Finite Element and Boundary Element methods for contact with adhesion

Von der Fakultät für Mathematik und Physik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades
Doktor der Naturwissenschaften
Dr. rer. nat.
genehmigte Dissertation
von

Leo Nesemann, M. Sc.

geboren am 17. 12. 1981 in Halle (Westf.)

2011

Referent:	Prof. Dr. Ernst P. Stephan, Leibniz Universität Hannover
Korreferent:	Prof. Dr. Joachim Gwinner, Universität der Bundeswehr, München
Tag der Promotion:	16. 12. 2010

Abstract

In this work, we consider mechanical problems with adhesive contact. Some contact problems, like the Signorini problem, may be represented through complementarity conditions. This does not hold true anymore for more involved surface laws, which are induced for example by adhesion or friction. In particular, these laws are in general not continuous. To circumvent this problem, we extend the boundary conditions of the original partial differential equation by terms that may be set-valued in certain points.
The PDE can now directly be converted into a variational formulation. We may also transfer the PDE into a boundary integral equation first and then retrieve a variational formulation on the boundary only. In both cases, we finally retrieve hemivariational inequalities due to the presence of nonsmooth surface laws.

Two schemes for the numerical solution of hemivariational inequalities are presented. For the first scheme, we use a minimization problem that is equivalent to the original inequality. We show that the objective function of this problem has a unique minimum under certain conditions, and that it is even convex under stronger assumptions.
By choosing finite-dimensional ansatz spaces in the domain or on the boundary, we retrieve a finite minimization problem with a non-differentiable objective function. This problem can be solved with the Bundle-Newton method.
For a second scheme, a primal-dual active set method on a membrane is considered. This method can be applied for certain surface laws. We describe a domain decomposition method to couple the membrane to the full mechanical system.

Finally, we demonstrate the applicability of these schemes in numerical experiments.

Keywords. Finite Element method, Boundary Element method, hemivariational inequality, active set strategy

Zusammenfassung

In dieser Arbeit betrachten wir mechanische Kontaktprobleme mit Adhäsion. Einige Kontaktprobleme, wie das Signorini-Problem, lassen sich durch Komplementaritätsbedingungen darstellen. Dies gilt nicht mehr für komplexere Oberflächengesetze, die zum Beispiel durch Adhäsion oder Reibung entstehen. Insbesondere sind diese Gesetze im Allgemeinen nicht mehr stetig. Um dieses Problem zu umgehen, erweitern wir die zugrunde liegenden Randbedingungen der partiellen Differentialgleichung durch Terme, die punktweise mengenwertig sein können.
Die Differentialgleichung kann nun entweder direkt in eine variationelle Formulierung überführt werden, oder das Problem wird zunächst mit Hilfe von Randintegralgleichungen dargestellt und dann in eine variationelle Formulierung auf dem Rand gebracht. In beiden Fällen erhalten wir durch die Beteiligung nichtglatter Materialgesetze schließlich hemivariationelle Ungleichungen.

Zur numerischen Lösung von hemivariationellen Ungleichungen zeigen wir zwei Verfahren. Für das erste Verfahren verwenden wir ein Minimierungsproblem, das zur ursprünglichen Ungleichung äquivalent ist. Wir zeigen, dass die Zielfunktion dieses Problems unter bestimmten Bedingungen ein eindeutiges Minimum besitzt und, unter strengeren Bedingungen, sogar konvex ist.
Indem wir endlich-dimensionale Ansatzräume im Gebiet oder auf dem Rand wählen, erhalten wir ein endliches Minimierungsproblem mit einer nichtdifferenzierbaren Zielfunktion, das mit dem Bundle-Newton-Verfahren gelöst werden kann.
In einem zweiten Verfahren wird eine primal-duale Active-Set-Methode auf einer Membran betrachtet. Diese Methode kann für bestimmte Oberflächengesetze verwendet werden. Wir geben ein Gebietszerlegungs-Verfahren an, das die Membran an das volle mechanische Problem koppelt.

Schließlich zeigen wir in numerischen Experimenten die Anwendbarkeit der Verfahren.

Schlagwörter. Finite-Elemente-Methode, Randelemente-Methode, Hemivariationsungleichung, Active-Set-Strategie

Acknowledgements

It is a pleasure for me to thank my advisor, Prof. Dr. E. P. Stephan, for his continuous support and various discussions about possible directions of my work. I am very grateful that he took some time for me when necessary, and left me some freedom where possible.

I would like to thank all members of the working group "Numerical Analysis". Especially, I thank Michael Andres for fruitful discussions and for being a fine company while we shared an office. Catalina Domínguez, Elke Ostermann, Florian Leydecker and Ricardo Prato helped me in various stages of my dissertation.
I would also like to thank my co-referee, Prof. Dr. J. Gwinner, for his readiness to examine this thesis. He and his whole working group made my research stay in Munich well-spent.

Furthermore, I am wholeheartedly grateful to my family for their support and their patience with me during the last years.

My work at the IfAM was made possible by a three-year scholarship of the German Research Foundation (DFG) in the Graduiertenkolleg 615, *Interaction of Modeling, Computation Methods and Software Concepts for Scientific-Technological Problems*. This research training group also enabled me to visit several conferences, and more importantly, provided a platform for discussions with PhD students from other areas of research.

