Numerische Mathematik (2020) 144:55–88
https://doi.org/10.1007/s00211-019-01084-0

Numerische Mathematik

First-order least-squares method for the obstacle problem

Thomas Führer[1]

Received: 29 October 2018 / Revised: 14 May 2019 / Published online: 22 October 2019

Abstract
We define and analyse a least-squares finite element method for a first-order reformulation of the obstacle problem. Moreover, we derive variational inequalities that are based on similar but non-symmetric bilinear forms. A priori error estimates including the case of non-conforming convex sets are given and optimal convergence rates are shown for the lowest-order case. We provide a posteriori bounds that can be used as error indicators in an adaptive algorithm. Numerical studies are presented.

Mathematics Subject Classification 65N30 · 65N12 · 49J40

1 Introduction

Many physical problems are of obstacle type, or more generally, described by variational inequalities [25,29]. In this article we consider, as a model problem, the classical obstacle problem where one seeks the equilibrium position of an elastic membrane constrained to lie over an obstacle. Another important example of an elliptic obstacle problem is the bending of a plate over an obstacle.

There exists already a long history of numerical methods, in particular finite element methods, see e.g., the books [16,17] for an overview on the topic. However, the literature on least-squares methods for obstacle problems is scarce. In fact, until the writing of this paper only [9] was available for the classical obstacle problem where the idea goes back to a Nitsche-based method for contact problems introduced and analyzed in [11]. An analysis of first-order least-squares finite element methods for Signorini problems can be found in [1] and more recently [26]. Let us also mention the pioneering work [14] for the a priori analysis of a classical finite element scheme. Newer articles include [18–20] where mixed and stabilized methods are considered.

Least-squares finite element methods are a widespread class of numerical schemes and their basic idea is to approximate the solution by minimizing the residual in some

✉ Thomas Führer
tofuhrer@mat.uc.cl

[1] Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile

given norm. Let us recall some important properties of least-squares finite element methods, a detailed list is given in the introduction of the overview article [5], see also the book [6].

- *Unconstrained stability* One feature of least-squares schemes is that the methods are stable for all pairings of discrete spaces.
- *Adaptivity* Another feature is that a posteriori error bounds are obtained by simply evaluating the least-squares functional. For instance, standard least-squares methods for the Poisson problem [6] are based on minimizing residuals in L^2 norms, which can be localized and, then, be used as error indicators in an adaptive algorithm.

The main purpose of this paper is to close the gap in the literature and define least-squares based methods for the obstacle problem. In particular, we want to study if the aforementioned properties transfer to the case of obstacle problems. Let us shortly describe the functional our method is based on. For simplicity assume a zero obstacle (the remainder of the paper deals with general non-zero obstacles). Then, the problem reads

$$-\Delta u \geq f, \quad u \geq 0, \quad (-\Delta u - f)u = 0$$

in some domain Ω and $u|_{\partial\Omega} = 0$. Introducing the Lagrange multiplier (or reaction force) $\lambda = -\Delta u - f$ and $\boldsymbol{\sigma} = \nabla u$, we rewrite the problem as a first-order system, see also [2,3,9,18],

$$-\operatorname{div}\boldsymbol{\sigma} - \lambda = f, \quad \boldsymbol{\sigma} - \nabla u = 0, \quad u \geq 0, \quad \lambda \geq 0, \quad \lambda u = 0.$$

Note that $f \in L^2(\Omega)$ does not imply more regularity for u so that $\lambda \in H^{-1}(\Omega)$ lives only in the dual space in general. However, observe that $\operatorname{div}\boldsymbol{\sigma} + \lambda = -f \in L^2(\Omega)$ and therefore the functional

$$J((u, \boldsymbol{\sigma}, \lambda); f) := \|\operatorname{div}\boldsymbol{\sigma} + \lambda + f\|^2 + \|\nabla u - \boldsymbol{\sigma}\|^2 + \langle \lambda, u \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes a duality pairing, is well-defined for $\operatorname{div}\boldsymbol{\sigma} + \lambda \in L^2(\Omega)$. We will show that minimizing J over a convex set with the additional constraints $u \geq 0$, $\lambda \geq 0$ is equivalent to solving the obstacle problem. We will consider the variational inequality associated to this problem with corresponding bilinear form $a(\cdot, \cdot)$. An issue that arises is that $a(\cdot, \cdot)$ is not necessarily coercive. However, as it turns out, a simple scaling of the first term in the functional ensures coercivity on the whole space. In view of the aforementioned properties, this means that our method is *unconstrained stable*. The recent work [18] based on a Lagrange formulation (without reformulation to a first-order system) considers augmenting the trial spaces with bubble functions (mixed method) resp. adding residual terms (stabilized method) to obtain stability. The authors extended their work also to plate-bending problems, see [20].

Another motivation of the proposed first-order reformulation is that it allows to simultaneously approximate displacements and stresses. In many problems of structural engineering the stress is usually the primary quantity of interest. For the present

problem of an elastic membrane the stress is directly related to the gradient and for the problem of bending a plate over an obstacle the physical quantities of interest are the bending moments.

Furthermore, we will see that the functional J evaluated at some discrete approximation $(u_h, \boldsymbol{\sigma}_h, \lambda_h)$ with $u_h, \lambda_h \geq 0$ is an upper bound for the error. Note that for $\lambda_h \in L^2(\Omega)$ the duality $\langle \lambda_h , u_h \rangle$ reduces to the L^2 inner product. Thus, all the terms in the functional can be localized and used as error indicators.

Additionally, we will derive and analyse other variational inequalities that are also based on the first-order reformulation. The resulting methods are quite similar to the least-squares scheme since they share the same residual terms. The only difference is that the complementary condition $\lambda u = 0$ is incorporated in a different, non-symmetric, way. We will present a uniform analysis that covers the least-squares formulation and the novel variational inequalities of the obstacle problem.

Finally, we point out that the use of adaptive schemes for obstacle problems is quite natural. First, the solutions may suffer from singularities stemming from the geometry, and second, the free boundary is a priori unknown. There exists plenty of literature on a posteriori estimators resp. adaptivity for finite elements methods for the obstacle problem, see e.g., [4,7,10,27,28,31,32] to name a few. Many of the estimators are based on the use of a discrete Lagrange multiplier which is obtained in a postprocessing step. In contrast, our proposed methods simultaneously approximate the Lagrange multiplier. This allows for a simple analysis of reliable a posteriori bounds.

1.1 Outline

The remainder of the paper is organized as follows. In Sect. 2 we describe the model problem, introduce the corresponding first-order system and based on that reformulation define our least-squares method. Then, Sect. 3 deals with the definition and analysis of different variational inequalities. In Sect. 4 we provide an a posteriori analysis and numerical studies are presented in Sect. 5. The appendix contains an example, which shows that $a(\cdot, \cdot)$ is not coercive in general, and proofs of some auxiliary results.

2 Least-squares method

In Sects. 2.1 and 2.2 we describe the model problem and introduce the reader to our notation. Then, Sect. 2.3 is devoted to the definition and analysis of a least-squares functional.

2.1 Model problem

Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$ denote a polygonal Lipschitz domain with boundary $\Gamma = \partial\Omega$. For given $f \in L^2(\Omega)$ and $g \in H^1(\Omega)$ with $g|_\Gamma \leq 0$ we consider the classical obstacle problem: find a solution u to

$$-\Delta u \geq f \quad \text{in } \Omega, \tag{1a}$$
$$u \geq g \quad \text{in } \Omega, \tag{1b}$$
$$(u-g)(-\Delta u - f) = 0 \quad \text{in } \Omega, \tag{1c}$$
$$u = 0 \quad \text{on } \Gamma. \tag{1d}$$

It is well-known that this problem admits a unique solution $u \in H_0^1(\Omega)$, and it can be equivalently characterized by the variational inequality: find $u \in H_0^1(\Omega)$, $u \geq g$ such that

$$\int_\Omega \nabla u \cdot \nabla(v-u)\,dx \geq \int_\Omega f(v-u)\,dx \quad \text{for all } v \in H_0^1(\Omega),\ v \geq g, \tag{2}$$

see [25]. For a more detailed description of the involved function spaces we refer to Sect. 2.2 below.

2.2 Notation and function spaces

We use the common notation for Sobolev spaces $H_0^1(\Omega)$, $H^s(\Omega)$ ($s > 0$). Let $(\cdot\,,\cdot)$ denote the $L^2(\Omega)$ inner product, which induces the norm $\|\cdot\|$. The dual of $H_0^1(\Omega)$ is denoted by $H^{-1}(\Omega) := (H_0^1(\Omega))^*$, where duality $\langle\cdot\,,\cdot\rangle$ is understood with respect to the extended $L^2(\Omega)$ inner product. We equip $H^{-1}(\Omega)$ with the dual norm

$$\|\lambda\|_{-1} := \sup_{0\neq v\in H_0^1(\Omega)} \frac{\langle \lambda\,,\, v\rangle}{\|\nabla v\|}.$$

Recall Friedrichs' inequality

$$\|u\| \leq C_F \|\nabla v\| \quad \text{for } v \in H_0^1(\Omega),$$

where $0 < C_F = C_F(\Omega) \leq \operatorname{diam}(\Omega)$. Thus, by definition we have $\|\lambda\|_{-1} \leq C_F\|\lambda\|$ for $\lambda \in L^2(\Omega)$.

Let $\operatorname{div} : \boldsymbol{L}^2(\Omega) := L^2(\Omega)^n \to H^{-1}(\Omega)$ denote the generalized divergence operator, i.e., $\langle \operatorname{div}\boldsymbol{\sigma}\,,\, u\rangle := -(\boldsymbol{\sigma}\,,\,\nabla u)$ for all $\boldsymbol{\sigma} \in \boldsymbol{L}^2(\Omega)$, $u \in H_0^1(\Omega)$. This operator is bounded,

$$\|\operatorname{div}\boldsymbol{\sigma}\|_{-1} = \sup_{0\neq v\in H_0^1(\Omega)} \frac{\langle \operatorname{div}\boldsymbol{\sigma}\,,\, v\rangle}{\|\nabla v\|} = \sup_{0\neq v\in H_0^1(\Omega)} \frac{-(\boldsymbol{\sigma}\,,\,\nabla v)}{\|\nabla v\|} \leq \|\boldsymbol{\sigma}\|.$$

Let $v \in H^1(\Omega)$. We say $v \geq 0$ if $v \geq 0$ a.e. in Ω. Moreover, $\lambda \geq 0$ for $\lambda \in H^{-1}(\Omega)$ means that $\langle \lambda\,, v\rangle \geq 0$ for all $v \in H_0^1(\Omega)$ with $v \geq 0$. We also interpret $v \geq w$ as $v - w \geq 0$ for $v, w \in H^1(\Omega)$.

Define the space

$$V := H_0^1(\Omega) \times \boldsymbol{L}^2(\Omega) \times H^{-1}(\Omega)$$

with norm

$$\|\boldsymbol{v}\|_V^2 := \|\nabla v\|^2 + \|\boldsymbol{\tau}\|^2 + \|\mu\|_{-1}^2 \quad \text{for } \boldsymbol{v} = (v, \boldsymbol{\tau}, \mu) \in V$$

and the space

$$U := \big\{(u, \boldsymbol{\sigma}, \lambda) \in V : \operatorname{div} \boldsymbol{\sigma} + \lambda \in L^2(\Omega)\big\}$$

with norm

$$\|\boldsymbol{u}\|_U^2 := \|\nabla u\|^2 + \|\boldsymbol{\sigma}\|^2 + \|\operatorname{div} \boldsymbol{\sigma} + \lambda\|^2 \quad \text{for } \boldsymbol{u} = (u, \boldsymbol{\sigma}, \lambda) \in U.$$

Observe that $\|\cdot\|_U$ is a stronger norm than $\|\cdot\|_V$, i.e.,

$$\begin{aligned}\|\nabla u\|^2 + \|\boldsymbol{\sigma}\|^2 + \|\lambda\|_{-1}^2 &\le \|\nabla u\|^2 + \|\boldsymbol{\sigma}\|^2 + 2\|\operatorname{div} \boldsymbol{\sigma} + \lambda\|_{-1}^2 + 2\|\operatorname{div} \boldsymbol{\sigma}\|_{-1}^2 \\ &\le \|\nabla u\|^2 + 3\|\boldsymbol{\sigma}\|^2 + 2C_F^2\|\operatorname{div} \boldsymbol{\sigma} + \lambda\|^2.\end{aligned}$$

Our first least-squares formulation will be based on the minimization over the non-empty, convex and closed subset

$$K^s := \big\{(u, \boldsymbol{\sigma}, \lambda) \in U : u - g \ge 0,\ \lambda \ge 0\big\},$$

where g is the given obstacle function. We will also derive and analyse variational inequalities based on non-symmetric bilinear forms that utilize the sets

$$\begin{aligned}K^0 &:= \big\{(u, \boldsymbol{\sigma}, \lambda) \in U : u - g \ge 0\big\}, \\ K^1 &:= \big\{(u, \boldsymbol{\sigma}, \lambda) \in U : \lambda \ge 0\big\}.\end{aligned}$$

Clearly, $K^s \subset K^j$ for $j = 0, 1$.

We write $A \lesssim B$ if there exists a constant $C > 0$, independent of quantities of interest, such that $A \le CB$. Analogously we define $A \gtrsim B$. If $A \lesssim B$ and $B \lesssim A$ hold then we write $A \simeq B$.

2.3 Least-squares functional

Let $u \in H_0^1(\Omega)$ denote the unique solution of the obstacle problem (1). Define $\lambda := -\Delta u - f \in H^{-1}(\Omega)$ and $\boldsymbol{\sigma} := \nabla u$. Problem (1) can equivalently be written as the first-order problem

$$-\operatorname{div} \boldsymbol{\sigma} - \lambda = f \quad \text{in } \Omega, \tag{3a}$$

$$\boldsymbol{\sigma} - \nabla u = 0 \quad \text{in } \Omega, \tag{3b}$$

$$u \ge g \quad \text{in } \Omega, \tag{3c}$$

$$\lambda \ge 0 \quad \text{in } \Omega, \tag{3d}$$

$$(u - g)\lambda = 0 \quad \text{in } \Omega, \tag{3e}$$

$$u = 0 \quad \text{on } \Gamma. \tag{3f}$$

Observe that $\operatorname{div}\boldsymbol{\sigma} + \lambda \in L^2(\Omega)$ and that the unique solution $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \lambda) \in U$ satisfies $\boldsymbol{u} \in K^s$. We consider the functional

$$J(\boldsymbol{u}; f, g) := \|\operatorname{div}\boldsymbol{\sigma} + \lambda + f\|^2 + \|\nabla u - \boldsymbol{\sigma}\|^2 + \langle \lambda, u - g\rangle$$

for $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \lambda) \in U$, $f \in L^2(\Omega)$, $g \in H^1_0(\Omega)$ and the minimization problem: find $\boldsymbol{u} \in K^s$ with

$$J(\boldsymbol{u}; f, g) = \min_{\boldsymbol{v} \in K^s} J(\boldsymbol{v}; f, g). \tag{4}$$

Note that the definition of the functional only makes sense if $g \in H^1_0(\Omega)$.

Theorem 1 *If $f \in L^2(\Omega)$, $g \in H^1_0(\Omega)$, then problems (3) and (4) are equivalent. In particular, there exists a unique solution $\boldsymbol{u} \in K^s$ of (4) and it holds that*

$$J(\boldsymbol{v}; f, g) \geq C_J \|\boldsymbol{v} - \boldsymbol{u}\|_U^2 \quad \textit{for all } \boldsymbol{v} \in K^s. \tag{5}$$

The constant $C_J > 0$ depends only on Ω.

