ be the solution of the reduced discrete problem (13). Then, the functions λ_{1h} and λ_{2h} defined by (14) coincide and we set $\lambda_h := \lambda_{1h} = \lambda_{2h} \in X_h^p$.*

Proof. We substract the first two equations of (14) taking $z_{1h} = z_{2h} = \psi_{h,\boldsymbol{x}_l}$ with $\boldsymbol{x}_l$ any internal Lagrange node to get

$$\langle\lambda_{1h} - \lambda_{2h}, \psi_{h,\boldsymbol{x}_l}\rangle_h = (\mu_1 \boldsymbol{\nabla} u_{1h} + \mu_2 \boldsymbol{\nabla} u_{2h}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{x}_l})_\Omega - (f_1 + f_2, \psi_{h,\boldsymbol{x}_l})_\Omega \quad \forall l = 1 \ldots \mathcal{N}_{\rm d}^{p,\rm int}.$$

Taking $v_{1h} := u_{1h} + \psi_{h,\boldsymbol{x}_l}$ and $v_{2h} := u_{2h} + \psi_{h,\boldsymbol{x}_l}$ in (13) and noting that $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ where $\psi_{h,\boldsymbol{x}_l}$ is the Lagrange basis function associated to the internal node $\boldsymbol{x}_l$, we get

$$\langle\lambda_{1h} - \lambda_{2h}, \psi_{h,\boldsymbol{x}_l}\rangle_h \geq 0 \quad \forall l = 1 \ldots \mathcal{N}_{\rm d}^{p,\rm int}. \tag{17}$$

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

$$\langle \lambda_{1h} - \lambda_{2h}, \psi_{h,\boldsymbol{x}_l} \rangle_h \leq 0 \quad \forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}. \tag{18}$$

Finally, combining (17) and (18) with the last two equations of (14) 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 2.2. *Let $(u_{1h}, u_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ be the solution of the reduced problem (13) and let λ_h be defined by (14). Then, there holds*

$$\langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}. \tag{19}$$

Proof. Let $\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$ be an internal node. First observe that $(v_{1h}, v_{2h}) := (u_{1h} + \psi_{h,\boldsymbol{x}_l}, u_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$. The conclusion follows from the reduced problem (13), the characterization (14), and Lemma 2.1 giving

$$\mu_1 \left(\boldsymbol{\nabla} u_{1h}, \boldsymbol{\nabla} \psi_{h,\boldsymbol{x}_l}\right)_\Omega - \left(f_1, \psi_{h,\boldsymbol{x}_l}\right)_\Omega = \langle \lambda_{1h}, \psi_{h,\boldsymbol{x}_l} \rangle_h = \langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall l = 1 \dots \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}.$$

□

Following Lemma 2.2 we suggest to define the discrete convex set for λ_h by

$$\Lambda_h^p := \left\{ v_h \in X_h^p;\ \langle v_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0\ \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}},\ \langle v_h, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0\ \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{ext}} \right\}. \tag{20}$$

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

$$\Lambda_h^1 = \left\{ v_h \in X_{0h}^1;\ v_h(\boldsymbol{a}) \geq 0 \quad \forall \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}} \right\} \subset \Lambda, \tag{21}$$

which is the same as in [10, Section 4]. Note that when $\chi_h \in \Lambda_h^p$ and $\boldsymbol{v}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$,

$$\langle \chi_h, v_{1h} - v_{2h} \rangle_h = \sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}} \left(v_{1h} - v_{2h}\right)(\boldsymbol{x}_l) \left\langle \chi_h, \psi_{h,\boldsymbol{x}_l} \right\rangle_h \geq 0, \tag{22}$$

A discrete formulation, built by the Galerkin method corresponding to problem (9) consists in: find $(u_{1h}, u_{2h}, \lambda_h) \in X_{gh}^p \times X_{0h}^p \times \Lambda_h^p$ such that

$$\begin{aligned} &\sum_{\alpha=1}^{2} \mu_\alpha \left(\boldsymbol{\nabla} u_{\alpha h}, \boldsymbol{\nabla} z_{\alpha h}\right)_\Omega - \langle \lambda_h, z_{1h} - z_{2h} \rangle_h = \sum_{\alpha=1}^{2} \left(f_\alpha, z_{\alpha h}\right)_\Omega \quad \forall (z_{1h}, z_{2h}) \in [X_{0h}^p]^2, \\ &\langle \chi_h - \lambda_h, u_{1h} - u_{2h} \rangle_h \geq 0 \quad \forall \chi_h \in \Lambda_h^p. \end{aligned} \tag{23}$$

Lemma 2.3. *For any solution $(u_{1h}, u_{2h}, \lambda_h)$ of problem (23), the pair (u_{1h}, u_{2h}) is a solution of problem (13). Conversely, for any solution (u_{1h}, u_{2h}) of problem (13), defining the function $\lambda_h = \lambda_{\alpha h}$, $\alpha = 1, 2$ by (14), the triple $(u_{1h}, u_{2h}, \lambda_h)$ is a solution to problem (23).*

Proof. For the case $p = 1$, the proof is given in [11, Lemma 13]. Let $p \geq 2$ and let $(u_{1h}, u_{2h}, \lambda_h)$ be the solution of problem (23). Decomposing $u_{1h} - u_{2h}$ in the Lagrange basis we obtain

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

Next, note that for $\chi_h, \lambda_h \in \Lambda_h^p$, $\chi_h + \lambda_h \in \Lambda_h^p$. Take $\chi_h = \chi_h + \lambda_h$ as a test function in (23), we get

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

and then

$$\sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^p} (u_{1h} - u_{2h})(\boldsymbol{x}_l) \langle \chi_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0 \quad \forall \chi_h \in \Lambda_h^p. \tag{24}$$

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

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

Note that each vector of the dual basis $\Theta_{h,\boldsymbol{x}_l}$ can be determined by inverting the finite element mass matrix; importantly, all $\Theta_{h,\boldsymbol{x}_l}$, $1 \leq l \leq \mathcal{N}_{\mathrm{d}}^p$ belong to Λ_h^p. Finally, taking in (24) $\chi_h = \Theta_{h,\boldsymbol{x}_l^*}$, 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 (13). 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 (23) provides

$$\langle \lambda_h, v_{1h} - v_{2h} \rangle_h - \langle \lambda_h, u_{1h} - u_{2h} \rangle_h = a(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) - l(\boldsymbol{v}_h - \boldsymbol{u}_h). \tag{25}$$

Using (22) with $\lambda_h \in \Lambda_h^p$ and $\boldsymbol{v}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$ and taking $\chi_h = 0$ in (23) gives

$$\langle \lambda_h, v_{1h} - v_{2h} \rangle_h \geq 0, \quad \langle -\lambda_h, u_{1h} - u_{2h} \rangle_h \geq 0. \tag{26}$$

Combining (25), and (26) provides (13).

Conversely, let $(u_{1h}, u_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$ be the solution of the reduced problem (13) and let $(z_{1h}, z_{2h}) \in X_{0h}^p \times X_{0h}^p$ be arbitrary. The characterization (14) combined with Lemma 2.1 and Lemma 2.2 yields $\lambda_h \in \Lambda_h^p$. Next, subtracting the two equations of (14) gives the first line of (23). It remains to prove the second line of (23). Let now $(v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p$. The first line in (23) now implies (25) and the reduced problem (13) yields

$$-\langle \lambda_h, u_{1h} - u_{2h} \rangle_h + \langle \lambda_h, v_{1h} - v_{2h} \rangle_h \geq 0 \quad \forall (v_{1h}, v_{2h}) \in \boldsymbol{\mathcal{K}}_{gh}^p. \tag{27}$$

For $v_{1h} := u_{1h} - \sum_{\boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}} u_{1h}(\boldsymbol{x}_l) \psi_{h,\boldsymbol{x}_l} \in X_{gh}^p$ 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 $\langle \lambda_h, v_{1h} - v_{2h} \rangle_h = 0$ and the inequality (27) yields

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

To conclude the proof, we use (22) with $\boldsymbol{u}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$ and for any $\chi_h \in \Lambda_h^p$. □

Remark 2.4. *Taking $\chi_h = 0$ and then $\chi_h = 2\lambda_h \in \Lambda_h^p$ in the second line of (23) yields*

$$\langle \lambda_h, u_{1h} - u_{2h} \rangle_h = 0. \tag{28}$$

Combining (28) with the definition of Λ_h^p and the fact that the solution $\boldsymbol{u}_h \in \boldsymbol{\mathcal{K}}_{gh}^p$ give the complementarity conditions

$$\begin{aligned}
&(u_{1h} - u_{2h})(\boldsymbol{x}_l) \geq 0, \quad \langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h \geq 0, \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}, \quad \langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{ext}}, \\
&\langle \lambda_h, u_{1h} - u_{2h} \rangle_h = 0.
\end{aligned} \tag{29}$$

Note that when linear finite elements are employed ($p = 1$), (29) reduces to

$$(u_{1h} - u_{2h})(\boldsymbol{a}) \geq 0, \quad \lambda_h(\boldsymbol{a}) \geq 0, \quad \lambda_h(\boldsymbol{a})(u_{1h} - u_{2h})(\boldsymbol{a}) = 0 \quad \forall \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}. \tag{30}$$

Section 2.3-2.4 are dedicated, when $p = 1$, to formulate matricially problem (23) as a prelude for implementation. From the previous analysis, we deduce

Lemma 2.5. *The formulation (23) is well posed for any polynomial degree $p \geq 1$.*

Remark 2.6. *Problems (13) and (23) are equivalent in the sense of Lemma 2.3. 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.*