Contents

1. Introduction

The mathematical modelling and numerical analysis of solid mechanics is a field of particular importance. The range of industrial applications goes from small, linear-elastic benchmark problems to fully transient simulations of whole cars in crash tests. As computer simulations gain more weight in product design processes, there grows a need for mathematical simulation tools that are reliable, efficient and applicable in a large group of problems. Even with the speed of computers increasing from year to year, and new computing concepts like parallelization available, we need these algorithms and results to allow solutions of more and more complex problems.

While the modelling process usually describes solutions in terms of a partial differential equation, variational formulations are in general used for further treatment. If a free contact boundary occurs, the framework of variational inequalities is typically used [16]. Here the Finite Element method, which can be derived from the variational formulation, has established itself as one of the most important numerical approximation methods. However, a detour via integral equations is possible and results in the Boundary Element method, which has also successfully been applied to a number of mechanical problems.

Some nonsmooth problem classes have already been studied extensively. One example are contact problems with monotone friction. Here, a convex functional $j(\cdot)$ is introduced into the variational formulation,

$$a(u, v-u) \; + \; j(v) - j(u) \; \geq \; f(v-u) \qquad \forall v \in \mathcal{K} \,, \tag{1.1}$$

which is lower semicontinuous, but may be nonsmooth. Kikuchi and Oden [23, chapters 10,11,13] give an overview of common cases and provide solution methods. Many further specializations to different types of friction are possible. As a particular example, we refer to the work of Chernov and Stephan [9], where boundary elements are used instead of finite elements.
In this thesis, we will demonstrate modelling and numerical solution methods for some types of nonlinear, nonsmooth behavior. We will still get a variational formulation of the type (1.1), but the functional $j(\cdot)$ is not convex anymore. The inequality (1.1) can then alternatively be expressed in the form

$$\begin{aligned} a(u, v-u) \; + \; \langle \xi, v-u \rangle \; &\geq f(v-u) \qquad && \forall v \in \mathcal{K} \\ \xi(x) \; &\in \hat{b}(u\big(x\big)) \qquad && \text{a.e. } x \in \omega \subset \Omega \,, \end{aligned} \tag{1.2}$$

using an auxiliary function ξ. Systems of this type are called hemivariational inequalities. That term was coined by Panagiotopoulos in the 1990s, see e.g. [32] for

an overview. As $j(\cdot)$ is not convex, we can not assume uniqueness of a solution anymore. Example problems are contact problems with adhesion or generalized friction, where the physical boundary laws may be discontinuous and nonmonotone. Two benchmarks are given by Baniotopoulos et al. [4].

Hemivariational inequalities are needed if discontinuous, nonconvex laws need to be included directly. The drawback of this approach is that it is tedious to find and apply appropriate numerical methods. If the material laws are at least continuous, semismooth Newton methods can be used, see e.g. Kunisch and Stadler [26] or Hager and Wohlmuth [18].
This thesis demonstrates two numerical methods for the discontinuous case, a Newton-like minimization method and an active set method that directly works on the hemivariational inequality. It is worth to mention that these problems can alternatively be treated with a regularization of the nonsmooth functional. The disadvantages of this procedure are that the involved linear equation systems in the solution process may be ill-conditioned, and that an additional approximation error needs to be controlled. The clear advantage is that a Newton method can directly be applied on the regularized problem.

A model problem: Unilateral contact with adhesion

The model problem stated in this section is only one example for the very broad class of hemivariational inequalities. We restrict the nonsmooth behavior to the boundary here.

Our benchmark problem is a two- or three-dimensional linear-elastic block under a given exterior load. The block is fixed on one side. On the bottom, the block is initially glued to a fixed flat obstacle and in contact along the full boundary. We are looking for a displacement field inside this body.

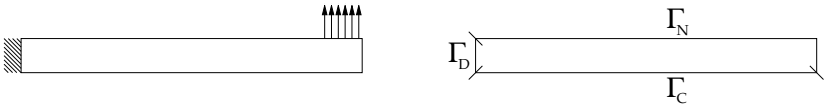


Figure 1.1.: Reference configuration for the 2D benchmark with load distribution; boundary decomposition

The thin layer of adhesive along Γ_C is not modelled geometrically. It may however expose discontinuous adhesion forces, which pull the body back to the obstacle and so make the surfaces stick to each other. These forces are dependent on the gap size, and thus also on the displacement field itself.

Two example adhesion laws are given in Figure 1.3. These laws are to be read as follows: Select an arbitrary material point x on the contact boundary. The abscissa

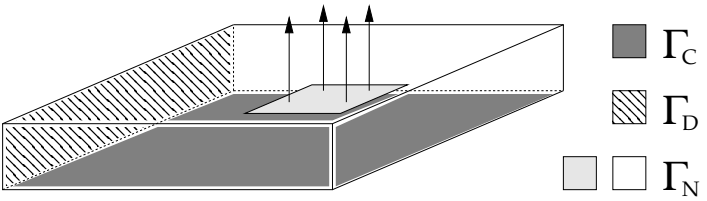


Figure 1.2.: Reference configuration for the 3D benchmark with force distribution