Proof Let $\boldsymbol{u} := (u, \boldsymbol{\sigma}, \lambda) = (u, \nabla u, -\Delta u - f) \in K^s$ denote the unique solution of (3). Observe that $J(\boldsymbol{v}; f, g) \geq 0$ for all $\boldsymbol{v} \in K^s$ and $J(\boldsymbol{u}; f, g) = 0$, thus, $\boldsymbol{u}$ minimizes the functional. Suppose (5) holds and that $\boldsymbol{u}^* \in K^s$ is another minimizer. Then, (5) proves that $\boldsymbol{u} = \boldsymbol{u}^*$. It only remains to show (5). Let $\boldsymbol{v} = (v, \boldsymbol{\tau}, \mu) \in K^s$. Note that all terms in $J(\boldsymbol{v}; f, g)$ are non-negative. Since $f = -\operatorname{div}\boldsymbol{\sigma} - \lambda$ and $\nabla u - \boldsymbol{\sigma} = 0$ we have with the constant $C_F > 0$ that

$$\begin{aligned}
J(\boldsymbol{v}; f, g) &= \|\operatorname{div}(\boldsymbol{\tau} - \boldsymbol{\sigma}) + (\mu - \lambda)\|^2 + \|\nabla(v - u) - (\boldsymbol{\tau} - \boldsymbol{\sigma})\|^2 + \langle \mu, v - g\rangle \\
&= \frac{1}{1 + C_F^2}\Big((1 + C_F^2)\|\operatorname{div}(\boldsymbol{\tau} - \boldsymbol{\sigma}) + (\mu - \lambda)\|^2 \\
&\quad + (1 + C_F^2)\|\nabla(v - u) - (\boldsymbol{\tau} - \boldsymbol{\sigma})\|^2 \\
&\quad + (1 + C_F^2)\langle \mu, v - g\rangle \Big) \\
&\geq \frac{1}{1 + C_F^2}\Big((1 + C_F^2)\|\operatorname{div}(\boldsymbol{\tau} - \boldsymbol{\sigma}) + (\mu - \lambda)\|^2 \\
&\quad + \|\nabla(v - u) - (\boldsymbol{\tau} - \boldsymbol{\sigma})\|^2 + \langle \mu, v - g\rangle \Big).
\end{aligned}$$

Moreover, $\langle \lambda, u - g\rangle = 0$ and $\langle \lambda, v - g\rangle \geq 0$, $\langle \mu, u - g\rangle \geq 0$. We estimate

$$\begin{aligned}
\langle \mu, v - g\rangle &= \langle \mu, v - u\rangle + \langle \mu, u - g\rangle + \langle \lambda, u - g\rangle \\
&\geq \langle \mu, v - u\rangle + \langle \lambda, u - g\rangle + \langle \lambda, g - v\rangle \\
&= \langle \mu, v - u\rangle + \langle \lambda, u - v\rangle = \langle \mu - \lambda, v - u\rangle.
\end{aligned}$$

Define $\boldsymbol{w} := (w, \boldsymbol{\chi}, \nu) := \boldsymbol{v} - \boldsymbol{u}$. Then, the Cauchy–Schwarz inequality, Young's inequality and the definition of the divergence operator yield

$$\begin{aligned} J(\boldsymbol{v}; f, g) &\gtrsim (1 + C_F^2)\| \operatorname{div}(\boldsymbol{\tau} - \boldsymbol{\sigma}) + (\mu - \lambda)\|^2 \\ &\quad + \|\nabla(v - u) - (\boldsymbol{\tau} - \boldsymbol{\sigma})\|^2 + \langle \mu \,,\, v - g\rangle \\ &\geq (1 + C_F^2)\| \operatorname{div} \boldsymbol{\chi} + \nu\|^2 + \|\nabla w - \boldsymbol{\chi}\|^2 + \langle \nu \,,\, w\rangle \\ &= (1 + C_F^2)\| \operatorname{div} \boldsymbol{\chi} + \nu\|^2 + \|\nabla w\|^2 \\ &\quad + \|\boldsymbol{\chi}\|^2 - (\nabla w \,,\, \boldsymbol{\chi}) + \langle \operatorname{div} \boldsymbol{\chi} \,,\, w\rangle + \langle \nu \,,\, w\rangle \\ &\geq (1 + C_F^2)\| \operatorname{div} \boldsymbol{\chi} + \nu\|^2 + \tfrac{1}{2}\|\nabla w\|^2 + \tfrac{1}{2}\|\boldsymbol{\chi}\|^2 + \langle \operatorname{div} \boldsymbol{\chi} + \nu \,,\, w\rangle. \end{aligned}$$

Application of the Cauchy–Schwarz inequality, Friedrichs' inequality and Young's inequality gives us for the last term and $\delta > 0$

$$|\langle \operatorname{div} \boldsymbol{\chi} + \nu \,,\, w\rangle| \leq C_F \| \operatorname{div} \boldsymbol{\chi} + \nu\| \|\nabla w\| \leq C_F^2 \frac{\delta^{-1}}{2} \| \operatorname{div} \boldsymbol{\chi} + \nu\|^2 + \frac{\delta}{2}\|\nabla w\|^2.$$

Putting altogether and choosing $\delta = \frac{1}{2}$ we end up with

$$\begin{aligned} J(\boldsymbol{v}; f, g) &\gtrsim (1 + C_F^2)\| \operatorname{div} \boldsymbol{\chi} + \nu\|^2 + \|\nabla w - \boldsymbol{\chi}\|^2 + \langle \mu \,,\, v - g\rangle \\ &\geq (1 + C_F^2)\| \operatorname{div} \boldsymbol{\chi} + \nu\|^2 + \|\nabla w - \boldsymbol{\chi}\|^2 + \langle \nu \,,\, w\rangle \\ &\geq \| \operatorname{div} \boldsymbol{\chi} + \nu\|^2 + \tfrac{1}{4}\|\nabla w\|^2 + \tfrac{1}{2}\|\boldsymbol{\chi}\|^2 \simeq \|\boldsymbol{w}\|_U^2 = \|\boldsymbol{v} - \boldsymbol{u}\|_U^2, \end{aligned}$$

which finishes the proof. □

Remark 2 Note that (5) measures the error of any function $\boldsymbol{v} \in K^s$, in particular, it can be used as a posteriori error estimator when $\boldsymbol{v} \in K_h^s \subset K^s$ is a discrete approximation. However, in practice the condition $K_h^s \subset K^s$ is often hard to realize. Below we introduce a simple scaling of the first term in the least-squares functional that allows us to prove coercivity of the associated bilinear form on the whole space U.

For given $f \in L^2(\Omega)$, $g \in H_0^1(\Omega)$, and fixed parameter $\beta > 0$ define the bilinear form $a_\beta \colon U \times U \to \mathbb{R}$ and the functional $F_\beta \colon U \to \mathbb{R}$ by

$$a_\beta(\boldsymbol{u}, \boldsymbol{v}) := \beta(\operatorname{div} \boldsymbol{\sigma} + \lambda \,,\, \operatorname{div} \boldsymbol{\tau} + \mu) + (\nabla u - \boldsymbol{\sigma} \,,\, \nabla v - \boldsymbol{\tau}) + \tfrac{1}{2}(\langle \mu \,,\, u\rangle + \langle \lambda \,,\, v\rangle), \tag{6}$$

$$F_\beta(\boldsymbol{v}) := -\beta(f \,,\, \operatorname{div} \boldsymbol{\tau} + \mu) + \tfrac{1}{2}\langle \mu \,,\, g\rangle \tag{7}$$

for all $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \lambda)$, $\boldsymbol{v} = (v, \boldsymbol{\tau}, \mu) \in U$. We stress that $a_1(\cdot, \cdot)$ and $F_1(\cdot)$ induce the functional $J(\cdot; \cdot)$, i.e.,

$$J(\boldsymbol{u}; f, g) = a_1(\boldsymbol{u}, \boldsymbol{u}) - 2F_1(\boldsymbol{u}) + (f \,,\, f).$$

Since J is differentiable it is well-known that the solution $\boldsymbol{u} \in K^s$ of (4) satisfies the variational inequality

$$a_1(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq F_1(\boldsymbol{v} - \boldsymbol{u}) \quad \text{for all } \boldsymbol{v} \in K^s. \tag{8}$$

Conversely, if J is also convex in K^s, then any solution of (8) solves (4). However, J is convex on K^s iff $a_1(\boldsymbol{v} - \boldsymbol{w}, \boldsymbol{v} - \boldsymbol{w}) \geq 0$ for all $\boldsymbol{v}, \boldsymbol{w} \in K^s$, which is not true in general, see the example from "Appendix A". In Sect. 3 below we will show that for sufficiently large $\beta > 1$ the bilinear form $a_\beta(\cdot, \cdot)$ is coercive, even on the whole space U. This has the advantage that we can prove unique solvability of the continuous problem and its discretization simultaneously. More important, in practice this allows the use of non-conforming subsets $K_h^s \not\subseteq K^s$.

3 Variational inequalities

In this section we introduce and analyse different variational inequalities. The idea of including the complementary condition in different ways has also been used in [15] to derive DPG methods for contact problems.

We define the bilinear forms $b_\beta, c_\beta \colon U \times U \to \mathbb{R}$ and functionals G_β, H_β by

$$\begin{aligned}
b_\beta(\boldsymbol{u}, \boldsymbol{v}) &:= \beta(\operatorname{div} \boldsymbol{\sigma} + \lambda \,, \operatorname{div} \boldsymbol{\tau} + \mu) + (\nabla u - \boldsymbol{\sigma} \,, \nabla v - \boldsymbol{\tau}) + \langle \lambda \,, v \rangle, \\
c_\beta(\boldsymbol{u}, \boldsymbol{v}) &:= \beta(\operatorname{div} \boldsymbol{\sigma} + \lambda \,, \operatorname{div} \boldsymbol{\tau} + \mu) + (\nabla u - \boldsymbol{\sigma} \,, \nabla v - \boldsymbol{\tau}) + \langle \mu \,, u \rangle, \\
G_\beta(\boldsymbol{v}) &:= -\beta(f \,, \operatorname{div} \boldsymbol{\tau} + \mu) \\
H_\beta(\boldsymbol{v}) &:= -\beta(f \,, \operatorname{div} \boldsymbol{\tau} + \mu) + \langle \mu \,, g \rangle.
\end{aligned}$$

Let $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \lambda) \in K^s \subset K^j$ $(j = 0, 1)$ denote the unique solution of (3) with $f \in L^2(\Omega)$, $g \in H_0^1(\Omega)$. Recall that $\operatorname{div} \boldsymbol{\sigma} + \lambda = -f$. Testing this identity with $\operatorname{div} \boldsymbol{\tau} + \mu$, multiplying with $(\beta - 1)$ and adding it to (8) we see that the solution $\boldsymbol{u} \in K^s$ satisfies the variational inequality

$$a_\beta(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq F_\beta(\boldsymbol{v} - \boldsymbol{u}) \quad \text{for all } \boldsymbol{v} \in K^s. \tag{VIa}$$

For the derivation of our second variational inequality let $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \lambda) \in K^0$ denote the unique solution of (3) with $f \in L^2(\Omega)$, $g \in H^1(\Omega)$, $g|_\Gamma \leq 0$. Recall that $\lambda = -\Delta u - f$. By (2) we have that

$$\langle \lambda \,, v - u \rangle = (\nabla u \,, \nabla(v - u)) - (f \,, v - u) \geq 0$$

for all $v \in H_0^1(\Omega)$, $v \geq g$. Thus, $\boldsymbol{u} \in K^0$ satisfies the variational inequality

$$b_\beta(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq G_\beta(\boldsymbol{v} - \boldsymbol{u}) \quad \text{for all } \boldsymbol{v} \in K^0. \tag{VIb}$$

Our final method is based on the observation that for $\mu \geq 0$, we have that $\langle \mu \,, u - g \rangle \geq 0$ for $u \geq g \in H_0^1(\Omega)$. Together with $\langle \lambda \,, u - g \rangle = 0$ we conclude $\langle \mu - \lambda \,, u - g \rangle \geq 0$. Thus, $\boldsymbol{u} \in K^1$ satisfies the variational inequality

$$c_\beta(\boldsymbol{u}, \boldsymbol{v}-\boldsymbol{u}) \geq H_\beta(\boldsymbol{v}-\boldsymbol{u}) \quad \text{for all } \boldsymbol{v} \in K^1. \tag{VIc}$$

Note that a_β is symmetric, whereas b_β, c_β are not.

3.1 Solvability

In what follows we analyse the (unique) solvability of the variational inequalities (VIa)–(VIc) in a uniform manner (including discretizations).

Lemma 3 *Suppose $\beta > 0$. Let $A \in \{a_\beta, b_\beta, c_\beta\}$. There exists $C_\beta > 0$ depending only on $\beta > 0$ and Ω such that*

$$|A(\boldsymbol{u}, \boldsymbol{v})| \leq C_\beta \|\boldsymbol{u}\|_U \|\boldsymbol{v}\|_U \quad \textit{for all } \boldsymbol{u}, \boldsymbol{v} \in U.$$

If $\beta \geq 1 + C_F^2$, then A is coercive, i.e.,

$$C\|\boldsymbol{u}\|_U^2 \leq A(\boldsymbol{u}, \boldsymbol{u}) \quad \textit{for all } \boldsymbol{u} \in U.$$

The constant $C > 0$ is independent of β and Ω.

Proof We prove boundedness of $A = a_\beta$. Let $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \lambda)$, $\boldsymbol{v} = (v, \boldsymbol{\tau}, \mu) \in U$ be given. The Cauchy–Schwarz inequality together with the Friedrichs' inequality and boundedness of the divergence operator yields

$$\begin{aligned}
|a_\beta(\boldsymbol{u}, \boldsymbol{v})| &\leq \beta \|\operatorname{div}\boldsymbol{\sigma} + \lambda\| \|\operatorname{div}\boldsymbol{\tau} + \mu\| + \|\nabla u - \boldsymbol{\sigma}\| \|\nabla v - \boldsymbol{\tau}\| \\
&\quad + \tfrac{1}{2}(\langle \operatorname{div}\boldsymbol{\tau} + \mu\,, u\rangle - \langle \operatorname{div}\boldsymbol{\tau}\,, u\rangle + \langle \operatorname{div}\boldsymbol{\sigma} + \lambda\,, v\rangle - \langle \operatorname{div}\boldsymbol{\sigma}\,, v\rangle) \\
&\leq \beta \|\operatorname{div}\boldsymbol{\sigma} + \lambda\| \|\operatorname{div}\boldsymbol{\tau} + \mu\| + \|\nabla u - \boldsymbol{\sigma}\| \|\nabla v - \boldsymbol{\tau}\| \\
&\quad + \tfrac{1}{2}\big((C_F\|\operatorname{div}\boldsymbol{\tau} + \mu\| + \|\boldsymbol{\tau}\|)\|\nabla u\| + (C_F\|\operatorname{div}\boldsymbol{\sigma} + \lambda\| + \|\boldsymbol{\sigma}\|)\|\nabla v\|\big).
\end{aligned}$$

This shows boundedness of $a_\beta(\cdot,\cdot)$. Similarly, one concludes boundedness of $b_\beta(\cdot,\cdot)$ and $c_\beta(\cdot,\cdot)$.

For the proof of coercivity, observe that $a_\beta(\boldsymbol{w}, \boldsymbol{w}) = b_\beta(\boldsymbol{w}, \boldsymbol{w}) = c_\beta(\boldsymbol{w}, \boldsymbol{w})$ for all $\boldsymbol{w} \in U$. We stress that coercivity directly follows from the arguments given in the proof of Theorem 1. Note that the choice of β yields

$$A(\boldsymbol{w}, \boldsymbol{w}) \geq (1 + C_F^2)\|\operatorname{div}\boldsymbol{\chi} + \nu\|^2 + \|\nabla w - \boldsymbol{\chi}\|^2 + \langle \nu\,, w\rangle$$

for $\boldsymbol{w} = (w, \boldsymbol{\chi}, \nu) \in U$. The right-hand side can be further estimated following the argumentation as in the proof of Theorem 1 which gives us

$$(1 + C_F^2)\|\operatorname{div}\boldsymbol{\chi} + \nu\|^2 + \|\nabla w - \boldsymbol{\chi}\|^2 + \langle \nu\,, w\rangle \gtrsim \|\boldsymbol{w}\|_U^2.$$

This finishes the proof. □

Remark 4 Recall that $C_F \leq \operatorname{diam}(\Omega)$. Therefore, we can always choose $\beta = 1 + \operatorname{diam}(\Omega)^2$ to ensure coercivity of our bilinear forms. We stress that a choice of β of order $\operatorname{diam}(\Omega)$ is not only sufficient to ensure coercivity but also necessary in general as the example from "Appendix A" shows. Another possibility is to rescale Ω such that $\operatorname{diam}(\Omega) \leq 1$ which implies that we can choose $\beta = 2$. Furthermore, observe that a scaling of Ω transforms (1) to an equivalent obstacle problem (with appropriate redefined functions f, g). To be more precise, define $\widetilde{u}(x) := u(dx)$ with $d := \operatorname{diam}(\Omega) > 0$ and $u \in H^1_0(\Omega)$ the solution of (1). Moreover, set $\widetilde{f}(x) = d^2 f(dx)$, $\widetilde{g}(x) := g(dx)$. Then, $\widetilde{u}$ solves (1) in $\widetilde{\Omega} := \{x/d : x \in \Omega\}$ with f, g replaced by $\widetilde{f}, \widetilde{g}$.