2.3 Discrete complementarity problem for linear finite elements, $p = 1$

We now focus on the case $p = 1$ and express the discrete problem (23)-(30) under an algebraic form using the basis $(\psi_{h,\boldsymbol{a}})_{\boldsymbol{a}\in\mathcal{V}_h^{\rm int}}$ of X_{0h}^1. Note that the functions of Λ_h^1 can also be expressed in the same basis. The affine space X_{gh}^1 is decomposed into $X_{gh}^1 = X_{0h}^1 + g$, where we recall that $g > 0$ is constant. Therefore, $u_{1h} = \hat{u}_{1h} + g$, where $\hat{u}_{1h} \in X_{0h}^1$. The components of the solution $(\hat{u}_{1h}, u_{2h}, \lambda_h)$ in the basis $(\psi_{h,\boldsymbol{a}})_{\boldsymbol{a}\in\mathcal{V}_h^{\rm int}}$ are gathered by blocks as $\boldsymbol{X}_{1h} := (\hat{u}_{1h}(\boldsymbol{a}_l))_{l=1,\dots,N_h^{\rm int}}$, $\boldsymbol{X}_{2h} := (u_{2h}(\boldsymbol{a}_l))_{l=1,\dots,N_h^{\rm int}}$, $\boldsymbol{X}_{3h} := (\lambda_h(\boldsymbol{a}_l))_{l=1,\dots,N_h^{\rm int}}$, such that $\boldsymbol{X}_h^T := (\boldsymbol{X}_{1h}, \boldsymbol{X}_{2h}, \boldsymbol{X}_{3h})^T$; Taking successively $v_{1h} = \psi_{h,\boldsymbol{a}'}$, $v_{2h} = 0$, and $v_{1h} = 0$, $v_{2h} = \psi_{h,\boldsymbol{a}'}$ for all $\boldsymbol{a}' \in \mathcal{V}_h^{\rm int}$, the first line of (23) reads

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

where the rectangular matrix $\mathbb{E} \in \mathbb{R}^{2N_h^{\rm int}, 3N_h^{\rm int}}$ is defined by

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

with the stiffness matrix $\mathbb{S} \in \mathbb{R}^{N_h^{\rm int}, N_h^{\rm int}}$ and the diagonal mass lumped matrix $\mathbb{D} \in \mathbb{R}^{N_h^{\rm int}, N_h^{\rm int}}$ respectively defined by

$$\mathbb{S}_{l,m} := (\boldsymbol{\nabla}\psi_{h,\boldsymbol{a}_m}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{a}_l})_\Omega \quad \forall 1 \le l, m \le N_h^{\rm int}, \quad \mathbb{D}_{l,l} := M_{\boldsymbol{a}_l} \quad \forall 1 \le l \le N_h^{\rm int}.$$

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

$$(\boldsymbol{F}_1)_l := (f_1, \psi_{h,\boldsymbol{a}_l})_\Omega \quad \text{and} \quad (\boldsymbol{F}_2)_l := (f_2, \psi_{h,\boldsymbol{a}_l})_\Omega, \quad \forall 1 \le l \le N_h^{\rm int}.$$

Problem (23), taking into account that (23) implies (30), can then be written under the compact form

$$\begin{array}{rl} \mathbb{E}\boldsymbol{X}_h & = \boldsymbol{F}, \\ \boldsymbol{X}_{1h} + g - \boldsymbol{X}_{2h} & \ge \mathbf{0}, \quad \boldsymbol{X}_{3h} \ge \mathbf{0}, \quad (\boldsymbol{X}_{1h} + g - \boldsymbol{X}_{2h}) \perp \boldsymbol{X}_{3h}. \end{array} \tag{31}$$

Consequently, denoting $\boldsymbol{K}(\boldsymbol{X}_h) := \boldsymbol{X}_{1h} + g - \boldsymbol{X}_{2h}$ and $\boldsymbol{G}(\boldsymbol{X}_h) := \boldsymbol{X}_{3h}$ which are respectively affine and linear, (31) fits the abstract class of problems (1) of the introduction.

2.4 C-functions

We now express the complementarity constraints in (31) via non-differentiable equations. Let us recall that a function $f : (\mathbb{R}^m)^2 \to \mathbb{R}^m$ ($m \ge 1$) is a C-function or a complementarity function if

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

Examples of C-functions are the min function

$$(\min\{\boldsymbol{x}, \boldsymbol{y}\})_l := \min\{\boldsymbol{x}_l, \boldsymbol{y}_l\} \qquad l = 1, \dots, m, \tag{32}$$

the Fischer–Burmeister function

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

or the Mangasarian function

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

where $\xi : \mathbb{R} \mapsto \mathbb{R}$ is an increasing function satisfying $\xi(0) = 0$. For more details on C-functions see [38, 39]. Let $\tilde{\mathbf{C}}$ be any C-function, *i.e.*, satisfying (for $m = N_h^{\rm int}$) $\tilde{\mathbf{C}}(\boldsymbol{X}_{1h}+g-\boldsymbol{X}_{2h}, \boldsymbol{X}_{3h}) = 0 \iff \boldsymbol{X}_{1h}+g-\boldsymbol{X}_{2h} \ge 0$, $\boldsymbol{X}_{3h} \ge 0$, and $(\boldsymbol{X}_{1h} + g - \boldsymbol{X}_{2h})\cdot\boldsymbol{X}_{3h} = 0$. Then, introducing the function $\mathbf{C} : \mathbb{R}^{3N_h^{\rm int}} \to \mathbb{R}^{N_h^{\rm int}}$ defined as $\mathbf{C}(\boldsymbol{X}_h) = \tilde{\mathbf{C}}(\boldsymbol{X}_{1h} + g - \boldsymbol{X}_{2h}, \boldsymbol{X}_{3h})$, the problem (31) can be equivalently rewritten as

$$\left\{ \begin{array}{ll} \mathbb{E}\boldsymbol{X}_h & = \boldsymbol{F}, \\ \mathbf{C}(\boldsymbol{X}_h) & = \mathbf{0}. \end{array} \right. \tag{33}$$

The C-functions that are commonly used are locally Lipschitz and continuous, thus differentiable almost everywhere as a result of the Rademacher Theorem (see [28, 38]). 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 [16, 38, 39]). For instance we can employ the Newton-min, the Newton–Fischer–Burmeister or the Newton–Mangasarian algorithms.

3 Inexact semismooth Newton methods

We address in this section the numerical solution of the system of nonlinear algebraic inequalities corresponding to (23). We assume that an iterative linearization procedure is applied such that for a given initial vector $\boldsymbol{X}_h^0 \in \mathbb{R}^{2\mathcal{N}_{\rm d}^{p,\rm int}+\mathcal{N}_{\rm d}^p}$, on step $k \geq 1$ one looks for $\boldsymbol{X}_h^k \in \mathbb{R}^{2\mathcal{N}_{\rm d}^{p,\rm int}+\mathcal{N}_{\rm d}^p}$ such that

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^k = \boldsymbol{B}^{k-1}, \tag{34}$$

where the square matrix $\mathbb{A}^{k-1}$ and the right-hand side vector are given. Let us now give details and examples in the case $p = 1$.

3.1 Example of the semismooth method: case $p = 1$

For $\boldsymbol{X}_h^0 \in \mathbb{R}^{3N_h^{\rm int}}$ given, a semismooth Newton method to solve (33) searches, on step $k \geq 1$, $\boldsymbol{X}_h^k \in \mathbb{R}^{3N_h^{\rm int}}$ such that

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

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

$$\mathbb{A}^{k-1} := \begin{bmatrix} \mathbb{E} \\ \mathbf{J_C}(\boldsymbol{X}_h^{k-1}) \end{bmatrix}, \quad \boldsymbol{B}^{k-1} := \begin{bmatrix} \boldsymbol{F} \\ \mathbf{J_C}(\boldsymbol{X}_h^{k-1})\boldsymbol{X}_h^{k-1} - \mathbf{C}(\boldsymbol{X}_h^{k-1}) \end{bmatrix}. \tag{35}$$

Note that since the first line of (33) is linear, the corresponding Jacobian is constant and equal to $\mathbb{E}$. The semismooth nonlinearity occurs in the second line of (33). The Clarke subdifferential of the semismooth C-function $\mathbf{C}$ at $\boldsymbol{X}_h^{k-1}$ is a set composed of $2^{N_h^{\rm int}}$ Jacobians (*cf.* [38, 39]). For example, if we consider the semismooth function min (32),

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

and if we define the block matrices $\mathbb{K}$ and $\mathbb{G}$ in $\mathbb{R}^{N_h^{\rm int},3N_h^{\rm int}}$ respectively by

$$\mathbb{K} := \left[\mathbf{Id}_{N_h^{\rm int}\times N_h^{\rm int}}, -\mathbf{Id}_{N_h^{\rm int}\times N_h^{\rm int}}, \mathbf{0}_{N_h^{\rm int}\times N_h^{\rm int}}\right], \quad \mathbb{G} := \left[\mathbf{0}_{N_h^{\rm int}\times N_h^{\rm int}}, \mathbf{0}_{N_h^{\rm int}\times N_h^{\rm int}}, \mathbf{Id}_{N_h^{\rm int}\times N_h^{\rm int}}\right],$$

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

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 (34). Given an initial vector $\boldsymbol{X}_h^{k,0} \in \mathbb{R}^{2\mathcal{N}_{\rm d}^{p,\rm int}+\mathcal{N}_{\rm d}^p}$, 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

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^{k,i} = \boldsymbol{B}^{k-1} - \boldsymbol{R}_h^{k,i}, \tag{36}$$

where $\boldsymbol{R}_h^{k,i} := \boldsymbol{B}^{k-1} - \mathbb{A}^{k-1}\boldsymbol{X}_h^{k,i} \in \mathbb{R}^{2\mathcal{N}_{\rm d}^{p,\rm int}+\mathcal{N}_{\rm d}^p}$ 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 (23) (with $v_{2h} = 0$), $\boldsymbol{R}_{2h}^{k,i}$ corresponds to the test functions v_{2h} in the first line of (23) (with $v_{1h} = 0$), and issues from the second line of (23) the complementarity constraints (29). Following [52], we associate respectively with $\boldsymbol{R}_{1h}^{k,i}$ and $\boldsymbol{R}_{2h}^{k,i}$ discontinuous elementwise 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 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 $\psi_{h,\boldsymbol{x}_l}, \boldsymbol{x}_l \in \mathcal{V}_{\mathrm{d}}^{p,\mathrm{int}}$ nonzero on K. It is easily seen that the first $2\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}$ lines of (36) read, *cf.*(23) and (14),

$$\begin{aligned} \mu_1 \left(\boldsymbol{\nabla} u_{1h}^{k,i}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{x}_l}\right)_\Omega &= \left(f_1 + \tilde{\lambda}_{h,l}^{k,i} - r_{1h}^{k,i}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega \quad \forall l = 1,\dots,\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}, \\ \mu_2 \left(\boldsymbol{\nabla} u_{2h}^{k,i}, \boldsymbol{\nabla}\psi_{h,\boldsymbol{x}_l}\right)_\Omega &= \left(f_2 - \tilde{\lambda}_{h,l}^{k,i} - r_{2h}^{k,i}, \psi_{h,\boldsymbol{x}_l}\right)_\Omega \quad \forall l = 1,\dots,\mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}}, \end{aligned} \tag{37}$$

where

$$\tilde{\lambda}_{h,l}^{k,i} = \begin{cases} \lambda_h^{k,i}(\boldsymbol{x}_l) \text{ (real number given by the vertex value of } \lambda_h^{k,i}) \text{ if } p = 1 \text{ and } \mathcal{N}_{\mathrm{d}}^{p,\mathrm{int}} = \mathcal{V}_h^{\mathrm{int}} \\ \lambda_h^{k,i} \text{ (function } \lambda_h^{k,i}\text{, the index } l \text{ is discarded) if } p \geq 2 \end{cases}. \tag{38}$$

In the sequel, we also use the shorthand notation

$$\tilde{\lambda}_{h,\boldsymbol{a}}^{k,i} = \begin{cases} \lambda_h^{k,i}(\boldsymbol{a}) & \text{if} \quad p = 1 \\ \lambda_h^{k,i} & \text{if} \quad p \geq 2 \end{cases}. \tag{39}$$

Using the representations $r_{\alpha h}^{k,i}$ of the algebraic vectors and (37) will be useful to formulate our a posteriori estimators in Section 5.