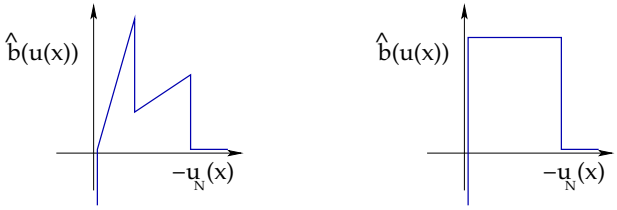


Figure 1.3.: Nonsmooth adhesion laws

describes the negative normal displacement, this is the gap that opens between the obstacle and the body. With a growing gap, a reaction force pulls the material back to the obstacle. That force is given on the ordinate; it may be as simple as in the right graph or more complicated, as in the left graph. At some point, the gap has grown so large that the adhesive is damaged. The adhesion force is now zero.
Note that these laws are set-valued in some points, where a discontinuity gap had to be filled out. A special case appears in the contact point, where the normal displacement is zero: The reaction force here becomes the contact pressure in that material point, which is allowed to grow as large as necessary in the negative direction. The construction of set-valued laws by filling out discontinuity gaps follows a classic procedure when encountering nonsmooth boundary conditions, see e.g. the monograph by Filippov [15] or the earlier work by Rauch [37].

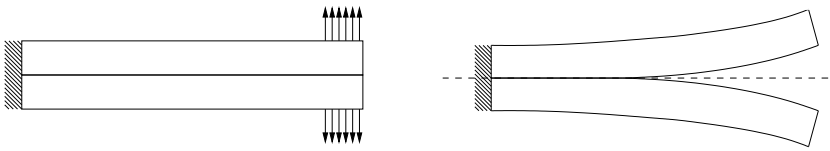


Figure 1.4.: Symmetric delamination problem: Reference configuration; deformed configuration with symmetry axis

This configuration can also be used to describe a delamination benchmark: If two

elastic, bonded blocks are subjected to a given force such that the problem is symmetric along Γ_C, this symmetry can be exploited. We only need to compute a solution for the upper block to retrieve the deformation of two layers of finite thickness, see Figure 1.4.

Structure

This thesis is organized as follows:
In Chapter 2, we repeat some concepts, definitions and notations. We assume that the reader is familiar with basic concepts of the Finite Element method. We introduce generalizations of derivatives that are needed later. Further, the classical problem of linear-elastic deformation is derived from a mechanical point of view. It is then generalized to incorporate contact and adhesion conditions. We state the weak formulation for the Signorini problem. Finally, spaces and operators are shortly introduced that are needed for a boundary integral representation.

Minimization methods have been mainly used as numerical solvers for hemivariational inequalities. The nonsmooth behavior on the boundary however implicates that the constructed objective function is not differentiable anymore. Chapter 3 reviews the approximation of hemivariational inequalities with finite elements and boundary elements.
In a first step, the hemivariational inequalities for contact with adhesion are deduced. Further, an equivalent minimization problem is stated. A basic, one-dimensional example underlines that the objective function is no more differentiable. This example also exposes areas with multiple solutions due to a nonconvex objective function.
Subsequently, we give criteria for the uniqueness of a solution for differentiable and non-differentiable potentials. Theorem 3.10 gives a first criterion for strict convexity of the objective function. The second, more general proposition is Theorem 3.16. Its criteria return uniqueness for non-differentiable potentials.
We state the Bundle-Newton method by Lukšan and Vlček [29], which is widely used to solve hemivariational inequalities.

A different, new ansatz for numerical solvers is the direct treatment of the hemivariational inequality with additional Lagrange multipliers. An active set method has recently been conveyed by Hintermüller, Kovtunenko and Kunisch in [20] to solve a 2D membrane problem under certain conditions. Chapter 4 is dedicated to employ this method for problems in linear elasticity.
The original active set method relies on Hopf's maximum principle and can not be directly applied to problems in elasticity. Instead, we split off the normal displacements on the contact boundary, effectively retreating to an iterative method on three subproblems: Two of these subproblems are linear elastic problems, and the third one reduces to a 2D membrane problem, which can now be treated with the original active set method.

We focus on the solution algorithm and several implementation details, including computational difficulties that arise from the decoupled formulation.

Finally, we demonstrate some numerical experiments in Chapter 5. These experiments introduce a heuristic, residual-based error indicator for Finite Element computations. We present some 2D FE computations and discuss the qualitative behavior of the solutions. Further computations show that FE and BE methods are also applicable for the 3D case.
A last benchmark is performed with the active set iteration from Chapter 4. The method converges monotonically for this benchmark.
We discuss parallelization issues and the implementation efficiency for both the minimization and the active set strategy.

Additional implementation details are given in the appendix. The result on M-matrices in Chapter 4 relies on a definition of the reference element, so the local basis functions and transformations are reestablished.
We present a construction scheme for regular quadrature rules in triangles and tetrahedra; these rules are useful for Finite Element implementations. As the arising integrands for the Boundary Element method are singular, we also introduce adaptive quadrature rules for the 3D implementation on triangles.
The appendix closes with a section on adaptive refinement. We summarize the refinement algorithms that were used for the numerical experiments; further, we specify the necessary element decompositions.