The variational inequalities (VIa)–(VIc) are of the first kind and we use a standard framework for the analysis (Lions–Stampacchia theorem), see [16,17,25].

Theorem 5 *Suppose $\beta \geq 1 + C_F^2$. Let $A \in \{a_\beta, b_\beta, c_\beta\}$ and let $F : U \to \mathbb{R}$ denote a bounded linear functional. If $K \subseteq U$ is a non-empty convex and closed subset, then the variational inequality*

$$\text{Find } \boldsymbol{u} \in K \text{ s.t. } A(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq F(\boldsymbol{v} - \boldsymbol{u}) \quad \text{for all } \boldsymbol{v} \in K \tag{9}$$

admits a unique solution.

In particular, for $f \in L^2(\Omega)$, $g \in H^1_0(\Omega)$ each of the problems (VIa)–(VIc) has a unique solution and the problems are equivalent to (3).

Proof With the assumption on β, Lemma 3 proves that the bilinear forms are coercive and bounded. Then, unique solvability of (9) follows from the Lions–Stampacchia theorem, see e.g., [16,17,25].

Unique solvability of (VIa)–(VIc) follows since the functionals F_β, G_β, H_β are linear and bounded: For example, boundedness of F_β can be seen from

$$\begin{aligned} |F_\beta(\boldsymbol{v})| &= |-\beta(f, \operatorname{div} \boldsymbol{\tau} + \mu) + \tfrac{1}{2}(\operatorname{div} \boldsymbol{\tau} + \mu, g) - \tfrac{1}{2}\langle \operatorname{div} \boldsymbol{\tau}, g\rangle| \\ &\leq \beta \|f\| \|\operatorname{div} \boldsymbol{\tau} + \mu\| + \tfrac{1}{2}\|\operatorname{div} \boldsymbol{\tau} + \mu\| \|g\| + \tfrac{1}{2}\|\boldsymbol{\tau}\| \|\nabla g\| \\ &\lesssim (\|f\| + \|\nabla g\|)\|\boldsymbol{v}\|_U. \end{aligned}$$

The same arguments prove that G_β and H_β are bounded.

Finally, equivalence to (3) follows since all problems admit unique solutions and by construction the solution of (3) also solves each of the problems (VIa)–(VIc). □

Remark 6 We stress that the assumption $g \in H^1_0(\Omega)$ is necessary. If $g \in H^1(\Omega)$ then the term $\langle \mu, g\rangle$ in F_β, H_β is not well-defined. However, this term does not appear in G_β and therefore the variational inequality in (VIb) admits a unique solution if we only assume $g \in H^1(\Omega)$ with $g|_\Gamma \leq 0$.

Remark 7 The variational inequality (VIa) corresponds to a least-squares finite element method with convex functional

$$J_\beta(\boldsymbol{u}; f, g) := a_\beta(\boldsymbol{u}, \boldsymbol{u}) - 2F_\beta(\boldsymbol{u}) + \beta(f, f).$$

Then, Theorem 5 proves that the problem

$$J_\beta(\boldsymbol{u};\, f, g) = \min_{\boldsymbol{v}\in K} J_\beta(\boldsymbol{v};\, f, g)$$

admits a unique solution for all non-empty convex and closed sets $K \subseteq U$. Moreover, $J_\beta(\boldsymbol{u};\, f, g) \simeq J(\boldsymbol{u};\, f, g)$ for $\boldsymbol{u} \in K^s$, so that this problem is equivalent to (4) for $K = K^s$.

3.2 A priori analysis

The following three results provide general bounds on the approximation error. The proofs are based on standard arguments, see e.g., [14]. We give details for the proof of the first result, the others follow the same lines of argumentation and are left to the reader.

Theorem 8 *Suppose $\beta \geq 1 + C_F^2$. Let $\boldsymbol{u} \in K^s$ denote the solution of (VIa), where $f \in L^2(\Omega)$, $g \in H^1_0(\Omega)$. Let $K_h \subset U$ denote a non-empty convex and closed subset and let $\boldsymbol{u}_h \in K_h$ denote the solution of (9) with $A = a_\beta$, $F = F_\beta$ and $K = K_h$. It holds that*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2 \leq C_{\mathrm{opt}}\Big(\inf_{\boldsymbol{v}_h\in K_h} \big(\|\boldsymbol{u} - \boldsymbol{v}_h\|_U^2 + |\langle \lambda\,,\, v_h - u\rangle + \langle \mu_h - \lambda\,,\, u - g\rangle|\big) + \inf_{\boldsymbol{v}\in K^s} |\langle \lambda\,,\, v - u_h\rangle + \langle \mu - \lambda_h\,,\, u - g\rangle|\Big).$$

The constant $C_{\mathrm{opt}} > 0$ depends only on β and Ω.

Proof Throughout let $\boldsymbol{v} = (v, \boldsymbol{\tau}, \mu) \in K^s$, $\boldsymbol{v}_h = (v_h, \boldsymbol{\tau}_h, \mu_h) \in K_h$ and let $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \lambda) \in K^s$ denote the exact solution of (VIa). Thus, $\operatorname{div}\boldsymbol{\sigma} + \lambda + f = 0$ and $\nabla u - \boldsymbol{\sigma} = 0$. For arbitrary $\boldsymbol{w} = (w, \boldsymbol{\chi}, \nu) \in U$ it holds that

$$\begin{aligned} a_\beta(\boldsymbol{u}, \boldsymbol{w}) &= \beta(\operatorname{div}\boldsymbol{\sigma} + \lambda\,,\, \operatorname{div}\boldsymbol{\chi} + \nu) + (\nabla u - \boldsymbol{\sigma}\,,\, \nabla w - \boldsymbol{\chi}) + \tfrac{1}{2}(\langle \lambda\,,\, w\rangle + \langle \nu\,,\, u\rangle) \\ &= -\beta(f\,,\, \operatorname{div}\boldsymbol{\chi} + \nu) + \tfrac{1}{2}\langle \nu\,,\, g\rangle + \tfrac{1}{2}(\langle \lambda\,,\, w\rangle + \langle \nu\,,\, u - g\rangle) \\ &= F_\beta(\boldsymbol{w}) + \tfrac{1}{2}(\langle \lambda\,,\, w\rangle + \langle \nu\,,\, u - g\rangle). \end{aligned} \tag{10}$$

Using coercivity of $a_\beta(\cdot,\cdot)$, identity (10) and the fact that $\boldsymbol{u}_h$ solves the discretized variational inequality (on K_h) shows that

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2 &\lesssim a_\beta(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{u}_h) \\ &= a_\beta(\boldsymbol{u}, \boldsymbol{u} - \boldsymbol{u}_h) - a_\beta(\boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{v}_h) - a_\beta(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) \\ &\leq F_\beta(\boldsymbol{u} - \boldsymbol{u}_h) + \tfrac{1}{2}(\langle \lambda\,,\, u - u_h\rangle + \langle \lambda - \lambda_h\,,\, u - g\rangle) \\ &\quad - a_\beta(\boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{v}_h) - F_\beta(\boldsymbol{v}_h - \boldsymbol{u}_h) \\ &= F_\beta(\boldsymbol{u} - \boldsymbol{v}_h) + \tfrac{1}{2}(\langle \lambda\,,\, u - u_h\rangle + \langle \lambda - \lambda_h\,,\, u - g\rangle) - a_\beta(\boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{v}_h). \end{aligned}$$

Note that $0 = \langle \lambda , u - g\rangle \leq \langle \lambda , v - g\rangle$ and $\langle \lambda , u - g\rangle \leq \langle \mu , u - g\rangle$. Hence,

$$\begin{aligned}\langle \lambda , u - u_h\rangle + \langle \lambda - \lambda_h , u - g\rangle &= \langle \lambda , u - g + g - u_h\rangle + \langle \lambda - \lambda_h , u - g\rangle \\ &\leq \langle \lambda , v - g + g - u_h\rangle + \langle \mu - \lambda_h , u - g\rangle.\end{aligned}$$

This and identity (10) with $\boldsymbol{w} = \boldsymbol{u} - \boldsymbol{v}_h$ imply that

$$\begin{aligned}&F_\beta(\boldsymbol{u} - \boldsymbol{v}_h) - a_\beta(\boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{v}_h) + \tfrac{1}{2}(\langle \lambda , u - u_h\rangle + \langle \lambda - \lambda_h , u - g\rangle) \\ &\quad \leq a_\beta(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{v}_h) - \tfrac{1}{2}(\langle \lambda , u - v_h\rangle + \langle \lambda - \mu_h , u - g\rangle) \\ &\qquad + \tfrac{1}{2}(\langle \lambda , v - u_h\rangle + \langle \mu - \lambda_h , u - g\rangle).\end{aligned}$$

Putting altogether, boundedness of $a_\beta(\cdot, \cdot)$ and an application of Young's inequality with parameter $\delta > 0$ show that

$$\begin{aligned}\|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2 \lesssim \frac{\delta}{2}\|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2 + \frac{\delta^{-1}}{2}\|\boldsymbol{u} - \boldsymbol{v}_h\|_U^2 + |\langle \lambda , v_h - u\rangle + \langle \mu_h - \lambda , u - g\rangle| \\ + |\langle \lambda , v - u_h\rangle + \langle \mu - \lambda_h , u - g\rangle|.\end{aligned}$$

Subtracting the term $\delta/2\|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2$ for some sufficiently small $\delta > 0$ finishes the proof since $\boldsymbol{v} \in K^s$, $\boldsymbol{v}_h \in K_h$ are arbitrary. □

Theorem 9 *Suppose $\beta \geq 1 + C_F^2$. Let $\boldsymbol{u} \in K^0$ denote the solution of (VIb), where $f \in L^2(\Omega)$, $g \in H^1(\Omega)$ with $g|_\Gamma \leq 0$. Let $K_h \subset U$ denote a non-empty convex and closed subset and let $\boldsymbol{u}_h \in K_h$ denote the solution of (9) with $A = b_\beta$, $F = G_\beta$, and $K = K_h$. It holds that*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2 \leq C_{\mathrm{opt}}\Big(\inf_{\boldsymbol{v}_h \in K_h}\big(\|\boldsymbol{u} - \boldsymbol{v}_h\|_U^2 + |\langle \lambda , v_h - u\rangle|\big) + \inf_{\boldsymbol{v} \in K^0}|\langle \lambda , v - u_h\rangle|\Big).$$

The constant $C_{\mathrm{opt}} > 0$ depends only on β and Ω.

Theorem 10 *Suppose $\beta \geq 1 + C_F^2$. Let $\boldsymbol{u} \in K^1$ denote the solution of (VIc), where $f \in L^2(\Omega)$, $g \in H_0^1(\Omega)$. Let $K_h \subset U$ denote a non-empty convex and closed subset and let $\boldsymbol{u}_h \in K_h$ denote the solution of (9) with $A = c_\beta$, $F = H_\beta$, and $K = K_h$. It holds that*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2 \leq C_{\mathrm{opt}}\Big(\inf_{\boldsymbol{v}_h \in K_h}\big(\|\boldsymbol{u} - \boldsymbol{v}_h\|_U^2 + |\langle \mu_h - \lambda , u - g\rangle|\big) + \inf_{\boldsymbol{v} \in K^1}|\langle \mu - \lambda_h , u - g\rangle|\Big).$$

The constant $C_{\mathrm{opt}} > 0$ depends only on β and Ω.

3.3 Discretization

Let $\mathcal{T}$ denote a regular triangulation of Ω, $\bigcup_{T\in\mathcal{T}} \overline{T} = \overline{\Omega}$. We assume that $\mathcal{T}$ is κ-shape regular, i.e.,

$$\sup_{T\in\mathcal{T}} \frac{\operatorname{diam}(T)^n}{|T|} \leq \kappa < \infty.$$

Moreover, let $\mathcal{V}$ denote the vertices of the mesh $\mathcal{T}$ and $\mathcal{V}_0 := \mathcal{V}\backslash\Gamma$. Let $h_{\mathcal{T}} \in L^\infty(\Omega)$ denote the mesh-size function, $h_{\mathcal{T}}|_T := h_T := \operatorname{diam}(T)$ for $T \in \mathcal{T}$. Set $h := \max_{T\in\mathcal{T}} \operatorname{diam}(T)$. We use standard finite element spaces for the discretization. Let $\mathcal{P}^p(\mathcal{T})$ denote the space of $\mathcal{T}$-elementwise polynomials of degree less or equal than $p \in \mathbb{N}_0$. Let $\mathcal{RT}^p(\mathcal{T})$ denote the Raviart–Thomas space of degree $p \in \mathbb{N}_0$, $\mathcal{S}_0^{p+1}(\mathcal{T}) := \mathcal{P}^{p+1}(\mathcal{T}) \cap H_0^1(\Omega)$, and

$$U_{hp} := \mathcal{S}_0^{p+1}(\mathcal{T}) \times \mathcal{RT}^p(\mathcal{T}) \times \mathcal{P}^p(\mathcal{T}).$$

Clearly, $U_{hp} \subset U$. We stress that the polynomial degree is chosen, so that the best approximation in the norm $\|\cdot\|_U$ is of order h^{p+1}.

To define admissible convex sets for the discrete variational inequalities we need to put constraints on functions from the space $\mathcal{S}_0^{p+1}(\mathcal{T})$ or from $\mathcal{P}^p(\mathcal{T})$ or both. Let us remark that for a polynomial degree ≥ 2 such constraints are not straightforward to implement. One possibility would be to impose such constraints pointwise and then analyse the consistency error. We comment on the case $p = 1$ and $n = 2$ below. For some hp-FEM method for elliptic obstacle problems we refer to [2,3]. In order to avoid such quite technical treatments and for a simpler representation of the basic ideas we consider from now on the lowest-order case only, where the linear constraints can easily be built in. To that end define the non-empty convex subsets

$$K_h^s := \big\{(v_h, \boldsymbol{\tau}_h, \mu_h) \in U_{h0} \colon \mu_h \geq 0,\ v_h(x) \geq g(x) \text{ for all } x \in \mathcal{V}_0\big\}, \tag{11a}$$

$$K_h^0 := \big\{(v_h, \boldsymbol{\tau}_h, \mu_h) \in U_{h0} \colon v_h(x) \geq g(x) \text{ for all } x \in \mathcal{V}_0\big\}, \tag{11b}$$

$$K_h^1 := \big\{(v_h, \boldsymbol{\tau}_h, \mu_h) \in U_{h0} \colon \mu_h \geq 0\big\}. \tag{11c}$$

In the definition of K_h^s, K_h^0 we assume $g \in H^1(\Omega)\cap C^0(\overline{\Omega})$ so that the point evaluation is well-defined.

Let us shortly comment on how to incorporate the constraints for the higher-order space U_{h1} and $n = 2$. Let $\mathcal{V}_m$ denote the midpoints of interior edges of the triangulation $\mathcal{T}$. Then, a choice for the discrete convex set is

$$K_{h1}^s := \big\{(v_h, \boldsymbol{\tau}_h, \mu_h) \in U_{h1} \colon \mu_h \geq 0,\ v_h(z) \geq g(z) \quad \text{for all } z \in \mathcal{V}_0 \cup \mathcal{V}_m\big\}.$$

In the same manner one defines K_{h1}^0 resp. K_{h1}^1.