4 Flux reconstructions

This section introduces flux reconstructions that will be central in our a posteriori error analysis, following some general concepts in [19, 32, 34], 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 $\boldsymbol{\Pi}_{\mathbb{P}_p}$ 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 4.1. *There exist* $\boldsymbol{\sigma}_{\alpha h}^{k,i} \in \mathbf{H}(\mathrm{div}, \Omega)$, $\alpha \in \{1,2\}$, *such that*

$$\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i} = \boldsymbol{\Pi}_{\mathbb{P}_p}(f_\alpha) - (-1)^\alpha \lambda_h^{k,i} \quad \in \mathbb{P}_p(\mathcal{T}_h). \tag{40}$$

Recall that by definition, the strong form (40) implies

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

Second, following [34], in order to distinguish the algebraic and discretization error components, we make:

Assumption 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 4.3. *The construction of the fluxes is based on the first two diffusion equations in* (5) *that are linear. Thus we do not need to construct any linearization fluxes as in [34]. 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.*

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 $p \geq 1$ [57, 20, 26, 54] 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

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

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

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

Definition 4.4. *Let $\left(u_{1h}^{k,i}, u_{2h}^{k,i}, \lambda_h^{k,i}\right)$ be the approximate solution given by (36), verifying in particular (37). For each vertex $\boldsymbol{a} \in \mathcal{V}_h$, define $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}} \in \mathbf{V}_h^{\boldsymbol{a}}$ and $\gamma_{\alpha h}^{k,i,\boldsymbol{a}} \in Q_h^{\boldsymbol{a}}$, by solving:*

$$\begin{array}{lcll} \left(\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, \boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} - \left(\gamma_{\alpha h}^{k,i,\boldsymbol{a}}, \boldsymbol{\nabla}{\cdot}\boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} & = & -\left(\mu_\alpha \psi_{h,\boldsymbol{a}} \boldsymbol{\nabla} u_{\alpha h}^{k,i}, \boldsymbol{\tau}_h\right)_{\omega_h^{\boldsymbol{a}}} & \forall \boldsymbol{\tau}_h \in \mathbf{V}_h^{\boldsymbol{a}}, \\ \left(\boldsymbol{\nabla}{\cdot}\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} & = & \left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, q_h\right)_{\omega_h^{\boldsymbol{a}}} & \forall q_h \in Q_h^{\boldsymbol{a}}, \end{array} \tag{41}$$

where the right-hand sides are defined by

$$\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}} := \left(f_\alpha - (-1)^\alpha \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i} - r_{\alpha h}^{k,i}\right)\psi_{h,\boldsymbol{a}} - \mu_\alpha \boldsymbol{\nabla} u_{\alpha h}^{k,i}{\cdot}\boldsymbol{\nabla}\psi_{h,\boldsymbol{a}} \quad \forall \boldsymbol{a} \in \mathcal{V}_h. \tag{42}$$

where we recall the notation (38). Then set

$$\boldsymbol{\sigma}_{1h,\mathrm{disc}}^{k,i} := \sum_{\boldsymbol{a}\in\mathcal{V}_h} \boldsymbol{\sigma}_{1h,\mathrm{disc}}^{k,i,\boldsymbol{a}} \quad \textit{and} \quad \boldsymbol{\sigma}_{2h,\mathrm{disc}}^{k,i} := \sum_{\boldsymbol{a}\in\mathcal{V}_h} \boldsymbol{\sigma}_{2h,\mathrm{disc}}^{k,i,\boldsymbol{a}}. \tag{43}$$

When $\forall \boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$, using (37), there holds $\left(\tilde{g}_{1h}^{k,i,\boldsymbol{a}}, 1\right)_{\omega_h^{\boldsymbol{a}}} = 0$, and we conclude that the Neumann compatibility condition is satisfied for problems (41). Consequently, the second line of (41) holds true for all $q_h \in \mathbb{P}_p(\mathcal{T}_h|_{\omega_h^{\boldsymbol{a}}})$ (and not only on $Q_h^{\boldsymbol{a}}$). Since the functions $\psi_{h,\boldsymbol{a}}$ form a partition of unity and since $r_{\alpha h}^{k,i}|_K$ and $\lambda_h^{k,i}|_K$ belong to $\mathbb{P}_p(K)$ for all $K \in \mathcal{T}_h$, we immediately have from [35, Lemma 3.5] that, for $\alpha \in \{1,2\}$,

$$\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i} \in \mathbf{H}(\mathrm{div},\Omega) \ \text{ with } \ \boldsymbol{\nabla}{\cdot}\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i} = \boldsymbol{\Pi}_{\mathbb{P}_p}(f_\alpha) - (-1)^\alpha \lambda_h^{k,i} - r_{\alpha h}^{k,i}. \tag{44}$$

Note that we use in particular $\sum_{\boldsymbol{a}\in\mathcal{V}_K} \left(\lambda_h^{k,i}\psi_{h,\boldsymbol{a}}\right)|_K = \lambda_h^{k,i}$ for $p \geq 2$ by the partition of unity by $\sum_{\boldsymbol{a}\in\mathcal{V}_h} \psi_{h,\boldsymbol{a}}|_K = 1$, whereas $\sum_{\boldsymbol{a}\in\mathcal{V}_h} \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\psi_{h,\boldsymbol{a}} = \sum_{\boldsymbol{a}\in\mathcal{V}_h} \lambda_h^{k,i}(\boldsymbol{a})\psi_{h,\boldsymbol{a}} = \lambda_h^{k,i}$ by the definition of the Lagrange basis for $p = 1$. Furthermore, (44) implies that Definition 4.4 combined with Assumption 4.2 satisfies Assumption 4.1.

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 [52]. 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 $\mathcal{T}_{j-1}, 1 \leq j \leq J$,

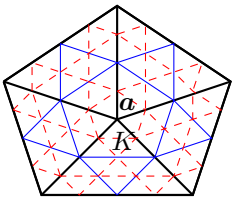

Figure 1: Example of nested meshes with $J = 2$. Coarsest mesh $\mathcal{T}_0$ with $\boldsymbol{a} \in \mathcal{V}_0$ and $\omega_0^{\boldsymbol{a}}$ constituted by 5 elements (black, thick), first refined mesh $\mathcal{T}_1$ (blue, thin), and second refined mesh $\mathcal{T}_2 = \mathcal{T}_h$ (red,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

$\mathcal{T}_h = \mathcal{T}_J$, see Figure 1. The set of vertices of $\mathcal{T}_j$ is denoted by $\mathcal{V}_j$ and it is partitioned into interior vertices $\mathcal{V}_j^{\rm int}$ and boundary vertices $\mathcal{V}_j^{\rm ext}$. 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 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

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

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

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

when $\boldsymbol{a} \in \mathcal{V}_{j-1}^{\rm ext}$. Denote also by $\boldsymbol{\Pi}_{j-1}$ the $L^2(\Omega)$-orthogonal projection onto the broken space $\mathbb{P}_p(\mathcal{T}_{j-1})$. With an abuse of notation, we set $\boldsymbol{\Pi}_0 := 0$. Following [52], the reconstructions $\boldsymbol{\sigma}_{1j,\rm alg}^{k,i}$ and $\boldsymbol{\sigma}_{2j,\rm alg}^{k,i}$ is obtained as follows:

Definition 4.5. *Let $(u_{1h}^{k,i}, u_{2h}^{k,i}, \lambda_h^{k,i})$ be the approximate solution given by (36), verifying in particular (37). Let $1 \le j \le J$. For any $\boldsymbol{a} \in \mathcal{V}_{j-1}^{\rm int}$, we prescribe $\boldsymbol{\sigma}_{\alpha j,\rm alg}^{k,i,\boldsymbol{a}} \in \mathbf{V}_{j,j-1}^{\boldsymbol{a}}$ and $\gamma_{\alpha j}^{k,i,\boldsymbol{a}} \in Q_{j,j-1}^{\boldsymbol{a}}$ by solving*

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

where the right-hand sides are defined by

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

Then set

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

In this definition, each flux $\boldsymbol{\sigma}_{\alpha j,\rm alg}^{k,i,\boldsymbol{a}}$ 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, following [52], $\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h,\rm alg}^{k,i} = r_{\alpha h}^{k,i}$. Consequently, it is immediate to check that the flux reconstructions Definitions 4.4 and 4.5 satisfy Assumption 4.2.