2. Fundamentals

In this chapter, we will give a brief overview of some mathematical constructs that we use. Further, we derive the differential operator for linear elasticity, which allows us to state the model problem for adhesion from the introduction as a system of partial differential equations. We state Korn's inequality, which renders the bilinear form for the FE formulation coercive. Finally, Somigliana's identity is recalled, resulting in the Steklov-Poincaré operator with the appropriate boundary integral operators. This operator will replace the bilinear form $a(\cdot,\cdot)$ in the next chapter if a discretization by boundary elements is chosen instead of a finite element discretization.

In Chapter 4, we will make use of indexed functions. The Einstein summation convention becomes handy:

Remark 2.1: *If an index appears double, a sum over this index is implied. As we deal with two- and three-dimensional objects, the upper sum limit is* 2 *or* 3*, respectively.*
Examples are the matrix-vector multiplication

$$\left(A \cdot \vec{b}\right)_i \;=\; \sum_{j=1}^{3} a_{ij} b_j \;=\; a_{ij} b_j$$

or the trace of a tensor of rank 2

$$\operatorname{tr} \mathbf{b} \;=\; \sum_{i=1}^{3} b_{ii} \;=\; b_{ii}\,.$$

Further, we introduce a shorthand notation for partial derivatives by

$$a_{,i} \;:=\; \frac{\partial a}{\partial x_i}\,.$$

We can now write e.g. the gradient of a vector-valued function and the divergence of a matrix-valued function by

$$(\nabla \mathbf{u})_{ij} \;=\; u_{i,j}\,; \quad (\operatorname{div} \mathbf{b})_i \;=\; b_{ij,j}\,.$$

2.1. Derivatives

One method to treat contact problems with nonsmooth laws is the minimization of a non-differentiable function. For the original problem, this function is defined on a convex subset of a Sobolev space; for a discretized problem, the function is defined on a convex subset of a finite-dimensional subspace, or equivalently on a convex subset of $\mathbb{R}^n$. To deal with function spaces and non-differentiable behaviour, we need to introduce generalized derivatives.

There exists a large variety of generalized derivatives, see e.g. [11] or [21]. The derivatives we will effectively use are the Gâteaux, second order Dini and Clarke differentials. In the following definitions, let $f : X \to R$ be a locally Lipschitz continuous function, where X is a Banach space and $x, v \in X$.

Definition 2.2:
If the limit

$$f^D(x; v) := \lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t} \tag{2.1}$$

exists, it is called the *(upper) Dini-directional derivative* of f in x in the direction v.
If the limit exists for all $v \in X$, we can find an element $DF(x) \in X'$ such that

$$\lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t} = \left\langle DF(x), v \right\rangle \quad \forall v \in X\,. \tag{2.2}$$

$DF(x)$ is then called the *Gâteaux derivative* of f in x.

Definition 2.3:
The limit

$$f^0(x; v) := \limsup_{y \to x;\, t \downarrow 0} \frac{f(y + tv) - f(y)}{t} \tag{2.3}$$

defines the *generalized directional derivative* of f in x in direction v.
The *Clarke derivative* of f in x is defined as

$$\bar{\partial} f(x) := \left\{ y \in X' : f^0(x; v) \geq \langle y, v \rangle \ \forall v \in X \right\}. \tag{2.4}$$

The Clarke derivative is a subset of the dual space X'. For finite-dimensional X and continuously differentiable f, the classical derivative $f'(x)$ will be the only element in $\bar{\partial} f(x)$. As an example, the Clarke derivative of the function $f(x) = |x|$ is $\{-1\}$ for $x < 0$ and $\{+1\}$ for $x > 0$. In the point $x = 0$, one can show that $\bar{\partial} f(0) = [-1, 1]$, which fills the gap between the limits ± 1 (see [11, Example 2.1.3]).

Definition 2.4:
Let $f : X \to \mathbb{R}$ be continuously Gâteaux differentiable. Then the second order (upper) Dini-directional derivative of f in x in the direction v is defined by

$$f^{DD}(x; v) := \limsup_{t \downarrow 0} \frac{\langle DF(x - tv), v\rangle - \langle DF(x), v\rangle}{t} . \tag{2.5}$$

Note that the same direction v is chosen here for the first and second derivative, and that $\langle DF(x - tv), v\rangle$ instead of $\langle DF(x + tv), v\rangle$ is used. This is to keep the notation consistent with Yang [48].

Definition 2.5:
Let $\mathcal{K}$ be a closed convex subset of X. Then the *normal cone* is defined as

$$N_{\mathcal{K}}(x) := \{y \in X' : \langle y, v - x\rangle \leq 0 \ \forall v \in \mathcal{K}\} . \tag{2.6}$$

The subdifferential of the extended real indicator function

$$\chi_{\mathcal{K}}(x) := \begin{cases} 0 , & x \in \mathcal{K} \\ +\infty , & x \notin \mathcal{K} \end{cases} \tag{2.7}$$

is the normal cone of $\mathcal{K}$ in x. It coincides with the Clarke subdifferential $\bar{\partial}\chi_{\mathcal{K}}(x)$ (see [11, Proposition 2.4.12]).

2.2. Mechanical foundations

Static and quasi-static deformations of mechanical bodies are considered throughout this thesis. This section aims to describe the underlying physical laws.