3.4 Auxiliary results

For the analysis of the convergence rates we use the nodal interpolation operator $I_h\colon H^2(\Omega) \to \mathcal{S}^1(\mathcal{T}) := \mathcal{P}^1(\mathcal{T}) \cap C^0(\overline{\Omega})$, the Raviart–Thomas projector $\Pi_h^{\mathrm{div}}\colon H^1(\Omega)^n \to \mathcal{RT}^0(\mathcal{T})$, and the $L^2(\Omega)$ projector $\Pi_h\colon L^2(\Omega) \to \mathcal{P}^0(\mathcal{T})$. Observe that with $v \geq 0$, $\mu \geq 0$ we have (with sufficient regularity) that $I_h v \geq 0$, $\Pi_h \mu \geq 0$. Moreover, recall the commutativity property $\operatorname{div} \Pi_h^{\mathrm{div}} = \Pi_h \operatorname{div}$, as well as the approximation properties

$$\|v - I_h v\| + h\|\nabla(v - I_h v)\| \lesssim h^2 \|D^2 v\|, \tag{12}$$

$$\|\boldsymbol{\tau} - \Pi_h^{\mathrm{div}} \boldsymbol{\tau}\| \lesssim h \|\nabla \boldsymbol{\tau}\|, \tag{13}$$

$$\|\mu - \Pi_h \mu\| \lesssim \|h_{\mathcal{T}} \nabla_{\mathcal{T}} \mu\|. \tag{14}$$

Here, $\nabla \boldsymbol{\tau}$ is understood componentwise, $\nabla_{\mathcal{T}} \mu$ denotes the $\mathcal{T}$-elementwise gradient of $\mu \in H^1(\mathcal{T}) := \{v \in L^2(\Omega)\colon v|_T \in H^1(T),\ T \in \mathcal{T}\}$. Set $\|v\|_{H^1(\mathcal{T})}^2 := \|v\|^2 + \|\nabla_{\mathcal{T}} v\|^2$. The involved constants depend only on the κ-shape regularity of $\mathcal{T}$ but are otherwise independent of $\mathcal{T}$. Furthermore, for $\mu \in L^2(\Omega)$, it also holds that

$$\|\mu - \Pi_h \mu\|_{-1} \lesssim \|h_{\mathcal{T}}(\mu - \Pi_h \mu)\|,$$

which follows from the definition of the dual norm, the projection and approximation property of Π_h.

The proof of optimal a priori convergence rates will also rely on the following two results. Scaling arguments and the continuous embedding $H^2(T_{\mathrm{ref}}) \hookrightarrow C^0(\overline{T_{\mathrm{ref}}})$ show the next result. Here, $\overline{T_{\mathrm{ref}}}$ denotes some reference element.

Lemma 11 *There exists a constant $C > 0$ depending only on T_{ref} and κ-shape regularity of the triangulation such that*

$$\|v\|_{L^\infty(T)} \leq C|T|^{-1/2}\big(\|v\|_T + h_T\|\nabla v\|_T + h_T^2\|D^2 v\|_T\big) \quad \text{for all } v \in H^2(T). \tag{15}$$

The next result is proven along the lines of [12, Lemma 7]. For completeness we present the proof of the nontrivial result adapted to our situation in "Appendix B". For each element $T \in \mathcal{T}$ and $v \in H^2(T)$ we define the level set

$$T_{\mathrm{C}}(v) := \{x \in T\colon v(x) = 0\} \quad \text{as well as the set} \quad T_{\mathrm{NC}}(v) := \{x \in T\colon v(x) \neq 0\}.$$

Note that $v \in H^2(T)$ implies that these sets are measurable. Moreover, $|T_{\mathrm{C}}(v)| + |T_{\mathrm{NC}}(v)| = |T|$.

Lemma 12 *Let $v \in H^2(T)$. Assume $|T_{\mathrm{C}}(v)| > 0$. Then,*

$$\|v\|_T \leq C h_T \frac{|T|^{1/2}}{|T_{\mathrm{C}}(v)|^{1/2}} \|\nabla v\|_T$$

and in particular

$$\|\nabla v\|_T \le C h_T \frac{|T|^{1/2}}{|T_{\mathrm{C}}(v)|^{1/2}} \|D^2 v\|_T.$$

Here, $C = \sqrt{n}$ for $n = 2, 3$.

3.5 Optimal a priori convergence rates

Theorem 13 *Suppose $\beta \ge 1 + C_F^2$. Let $\boldsymbol{u} \in K^s$ denote the solution of (VIa) with data $f \in L^2(\Omega)$, $g \in H_0^1(\Omega)$. Let K_h^s denote the set defined in (11a) and let $\boldsymbol{u}_h \in K_h^s$ denote the solution of (9) with $A = a_\beta$, $F = F_\beta$, and $K = K_h^s$. If $u \in H^2(\Omega)$, $g \in H^2(\Omega)$ and $f \in H^1(\mathcal{T})$, then*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U \le C_{\mathrm{app}} h (\|u\|_{H^2(\Omega)} + \|\nabla_{\mathcal{T}} f\| + \|\lambda\| + \|g\|_{H^2(\Omega)}).$$

The constant $C_{\mathrm{app}} > 0$ depends only on β, Ω, and κ-shape regularity of $\mathcal{T}$.

Proof Choose $\boldsymbol{v}_h = (I_h u, \Pi_h^{\mathrm{div}} \boldsymbol{\sigma}, \Pi_h \lambda) \in K_h^s$. The commutativity property of Π_h^{div} shows that

$$\mathrm{div}(\boldsymbol{\sigma} - \Pi_h^{\mathrm{div}} \boldsymbol{\sigma}) + \lambda - \Pi_h \lambda = (1 - \Pi_h)(\mathrm{div}\, \boldsymbol{\sigma} + \lambda) = (1 - \Pi_h) f.$$

Therefore, using the approximation properties of the involved operators proves

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{v}_h\|_U &\le \|(1 - \Pi_h) f\| + \|\boldsymbol{\sigma} - \Pi_h^{\mathrm{div}} \boldsymbol{\sigma}\| + \|\nabla(u - I_h u)\| \\ &\lesssim h \|\nabla_{\mathcal{T}} f\| + h \|u\|_{H^2(\Omega)}. \end{aligned}$$

Moreover,

$$|\langle \lambda\,,\, I_h u - u \rangle| \le \|\lambda\| h^2 \|D^2 u\| \lesssim h^2 (\|u\|_{H^2(\Omega)}^2 + \|\lambda\|^2).$$

We have to estimate the term

$$|\langle \Pi_h \lambda - \lambda\,,\, u - g \rangle|.$$

Define $T_{\mathrm{C}} := T_{\mathrm{C}}(u - g)$ and $T_{\mathrm{NC}} := T_{\mathrm{NC}}(u - g)$. Note that these two sets are measurable and we have that $|T_{\mathrm{C}}| + |T_{\mathrm{NC}}| = |T|$. We consider three cases: First, assume that $|T_{\mathrm{C}}| = 0$. This implies that $u - g > 0$ a.e. in T but since $(u - g)\lambda = 0$ we infer that $\lambda = 0$ a.e. in T. Therefore, $\Pi_h \lambda|_T = 0$ and we have that $\langle \Pi_h \lambda - \lambda\,,\, u - g \rangle_T = 0$. Second, assume that $|T_{\mathrm{NC}}| = 0$. But then, $u - g = 0$ a.e. in T and we have again $\langle \Pi_h \lambda - \lambda\,,\, u - g \rangle_T = 0$. The final case to be considered is $|T_{\mathrm{NC}}| > 0$, $|T_{\mathrm{C}}| > 0$: We have that

$$|\langle \Pi_h \lambda - \lambda\,,\, u - g \rangle_T| = |\langle \lambda\,,\, (\Pi_h - 1)(u - g) \rangle_T|$$

$$\leq \|\lambda\|_{L^1(T)} \|(1-\Pi_h)(u-g)\|_{L^\infty(T)}.$$

Note that $\lambda|_{T_{\mathrm{NC}}} = 0$. Thus, $\|\lambda\|_{L^1(T)} = \|\lambda\|_{L^1(T_{\mathrm{C}})} \leq |T_{\mathrm{C}}|^{1/2} \|\lambda\|_T$. For the second term we apply Lemma 11 with $v = (1-\Pi_h)(u-g)$ and together with the approximation property of Π_h we get the estimate

$$\begin{aligned}\|(1-\Pi_h)(u-g)\|_{L^\infty(T)} &\lesssim |T|^{-1/2} \big(\|(1-\Pi_h)(u-g)\|_T \\ &\quad + h_T \|\nabla(u-g)\|_T + h_T^2 \|D^2(u-g)\|_T \big) \\ &\lesssim |T|^{-1/2} \big(h_T \|\nabla(u-g)\|_T + h_T^2 \|D^2(u-g)\|_T \big).\end{aligned}$$

We can estimate the gradient term by applying the second inequality of Lemma 12 which gives us

$$\|\nabla(u-g)\|_T \lesssim \frac{h_T |T|^{1/2}}{|T_{\mathrm{C}}|^{1/2}} \|D^2(u-g)\|_T.$$

Clearly $|T_{\mathrm{C}}|^{1/2} \leq |T|^{1/2}$, thus $|T|^{-1/2} \leq |T_{\mathrm{C}}|^{-1/2}$ and we conclude that

$$\|(1-\Pi_h)(u-g)\|_{L^\infty(T)} \lesssim \frac{h_T^2}{|T_{\mathrm{C}}|^{1/2}} \|D^2(u-g)\|_T.$$

Using $\|\lambda\|_{L^1(T)} \leq |T_{\mathrm{C}}|^{1/2} \|\lambda\|_T$ then yields that

$$\begin{aligned}|\langle \Pi_h \lambda - \lambda \,,\, u-g \rangle_T| &\lesssim |T_{\mathrm{C}}|^{1/2} \|\lambda\|_T \frac{h_T^2}{|T_{\mathrm{C}}|^{1/2}} \|D^2(u-g)\|_T \\ &\leq h_T^2 \left(\|\lambda\|_T^2 + \|u\|_{H^2(T)}^2 + \|g\|_{H^2(T)}^2 \right).\end{aligned}$$

Summing up we have that

$$\begin{aligned}&\inf_{\boldsymbol{v}_h \in K_h^s} \left(\|\boldsymbol{u} - \boldsymbol{v}_h\|_U^2 + |\langle \lambda \,,\, v_h - u \rangle + \langle \mu_h - \lambda \,,\, u-g \rangle| \right) \\ &\qquad \lesssim h^2 \big(\|u\|_{H^2(\Omega)}^2 + \|\nabla_{\mathcal{T}} f\|^2 + \|\lambda\|^2 + \|g\|_{H^2(\Omega)}^2 \big).\end{aligned}$$

Therefore, in view of Theorem 8 it only remains to estimate the consistency error

$$\inf_{\boldsymbol{v} \in K^s} |\langle \lambda \,,\, v - u_h \rangle + \langle \mu - \lambda_h \,,\, u-g \rangle|.$$

Define $\boldsymbol{v} := (v, \boldsymbol{\chi}, \mu) := (v, 0, \lambda_h) \in U$ with $v := \sup\{u_h, g\}$ and observe that $\boldsymbol{v} \in K^s$. This directly leads to $\langle \mu - \lambda_h \,,\, u-g \rangle = 0$. For the remaining term we follow the seminal work [14] of Falk. The same lines as in the proof of [14, Lemma 4] show that

$$|\langle \lambda \,,\, v - u_h \rangle| \leq \|\lambda\| \|v - u_h\| \leq \|\lambda\| \|g - I_h g\| \lesssim h^2 \|g\|_{H^2(\Omega)} \|\lambda\|.$$

This finishes the proof. □

The proof of the following result can be obtained in the same fashion as the previous one and is therefore omitted.

Theorem 14 *Suppose $\beta \geq 1 + C_F^2$. Let $\boldsymbol{u} \in K^0$ denote the solution of (VIb) with data $f \in L^2(\Omega)$, $g \in H^1(\Omega)$, $g|_\Gamma \leq 0$. Let $\boldsymbol{u}_h \in K_h$ denote the solution of (9) with $A = b_\beta$, $F = G_\beta$, and $K = K_h$, where either $K_h = K_h^s$ or $K_h = K_h^0$. If $u \in H^2(\Omega)$, $g \in H^2(\Omega)$ and $f \in H^1(\mathcal{T})$, then*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U \leq C_{\mathrm{app}} h (\|u\|_{H^2(\Omega)} + \|\nabla_{\mathcal{T}} f\| + \|\lambda\| + \|g\|_{H^2(\Omega)}).$$

The constant $C_{\mathrm{app}} > 0$ depends only on β, Ω, and κ-shape regularity of $\mathcal{T}$.

Finally, we show convergence rates for problem (VIc) and its approximation. Note that for the sets K_h^1, K_h^s defined in (11c), (11a) it holds that $K_h^s \subset K_h^1 \subset K^1$ and thus the consistency error, see Theorem 10, vanishes. The proof is similar to the one of Theorem 13 and is therefore left to the reader.

Theorem 15 *Suppose $\beta \geq 1 + C_F^2$. Let $\boldsymbol{u} \in K^1$ denote the solution of (VIc) with data $f \in L^2(\Omega)$, $g \in H_0^1(\Omega)$. Let $\boldsymbol{u}_h \in K_h$ denote the solution of (9) with $A = c_\beta$, $F = H_\beta$, and $K = K_h$, where either $K_h = K_h^s$ or $K_h = K_h^1$. If $u \in H^2(\Omega)$, $g \in H^2(\Omega)$ and $f \in H^1(\mathcal{T})$, then*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U \leq C_{\mathrm{app}} h \left(\|u\|_{H^2(\Omega)} + \|\nabla_{\mathcal{T}} f\| + \|\lambda\| + \|g\|_{H^2(\Omega)}\right).$$

The constant $C_{\mathrm{app}} > 0$ depends only on β, Ω, and κ-shape regularity of $\mathcal{T}$.

4 A posteriori analysis

In this section we derive reliable error bounds that can be used as an a posteriori estimator. We define

$$\mathrm{osc} := \mathrm{osc}(f) := \|(1 - \Pi_h) f\|.$$

The estimator below includes the residual term

$$\eta^2 := \eta(\boldsymbol{u}_h, f)^2 := \|\operatorname{div} \boldsymbol{\sigma}_h + \lambda_h + \Pi_h f\|^2 + \|\nabla u_h - \boldsymbol{\sigma}_h\|^2,$$

which can be localized. The derivation of our estimators is quite simple and is based on the following observation. Let $\boldsymbol{u} \in K^s \subset K^j$ denote the unique solution of (3) and let $\boldsymbol{u}_h \in U_{h0}$ be arbitrary. Take $\beta = 1 + C_F^2$ and recall that by Lemma 3 it holds that $a_\beta(\boldsymbol{v}, \boldsymbol{v}) = b_\beta(\boldsymbol{v}, \boldsymbol{v}) = c_\beta(\boldsymbol{v}, \boldsymbol{v}) \gtrsim \|\boldsymbol{v}\|_U^2$ for all $\boldsymbol{v} \in U$. Then, together with the Pythagoras theorem $\|\mu\|^2 = \|(1 - \Pi_h)\mu\|^2 + \|\Pi_h \mu\|^2$ for $\mu \in L^2(\Omega)$ and using $\operatorname{div} \boldsymbol{\sigma} + \lambda + f = 0$, $\nabla u = \boldsymbol{\sigma}$, $\operatorname{div} \boldsymbol{\sigma}_h + \lambda_h \in \mathcal{P}^0(\mathcal{T})$, it follows that

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2 \lesssim \beta \|\operatorname{div} \boldsymbol{\sigma}_h + \lambda_h + f\|^2 + \|\nabla u_h - \boldsymbol{\sigma}_h\|^2 + \langle \lambda_h - \lambda, u_h - u \rangle$$

$$
\begin{aligned}
&= \beta \| \operatorname{div} \boldsymbol{\sigma}_h + \lambda_h + \Pi_h f \|^2 + \beta \operatorname{osc}^2 \\
&\quad + \| \nabla u_h - \boldsymbol{\sigma}_h \|^2 + \langle \lambda_h - \lambda \,,\, u_h - u \rangle \\
&\leq \beta(\eta^2 + \operatorname{osc}^2) + \langle \lambda_h - \lambda \,,\, u_h - u \rangle. \qquad (16)
\end{aligned}
$$

The remaining results in this section are proved by estimating the duality term $\langle \lambda_h - \lambda \,,\, u_h - u \rangle$ from (16). In particular, the proof of the next result employs only $\lambda_h \geq 0$ We will need the positive resp. negative part of a function $v \colon \Omega \to \mathbb{R}$,

$$
v_+ := \max\{0, v\}, \quad v_- := -\min\{0, v\}.
$$

This definition implies that $v = v_+ - v_-$. The ideas of estimating the duality term are similar as in [18,31] and references therein, see also [15] for a related estimate for Signorini-type problems. Note that we do not need to assume $g \in H^1_0(\Omega)$.