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 (36), (37). The main difficulty is located in the treatment of the constraints: the conditions $u_{1h}^{k,i} - u_{2h}^{k,i} \geq 0$, and $\lambda_h \geq 0$ as well as $\langle \lambda_h, \psi_{h,\boldsymbol{x}_l} \rangle_h = 0 \quad \forall \boldsymbol{x}_l \in \mathcal{V}_{\rm d}^{p,\rm ext}$ for $p \geq 2$ do not necessarily hold before convergence of both solvers. For $p \geq 2$, moreover, the constraints $u_{1h} - u_{2h} \geq 0$ and $\lambda_h \geq 0$ are not even satisfied at convergence. In this respect, it will be useful to introduce positive and negative parts of $\lambda_h^{k,i}$,

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

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

$$\widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p := \left\{ (v_{1h}, v_{2h}) \in X_{gh}^p \times X_{0h}^p, \ v_{1h} - v_{2h} \geq 0 \right\} \subset \boldsymbol{\mathcal{K}}_g. \tag{45}$$

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 $\eta_{\cdot}^{k,i} := \left\{ \sum_{K \in \mathcal{T}_h} \left(\eta_{\cdot,K}^{k,i} \right)^2 \right\}^{\frac{1}{2}}$. The first main result of this article is:

Theorem 5.1. *Let $\boldsymbol{u} = (u_1, u_2) \in \boldsymbol{\mathcal{K}}_g$ be the solution of the continuous reduced problem (11). 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 (36) for any $p \geq 1$, verifying in particular (37). Let $\boldsymbol{\sigma}_{1h}^{k,i}$ and $\boldsymbol{\sigma}_{2h}^{k,i}$ be equilibrated flux reconstructions satisfying Assumptions 4.1 and 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*

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

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

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

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

Then, the following a posteriori error estimate holds:

$$\left| \left| \left| \boldsymbol{u} - \boldsymbol{u}_h^{k,i} \right| \right| \right| \leq \eta^{k,i} := \left\{ \left(\eta_1^{k,i} + \eta_{{\rm nonc},1}^{k,i} + \eta_{{\rm nonc},2}^{k,i} \right)^2 + \eta_{{\rm nonc},3}^{k,i} + \sum_{K \in \mathcal{T}_h} \eta_{{\rm C},K}^{k,i,\rm pos} \right\}^{\frac{1}{2}}. \tag{46}$$

Remark 5.2. *The estimate (46) 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 5.1 reflect various violations of physical properties of the approximate solution $\boldsymbol{u}_h^{k,i}$: $\eta_{{\rm F},K,\alpha}^{k,i}$ and $\eta_{{\rm osc},K,\alpha}$ represent the nonconformity of the flux,* i.e., *the fact that $-\mu_\alpha \boldsymbol{\nabla} u_{\alpha h}^{k,i} \notin \mathbf{H}(\mathrm{div}, \Omega)$; $\eta_{{\rm C},K}^{k,i,\rm pos}$ 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 $\eta_{{\rm nonc},1,K}^{k,i}$, $\eta_{{\rm nonc},2,K}^{k,i}$, and $\eta_{{\rm nonc},3,K}^{k,i}$ 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 (12) and (21)). 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 61 below. In Corollary 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 5.3. *In [12], 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 (13) for* $p = 1$ *not taking into account inexact nonlinear and linear solvers was derived. The estimate of [12, Lemma 3.3] has the form*

$$|||\boldsymbol{u} - \boldsymbol{u}_h||| \leq \left\{ \sum_{K\in\mathcal{T}_h} \left\{ \sum_{\alpha=1}^{2} (\eta_{\mathrm{F},K,\alpha}^{\infty,\infty} + \eta_{\mathrm{osc},K,\alpha})^2 + \eta_{\mathrm{C},K}^{\infty,\infty,\mathrm{pos}} \right\} \right\}^{\frac{1}{2}}, \tag{47}$$

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 (46) for* $p = 1$ *takes the same form as (47) at convergence, which shows the consistency of our approach.*

Proof of Theorem 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

$$a(\boldsymbol{s}, \boldsymbol{v} - \boldsymbol{s}) \geq a(\boldsymbol{u}_h^{k,i}, \boldsymbol{v} - \boldsymbol{s}) \quad \forall \boldsymbol{v} \in \boldsymbol{\mathcal{K}}_g, \tag{48}$$

where we recall that the bilinear symmetric form a was defined in (10). Problem (48) is well-posed thanks to the Lions–Stampacchia theorem [48], because a defines a scalar product on $[H_0^1(\Omega)]^2$. Developing the square, the projection $\boldsymbol{s}$ satisfies for each $\boldsymbol{v} \in \boldsymbol{\mathcal{K}}_g$

$$\left|\left|\left| \boldsymbol{v} - \boldsymbol{u}_h^{k,i} \right|\right|\right|^2 = |||\boldsymbol{v} - \boldsymbol{s}|||^2 + 2a(\boldsymbol{v} - \boldsymbol{s}, \boldsymbol{s} - \boldsymbol{u}_h^{k,i}) + \left|\left|\left| \boldsymbol{s} - \boldsymbol{u}_h^{k,i} \right|\right|\right|^2. \tag{49}$$

Since $a(\boldsymbol{v}-\boldsymbol{s}, \boldsymbol{s}-\boldsymbol{u}_h^{k,i}) \geq 0$ from (48), taking successively in (49) $\boldsymbol{v} = \boldsymbol{u}$ 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

$$|||\boldsymbol{u} - \boldsymbol{s}||| \leq \left|\left|\left| \boldsymbol{u} - \boldsymbol{u}_h^{k,i} \right|\right|\right|, \tag{50}$$

$$\left|\left|\left| \boldsymbol{s} - \boldsymbol{u}_h^{k,i} \right|\right|\right| \leq \left|\left|\left| \boldsymbol{s}_h^{k,i} - \boldsymbol{u}_h^{k,i} \right|\right|\right| = \eta_{\mathrm{nonc},1}^{k,i}. \tag{51}$$

Second, the energy norm of the error is decomposed as

$$\left|\left|\left| \boldsymbol{u} - \boldsymbol{u}_h^{k,i} \right|\right|\right|^2 = a(\boldsymbol{u} - \boldsymbol{u}_h^{k,i}, \boldsymbol{u} - \boldsymbol{u}_h^{k,i}) = a(\boldsymbol{u} - \boldsymbol{u}_h^{k,i}, \boldsymbol{u} - \boldsymbol{s}) + a(\boldsymbol{u} - \boldsymbol{u}_h^{k,i}, \boldsymbol{s} - \boldsymbol{u}_h^{k,i}). \tag{52}$$

We estimate both terms in (52) separately. The second one is bounded by the Cauchy–Schwarz inequality and (51),

$$a(\boldsymbol{u} - \boldsymbol{u}_h^{k,i}, \boldsymbol{s} - \boldsymbol{u}_h^{k,i}) \leq \left|\left|\left| \boldsymbol{u} - \boldsymbol{u}_h^{k,i} \right|\right|\right| \left|\left|\left| \boldsymbol{s} - \boldsymbol{u}_h^{k,i} \right|\right|\right| \leq \left|\left|\left| \boldsymbol{u} - \boldsymbol{u}_h^{k,i} \right|\right|\right| \eta_{\mathrm{nonc},1}^{k,i}. \tag{53}$$

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

The reduced problem (11) for $\boldsymbol{v} = \boldsymbol{s} \in \boldsymbol{\mathcal{K}}_g$ yields

$$a(\boldsymbol{u}, \boldsymbol{u} - \boldsymbol{s}) \leq l(\boldsymbol{u} - \boldsymbol{s}). \tag{54}$$

Setting $\boldsymbol{w} = \boldsymbol{u} - \boldsymbol{s}$, we estimate the first term in (52) using (54) and adding and substracting $b(\boldsymbol{w}, \lambda_h^{k,i})$ and employing the definitions of b and l of (10)

$$\begin{aligned} a(\boldsymbol{u} - \boldsymbol{u}_h^{k,i}, \boldsymbol{w}) &\leq l(\boldsymbol{w}) + b(\boldsymbol{w}, \lambda_h^{k,i}) - a(\boldsymbol{u}_h^{k,i}, \boldsymbol{w}) - b(\boldsymbol{w}, \lambda_h^{k,i}), \\ &= \sum_{\alpha=1}^{2} \left(f_\alpha - (-1)^\alpha \lambda_h^{k,i}, w_\alpha \right)_\Omega - \sum_{\alpha=1}^{2} \left(\mu_\alpha \boldsymbol{\nabla} u_{\alpha h}^{k,i}, \boldsymbol{\nabla} w_\alpha \right)_\Omega - b(\boldsymbol{w}, \lambda_h^{k,i}). \end{aligned} \tag{55}$$

Besides, as $\boldsymbol{\sigma}_{\alpha h}^{k,i} \in \mathbf{H}(\mathrm{div},\Omega)$ by Assumption 4.1 and since $w_\alpha \in H_0^1(\Omega)$, the Green formula gives

$$\left(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i}, w_\alpha\right)_\Omega = -\left(\boldsymbol{\sigma}_{\alpha h}^{k,i}, \boldsymbol{\nabla} w_\alpha\right)_\Omega \quad \forall \alpha \in \{1,2\}\,. \tag{56}$$

Then, using (56) in (55), one has

$$\begin{aligned} a(\boldsymbol{u}-\boldsymbol{u}_h^{k,i}, \boldsymbol{w}) \le \sum_{\alpha=1}^{2}\sum_{K\in\mathcal{T}_h} \Big\{ &\left(f_\alpha - (-1)^\alpha \lambda_h^{k,i} - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i}, w_\alpha\right)_K \\ &- \left(\mu_\alpha^{\frac12}\boldsymbol{\nabla} u_{\alpha h}^{k,i} + \mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h}^{k,i}, \mu_\alpha^{\frac12}\boldsymbol{\nabla} w_\alpha\right)_K \Big\} - b(\boldsymbol{w}, \lambda_h^{k,i}). \end{aligned} \tag{57}$$

It remains to bound each of the three terms in (57). Using the divergence property (40) of Assumption 4.1, the Cauchy–Schwarz and Poincaré–Wirtinger (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 $\alpha = 1, 2$

$$\left(f_\alpha - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{\alpha h}^{k,i} - (-1)^\alpha\lambda_h^{k,i}, w_\alpha\right)_K = \left(f_\alpha - \boldsymbol{\Pi}_{\mathbb{P}_p}(f_\alpha), w_\alpha - \overline{w}_{\alpha,K}\right)_K, \le \eta_{\mathrm{osc},K,\alpha}\left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla} w_\alpha\right\|_K. \tag{58}$$

Furthermore, by the Cauchy–Schwarz inequality

$$-\left(\mu_\alpha^{\frac12}\boldsymbol{\nabla} u_{\alpha h}^{k,i} + \mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h}^{k,i}, \mu_\alpha^{\frac12}\boldsymbol{\nabla} w_\alpha\right)_K \le \eta_{\mathrm{F},K,\alpha}^{k,i}\left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla} w_\alpha\right\|_K. \tag{59}$$