2.2.1. Linear-elastic material law

There exists a large variety of literature describing deformation processes of continua. As we will consider linear-elastic, isotropic small deformations in quasi-stationary processes, the introduction by Landau / Lifschitz [27] may be used. Alternatively, textbooks like Altenbach / Altenbach [3] provide access to the modelling of continua from an engineering point of view.

Denote the body of interest by $\mathcal{B}$. This body may have several configurations, each describing a point in the deformation process. In the case of quasi-static small deformations, we need only use two configurations:

We start from $\mathcal{B}_0$, which describes the *undeformed configuration* without any external loads or obstacle conditions.

The *deformed configuration* is denoted by $\mathcal{B}_{\text{def}}$. We can add a preliminary index t to denote the deformed configuration $\mathcal{B}^t_{\text{def}}$ at some specified point in time. Taking a particle with position $\mathbf{X} \in \mathcal{B}_0$, it will be displaced to a position $\mathbf{x} \in \mathcal{B}^t_{\text{def}}$ by some transformation function $\boldsymbol{\varphi}$:

$$\mathbf{x} \;=\; \mathbf{x}(\mathbf{X}, t) \;=\; \boldsymbol{\varphi}(\mathbf{X}, t)$$

We can equally write the deformation as the displacement $\mathbf{u}$ in each point of $\mathcal{B}$, $\mathbf{x} \;=\; \mathbf{X} + \mathbf{u}(\mathbf{X}, t)$, which finally reduces to

$$\mathbf{x} \;=\; \mathbf{X} + \mathbf{u}(\mathbf{X}) \qquad\qquad \text{or} \qquad\qquad \mathbf{u}(\mathbf{X}) \;=\; \mathbf{x} - \mathbf{X} \quad \forall \mathbf{X} \in \mathcal{B}_0$$

for a quasi-stationary process.

Only continuous deformations with finite displacements are allowed. If the deformation gradient is given by $\mathbf{F} := \partial \mathbf{x} / \partial \mathbf{X}$, the condition $\det \mathbf{F} > 0$ must hold.

The gradient of $\mathbf{u}$ is $\mathbf{H} := \partial \mathbf{u} / \partial \mathbf{X} \;=\; \mathbf{F} - 1$. (Differentiability of $\mathbf{u}$ is needed here, but will be weakened later.) The Green-Lagrange strain tensor is defined as $\mathbf{E} := \frac{1}{2}(\mathbf{H} + \mathbf{H}^\top + \mathbf{H}^\top \mathbf{H})$; for *small deformations*, we can linearize it and retrieve

$$\boldsymbol{\varepsilon}(\mathbf{u}) \approx \mathbf{E} \;\approx\; \frac{1}{2}\Big(\mathbf{H} + \mathbf{H}^\top\Big) \;=\; \frac{1}{2}\Big(\nabla \mathbf{u} + (\nabla \mathbf{u})^\top\Big)\,.$$

From the laws of thermodynamics, we can describe a derivative of the free energy W in an isothermal process:

$$\sigma_{ij} \;=\; \left(\frac{\partial W}{\partial \varepsilon_{ij}}\right)_T$$

The variable $\boldsymbol{\sigma}$ is called the Cauchy stress tensor. For an isotropic, homogeneous body, we can expand W into a series by a set of independent quadratic terms in $\boldsymbol{\varepsilon}$.

$$W \;=\; W_0 + \frac{\lambda}{2}(\operatorname{tr} \boldsymbol{\varepsilon})^2 + \mu(\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon}) \;+\; O(\varepsilon_{ij}^4)$$

Up to higher order terms, we get an explicit representation for $\boldsymbol{\sigma}$:

$$\sigma_{ij} \;=\; \lambda(\operatorname{tr} \boldsymbol{\varepsilon})\delta_{ij} + 2\mu\varepsilon_{ij} \tag{2.8}$$

In the case of linear, anisotropic material behavior, the more general relation

$$\boldsymbol{\sigma} \;=\; \mathbb{C} : \boldsymbol{\varepsilon} \tag{2.9}$$

can be derived with a symmetric fourth order tensor $\mathbb{C}$, the Hooke tensor.

The stress tensor $\boldsymbol{\sigma}$ corresponds to the stress vector $\mathbf{t}$ in the following way: On the surface of an arbitrary test volume $V^t \subset \mathcal{B}^t_{\text{def}}$, there holds

$$\mathbf{t} \;=\; \lim_{\Delta a \to 0} \frac{\Delta \mathbf{f}}{\Delta a} \;=\; \boldsymbol{\sigma} \cdot \mathbf{n}$$

in every point with the surface force $\mathbf{f}$ and area a.