Theorem 16 *Let $\boldsymbol{u} \in K^s$ denote the solution of (3). Let $\boldsymbol{u}_h \in K_h$, where $K_h \in \{K^s_h, K^1_h\}$, be arbitrary. The error satisfies*

$$
\| \boldsymbol{u} - \boldsymbol{u}_h \|^2_U \leq C_{\mathrm{rel}} \big(\eta^2 + \rho^2 + \operatorname{osc}^2 \big),
$$

where the estimator contribution ρ is given by

$$
\rho^2 := \langle \lambda_h \,,\, (u_h - g)_+ \rangle + \| \nabla (g - u_h)_+ \|^2.
$$

The constant $C_{\mathrm{rel}} > 0$ depends only on Ω.

Proof In view of estimate (16) we only have to tackle the term $\langle \lambda_h - \lambda \,,\, u_h - u \rangle$. Define $v_h := \max\{u_h, g\}$. Clearly, $v_h \geq g$ and $v_h \in H^1_0(\Omega)$. Note that $\lambda = -\Delta u - f \in H^{-1}(\Omega)$. Therefore, $\langle \lambda \,,\, v \rangle = (\nabla u \,,\, \nabla v) - (f \,,\, v)$ for all $v \in H^1_0(\Omega)$ and using the variational inequality for the exact solution (2) yields

$$
\begin{aligned}
-\langle \lambda \,,\, u_h - u \rangle &= -\langle \lambda \,,\, u_h - v_h \rangle - \langle \lambda \,,\, v_h - u \rangle \leq -\langle \lambda \,,\, u_h - v_h \rangle \\
&= \langle \lambda \,,\, (u_h - g)_- \rangle = \langle \lambda - \lambda_h \,,\, (u_h - g)_- \rangle + \langle \lambda_h \,,\, (u_h - g)_- \rangle \\
&\leq \frac{\delta}{2} \| \lambda - \lambda_h \|^2_{-1} + \frac{\delta^{-1}}{2} \| \nabla (u_h - g)_- \|^2 + \langle \lambda_h \,,\, (u_h - g)_- \rangle
\end{aligned}
$$

for all $\delta > 0$. Employing $\lambda_h \geq 0$, $g - u \leq 0$, and $v + v_- = v_+$ we further infer that

$$
\begin{aligned}
\langle \lambda_h - \lambda \,,\, u_h - u \rangle &\leq \langle \lambda_h \,,\, u_h - g + (u_h - g)_- \rangle + \langle \lambda_h \,,\, g - u \rangle \\
&\quad + \frac{\delta}{2} \| \lambda - \lambda_h \|^2_{-1} + \frac{\delta^{-1}}{2} \| \nabla (u_h - g)_- \|^2 \\
&\leq \langle \lambda_h \,,\, (u_h - g)_+ \rangle + \frac{\delta}{2} \| \lambda - \lambda_h \|^2_{-1} + \frac{\delta^{-1}}{2} \| \nabla (u_h - g)_- \|^2.
\end{aligned}
$$

Recall that $\| \lambda - \lambda_h \|_{-1} \leq \| \boldsymbol{u} - \boldsymbol{u}_h \|_V \lesssim \| \boldsymbol{u} - \boldsymbol{u}_h \|_U$, where the involved constant depends only on Ω. Thus, choosing $\delta > 0$ sufficiently small the proof is concluded with (16). □

We could derive a similar estimate if $\boldsymbol{u}_h \in K_h^0$ by changing the role of u_h and λ_h resp. u and λ in the proof. However, this leads to an estimator with a non-local term. To see this, suppose $g = 0$. Then, following the last proof we get

$$\langle \lambda_h - \lambda\,,\, u_h - u\rangle \leq \langle (\lambda_h)_+\,,\, u_h\rangle + \frac{\delta}{2}\|\nabla(u-u_h)\|^2 + \frac{\delta^{-1}}{2}\|(\lambda_h)_-\|_{-1}^2$$

for $\delta > 0$. For the total error this would yield

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2 \lesssim \eta^2 + \operatorname{osc}^2 + \langle (\lambda_h)_+\,,\, u_h\rangle + \|(\lambda_h)_-\|_{-1}^2.$$

The last term is not localizable and therefore it is not feasible to use this estimate as an a posteriori error estimator in an adaptive algorithm.

Remark 17 The derived estimator is efficient up to the term ρ, i.e.,

$$\eta^2 + \operatorname{osc}^2 \lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U^2.$$

To see this, we employ the Pythagoras theorem to obtain

$$\eta^2 + \operatorname{osc}^2 = \|\operatorname{div}\boldsymbol{\sigma}_h + \lambda_h + f\|^2 + \|\nabla u_h - \boldsymbol{\sigma}_h\|^2.$$

Then, $\operatorname{div}\boldsymbol{\sigma} + \lambda = -f$, $\nabla u = \boldsymbol{\sigma}$ and the triangle inequality prove the asserted estimate. The proof of the efficiency estimate $\rho \lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U$ (up to possible data resp. obstacle oscillations) is an open problem, see also the related works [1,18].

5 Examples

In this section we present numerical studies that demonstrate the performance of our proposed methods in different situations:

- In Sect. 5.3 we consider a problem on the unit square with smooth obstacle and known smooth solution.
- In Sect. 5.4 we consider the example from [4, Section 5.2] where the solution is known and exhibits a singularity.
- In Sect. 5.5 we consider a problem on an L-shaped domain with a pyramid-like obstacle and unknown solution.

Before we come to a detailed discussion on the numerical studies some remarks are in order. In all examples we choose $\beta = 1 + \operatorname{diam}(\Omega)^2$ to ensure coercivity of the bilinear forms (Lemma 3). This also implies that the Galerkin matrices associated to the bilinear forms a_β, b_β, and c_β are positive definite. Choosing standard basis functions for $\mathcal{S}_0^1(\mathcal{T})$ (nodal basis), $\mathcal{RT}^0(\mathcal{T})$ (lowest-order Raviart–Thomas basis) and $\mathcal{P}^0(\mathcal{T})$ (characteristic functions), the constraints in the discrete convex sets $K_h^\star$, where $\star = 0$, $\star = 1$ or $\star = s$, are straightforward to impose. The resulting discrete variational inequalities are then solved using a (primal-dual) active set strategy, see e.g., [21–23].

5.1 Active set method and discrete variational inequalities

In this section we first define and collect results on the *(primal-dual) active set method*. Then, we recall the variational inequalities (VIa)–(VIc) and write down their discrete variants.

5.1.1 Active set method

Let $\mathcal{N} = \{1, \dots, N\}$, $N \in \mathbb{N}$, and let $\mathcal{N}_\gamma \subseteq \mathcal{N}$ be a non-empty subset. We set $\mathcal{N}_\omega := \mathcal{N} \backslash \mathcal{N}_\gamma$. For a vector $\boldsymbol{x} \in \mathbb{R}^N$ we write $\boldsymbol{x} = 0$ if all components are equal to 0. Similarly, $\boldsymbol{x} \geq 0$ means that all components of $\boldsymbol{x}$ are ≥ 0. For a subset $\mathcal{I} \subseteq \mathcal{N}$, $\boldsymbol{x}_\mathcal{I} = 0$ means $\boldsymbol{x}_i = 0$ for all $i \in \mathcal{I}$. We also use the notation $\boldsymbol{x}_\mathcal{I} \geq 0$, which means $\boldsymbol{x}_i \geq 0$ for all $i \in \mathcal{I}$ and $\boldsymbol{x}_\mathcal{I} \geq \boldsymbol{y}_\mathcal{I}$ stands for $\boldsymbol{x}_\mathcal{I} - \boldsymbol{y}_\mathcal{I} \geq 0$.

For $\boldsymbol{g} \in \mathbb{R}^N$ we consider the convex set

$$K := K_{\boldsymbol{g}} := \left\{ \boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x}_{\mathcal{N}_\gamma} \geq \boldsymbol{g}_{\mathcal{N}_\gamma} \right\}.$$

Let $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ denote a positive definite (but possibly non-symmetric) matrix, and $\boldsymbol{b} \in \mathbb{R}^N$ some arbitrary vector. We consider the variational inequality: find $\boldsymbol{x} \in K$, such that

$$\langle \boldsymbol{S}\boldsymbol{x}\,,\, \boldsymbol{y} - \boldsymbol{x} \rangle_2 \geq \langle \boldsymbol{b}\,,\, \boldsymbol{y} - \boldsymbol{x} \rangle_2 \quad \text{for all } \boldsymbol{y} \in K, \tag{17}$$

where $\langle \cdot\,,\, \cdot \rangle_2$ denotes the Euclidean inner product on $\mathbb{R}^N$. Since $\boldsymbol{S}$ is positive definite this problem admits a unique solution. It is well-known that problem (17) can be rewritten as follows: find $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathbb{R}^N \times \mathbb{R}^N$ such that

$$\boldsymbol{S}\boldsymbol{x} - \boldsymbol{\lambda} = \boldsymbol{b}, \tag{18a}$$

$$\boldsymbol{\lambda}_{\mathcal{N}_\omega} = 0, \tag{18b}$$

$$\boldsymbol{\lambda}_{\mathcal{N}_\gamma} = \max\{0, \boldsymbol{\lambda}_{\mathcal{N}_\gamma} - C(\boldsymbol{x}_{\mathcal{N}_\gamma} - \boldsymbol{g}_{\mathcal{N}_\gamma})\}, \tag{18c}$$

where $\max\{\cdot, \cdot\}$ denotes the componentwise maximum and $C > 0$ is some constant. Note that the solution is independent of C. Now following the seminal work [21] one defines a (semi-smooth) Newton method for solving (18). The same lines of argumentation as in [21] show that the method can be written as an active set strategy. The algorithm adapted to our situation is given in Algorithm 1.

The solution of the linear system in Line 8 of Algorithm 1 can be written (with $\mathcal{I} = \mathcal{I}^k$, $\mathcal{J} = \mathcal{J}^k$) as

$$\begin{pmatrix} \boldsymbol{S}_{\mathcal{I}\mathcal{I}} & \boldsymbol{S}_{\mathcal{I}\mathcal{J}} \\ \boldsymbol{S}_{\mathcal{J}\mathcal{I}} & \boldsymbol{S}_{\mathcal{J}\mathcal{J}} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_\mathcal{I} \\ \boldsymbol{x}_\mathcal{J} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\lambda}_\mathcal{I} \\ \boldsymbol{\lambda}_\mathcal{J} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}_\mathcal{I} \\ \boldsymbol{b}_\mathcal{J} \end{pmatrix}.$$

With the constraints $\boldsymbol{x}_\mathcal{J} = \boldsymbol{g}_\mathcal{J}$ and $\boldsymbol{\lambda}_\mathcal{I} = 0$ this reduces to the solution of the system

$$\boldsymbol{S}_{\mathcal{I}\mathcal{I}} \boldsymbol{x}_\mathcal{I} = \boldsymbol{b}_\mathcal{I} - \boldsymbol{S}_{\mathcal{I}\mathcal{J}} \boldsymbol{g}_\mathcal{J}$$

Algorithm 1 Active Set Method for solving (17)

Input: $\boldsymbol{x}^0, \boldsymbol{\lambda}^0$ and $C > 0$

1: Set $k = 0$

2: **while** TRUE **do**

3: $\widehat{\boldsymbol{\lambda}}^k_{\mathcal{N}_\gamma} \leftarrow \max\{0, \boldsymbol{\lambda}^k_{\mathcal{N}_\gamma} - C(\boldsymbol{x}^k_{\mathcal{N}_\gamma} - \boldsymbol{g}_{\mathcal{N}_\gamma})\}$ and $\widehat{\boldsymbol{\lambda}}^k_{\mathcal{N}_\omega} \leftarrow 0$

4: Determine set of *active* ($\mathcal{J}$) and *inactive* ($\mathcal{I}$) degrees of freedom

$$\mathcal{J}^k \leftarrow \{j \in \mathcal{N}_\gamma : \widehat{\boldsymbol{\lambda}}^k_j > 0\},$$
$$\mathcal{I}^k \leftarrow \mathcal{N} \backslash \mathcal{J}^k.$$

5: **if** $k \geq 1$ & $\mathcal{J}^k = \mathcal{J}^{k-1}$ **then**

6: **return** $\boldsymbol{x} = \boldsymbol{x}^k$

7: **end if**

8: Solve

$$\boldsymbol{S}\boldsymbol{x} - \boldsymbol{\lambda} = \boldsymbol{b},$$
$$\boldsymbol{\lambda}_{\mathcal{I}^k} = 0,$$
$$\boldsymbol{x}_{\mathcal{J}^k} = \boldsymbol{g}_{\mathcal{J}^k}.$$

9: Set $\boldsymbol{x}^{k+1} := \boldsymbol{x}$ and $\boldsymbol{\lambda}^{k+1} := \boldsymbol{\lambda}$

10: Increase counter $k \leftarrow k + 1$

11: **end while**

and the definition

$$\boldsymbol{\lambda}_{\mathcal{J}} := \boldsymbol{S}_{\mathcal{J}\mathcal{I}}\boldsymbol{x}_{\mathcal{I}} + \boldsymbol{S}_{\mathcal{J}\mathcal{J}}\boldsymbol{g}_{\mathcal{J}} - \boldsymbol{b}_{\mathcal{J}}.$$

Since $\boldsymbol{S}$ is positive definite the subblock $\boldsymbol{S}_{\mathcal{I}\mathcal{I}}$ is as well and thus $\boldsymbol{S}_{\mathcal{I}\mathcal{I}}$ is invertible.

Some remarks are in order. We can follow the analysis of [21] to see that the basic (local) convergence result holds true in our case as well.

Proposition 18 [21, Theorem 3.1] *If the initial guess* $(\boldsymbol{x}^0, \boldsymbol{\lambda}^0)$ *is sufficient close to the exact solution* $(\boldsymbol{x}, \boldsymbol{\lambda})$ *of* (18) *then the iterates* $(\boldsymbol{x}^k, \boldsymbol{\lambda}^k)$ *in Algorithm 1 converge superlinearly to* $(\boldsymbol{x}, \boldsymbol{\lambda})$.

The stopping criterion in Line 5 can be replaced by other criterions. Here, we choose $\mathcal{J}^k = \mathcal{J}^{k-1}$ because then we know that we have hit the exact solution of (17). The proof of the following result is a slight modification of the proof of [23, Lemma 3.1] and the interested reader can find it in "Appendix C".

Lemma 19 *If the stopping criterion in Line 5 of Algorithm 1 is satisfied, then* $\boldsymbol{x} = \boldsymbol{x}^k$ *is the solution of* (17).

5.1.2 Discrete variational inequalities

In this section we recall the discrete versions of the variational inequalities (VIa)–(VIc) and present them in matrix-vector form. They fit into the abstract framework given in Sect. 5.1.1.