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 (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 (6a) inequalities noting that both w_α and $(s_\alpha - u_{\alpha h}^{k,i})$ belong to $H_0^1(\Omega)$, and also employing (51) give

$$-b(\boldsymbol{w}, \lambda_h^{k,i}) \le \eta_{\mathrm{nonc},2}^{k,i}\,|||\boldsymbol{w}||| + \frac12\eta_{\mathrm{nonc},3}^{k,i} + \frac12\sum_{K\in\mathcal{T}_h}\eta_{\mathrm{C},K}^{k,i,\mathrm{pos}}. \tag{60}$$

Therefore, combining (52), (53), (57), (58), (59), (60), and (50), 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 (46). □

The construction of $\boldsymbol{s}_h^{k,i} \in \widetilde{\boldsymbol{\mathcal{K}}}_{gh}^p \subset \boldsymbol{\mathcal{K}}_g$ when the polynomial degree $p \ge 2$ 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 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 $\boldsymbol{a} \in \mathcal{V}_h^{\mathrm{int}}$

$$\boldsymbol{s}_h^{k,i}(\boldsymbol{a}) := \begin{cases} \boldsymbol{u}_h^{k,i}(\boldsymbol{a}) = \left(u_{1h}^{k,i}(\boldsymbol{a}), u_{2h}^{k,i}(\boldsymbol{a})\right) & \text{if } u_{1h}^{k,i}(\boldsymbol{a}) \ge u_{2h}^{k,i}(\boldsymbol{a}), \\ \left(\frac12\left(u_{1h}^{k,i}(\boldsymbol{a}) + u_{2h}^{k,i}(\boldsymbol{a})\right), \frac12\left(u_{1h}^{k,i}(\boldsymbol{a}) + u_{2h}^{k,i}(\boldsymbol{a})\right)\right) & \text{if } u_{1h}^{k,i}(\boldsymbol{a}) < u_{2h}^{k,i}(\boldsymbol{a}). \end{cases} \tag{61}$$

Concerning $\lambda_h^{k,i}$, an estimate is provided in (recall the definition given in (8)):

Theorem 5.4. *Assume the hypotheses and notations of Theorem 5.1 and let $\lambda \in \Lambda$ be the solution of problem (9). Then the following a posteriori estimate holds:*

$$\left|\!\left|\!\left|\lambda - \lambda_h^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\Omega)} \le \eta^{k,i} + \eta_1^{k,i}. \tag{62}$$

Proof. The proof follows the one in [12, Corollary 3.5]. We only give the essential elements. Let $\mu_{\rm m} := \max(\mu_1, \mu_2)$. Employing (8) and extending appropriately b,

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

Fix $\boldsymbol{\phi} \in [H_0^1(\Omega)]^2$ such that $\mu_{\rm m}\sum_{\alpha=1}^2 \|\boldsymbol{\nabla}\phi_\alpha\|^2 = 1$. Invoking (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 $-a(\boldsymbol{u} - \boldsymbol{u}_h^{k,i}, \boldsymbol{\phi}) \le \left|\!\left|\!\left|\boldsymbol{u} - \boldsymbol{u}_h^{k,i}\right|\!\right|\!\right|$, since $\|\boldsymbol{\phi}\| \le 1$. The first three terms are identical to the first three terms of (55) but with $\boldsymbol{\phi} \in [H_0^1(\Omega)]^2$ instead of $\boldsymbol{w}$. Thus, using the estimates (58) and (59), one gets

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

which combined with (46) gives the result. □

Remark 5.5. *At convergence, for $\mathbb{P}_1$ finite elements, estimate (62) reduces to (3.30) in [12] 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

$$\eta_{{\rm lin},1,K}^{k,i} := \eta_{{\rm nonc},1,K}^{k,i}, \quad \eta_{{\rm lin},2,K}^{k,i} := \eta_{{\rm nonc},2,K}^{k,i}, \quad \eta_{{\rm lin},3,K}^{k,i} := \eta_{{\rm nonc},3,K}^{k,i}. \tag{63}$$

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 $\eta_{\rm lin}^{k,i} \to 0$ and $\eta_{\rm alg}^{k,i} \to 0$ when $k \to 0$ and $i \to 0$. A similar distinction for the several case $p \ge 2$ is possible but a little longer to write down.

Corollary 5.6. *Consider the assumptions and notations of Theorem 5.1 in the case $p = 1$. Define for $\alpha \in \{1, 2\}$ and for $K \in \mathcal{T}_h$*

$$\eta_{{\rm alg},K,\alpha}^{k,i} := \left\|\mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h,{\rm alg}}^{k,i}\right\|_K, \quad \eta_{{\rm disc},K,\alpha}^{k,i} := \left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla}u_{\alpha h}^{k,i} + \mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h,{\rm disc}}^{k,i}\right\|_K, \tag{64a}$$

$$\eta_{\rm disc}^{k,i} := \left\{\sum_{\alpha=1}^2\sum_{K\in\mathcal{T}_h}\left(\eta_{{\rm disc},K,\alpha}^{k,i} + \eta_{{\rm osc},K,\alpha}\right)^2\right\}^{\frac12} + \left\{\sum_{K\in\mathcal{T}_h}\left|\eta_{{\rm C},K}^{k,i,{\rm pos}}\right|\right\}^{\frac12}, \tag{64b}$$

$$\eta_{\rm lin}^{k,i} := \eta_{{\rm lin},1}^{k,i} + \eta_{{\rm lin},2}^{k,i} + \left(\eta_{{\rm lin},3}^{k,i}\right)^{\frac12}, \quad \eta_{\rm alg}^{k,i} := \left\{\sum_{\alpha=1}^2\sum_{K\in\mathcal{T}_h}(\eta_{{\rm alg},K,\alpha}^{k,i})^2\right\}^{\frac12}. \tag{64c}$$

Then,

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

Proof. As for $(A, B) \in \mathbb{R}_+ \times \mathbb{R}_+$, $(A + B)^{\frac12} \le A^{\frac12} + B^{\frac12}$, we have

$$\eta^{k,i} \le \eta_1^{k,i} + \eta_{{\rm lin},1}^{k,i} + \eta_{{\rm lin},2}^{k,i} + \left(\eta_{{\rm lin},3}^{k,i}\right)^{\frac12} + \left(\sum_K |\eta_{{\rm C},K}^{k,i,{\rm pos}}|\right)^{\frac12}. \tag{65}$$

Next, the definition of $\eta_1^{k,i}$ combined with the triangle inequality to separate the algebraic estimators $\eta_{\mathrm{alg},K,\alpha}^{k,i}$ from the discretization estimators $\eta_{\mathrm{disc},K,\alpha}^{k,i}$ give

$$\eta_1^{k,i} \leq \left(\sum_{K\in\mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{disc},K,\alpha}^{k,i} + \eta_{\mathrm{osc},K,\alpha} \right)^2 \right)^{\frac{1}{2}} + \left(\sum_{K\in\mathcal{T}_h} \sum_{\alpha=1}^{2} \left(\eta_{\mathrm{alg},K,\alpha}^{k,i} \right)^2 \right)^{\frac{1}{2}}, \tag{66}$$

which concludes the proof. □

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 [4, 34, 43, 51], 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 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.

6.1 Stopping criteria

Recall that we employ a semismooth Newton method for the nonlinear problem (33), yielding on each step $k \geq 1$ and each step $i \geq 1$ the linear system (36). 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:

$$\text{(a) } \eta_{\mathrm{alg}}^{k,i} \leq \gamma_{\mathrm{alg}} \max\left\{ \eta_{\mathrm{disc}}^{k,i}, \eta_{\mathrm{lin}}^{k,i} \right\}, \qquad \text{(b) } \eta_{\mathrm{lin}}^{k,i} \leq \gamma_{\mathrm{lin}} \eta_{\mathrm{disc}}^{k,i}. \tag{67}$$

Remark 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 [34, 43] and the references therein, one can aim at the balance of all error components in each mesh cell in place of* (67)*, while simultaneously guaranteeing the global criteria* (67)*. These local criteria read, with*

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

$$\eta_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha}^{k,i} \leq \min_{K\subset\omega_h^{\boldsymbol{a}}} \left\{ \gamma_{\mathrm{alg},K} \max\left\{ \eta_{\mathrm{disc},K,\alpha}^{k,i}, \eta_{\mathrm{lin},K}^{k,i} \right\} \right\} \qquad \forall \alpha \in \{1,2\}, \tag{68a}$$

$$\eta_{\mathrm{lin},K}^{k,i} \leq \min_{\alpha\in\{1,2\}} \left\{ \gamma_{\mathrm{lin},K} \eta_{\mathrm{disc},K,\alpha}^{k,i} \right\}, \tag{68b}$$

where

$$\eta_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha}^{k,i} := \left\{ \sum_{K\subset\omega_h^{\boldsymbol{a}}} \left(\eta_{\mathrm{alg},K,\alpha}^{k,i} \right)^2 \right\}^{\frac{1}{2}} = \left\| \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i} \right\|_{\omega_h^{\boldsymbol{a}}}. \tag{69}$$

The (complicated) form of the term $\eta_{\mathrm{lin},K}^{k,i}$ ensures that local criteria (68) *imply the global criteria* (67)*, 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 5.1. In particular, local efficiency will be proven below based on* (68)*.*

Remark 6.2. *When $p \geq 2$, we will prove below the local efficiency of the leading estimators from Theorem 5.1 directly (recall that we have only introduced Corollary 5.6 for $p = 1$). Then, the analogue of the local stopping criterion (68a) will be*

$$\eta^{k,i}_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha} \leq \min_{K\subset\omega_h^{\boldsymbol{a}}} \left\{ \gamma_{\mathrm{alg},K} \eta^{k,i}_{\mathrm{disc},K,\alpha} \right\} \qquad \forall \alpha \in \{1,2\}, \tag{70}$$

where $\eta^{k,i}_{\mathrm{alg},K,\alpha}$ and $\eta^{k,i}_{\mathrm{disc},K,\alpha}$ are given by (64a) and $\eta^{k,i}_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha}$ is given by (69).