We can now derive the governing equations from the conservation laws of momentum and mass. Let ρ_0 denote the mass density in a point $\mathbf{X} \in \mathcal{B}^0$. Taking test volumes $V^0 \subset \mathcal{B}^0$ and $V^t \subset \mathcal{B}^t_{\text{def}}$ of the reference and deformed body, we first deduce

$$\int_{V^0} (\rho_0 - \rho \det(\mathbf{F}))\, \mathrm{d}V^0 \;=\; 0\,.$$

Mass conservation now yields

$$0 \;=\; \frac{\mathrm{d}m}{\mathrm{d}t} \;=\; \frac{\mathrm{d}}{\mathrm{d}t} \int_{V^t} \rho \, \mathrm{d}V^t \;=\; \int_{V^t} \Big(\dot{\rho} + \rho \operatorname{div}(\dot{\boldsymbol{\varphi}})\Big)\, \mathrm{d}V^t\,. \tag{2.10}$$

The time derivative of the momentum must now balance the sum of external forces. If we assume a volume force $\mathbf{f}_V$ and a boundary traction $\mathbf{t}$ acting on a test volume, we get

$$\begin{aligned} &\int_{V^t} \rho \mathbf{f}_V \, \mathrm{d}V^t \;+\; \int_{\partial V^t} \mathbf{t}\, \mathrm{d}s_t \;=\; \frac{\mathrm{d}}{\mathrm{d}t} \int_{V^t} \rho \dot{\boldsymbol{\varphi}} \, \mathrm{d}V^t \;=\; \frac{\mathrm{d}}{\mathrm{d}t} \int_{V^0} \rho \dot{\boldsymbol{\varphi}} \det(\mathbf{F})\, \mathrm{d}V^0 \\ =\; &\int_{V^0} \rho \ddot{\boldsymbol{\varphi}} \det(\mathbf{F})\, \mathrm{d}V^0 \;+\; \int_{V^0} \Big(\dot{\rho} \dot{\boldsymbol{\varphi}} \det(\mathbf{F}) + \rho \dot{\boldsymbol{\varphi}} \operatorname{div}(\dot{\boldsymbol{\varphi}}) \det(\mathbf{F})\Big)\, \mathrm{d}V^0\,, \end{aligned}$$

and the mass conservation law (2.10) cancels out the last integral.
We finally have on an arbitrary test volume

$$\begin{aligned} \int_{V^t} \rho \ddot{\boldsymbol{\varphi}} \, \mathrm{d}V^t \;&=\; \int_{V^t} \rho \mathbf{f}_V \, \mathrm{d}V^t \;+\; \int_{\partial V^t} \mathbf{t}\, \mathrm{d}s_t\,, \\ \text{or} \quad \int_{V^t} \rho \ddot{\boldsymbol{\varphi}} \, \mathrm{d}V^t \;&=\; \int_{V^t} \Big(\rho \mathbf{f}_V \;+\; \operatorname{div} \boldsymbol{\sigma}\Big)\, \mathrm{d}V^t \end{aligned} \tag{2.11}$$

applying the Gauss theorem.

Using $\ddot{\boldsymbol{\varphi}} = \ddot{\mathbf{u}}$, the local form of momentum conservation is then

$$-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}) + \rho \ddot{\mathbf{u}} \;=\; \rho \mathbf{f}_V\,. \tag{2.12}$$

Additionally, the acceleration term $\ddot{\mathbf{u}}$ may be neglected for a quasi-stationary process (the body $\mathcal{B}$ will be in a motionless equilibrium). The governing equation reduces to

$$-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}) \;=\; \rho \mathbf{f}_V\,. \tag{2.13}$$

Remark 2.6: *The complexity of many problems can be reduced to two dimensions. There are two specific cases of plane elasticity, plane strain and plane stress. Nečas and Hlavaček demonstrate in [33, Section 10.2] how these cases can be derived.*

The stress vector on the surface can be decomposed further: We get the scalar-valued *normal stress* σ_N and the vector-valued *shear stress* $\boldsymbol{\sigma}_T$ by

$$\begin{aligned} \sigma_N \;&:=\; \mathbf{n} \cdot \boldsymbol{\sigma} \cdot \mathbf{n} \\ \boldsymbol{\sigma}_T \;&:=\; \boldsymbol{\sigma} \cdot \mathbf{n} \;-\; \sigma_N \mathbf{n}\,. \end{aligned} \tag{2.14}$$

The classical problem of linear elasticity

We need to give additional boundary conditions to fully describe the classical problem of linear elasticity. Let the body $\mathcal{B}$ occupy a bounded, open domain $\Omega \subset \mathbb{R}^3$ or $\Omega \subset \mathbb{R}^2$. Decompose the boundary $\Gamma = \partial\Omega$ into boundary parts that do not intersect with positive measure. We can give Dirichlet boundary conditions on a boundary part Γ_D and Neumann boundary conditions on Γ_N. A simple problem then reads: Find $\mathbf{u} : \Omega \to \mathbb{R}^3$ such that

$$\begin{aligned} -\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}) &= \mathbf{f} \quad \text{in } \Omega \\ \mathbf{u} &= \mathbf{u}^0 \quad \text{on } \Gamma_D \\ \boldsymbol{\sigma}(\mathbf{u}) \cdot \mathbf{n} &= \mathbf{t}^0 \quad \text{on } \Gamma_N \,. \end{aligned} \tag{2.15}$$

More complex boundary conditions can be prescribed as well. If the deformed body must not penetrate an obstacle, Γ contains a contact boundary Γ_C. In fact, our benchmark example will not only introduce an obstacle on Γ_C, but also an adhesion force on Γ_C. This force function will be dependent on the displacement u_N and is in general not continuous.