Let us recall the discrete space from Sect. 3.3,

$$U_{h0} = \mathcal{S}_0^1(\mathcal{T}) \times \mathcal{RT}^0(\mathcal{T}) \times \mathcal{P}^0(\mathcal{T}).$$

Let $\mathcal{E}$ denote the set of edges ($n = 2$) resp. faces ($n = 3$). Then, $\dim(U_{h0}) = \#\mathcal{V}_0 + \#\mathcal{E} + \#\mathcal{T} =: N$. Numbering the nodes x_j of $\mathcal{V}_0$, the edges/faces E_j in $\mathcal{E}$ and the elements T_j in $\mathcal{T}$, we consider the following functions:

- For $j = 1, \dots, \#\mathcal{V}_0$ let v_j denote the nodal basis functions associated to the node $x_j \in \mathcal{V}_0$.
- For $j = 1, \dots, \#\mathcal{E}$ let $\boldsymbol{\tau}^{(j)}$ denote the Raviart–Thomas basis functions associated to the edge/face $E_j \in \mathcal{E}$.
- For $j = 1, \dots, \#\mathcal{T}$ let χ_j denote the characteristic function of the element $T_j \in \mathcal{T}$.

We define the basis $(\boldsymbol{\xi}^{(j)})_{j=1}^N$ for the space U_{h0} by

$$\begin{aligned}
\boldsymbol{\xi}^{(j)} &:= (v_j, 0, 0) && \text{for } j = 1, \dots, \#\mathcal{V}_0,\\
\boldsymbol{\xi}^{(\#\mathcal{V}_0 + j)} &:= (0, \boldsymbol{\tau}^{(j)}, 0) && \text{for } j = 1, \dots, \#\mathcal{E},\\
\boldsymbol{\xi}^{(\#\mathcal{V}_0 + \#\mathcal{E} + j)} &:= (0, 0, \chi_j) && \text{for } j = 1, \dots, \#\mathcal{T}.
\end{aligned}$$

Recall from Eq. 11 the discrete convex sets

$$\begin{aligned}
K_h^s &:= \big\{(v_h, \boldsymbol{\tau}_h, \mu_h) \in U_{h0} \colon \mu_h \geq 0,\ v_h(x) \geq g(x) \text{ for all } x \in \mathcal{V}_0\big\},\\
K_h^0 &:= \big\{(v_h, \boldsymbol{\tau}_h, \mu_h) \in U_{h0} \colon v_h(x) \geq g(x) \text{ for all } x \in \mathcal{V}_0\big\},\\
K_h^1 &:= \big\{(v_h, \boldsymbol{\tau}_h, \mu_h) \in U_{h0} \colon \mu_h \geq 0\big\}.
\end{aligned}$$

These convex subsets of U_{h0} correspond to convex subsets of $\mathbb{R}^N$ as follows: For given obstacle function $g \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$ define the vector $\boldsymbol{g} \in \mathbb{R}^N$ by

$$\boldsymbol{g}_j = \begin{cases} g(x_j) & \text{for } j = 1, \dots, \#\mathcal{V}_0,\\ 0 & \text{else} \end{cases}.$$

Let $\mathcal{N} = \{1, \dots, N\}$ and define $\mathcal{N}_\gamma^s, \mathcal{N}_\gamma^0, \mathcal{N}_\gamma^1$ by

$$\begin{aligned}
\mathcal{N}_\gamma^s &:= \{1, \dots, \#\mathcal{V}_0, \#\mathcal{V}_0 + \#\mathcal{E} + 1, \dots, N\} \subset \mathcal{N},\\
\mathcal{N}_\gamma^0 &:= \{1, \dots, \#\mathcal{V}_0\} \subset \mathcal{N},\\
\mathcal{N}_\gamma^1 &:= \{\#\mathcal{V}_0 + \#\mathcal{E} + 1, \dots, N\} \subset \mathcal{N}.
\end{aligned}$$

Then, the three sets K_h^s, K_h^0, K_h^1 correspond to the sets

$$\begin{aligned}
K_N^s &:= \big\{\boldsymbol{x} \in \mathbb{R}^N \colon \boldsymbol{x}_{\mathcal{N}_\gamma^s} \geq \boldsymbol{g}_{\mathcal{N}_\gamma^s}\big\},\\
K_N^0 &:= \big\{\boldsymbol{x} \in \mathbb{R}^N \colon \boldsymbol{x}_{\mathcal{N}_\gamma^0} \geq \boldsymbol{g}_{\mathcal{N}_\gamma^0}\big\},
\end{aligned}$$

$$K_N^1 := \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x}_{\mathcal{N}_\gamma^1} \geq 0\}.$$

With these definitions we can now state the algebraic forms of the discrete variational inequalities:

Discrete version of (VIa) with K_h^s

The discrete version of (VIa) with convex set K_h^s reads: find $\boldsymbol{u}_h \in K_h^s$ such that

$$a_\beta(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) \geq F_\beta(\boldsymbol{v}_h - \boldsymbol{u}_h) \quad \text{for all } \boldsymbol{v}_h \in K_h^s. \tag{19}$$

Let $\boldsymbol{S}^{(s)} \in \mathbb{R}^N \times \mathbb{R}^N$ denote the Galerkin matrix of the bilinear form $a_\beta(\cdot, \cdot)$ and let $\boldsymbol{b}^{(s)} \in \mathbb{R}^N$ denote the load vector, i.e.,

$$\boldsymbol{S}_{jk}^{(s)} = a_\beta(\boldsymbol{\xi}^{(k)}, \boldsymbol{\xi}^{(j)}), \quad \boldsymbol{b}_j^{(s)} = F_\beta(\boldsymbol{\xi}^{(j)})$$

for all $j, k = 1, \ldots, N$. Note that $\boldsymbol{S}^{(s)}$ is symmetric and positive definite. Problem (19) then reads in algebraic form as: find $\boldsymbol{x} \in K_N^s$ such that

$$\langle \boldsymbol{S}^{(s)} \boldsymbol{x}\,, \boldsymbol{y} - \boldsymbol{x} \rangle_2 \geq \langle \boldsymbol{b}^{(s)}\,, \boldsymbol{y} - \boldsymbol{x} \rangle_2 \quad \text{for all } \boldsymbol{y} \in K_N^s. \tag{20}$$

Discrete version of (VIb) with K_h^0

The discrete version of (VIb) with convex set K_h^0 reads: find $\boldsymbol{u}_h \in K_h^0$ such that

$$b_\beta(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) \geq G_\beta(\boldsymbol{v}_h - \boldsymbol{u}_h) \quad \text{for all } \boldsymbol{v}_h \in K_h^0. \tag{21}$$

Let $\boldsymbol{S}^{(0)} \in \mathbb{R}^N \times \mathbb{R}^N$ denote the Galerkin matrix of the bilinear form $b_\beta(\cdot, \cdot)$ and let $\boldsymbol{b}^{(0)} \in \mathbb{R}^N$ denote the load vector, i.e.,

$$\boldsymbol{S}_{jk}^{(0)} = b_\beta(\boldsymbol{\xi}^{(k)}, \boldsymbol{\xi}^{(j)}), \quad \boldsymbol{b}_j^{(0)} = G_\beta(\boldsymbol{\xi}^{(j)})$$

for all $j, k = 1, \ldots, N$. Note that $\boldsymbol{S}^{(0)}$ is non-symmetric and positive definite. Problem (21) then reads in algebraic form as: find $\boldsymbol{x} \in K_N^0$ such that

$$\langle \boldsymbol{S}^{(0)} \boldsymbol{x}\,, \boldsymbol{y} - \boldsymbol{x} \rangle_2 \geq \langle \boldsymbol{b}^{(0)}\,, \boldsymbol{y} - \boldsymbol{x} \rangle_2 \quad \text{for all } \boldsymbol{y} \in K_N^0. \tag{22}$$

Discrete version of (VIc) with K_h^1

The discrete version of (VIc) with convex set K_h^1 reads: find $\boldsymbol{u}_h \in K_h^1$ such that

$$c_\beta(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) \geq H_\beta(\boldsymbol{v}_h - \boldsymbol{u}_h) \quad \text{for all } \boldsymbol{v}_h \in K_h^1. \tag{23}$$

Let $\boldsymbol{S}^{(1)} \in \mathbb{R}^N \times \mathbb{R}^N$ denote the Galerkin matrix of the bilinear form $c_\beta(\cdot,\cdot)$ and let $\boldsymbol{b}^{(1)} \in \mathbb{R}^N$ denote the load vector, i.e.,

$$\boldsymbol{S}^{(1)}_{jk} = c_\beta(\boldsymbol{\xi}^{(k)}, \boldsymbol{\xi}^{(j)}), \quad \boldsymbol{b}^{(1)}_j = H_\beta(\boldsymbol{\xi}^{(j)})$$

for all $j,k = 1,\dots,N$. Note that $\boldsymbol{S}^{(1)}$ is non-symmetric and positive definite. Problem (23) then reads in algebraic form as: find $\boldsymbol{x} \in K^1_N$ such that

$$\langle \boldsymbol{S}^{(1)}\boldsymbol{x}\,,\, \boldsymbol{y}-\boldsymbol{x}\rangle_2 \geq \langle \boldsymbol{b}^{(1)}\,,\, \boldsymbol{y}-\boldsymbol{x}\rangle_2 \quad \text{for all } \boldsymbol{y} \in K^1_N. \tag{24}$$

Solver setup

The algebraic problems (20), (22) and (24) are then solved using Algorithm 1. The initial data $(\boldsymbol{x}^0, \boldsymbol{\lambda}^0)$ is chosen as the solution of

$$\begin{aligned} \boldsymbol{S}^{(\star)}\boldsymbol{x}^0 &= \boldsymbol{b}^{(\star)}, \\ \boldsymbol{x}^0_{\mathcal{N}^\star_\gamma} &= \boldsymbol{g}_{\mathcal{N}^\star_\gamma} \end{aligned}$$

and

$$\boldsymbol{\lambda}^0 := \max\{0, \boldsymbol{S}^{(\star)}\boldsymbol{x}^0 - \boldsymbol{b}^{(\star)}\},$$

where $\star = s$, $\star = 0$ or $\star = 1$. The constant C in Algorithm 1 is choosen as $C = 1$. The linear systems in Line 8 of Algorithm 1 are solved using the MATLAB backslash operator.

5.2 Error and estimator quantities

We define the error resp. total estimator by

$$\mathrm{err}_U := \|\boldsymbol{u} - \boldsymbol{u}_h\|_U, \quad \mathrm{est}^2 := \eta^2 + \rho^2 + \mathrm{osc}^2.$$

Note that the estimator can be decomposed into local contributions,

$$\begin{aligned} \mathrm{est}^2 = \sum_{T\in\mathcal{T}} \mathrm{est}(T)^2 := \sum_{T\in\mathcal{T}} \Big(&\|\operatorname{div}\boldsymbol{\sigma}_h + \lambda_h + \Pi_h f\|_T^2 + \|\nabla u_h - \boldsymbol{\sigma}_h\|_T^2 \\ &+ (\lambda_h\,,\,(u_h-g)_+)_T + \|\nabla(g-u_h)_+\|_T^2 + \|(1-\Pi_h)f\|_T^2\Big), \end{aligned}$$

where $\|\cdot\|_T$ denotes the $L^2(T)$ norm and $(\cdot,\cdot)_T$ the $L^2(T)$ inner product. Moreover, we will estimate the error in the weaker norm $\|\cdot\|_V$. To do so we consider an upper bound given by

$$\mathrm{err}_V^2 := \|\nabla(u-u_h)\|^2 + \|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|^2 + \|\lambda-\lambda_h\|^2_{-1,h},$$

where the evaluation of $\|\cdot\|_{-1,h}$ is based on the discrete $H^{-1}(\Omega)$ norm discussed in the seminal work [8]: Let $Q_h\colon L^2(\Omega) \to \mathcal{S}^1_0(\mathcal{T})$ denote the $L^2(\Omega)$ projector. Let $\mu \in L^2(\Omega)$. We stress that using the projection and local approximation property of Q_h yields

$$\|(1-Q_h)\mu\|_{-1} = \sup_{0\neq v\in H^1_0(\Omega)} \frac{\langle (1-Q_h)\mu\,,\,(1-Q_h)v\rangle}{\|\nabla v\|} \lesssim \|h_{\mathcal{T}}\mu\|,$$

where the involved constant depends on shape regularity of $\mathcal{T}$. Following [8] it holds that

$$\|\mu\|_{-1} \le \|(1-Q_h)\mu\|_{-1} + \|Q_h\mu\|_{-1} \lesssim \|h_{\mathcal{T}}\mu\| + \|\nabla u_h[\mu]\|$$

where $u_h[\mu] \in \mathcal{S}^1_0(\mathcal{T})$ is the solution of

$$(\nabla u_h[\mu]\,,\,\nabla v_h) = \langle \mu\,,\,v_h\rangle \quad \text{for all } v_h \in \mathcal{S}^1_0(\mathcal{T}).$$

Note that $\|\nabla u_h[\mu]\| \le \|\mu\|_{-1}$. The estimate $\|Q_h\mu\|_{-1} \lesssim \|\nabla u_h[\mu]\|$ depends on the stability of the projection Q_h in $H^1(\Omega)$, $\|\nabla Q_h v\| \lesssim \|\nabla v\|$ for $v \in H^1_0(\Omega)$, i.e.,

$$\begin{aligned}
\|Q_h\mu\|_{-1} &= \sup_{0\neq v\in H^1_0(\Omega)} \frac{\langle Q_h\mu\,,\,v\rangle}{\|\nabla v\|} = \sup_{0\neq v\in H^1_0(\Omega)} \frac{\langle \mu\,,\,Q_h v\rangle}{\|\nabla v\|} \\
&= \sup_{0\neq v\in H^1_0(\Omega)} \frac{(\nabla u_h[\mu]\,,\,\nabla Q_h v)}{\|\nabla v\|} \\
&\lesssim \sup_{0\neq v\in H^1_0(\Omega)} \frac{(\nabla u_h[\mu]\,,\,\nabla Q_h v)}{\|\nabla Q_h v\|} = \|\nabla u_h[\mu]\|.
\end{aligned}$$

Here, we use newest-vertex bisection [30] as refinement strategy where stability of the $L^2(\Omega)$ projection is known [24].

We use an adaptive algorithm that basically consists of iterating the four steps

$$\boxed{\mathit{SOLVE}} \to \boxed{\mathit{ESTIMATE}} \to \boxed{\mathit{MARK}} \to \boxed{\mathit{REFINE}},$$

where the marking step is done with the bulk criterion, i.e., we determine a set $\mathcal{M} \subseteq \mathcal{T}$ of (up to a constant) minimal cardinality with

$$\theta\,\mathrm{est}^2 \le \sum_{T\in\mathcal{M}} \mathrm{est}(T)^2.$$

For the experiments the marking parameter θ is set to $\frac{1}{4}$.

Convergence rates in the figures are indicated by triangles, where the number α besides the triangle denotes the experimental rate $\mathcal{O}((\#\mathcal{T})^{-\alpha})$. For uniform refinement we have $h^{2\alpha} \simeq \#\mathcal{T}^{-\alpha}$.