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 $\boldsymbol{X}_h^0 \in \mathbb{R}^{3N_h^{\mathrm{int}}}$ and set $k = 1$.

1. From $\boldsymbol{X}_h^{k-1}$ define $\mathbb{A}^{k-1} \in \mathbb{R}^{3N_h^{\mathrm{int}},3N_h^{\mathrm{int}}}$ and $\boldsymbol{B}^{k-1} \in \mathbb{R}^{3N_h^{\mathrm{int}}}$ by (35).

2. Consider the linear system

$$\mathbb{A}^{k-1}\boldsymbol{X}_h^k = \boldsymbol{B}^{k-1}. \tag{71}$$

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 (71), 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 5.6 and check the stopping criterion for the linear solver in the form (67)(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 (67)(b). If satisfied, return $\boldsymbol{X}_h = \boldsymbol{X}_h^k$. If not, set $k = k + 1$ and go back to **1**.

7 Efficiency

We prove in this section local efficiency of the a posteriori error estimators of Corollary 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 [18, 34, 35, 52] 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 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 (68). In the generic case $p \geq 2$, we do not address the inexact linearization solver and show that the leading term in Theorem 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 $\eta_{\mathrm{osc},K,\alpha} = 0$, $\forall \alpha \in \{1,2\}$. We do not treat here the "complementarity" estimators $\eta^{k,i,\mathrm{pos}}_{\mathrm{C},K}$ that are typically numerically very small. Their local efficiency could be proven, when $p = 1$, along the lines of [12, Proposition 3.9].

7.1 Continuous-level problems with hat functions on patches

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

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

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

$$\|\boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\|_{\omega_h^{\boldsymbol{a}}} \leq C_{\mathrm{cont,PF}} \|\boldsymbol{\nabla} v\|_{\omega_h^{\boldsymbol{a}}} \quad \forall v \in H^1_*(\omega_h^{\boldsymbol{a}}), \tag{72}$$

see Carstensen and Funken [25], Braess *et al.* [18], or Ern and Vohralík [35]. Then, for any $p \geq 1$ we have

Lemma 7.1. *Let (u_1, u_2, λ) be the solution of (9) and let $(u_{1h}^{k,i}, u_{2h}^{k,i}, \lambda_h^{k,i})$ be the approximation given by (36), verifying in particular (37). 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*

$$(\mu_\alpha \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}, \boldsymbol{\nabla} v)_{\omega_h^{\boldsymbol{a}}} = \left(-\mu_\alpha \psi_{h,\boldsymbol{a}} \boldsymbol{\nabla} u_{\alpha h}^{k,i}, \boldsymbol{\nabla} v\right)_{\omega_h^{\boldsymbol{a}}} + \left(\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}, v\right)_{\omega_h^{\boldsymbol{a}}} \quad \forall v \in H^1_*(\omega_h^{\boldsymbol{a}}), \tag{73}$$

where $\tilde{g}_{\alpha h}^{k,i,\boldsymbol{a}}$ is defined in (42). Let $\mu_{\mathrm{m}} := \max(\mu_1, \mu_2)$. Then, for $\alpha \in \{1,2\}$,

$$\left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}\right\|_{\omega_h^{\boldsymbol{a}}} \leq C_{\mathrm{cont,PF}} \left(\left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}} + \mu_{\mathrm{m}}^{\frac{1}{2}} \mu_\alpha^{-\frac{1}{2}} \left|\!\left|\!\left|\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})} + \left\|\mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}} \right). \tag{74}$$

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

$$\left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}\right\|_{\omega_h^{\boldsymbol{a}}} = \sup_{v \in H^1_*(\omega_h^{\boldsymbol{a}}), \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} v\right\|_{\omega_h^{\boldsymbol{a}}} = 1} \left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}, \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} v\right)_{\omega_h^{\boldsymbol{a}}}. \tag{75}$$

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 (73), using the definition (42) 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 (9), we obtain

$$\left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}, \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla} v\right)_{\omega_h^{\boldsymbol{a}}} = \left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right), \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\right)_{\omega_h^{\boldsymbol{a}}} + \left((-1)^\alpha \left(\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right) - r_{\alpha h}^{k,i}, \psi_{h,\boldsymbol{a}} v\right)_{\omega_h^{\boldsymbol{a}}}. \tag{76}$$

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 4.2, the Green formula and the Cauchy–Schwarz inequality give

$$\left|\left(r_{\alpha h}^{k,i}, \psi_{h,\boldsymbol{a}} v\right)_{\omega_h^{\boldsymbol{a}}}\right| = \left|-\left(\mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}, \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\right)_{\omega_h^{\boldsymbol{a}}}\right| \leq \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\right\|_{\omega_h^{\boldsymbol{a}}} \left\|\mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}}. \tag{77}$$

Multiplying and dividing $\left(\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}, \psi_{h,\boldsymbol{a}} v\right)_{\omega_h^{\boldsymbol{a}}}$ by $\left\|\mu_{\mathrm{m}}^{\frac{1}{2}} \boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\right\|_{\omega_h^{\boldsymbol{a}}}$ and using that $\psi_{h,\boldsymbol{a}} v \in H_0^1(\omega_h^{\boldsymbol{a}})$ which allows us to employ the definition (8), we get

$$\left|\left(\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}, \psi_{h,\boldsymbol{a}} v\right)_{\omega_h^{\boldsymbol{a}}}\right| \leq \left|\!\left|\!\left|\lambda - \tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})} \mu_{\mathrm{m}}^{\frac{1}{2}} \mu_\alpha^{-\frac{1}{2}} \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\right\|_{\omega_h^{\boldsymbol{a}}}. \tag{78}$$

Finally, the Cauchy–Schwarz inequality leads to

$$\left(\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right), \mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\right)_{\omega_h^{\boldsymbol{a}}} \leq \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\left(u_\alpha - u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}} \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}(\psi_{h,\boldsymbol{a}} v)\right\|_{\omega_h^{\boldsymbol{a}}}. \tag{79}$$

The result now follows by combining (77), (78), and (79) with (72) together with (75). □

7.2 Local efficiency of the estimators

Recall the definition of $\zeta_{\alpha,\boldsymbol{a}}$ from (73) in Lemma 7.1. Following [18, 35], there exists a constant $C_{\mathrm{st}} > 0$ only depending on the shape regularity of the mesh $\mathcal{T}_h$ such that the discretization flux reconstructions $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}$ of Definition 4.4 satisfy

$$\left\|\mu_\alpha^{\frac{1}{2}} \psi_{h,\boldsymbol{a}} \boldsymbol{\nabla} u_{\alpha h}^{k,i} + \mu_\alpha^{-\frac{1}{2}} \boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}\right\|_{\omega_h^{\boldsymbol{a}}} \leq C_{\mathrm{st}} \left\|\mu_\alpha^{\frac{1}{2}} \boldsymbol{\nabla}\zeta_{\alpha,\boldsymbol{a}}\right\|_{\omega_h^{\boldsymbol{a}}}. \tag{80}$$

Our second main result is:

Theorem 7.2. *Let the flux reconstructions* $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}$ *and* $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i}$ *be given respectively by Definitions 4.4 and 4.5 for* $p \geq 1$. *Let the local stopping criteria (68) be satisfied for the estimators of Corollary 5.6 for* $p = 1$. *Let also (70) holds for* $p \geq 2$ *for the estimators of Theorem 5.1. Let finally the algebraic parameters* $\gamma_{\mathrm{alg},K}$ *be such that*

$$\begin{aligned} \gamma_{\mathrm{alg},K} &\leq \frac{1}{6C_{\mathrm{st}}C_{\mathrm{cont,PF}}\max\{1,\gamma_{\mathrm{lin},K}\}} \quad \textit{if} \quad p=1, \\ \gamma_{\mathrm{alg},K} &\leq \frac{1}{6C_{\mathrm{st}}C_{\mathrm{cont,PF}}} \quad \textit{if} \quad p \geq 2. \end{aligned} \tag{81}$$

Setting

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

and

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

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

$$\eta_{\mathrm{disc},K,\alpha}^{k,i}+\eta_{\mathrm{lin},K}^{k,i}+\eta_{\mathrm{alg},K,\alpha}^{k,i} \leq \delta_K \sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla}\left(u_\alpha-u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}}+\mu_{\mathrm{m}}^{\frac12}\mu_\alpha^{-\frac12}\left|\!\left|\!\left|\lambda-\tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})}\right) \quad \textit{if} \quad p=1,$$

$$\begin{aligned} \eta_{\mathrm{F},K,\alpha}^{k,i} &= \left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla}u_{\alpha h}^{k,i}+\mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h}^{k,i}\right\|_K \\ &\leq \eta_{\mathrm{disc},K,\alpha}^{k,i}+\eta_{\mathrm{alg},K,\alpha}^{k,i} \\ &\leq \delta_K \sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla}\left(u_\alpha-u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}}+\mu_{\mathrm{m}}^{\frac12}\mu_\alpha^{-\frac12}\left|\!\left|\!\left|\lambda-\tilde{\lambda}_h^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})}\right) \quad \textit{if} \quad p \geq 2. \end{aligned}$$

Proof. We first treat the case $p = 1$. Let $\alpha \in \{1,2\}$. First, the local criteria (68a) and (68b) and the definition of δ_K yield

$$\eta_{\mathrm{disc},K,\alpha}^{k,i}+\eta_{\mathrm{lin},K}^{k,i}+\eta_{\mathrm{alg},K,\alpha}^{k,i} \leq \frac{\delta_K}{2C_{\mathrm{st}}C_{\mathrm{cont,PF}}}\eta_{\mathrm{disc},K,\alpha}^{k,i}. \tag{82}$$

Next, By the partition of unity $\sum_{\boldsymbol{a}\in\mathcal{V}_K}\psi_{h,\boldsymbol{a}}|_K = 1|_K$, definition (64a), (43) which implies $\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i}|_K = \sum_{\boldsymbol{a}\in\mathcal{V}_K}\boldsymbol{\sigma}_{\alpha h,\mathrm{disc}}^{k,i,\boldsymbol{a}}|_K$, stability (80), and energy lower bound (74), we have

$$\eta_{\mathrm{disc},K,\alpha}^{k,i} \leq C_{\mathrm{st}}C_{\mathrm{cont,PF}}\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left(\left\|\mu_\alpha^{\frac12}\boldsymbol{\nabla}\left(u_\alpha-u_{\alpha h}^{k,i}\right)\right\|_{\omega_h^{\boldsymbol{a}}}+\mu_{\mathrm{m}}^{\frac12}\mu_\alpha^{-\frac12}\left|\!\left|\!\left|\lambda-\tilde{\lambda}_{h,\boldsymbol{a}}^{k,i}\right|\!\right|\!\right|_{H_*^{-1}(\omega_h^{\boldsymbol{a}})}+\left\|\mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}}\right). \tag{83}$$