2.2.2. Classical contact problem with adhesion

Let the body $\mathcal{B}$ now be close to a rigid obstacle. If $\mathcal{B}$ is deformed, this obstacle can block further displacements. The boundary $\partial\Omega$ now also contains a part Γ_C of potential contact. The penetration of the obstacle can be described to occur from normal displacements on the boundary,

$$u_N(\mathbf{x}) \;:=\; \mathbf{u}(\mathbf{x}) \cdot \mathbf{n}\,, \quad \mathbf{x} \in \Gamma\,.$$

The gap between deformed body and obstacle can be given by a nonlinear relation, see Kikuchi and Oden [23, Section 2.3]. However, they show that for the case of small deformations, the contact condition can be linearized by dropping higher order terms. We can now give a gap function $\psi : \Gamma_C \to \mathbb{R}$, which simply describes the normal distance from the (undeformed) contact boundary to the rigid obstacle. We can now restrict the normal displacement by introducing the condition

$$u_N \le \psi \quad \text{on } \Gamma_C$$

into our equation system, which coincides with the second contact condition in the Signorini problem [23, system (2.31)].

The shear stress $\boldsymbol{\sigma}_T$ is assumed to be zero on Γ_C, as no external forces are applied. The normal stress σ_N needs more attention. First, we state the case without adhesion: When a material point $\mathbf{x} \in \Gamma_C$ is in contact, we can not have a positive normal stress. When a material point is not in contact, the normal stress is zero. This is expressed in the complementarity conditions

$$\sigma_N(\mathbf{x}) \;\le\; 0\,; \qquad u_N(\mathbf{x}) \;\le\; \psi(\mathbf{x})\,; \qquad \sigma_N(\mathbf{x})\Big(u_N(\mathbf{x}) - \psi(\mathbf{x})\Big) \;=\; 0\,. \tag{2.16}$$

Now we can add an adhesion force $\hat{b}(\cdot)$ to the formulation. Here, we assume that this force acts in normal direction. The set-valued function $\bar{b}$ is created as the "envelope" of a piecewise continuous function b:

$$\bar{b}(t) \;:=\; \Big[\liminf_{\tau\to 0} b(t+\tau)\,,\; \limsup_{\tau\to 0} b(t+\tau)\Big]\,,$$

where the limits are to be understood in the L^∞-essential limit sense.

In the case of no contact, we demand that

$$\sigma_N(\mathbf{x}) \;\in\; \bar{b}(u_N\Big(\mathbf{x}\Big))\,.$$

We can extend the function $\bar{b}(\cdot)$ by introducing a dependence on the gap function. We pass the negative normal displacement $-u_N(\mathbf{x})$ as a parameter. If this parameter arrives at the gap function value (i.e., if the material point is in contact), the value interval of $\hat{b}(\cdot)$ is extended to $-\infty$:

$$\hat{b}(t,\mathbf{x}) \;:=\; \begin{cases} \Big(-\infty, \limsup_{\tau\to 0} b(t+\tau)\Big]\,, & t = -\psi(\mathbf{x}) \\ \bar{b}(t)\,, & t > -\psi(\mathbf{x}) \end{cases}$$

We do not need to define this function for $t < -\psi(\mathbf{x})$, as this would correspond to the situation $u_N(\mathbf{x}) > \psi(\mathbf{x})$, which would violate the non-penetration condition. Physically, the function $\hat{b}(t,\mathbf{x})$ now takes the gap size in the point $\mathbf{x} \in \Gamma_C$ as argument t and returns a set of admissible reaction forces in the normal direction.

Chapters 1 and 2 in [23] describe the modelling of the adhesion-free case in more detail. For more details on the modelling with adhesion, see Chapter 2 in [19].

We finally get the boundary value problem for a contact problem with adhesion:
Find $\mathbf{u} : \Omega \to \mathbb{R}^3$ such that

$$\begin{aligned} -\operatorname{div}\boldsymbol{\sigma}(\mathbf{u}) &= \mathbf{f} \quad \text{in } \Omega \\ \mathbf{u} &= \mathbf{u}^0 \quad \text{on } \Gamma_D \\ \boldsymbol{\sigma}(\mathbf{u})\cdot\mathbf{n} &= \mathbf{t}^0 \quad \text{on } \Gamma_N \\ u_N(\mathbf{x}) &\le \psi(\mathbf{x})\,, \\ \boldsymbol{\sigma}_T(\mathbf{u}(\mathbf{x})) &= 0\,, \\ \boldsymbol{\sigma}_N(\mathbf{u}(\mathbf{x})) &\in \hat{b}(-u_N(\mathbf{x}),\,\mathbf{x}) \quad \text{a.e. } \mathbf{x} \in \Gamma_C\,. \end{aligned} \tag{2.17}$$

In the case of adhesion, the body $\mathcal{B}$ is in full contact along Γ_C in the reference configuration $\mathcal{B}_0$, and the deformation process may break up some of this contact. The gap function ψ is then just the zero function. We can remove the dependence of $\hat{b}(t,\mathbf{x})$ on $\mathbf{x}$ and use the function $\hat{b} : [0,\infty) \to 2^{\mathbb{R}}$.