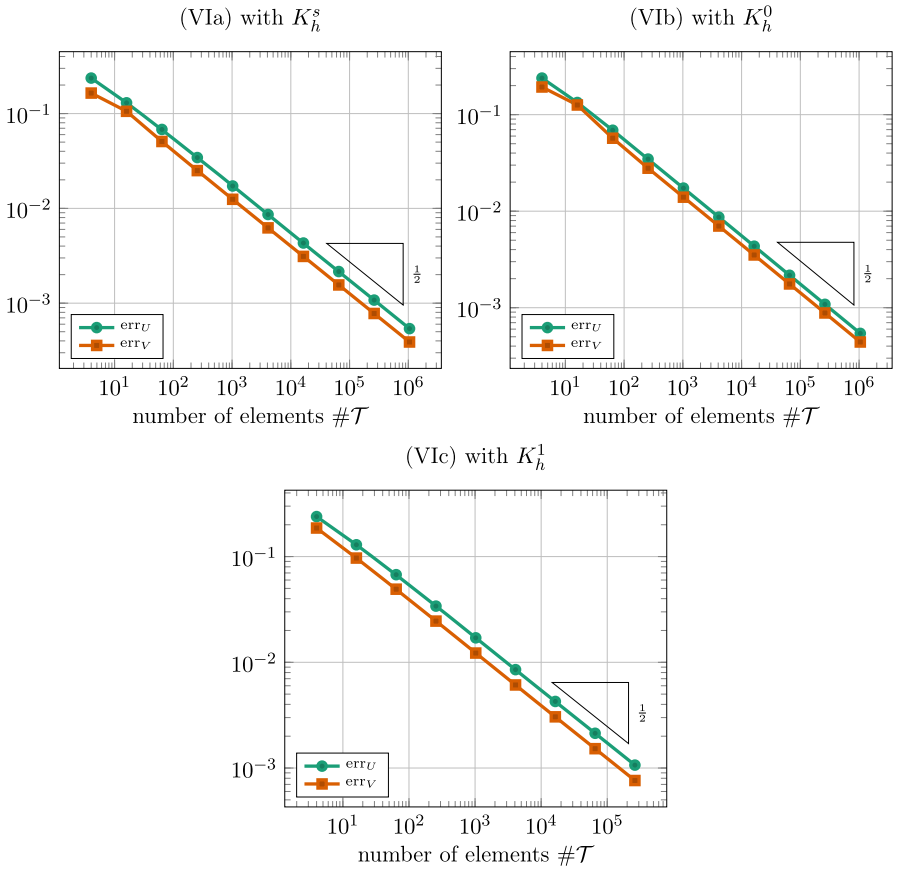

Fig. 1 Convergence rates for the problem from Sect. 5.3

5.3 Smooth solution

Let $\Omega = (0, 1)^2$, $u(x, y) = (1 - x)x(1 - y)y$,

$$f(x, y) := \begin{cases} 0 & x < \frac{1}{2} \\ -\Delta u(x, y) & x \geq \frac{1}{2} \end{cases}.$$

Then, u solves the obstacle problem (1) with data f and obstacle

$$g(x, y) = \begin{cases} (1 - x)x(1 - y)y & x \leq \frac{1}{2} \\ \widetilde{g}(x)(1 - y)y & x \in \left(\frac{1}{2}, \frac{3}{4}\right) \\ 0 & x \geq \frac{3}{4} \end{cases},$$

where $\widetilde{g}$ is the unique polynomial of degree 3 such that g and ∇g are continuous at the lines $x = \frac{1}{2}, \frac{3}{4}$. In particular, $g \in H^2(\Omega)$. Note that $\lambda = -\Delta u - f \in H^1(\mathcal{T})$. Figure 1

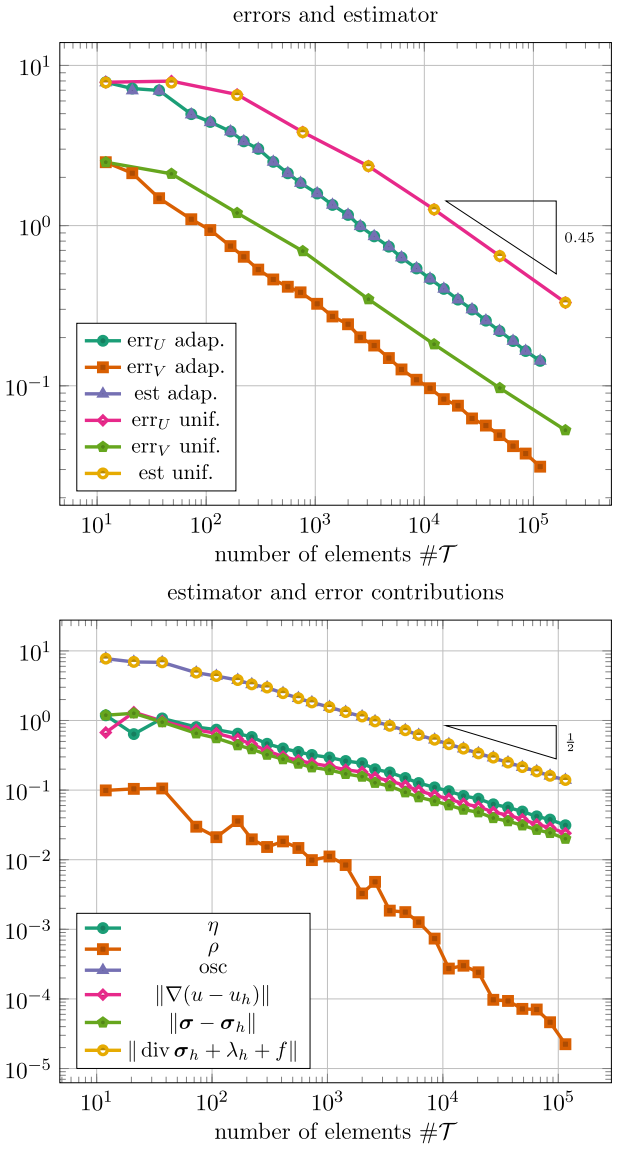

Fig. 2 Convergence rates for the problem from Sect. 5.4. The upper plot shows the total errors and estimators for uniform and adaptive refinement. The lower plot compares the error and estimator contributions in the case of adaptive refinements

shows that the convergence rates for the solutions of the discrete variational inequalities (VIa)–(VIc) based on the convex sets K_h^s, K_h^0, K_h^1 are optimal. This perfectly fits to our theoretic considerations in Theorems 13–15. Additionally, we plot err_V which is in all cases slightly smaller than err_U but of the same order. Note that since λ is a

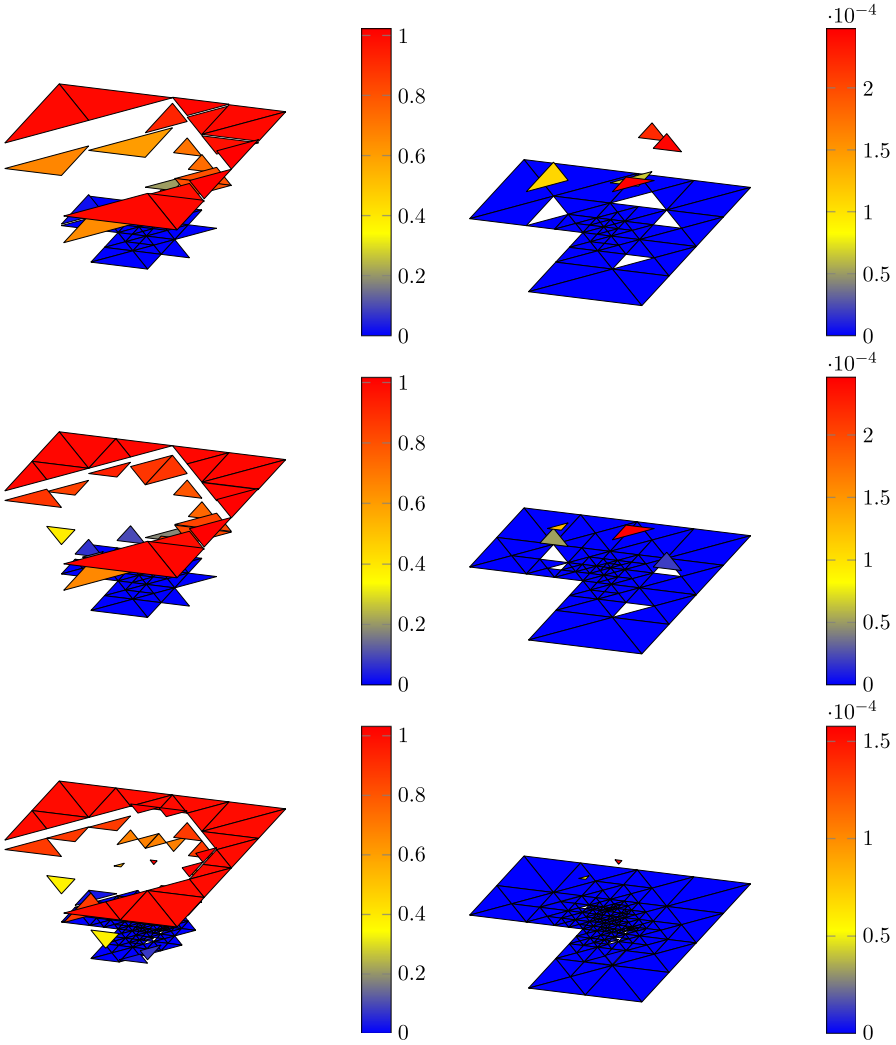

Fig. 3 Approximation λ_h (left) and distribution of the estimator contribution ρ^2 (right) for the example from Sect. 5.4

$\mathcal{T}$-elementwise polynomial, an inverse inequality shows that $h\|\lambda-\lambda_h\| \lesssim \|\lambda-\lambda_h\|_{-1}$ and thus err_V is equivalent to $\|\boldsymbol{u}-\boldsymbol{u}_h\|_V$.

5.4 Manufactured solution on L-shaped domain

We consider the same problem as given in [4, Section 5.2], where $g = 0$, $\Omega = (-2,2)^2\backslash[0,2]^2$ and

$$f(r,\varphi) := -r^{2/3}\sin(2/3\varphi)(\gamma'(r)/r+\gamma''(r)) - 4/3r^{-1/3}\gamma'(r)\sin(2/3\varphi) - \delta(r),$$

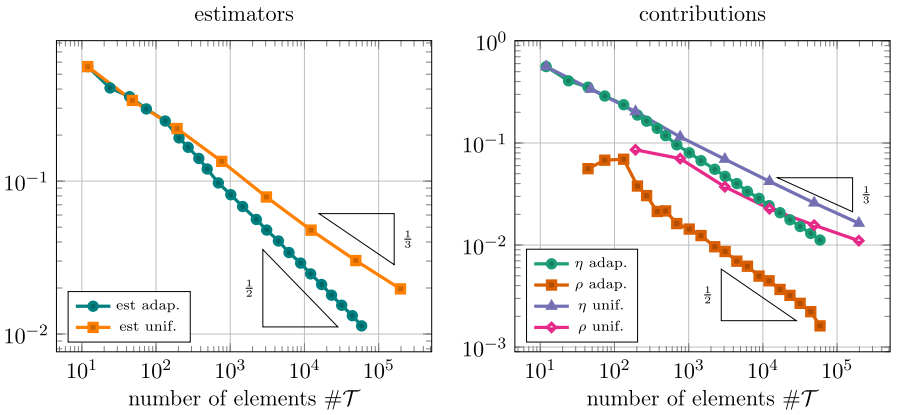

Fig. 4 Experimental convergence rates for the problem from Sect. 5.5

where (r, φ) denote polar coordinates. With $r_* = 2(r - 1/4)$, γ, δ are given by

$$\gamma(r) := \begin{cases} 1 & r_* < 0, \\ -6r_*^5 + 15r_*^4 - 10r_*^3 + 1 & 0 \le r_* < 1, \\ 0 & 1 \le r_*, \end{cases} \quad \delta(r) := \begin{cases} 0 & r \le 5/4, \\ 1 & r > 5/4. \end{cases}$$

The exact solution then reads $u(r, \varphi) = r^{2/3} \sin(2/3\varphi)\gamma(r)$. Note that u has a generic singularity at the reentrant corner. We consider the discrete version of (VIa), where solutions are sought in the convex set K_h^s. We conducted various tests with β between 1 and 100 and the results were in all cases comparable. For the results displayed here we have used $\beta = 3$. Figure 2 displays convergence rates in the case of uniform and adaptive mesh-refinement. We note that in the first plot the lines for err_U and est are almost identical. In the second plot we compare the contributions of the overall error and estimator in the adaptive case. The lines for osc and $\|\operatorname{div} \boldsymbol{\sigma}_h + \lambda_h + f\|$ are almost identical. This means that the estimator contribution $\|\operatorname{div} \boldsymbol{\sigma}_h + \lambda_h + \Pi_h f\|$ in η is negligible and osc is dominating the overall estimator. We observe from the first plot that err_V is much smaller than err_U but has the same rate of convergence. In the uniform case we see that the errors and estimators approximately converge at rate 0.45. One would expect a smaller rate due to the singularity. However, in this example the solution has a large gradient so that the algorithm first refines the regions where the gradient resp. f is large. This preasymptotic behavior was also observed in [4, Section 5.2]. Nevertheless, adaptivity yields a significant error reduction.

Figure 3 shows the approximation λ_h (left column) and the distribution of the estimator contribution ρ^2 (right column) on some adaptively refined meshes.

5.5 Unknown solution

For our final experiment, we choose $\Omega = (-1, 1)^2 \backslash [-1, 0]^2$, $f = 1$, and the pyramid-like obstacle $g(x) = \max\{0, \operatorname{dist}(x, \partial\Omega_u) - \frac{1}{4}\}$, where $\Omega_u = (0, 1)^2$. The solution

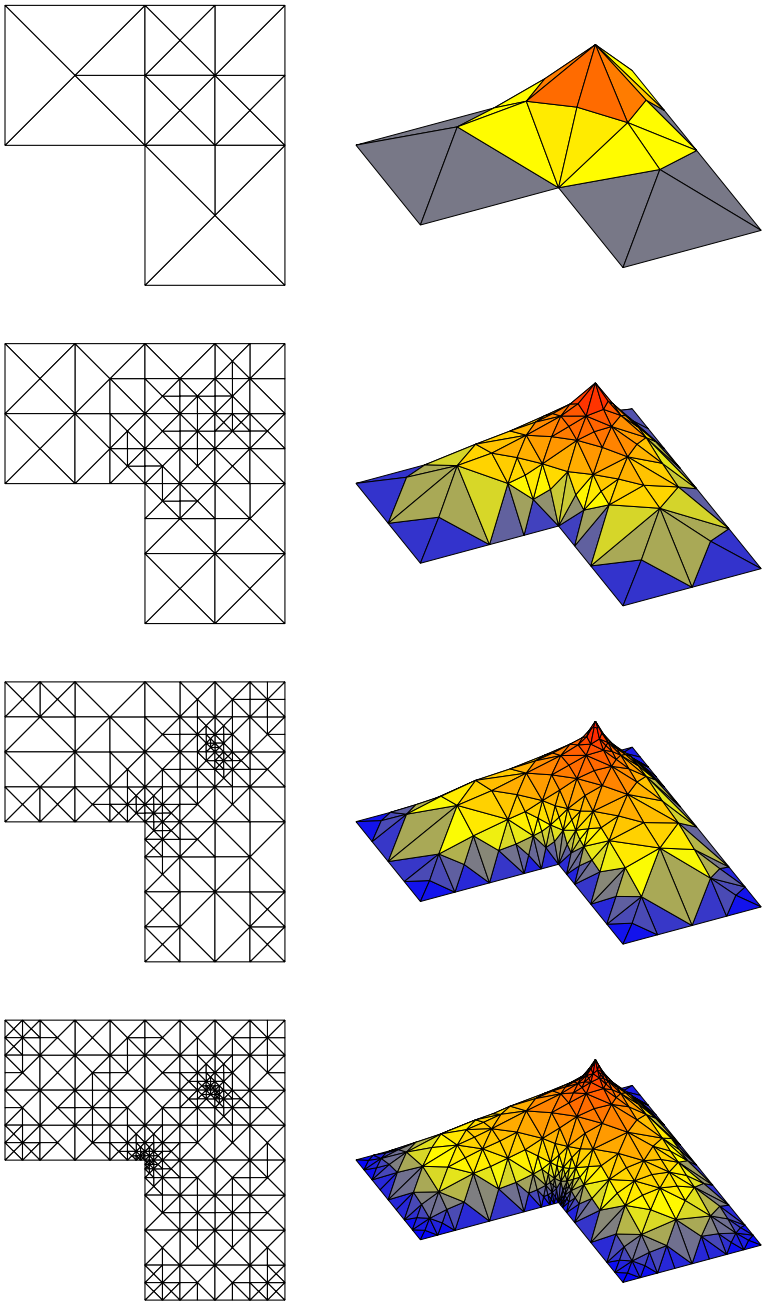

Fig. 5 Adaptively refined meshes and corresponding solution component u_h for the problem from Sect. 5.5

in this case is unknown. We solve the discrete version of (VIa) with convex set K_h^s. Since f is constant we have osc $= 0$. Figure 4 shows the overall estimator (left) and its contributions (right). We observe that uniform refinement leads to the reduced rate $\frac{1}{3}$, whereas for adaptive refinement we recover the optimal rate. Heuristically, we expect the solution to have a singularity at the reentrant corner as well as in the contact regions. This would explain the reduced rates. Figure 5 visualizes meshes produced by the adaptive algorithm and corresponding solution components u_h. We observe strong refinements towards the corner $(0, 0)$ and around the point $(\frac{1}{2}, \frac{1}{2})$, which coincides with the tip of the pyramid obstacle.