Using successively the local criteria (68), and since any triangle has three vertices

$$\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left\|\mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}} = \sum_{\boldsymbol{a}\in\mathcal{V}_K}\eta_{\mathrm{alg},\omega_h^{\boldsymbol{a}},\alpha}^{k,i} \leq 3\gamma_{\mathrm{alg},K}\max\left\{\eta_{\mathrm{disc},K,\alpha}^{k,i},\eta_{\mathrm{lin},K}^{k,i}\right\} \leq 3\gamma_{\mathrm{alg},K}\max\{1,\gamma_{\mathrm{lin},K}\}\,\eta_{\mathrm{disc},K,\alpha}^{k,i}. \tag{84}$$

Employing now crucially assumption (81), it follows that

$$C_{\mathrm{st}}C_{\mathrm{cont,PF}}\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left\|\mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}} \leq \frac{\eta_{\mathrm{disc},K,\alpha}^{k,i}}{2}. \tag{85}$$

Finally, combine (85) with (83) to bound $\eta_{\mathrm{disc},K,\alpha}^{k,i}$ without the term containing $\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}$, and conclude using (82).

If $p \geq 2$ the analogue of equation (82) reads

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

While, inequalities (83) and (85) remain the same, inequality (84) reads

$$\sum_{\boldsymbol{a}\in\mathcal{V}_K}\left\|\mu_\alpha^{-\frac12}\boldsymbol{\sigma}_{\alpha h,\mathrm{alg}}^{k,i}\right\|_{\omega_h^{\boldsymbol{a}}} \leq 3\gamma_{\mathrm{alg},K}\eta_{\mathrm{disc},K,\alpha}^{k,i}. \tag{86}$$

The conclusion follows immediately. □

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 [12] 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 (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 (32) combined with the GMRES linear solver for the system (34). For the computation of $\boldsymbol{\sigma}^{k,i}_{\alpha h,\mathrm{alg}}$, $\alpha = 1, 2$, following Section 4.2, we consider three levels of uniform mesh refinement ($J = 3$). We also define the linearization and algebraic residuals by

$$\boldsymbol{R}^{k,i}_{\mathrm{lin}} := \begin{pmatrix} \boldsymbol{F} - \mathbb{E}\boldsymbol{X}^{k,i}_h \\ -\mathbf{C}(\boldsymbol{X}^{k,i}_h) \end{pmatrix} \quad \text{and} \quad \boldsymbol{R}^{k,i}_{\mathrm{alg}} := \boldsymbol{B}^{k-1} - \mathbb{A}^{k-1}\boldsymbol{X}^{k,i}_h. \tag{87}$$

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 ***4b*** and **5** of Algorithm 1 by criteria on the relative residuals,

$$\text{(a)} \left\|\boldsymbol{R}^{k,i}_{\mathrm{alg}}\right\| / \left\|\boldsymbol{B}^{k-1}\right\| \le \varepsilon_{\mathrm{alg}}, \qquad \text{(b)} \left\|\boldsymbol{R}^{k,i}_{\mathrm{lin}}\right\| / \left\|\begin{pmatrix} \boldsymbol{F} \\ 0 \end{pmatrix}\right\| \le \varepsilon_{\mathrm{lin}}. \tag{88}$$

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

$$\text{(a)} \left\|\boldsymbol{R}^{k,i}_{\mathrm{alg}}\right\| / \left\|\boldsymbol{B}^{k-1}\right\| \le \alpha_{\mathrm{alg}} \left\|\boldsymbol{R}^{k,i}_{\mathrm{lin}}\right\| / \left\|\begin{pmatrix} \boldsymbol{F} \\ 0 \end{pmatrix}\right\|, \quad \text{(b)} \left\|\boldsymbol{R}^{k,i}_{\mathrm{lin}}\right\| / \left\|\begin{pmatrix} \boldsymbol{F} \\ 0 \end{pmatrix}\right\| \le \varepsilon_{\mathrm{lin}}. \tag{89}$$

3) The *adaptive inexact* Newton-min method (see Algorithm 1) that relies on the stopping criteria (67)(a) and (67)(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 $\boldsymbol{X}^0_h \in \mathbb{R}^{3N^{\mathrm{int}}_h}$ has its first N^{int}_h components equal to g and its next components equal to zero.

Figure 2 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 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 then shows the possible violation of the physical constraints during the iterations, before convergence is reached, see Remark 5.2. The figure on the left shows that $u^{k,i}_{1h} < u^{k,i}_{2h}$ can appear and the one on the right shows that $\lambda^{k,i}_h$ can be negative.

Figure 4 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 $\eta^{\overline{k},\overline{i}}$ (46) are almost identical for the three methods (exact, inexact, and adaptive inexact). Moreover, they have values close to $\eta^{\overline{k},\overline{i}}_{\mathrm{disc}}$, which is consistent with the fact that the error components from Newton-min and GMRES are relatively small. Next, $\eta^{\overline{k},\overline{i}}_{\mathrm{alg}}$ takes values below 10^{-11} for the exact semismooth Newton and below 10^{-8} for the inexact semismooth Newton, whereas $\eta^{\overline{k},\overline{i}}_{\mathrm{lin}}$ takes similar values in both cases (below 10^{-6}). The

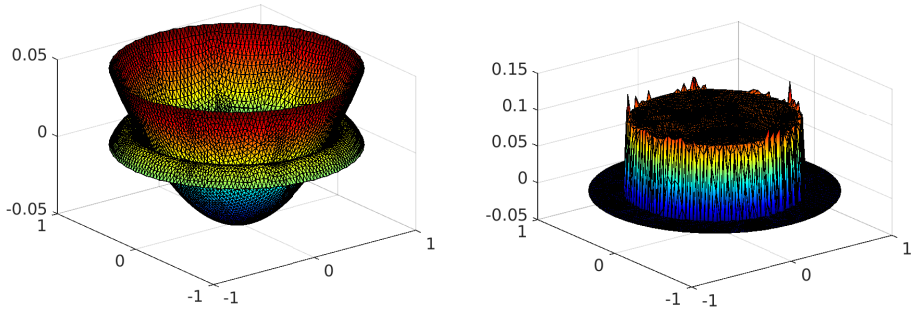

Figure 2: Solution at convergence for approximately 8000 elements. Left: position of the membranes (u_{1h}, u_{2h}). Right: discrete action (λ_h).

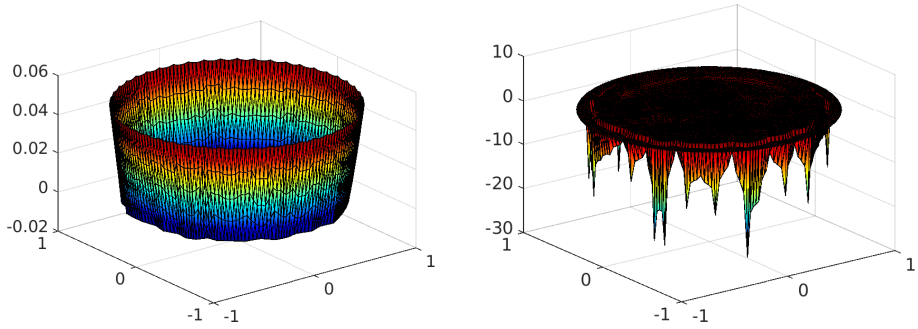

Figure 3: Left: $u_{1h}^{k,i} - u_{2h}^{k,i}$ at the second Newton-min step ($k = 2$, $i = 20$). Right: discrete action $\lambda_h^{k,i}$ for $k = 3$, $i = 20$.

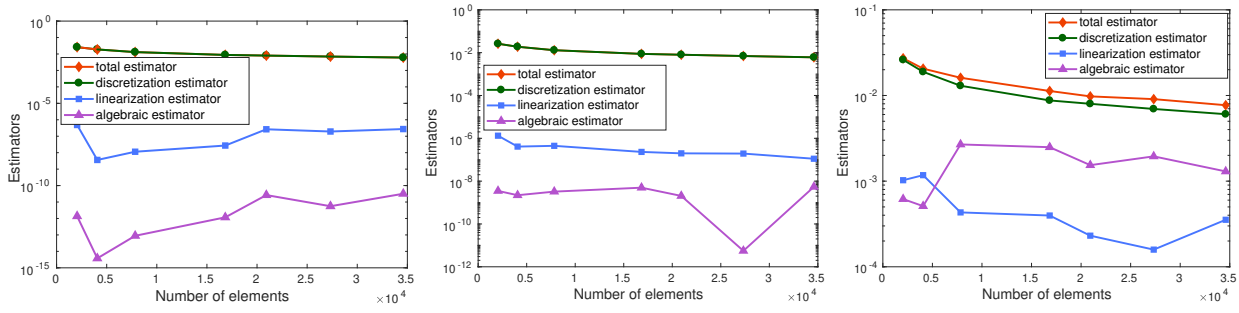

Figure 4: A posteriori estimators at convergence $(\eta^{\overline{k},\overline{i}},\ \eta_{\text{disc}}^{\overline{k},\overline{i}},\ \eta_{\text{lin}}^{\overline{k},\overline{i}},\ \eta_{\text{alg}}^{\overline{k},\overline{i}})$ as a function of the number of mesh elements. Exact (left), inexact (middle), and adaptive inexact (right) Newton-min methods with respectively the stopping criteria (88),(89), and (68). The log scales are different in each graph.

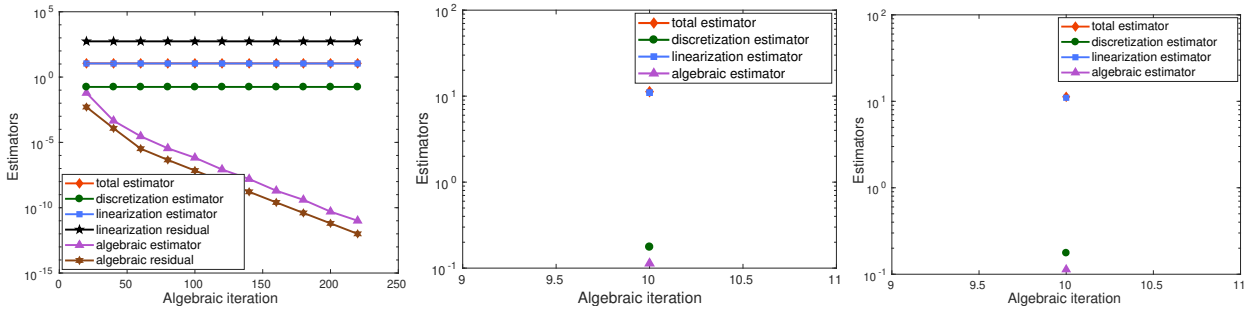

Figure 5: Estimators as a function of the algebraic iterations for $k = 1$. Exact (left), inexact (middle), and adaptive inexact (right) Newton-min methods.

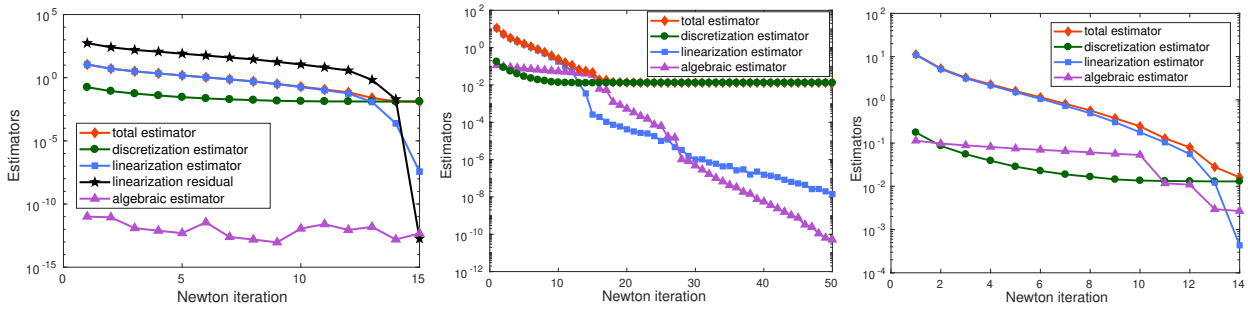

Figure 6: Estimators as a function of the Newton-min iterates k ($i = \bar{i}$). Exact (left), inexact (middle), and adaptive inexact (right) Newton-min methods.

adaptive inexact Newton method proposed here shows a different behavior: both $\eta_{\text{alg}}^{\bar{k},\bar{i}}$ and $\eta_{\text{lin}}^{\bar{k},\bar{i}}$ 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 $\boldsymbol{R}_{\text{lin}}^{k,i}$ is requested to lie below $\varepsilon_{\text{lin}} = 10^{-10}$ in both (88)(b) and (89)(b), the value of the present linearization estimator $\eta_{\text{lin}}^{k,i}$ still remains quite large, with values around 10^{-6} (see Figure 4, left and middle). Clearly, there is a huge difference between the l^2 size of the residual vector as expressed by $\left\|\boldsymbol{R}_{\text{lin}}^{k,i}\right\|$ and the size of its lifting back to the physical space expressed by $\eta_{\text{lin}}^{k,i}$.

Figure 5 shows the evolution of the various estimators and the behavior of $\|\boldsymbol{R}_{\text{lin}}^{k,i}\|$ and $\|\boldsymbol{R}_{\text{alg}}^{k,i}\|$ given by (87) during the algebraic iterations of the first 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 (88)(a), whereas in the inexact and adaptive inexact cases only 10 GMRES iterations are required to satisfy the stopping criteria (respectively (89)(a) and (67)(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 6 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 (89)(b), whereas the exact Newton-min and the adaptive inexact Newton-min methods require 15 and 14 iterations to achieve respectively the criteria (88)(b) and (67)(b).

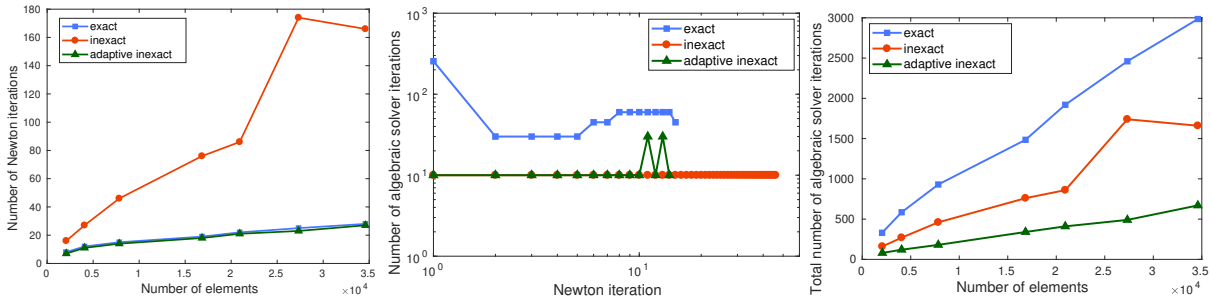

Figure 7: Number of Newton-min iterations per number of elements (left), number of algebraic solver iterations per Newton-min step for 8000 elements (middle), and total number of linear solver iterations per number of elements (right).

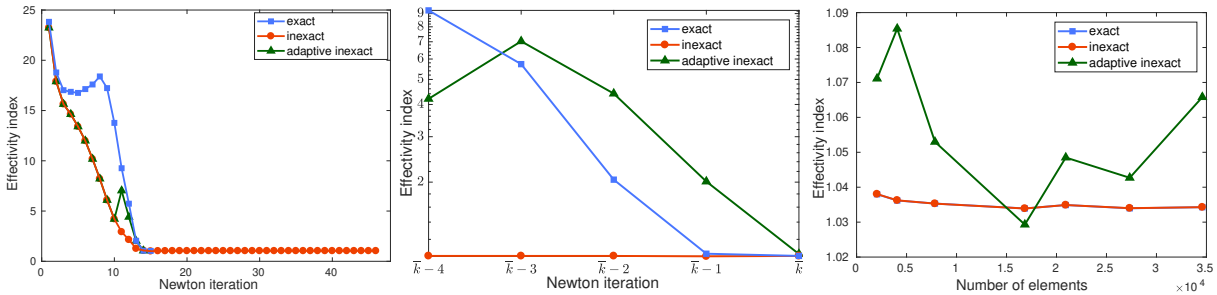

Figure 8: Effectivity index as a function of the Newton-min steps for three methods (left), effectivity indices for the last five semismooth Newton steps (middle), and effectivity indices as a function of the number of mesh elements (right); $\overline{k}$ stands for the last Newton-min step for each method ($\overline{k}$ = 15, 46, and 14 respectively for exact, inexact and adaptive inexact methods).

Figure 7 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 7 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 7. 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.

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 8 as a function of the Newton-min 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 8, we zoom on the last five semismooth Newton iterations for all the methods. In the right part of Figure 8, 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 9 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,

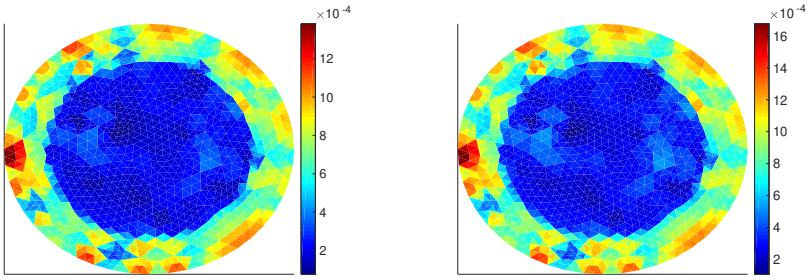

Figure 9: Error in energy norm (left) and total estimator (right), adaptive inexact Newton-min method, $p = 1$.

Table 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

$k = 3$, $i = \bar{i}$). We observe a very close agreement, even in the presence of algebraic and linearization errors.

Finally, Table 1 tests the dependency of our adaptive inexact methodology on the coefficients γ_{lin} and γ_{alg} in the algebraic and linearization stopping criteria (67)(a) and (67)(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 (67)(b) and on the second one the number of algebraic iterations required to obtain the stopping criterion (67)(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 iterations remains (much) smaller than for the exact and inexact semismooth Newton methods.

9 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 as well as the construction of a posteriori estimates for parabolic variational inequalities are under investigation.

References

[1] M. Aganagić, *Newton's method for linear complementarity problems*, Math. Programming, 28 (1984), pp. 349–362.

[2] 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.

[3] 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.

[4] M. Arioli, E. H. Georgoulis, and D. Loghin, *Stopping criteria for adaptive finite element solvers*, SIAM J. Sci. Comput., 35 (2013), pp. A1537–A1559.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] I. Ben Gharbia and J. Gilbert, *An algorithmic characterization of P-matricity II: corrections, refinements, and validation*. HAL Preprint 01672197, submitted for publication, 2017, https://hal.archives-ouvertes.fr/hal-01672197/.

[14] 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.

[15] I. Ben Gharbia and J. C. Gilbert, *An algorithmic characterization of* $\mathbf{P}$*-matricity*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 904–916, https://doi.org/10.1137/120883025.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Z. Chen, *Finite element methods and their applications*, Scientific Computation, Springer-Verlag, Berlin, 2005.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] F. Facchinei, C. Kanzow, and S. Sagratella, *Solving quasi-variational inequalities via their KKT conditions*, Math. Program., 144 (2014), pp. 369–412.

[38] 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.

[39] 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.

[40] Z. Ge, Q. Ni, and X. Zhang, *A smoothing inexact Newton method for variational inequalities with nonlinear constraints*, J. Inequal. Appl., (2017), pp. Paper No. 160, 12.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] R. Kornhuber, *A posteriori error estimates for elliptic variational inequalities*, Comput. Math. Appl., 31 (1996), pp. 49–60.

[48] J.-L. Lions and G. Stampacchia, *Variational inequalities*, Comm. Pure Appl. Math., 20 (1967), pp. 493–519.

[49] 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.

[50] 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).

[51] 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.

[52] 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/.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of numerical analysis, Vol. II, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991, pp. 523–639.

[58] 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.

[59] 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.

[60] 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.

[61] 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.