Acknowledgements This work was supported by CONICYT through FONDECYT project "Least-squares methods for obstacle problems" under Grant 11170050.

Appendix A: Non-convexity of functional *J*

Recall that the functional $J(\cdot; f, g)$ is convex if and only if

$$a_1(\boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v}) \geq 0 \quad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in K^s := \big\{(w, \boldsymbol{\chi}, \nu) \in U : w \geq g,\ \nu \geq 0\big\}.$$

In the following we construct a simple example that shows that the above inequality does not hold in general, thus J is not convex resp. $a(\cdot, \cdot) = a_1(\cdot, \cdot)$ is not coercive.

To that end, let $u \in H_0^1(\Omega)$ denote the solution of $\Delta u = 1$ in the square domain $\Omega = (0, d)^2$. Then, $u \leq 0$ in Ω. Choose the obstacle as $g = u$ (or $g \leq u$). Note that $\boldsymbol{v} := (0, 0, 0) \in K^s$ and that $\boldsymbol{u} := (u, \boldsymbol{\sigma}, \Delta u) := (u, \nabla u, \Delta u) \in K^s$. We have that

$$\|1\|_{-1} = \sup_{0 \neq v \in H_0^1(\Omega)} \frac{\langle 1\,,\, v\rangle}{\|\nabla v\|} = \sup_{0 \neq v \in H_0^1(\Omega)} \frac{-(\nabla u\,,\, \nabla v)}{\|\nabla v\|} = \|\nabla u\|.$$

Therefore $\|1\|_{-1}^2 = \|\nabla u\|^2 = -\langle 1\,,\, u\rangle$. Using this we infer that

$$a_1(\boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v}) = \|\operatorname{div} \boldsymbol{\sigma} + 1\|^2 + \|\nabla u - \boldsymbol{\sigma}\|^2 + \langle 1\,,\, u\rangle = \|2\|^2 - \|1\|_{-1}^2.$$

Hence, $a_1(\boldsymbol{u}-\boldsymbol{v}, \boldsymbol{u}-\boldsymbol{v}) < 0$ if and only if $\|2\| < \|1\|_{-1}$. Clearly, $\|2\| = 2|\Omega|^{1/2} = 2d$. We investigate the scaling of the negative order norm. Let $\widehat{\Omega}$ denote the unit square $(0, 1)^2$. Then,

$$\|1\|_{-1} = \sup_{0 \neq v \in H_0^1(\Omega)} \frac{\langle 1\,,\, v\rangle}{\|\nabla v\|} = |\Omega| \sup_{0 \neq \widehat{v} \in H_0^1(\widehat{\Omega})} \frac{(1\,,\, \widehat{v})_{\widehat{\Omega}}}{\|\nabla \widehat{v}\|_{\widehat{\Omega}}} = |\Omega|\, \|1\|_{-1,\widehat{\Omega}} =: |\Omega|\, C.$$

Finally, $\|2\| < \|1\|_{-1}$ if

$$2d < Cd^2$$

which holds for sufficiently large d.

Appendix B: Proof of Lemma 12

We use $T_{\mathrm{C}} := T_{\mathrm{C}}(v)$ and $T_{\mathrm{NC}} := T_{\mathrm{NC}}(v)$. If $|T_{\mathrm{NC}}| = 0$, then $v = 0$ on T and the first inequality is trivial. Note that also $\nabla v = 0$. Therefore, the second inequality is trivial as well. From now on we thus assume $|T_{\mathrm{NC}}| > 0$. Using that $v = 0$ on T_{C} and the identity

$$v(\boldsymbol{x}) - v(\boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y}) \cdot \int_0^1 \nabla v(s\boldsymbol{x} + (1-s)\boldsymbol{y})\, ds \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in T$$

we obtain that

$$\begin{aligned}
\|v\|_T^2 = \|v\|_{T_{\mathrm{NC}}}^2 &= \int_{T_{\mathrm{NC}}} v(\boldsymbol{x})^2\, d\boldsymbol{x} = \frac{1}{|T_{\mathrm{C}}|} \int_{T_{\mathrm{NC}}} \int_{T_{\mathrm{C}}} (v(\boldsymbol{x}) - v(\boldsymbol{y}))^2\, d\boldsymbol{y}\, d\boldsymbol{x} \\
&= \frac{1}{|T_{\mathrm{C}}|} \int_{T_{\mathrm{NC}}} \int_{T_{\mathrm{C}}} \Big((\boldsymbol{x} - \boldsymbol{y}) \cdot \int_0^1 \nabla v(s\boldsymbol{x} + (1-s)\boldsymbol{y})\, ds \Big)^2 \\
&\le \frac{h_T^2}{|T_{\mathrm{C}}|} \int_T \int_T \int_0^1 |\nabla v(s\boldsymbol{x} + (1-s)\boldsymbol{y})|^2\, ds\, d\boldsymbol{y}\, d\boldsymbol{x} \\
&= 2\frac{h_T^2}{|T_{\mathrm{C}}|} \int_T \int_{1/2}^1 \int_T |\nabla v(s\boldsymbol{x} + (1-s)\boldsymbol{y})|^2\, d\boldsymbol{x}\, ds\, d\boldsymbol{y},
\end{aligned}$$

where in the ultimate step we have used symmetry in $\boldsymbol{x}$ and $\boldsymbol{y}$. With the substitution $\boldsymbol{z} = \phi(\boldsymbol{x}) = s\boldsymbol{x} + (1-s)\boldsymbol{y} \in T$ we further get that

$$\begin{aligned}
\int_T \int_{1/2}^1 \int_T |\nabla v(s\boldsymbol{x} + (1-s)\boldsymbol{y})|^2\, d\boldsymbol{x}\, ds\, d\boldsymbol{y} &= \int_T \int_{1/2}^1 s^{-n} \int_{\phi(T)} |\nabla v(\boldsymbol{z})|^2\, d\boldsymbol{z}\, ds\, d\boldsymbol{y} \\
&\le \int_T \int_{1/2}^1 s^{-n} \int_T |\nabla v(\boldsymbol{z})|^2\, d\boldsymbol{z}\, ds\, d\boldsymbol{y} \\
&= |T| \frac{1 - (1/2)^{1-n}}{1-n} \|\nabla v\|_T^2.
\end{aligned}$$

Putting altogether this proves the first inequality.

For the second inequality we use the fact that the gradient on level sets vanishes, see [13, Theorem 3.3]. This means that $\nabla v = 0$ a.e. in T_{C}. Then, the same lines of proof as above (with v replaced by the components of ∇v) show the second inequality, which finishes the proof.

Appendix C: Proof of Lemma 19

We consider the decompositions

$$\mathcal{J}^{k-1} = \underbrace{\big\{ j \in \mathcal{J}^{k-1} : \boldsymbol{\lambda}_j^k > 0 \big\}}_{=:\mathcal{J}_1^{k-1}} \cup \underbrace{\big\{ j \in \mathcal{J}^{k-1} : \boldsymbol{\lambda}_j^k = 0 \big\}}_{=:\mathcal{J}_2^{k-1}},$$

$$\mathcal{I}^{k-1} = \underbrace{\left\{i \in \mathcal{I}^{k-1} : i \in \mathcal{N}_\omega \text{ or } i \in \mathcal{N}_\gamma \text{ with } \boldsymbol{x}_i^k \geq \boldsymbol{g}_i\right\}}_{=:\mathcal{I}_1^{k-1}} \cup \underbrace{\left\{i \in \mathcal{I}^{k-1} \cap \mathcal{N}_\gamma : \boldsymbol{x}_i^k < \boldsymbol{g}_i\right\}}_{=:\mathcal{I}_2^{k-1}}.$$

For the decomposition of $\mathcal{J}^k$ note that from Lines 8–9 of Algorithm 1 we have that $\boldsymbol{\lambda}_{\mathcal{N}_\gamma \cap \mathcal{I}^{k-1}} = 0$ and $\boldsymbol{x}_{\mathcal{J}^k} = \boldsymbol{g}_{\mathcal{J}^k}$. This yields

$$\mathcal{J}^k = \left\{j \in \mathcal{J}^{k-1} : \boldsymbol{\lambda}_j^k > 0\right\} \cup \left\{i \in \mathcal{I}^{k-1} \cap \mathcal{N}_\gamma : \boldsymbol{x}_i^k < \boldsymbol{g}_i\right\} = \mathcal{J}_1^{k-1} \cup \mathcal{I}_2^{k-1}$$

If $\mathcal{J}^k = \mathcal{J}^{k-1}$, then, since all decompositions are disjoint,

$$\mathcal{I}_2^{k-1} = \emptyset = \mathcal{J}_2^{k-1}.$$

This also means that $\widehat{\boldsymbol{\lambda}}_j^k > 0$ if and only if $\boldsymbol{\lambda}_j^k > 0$ with $\boldsymbol{x}_j^k = \boldsymbol{g}_j$. Thus,

$$\begin{aligned} \boldsymbol{\lambda}_{\mathcal{N}_\gamma}^k &= \max\{0, \boldsymbol{\lambda}_{\mathcal{N}_\gamma}^k - C(\boldsymbol{x}_{\mathcal{N}_\gamma} - \boldsymbol{g}_{\mathcal{N}_\gamma})\}, \\ \boldsymbol{\lambda}_{\mathcal{N}_\omega}^k &= 0, \end{aligned}$$

which implies that (18) is satisfied for $\boldsymbol{x} = \boldsymbol{x}^k$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}^k$ or equivalently $\boldsymbol{x} = \boldsymbol{x}^k$ solves (17).

References

1. Attia, F.S., Cai, Z., Starke, G.: First-order system least squares for the Signorini contact problem in linear elasticity. SIAM J. Numer. Anal. **47**(4), 3027–3043 (2009)
2. Banz, L., Schröder, A.: Biorthogonal basis functions in *hp*-adaptive FEM for elliptic obstacle problems. Comput. Math. Appl. **70**(8), 1721–1742 (2015)
3. Banz, L., Stephan, E.P.: A posteriori error estimates of *hp*-adaptive IPDG-FEM for elliptic obstacle problems. Appl. Numer. Math. **76**, 76–92 (2014)
4. Bartels, S., Carstensen, C.: Averaging techniques yield reliable a posteriori finite element error control for obstacle problems. Numer. Math. **99**(2), 225–249 (2004)
5. Bochev, P., Gunzburger, M.: Least-squares finite element methods. In: International Congress of Mathematicians, vol. III, pp. 1137–1162. Eur. Math. Soc., Zürich (2006)
6. Bochev, P.B., Gunzburger, M.D.: Least-Squares Finite Element Methods. Applied Mathematical Sciences, vol. 166. Springer, New York (2009)
7. Braess, D.: A posteriori error estimators for obstacle problems—another look. Numer. Math. **101**(3), 415–421 (2005)
8. Bramble, J.H., Lazarov, R.D., Pasciak, J.E.: A least-squares approach based on a discrete minus one inner product for first order systems. Math. Comput. **66**(219), 935–955 (1997)
9. Burman, E., Hansbo, P., Larson, M.G., Stenberg, R.: Galerkin least squares finite element method for the obstacle problem. Comput. Methods Appl. Mech. Eng. **313**, 362–374 (2017)
10. Chen, Z., Nochetto, R.H.: Residual type a posteriori error estimates for elliptic obstacle problems. Numer. Math. **84**(4), 527–548 (2000)
11. Chouly, F., Hild, P.: A Nitsche-based method for unilateral contact problems: numerical analysis. SIAM J. Numer. Anal. **51**(2), 1295–1307 (2013)
12. Drouet, G., Hild, P.: Optimal convergence for discrete variational inequalities modelling Signorini contact in 2D and 3D without additional assumptions on the unknown contact set. SIAM J. Numer. Anal. **53**(3), 1488–1507 (2015)

13. Evans, L.C., Gariepy, R.F.: Measure Theory and Fine Properties of Functions. Textbooks in Mathematics, revised edn. CRC Press, Boca Raton (2015)
14. Falk, R.S.: Error estimates for the approximation of a class of variational inequalities. Math. Comput. **28**, 963–971 (1974)
15. Führer, T., Heuer, N., Stephan, E.P.: On the DPG method for Signorini problems. IMA J. Numer. Anal. **38**(4), 1893–1926 (2018)
16. Glowinski, R.: Numerical methods for nonlinear variational problems. In: Scientific Computation. Springer, Berlin (2008). Reprint of the 1984 original
17. Glowinski, R., Lions, J.-L., Trémolières, R.: Numerical Analysis of Variational Inequalities, Volume 8 of Studies in Mathematics and Its Applications. North-Holland Publishing Co., Amsterdam (1981). Translated from the French
18. Gustafsson, T., Stenberg, R., Videman, J.: Mixed and stabilized finite element methods for the obstacle problem. SIAM J. Numer. Anal. **55**(6), 2718–2744 (2017)
19. Gustafsson, T., Stenberg, R., Videman, J.: On finite element formulations for the obstacle problem—mixed and stabilised methods. Comput. Methods Appl. Math. **17**(3), 413–429 (2017)
20. Gustafsson, T., Stenberg, R., Videman, J.: A stabilised finite element method for the plate obstacle problem. BIT **59**(1), 97–124 (2019)
21. Hintermüller, M., Ito, K., Kunisch, K.: The primal–dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**(3), 865–888 (2003), 2002
22. Hoppe, R.H.W., Kornhuber, R.: Adaptive multilevel methods for obstacle problems. SIAM J. Numer. Anal. **31**(2), 301–323 (1994)
23. Kärkkäinen, T., Kunisch, K., Tarvainen, P.: Augmented Lagrangian active set methods for obstacle problems. J. Optim. Theory Appl. **119**(3), 499–533 (2003)
24. Karkulik, M., Pavlicek, D., Praetorius, D.: On 2D newest vertex bisection: optimality of mesh-closure and H^1-stability of L_2-projection. Constr. Approx. **38**(2), 213–234 (2013)
25. Kinderlehrer, D., Stampacchia, G.: An Introduction to Variational Inequalities and Their Applications, Volume 31 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2000). Reprint of the 1980 original
26. Krause, R., Müller, B., Starke, G.: An adaptive least-squares mixed finite element method for the Signorini problem. Numer. Methods Partial Differ. Equ. **33**(1), 276–289 (2017)
27. Nochetto, R.H., Siebert, K.G., Veeser, A.: Pointwise a posteriori error control for elliptic obstacle problems. Numer. Math. **95**(1), 163–195 (2003)
28. Nochetto, R.H., Siebert, K.G., Veeser, A.: Fully localized a posteriori error estimators and barrier sets for contact problems. SIAM J. Numer. Anal. **42**(5), 2118–2135 (2005)
29. Rodrigues, J.-F.: Obstacle Problems in Mathematical Physics, Volume 134 of North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam (1987). Notas de Matemática [Mathematical Notes], 114
30. Stevenson, R.: The completion of locally refined simplicial partitions created by bisection. Math. Comput. **77**(261), 227–241 (2008)
31. Veeser, A.: Efficient and reliable a posteriori error estimators for elliptic obstacle problems. SIAM J. Numer. Anal. **39**(1), 146–167 (2001)
32. Weiss, A., Wohlmuth, B.I.: A posteriori error estimator for obstacle problems. SIAM J. Sci. Comput. **32**(5), 2627–2658 (2010)