2.3. The Ritz-Galerkin method

If we have a simple problem (2.15) with mixed Neumann and homogeneous Dirichlet boundary conditions, we can readily give a variational formulation. For that, we multiply by a test function and integrate by parts. The test space is chosen to be $V := \left[H^1_D(\Omega)\right]^3$, where the trace of functions is fixed to 0 on the Dirichlet boundary. The space V is equipped with the usual norm $\|\cdot\|_V$ (see [33, p.72]). We then look for a function $\mathbf{u} \in V$ with

$$\int_\Omega \sigma(\mathbf{u}) : \varepsilon(\mathbf{v})\,\mathrm{d}\mathbf{x} \;=\; \int_\Omega \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}\mathbf{x} \;+\; \int_{\Gamma_N} \mathbf{t}\cdot\mathbf{v}\,\mathrm{d}s_x \quad \forall \mathbf{v} \in V\,. \tag{2.18}$$

It is a well-known result that if $\mathbf{u}$ is smooth, the classical and the variational formulation are equivalent, see e.g. [17, Section 3.3].
If we have additional inequality constraints on the boundary displacements, we need to reduce the test space to a convex set

$$\mathcal{K} \;:=\; \left\{\mathbf{v} \in V \;:\; v_N(\mathbf{x}) \le \psi(\mathbf{x}) \text{ a.e. } \mathbf{x} \in \Gamma_C\right\} \subset V\,.$$

We get a variational inequality,

$$\int_\Omega \sigma(\mathbf{u}) : \varepsilon(\mathbf{v}-\mathbf{u})\,\mathrm{d}\mathbf{x} \;\ge\; \int_\Omega \mathbf{f}\cdot(\mathbf{v}-\mathbf{u})\,\mathrm{d}\mathbf{x} \;+\; \int_{\Gamma_N} \mathbf{t}\cdot(\mathbf{v}-\mathbf{u})\,\mathrm{d}s_x \quad \forall \mathbf{v} \in \mathcal{K}\,. \tag{2.19}$$

One important theorem is Korn's inequality. Later sections will also rely on two corollaries we will state here. For the proofs, we refer to Duvaut and Lions [12]; these results can be found there as Theorem 3.1, Theorem 3.3 and Theorem 3.4 in Part III (p. 110ff).
Define the inner product

$$a^0(\mathbf{u},\mathbf{v}) \;:=\; \int_\Omega \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v})\,\mathrm{d}\mathbf{x}\,.$$

Theorem 2.7: Korn's inequality
Let Ω be a bounded open set with regular boundary.
Then there holds the following equivalence relation:

$$c\|\mathbf{v}\|_V \;\le\; \left(a^0(\mathbf{v},\mathbf{v}) \;+\; \|\mathbf{v}\|^2_{L^2}\right)^{\frac{1}{2}} \;\le\; C\|\mathbf{v}\|_V \qquad \forall \mathbf{v} \in V\,, \tag{2.20}$$

where c and C are positive constants that only depend on Ω.

Corollary 2.8:
If the Dirichlet boundary Γ_D has positive measure, $a^0(.,.)$ is elliptic on V, i.e. there exists a $c > 0$ such that

$$a^0(\mathbf{v},\mathbf{v}) \;\ge\; c\|\mathbf{v}\|^2_V \qquad \forall \mathbf{v} \in V\,. \tag{2.21}$$

Corollary 2.9:
If the Dirichlet boundary Γ_D has measure zero, $a^0(.,.)$ induces a semi-norm on V. $a^0(.,.)$ is elliptic on the quotient space $V/\mathbb{R}$.

Remark 2.10: *In the general case, the integrand is not* $\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v})$*, but* $\boldsymbol{\varepsilon}(\mathbf{u}) : \mathbb{C} : \boldsymbol{\varepsilon}(\mathbf{v})$ *with an appropriate Hooke tensor* $\mathbb{C}_{ijkl}$*.*
We demand that

$$\begin{aligned} \mathbb{C}_{ijkl} &= \mathbb{C}_{klij} = \mathbb{C}_{jikl} \\ \mathbb{C}_{ijkl}\varepsilon_{ij}\varepsilon_{kl} &\geq m\varepsilon_{ij}\varepsilon_{ij} \qquad \text{for some } m > 0 \,. \end{aligned} \tag{2.22}$$

Then

$$a(\mathbf{u}, \mathbf{u}) := \int_\Omega \boldsymbol{\varepsilon}(\mathbf{u}) : \mathbb{C} : \boldsymbol{\varepsilon}(\mathbf{u}) \geq m\, a^0(\mathbf{u}, \mathbf{u}) \,,$$

transferring ellipticity properties of $a^0(.,.)$ *to* $a(.,.)$*.*
The isotropic, homogeneous material (λ, μ) *is a special case of this:*

$$\mathbb{C}_{ijkl}(\lambda, \mu) := \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) \,.$$

Condition (2.22) *is easy to check, as*

$$\lambda\varepsilon_{ii}(\mathbf{u})\varepsilon_{kk}(\mathbf{u}) + 2\mu\varepsilon_{ij}(\mathbf{u})\varepsilon_{ij}(\mathbf{u}) \geq 2\mu\varepsilon_{ij}(\mathbf{u})\varepsilon_{ij}(\mathbf{u}) \,.$$

In Chapter 3, we will use a minimization formulation for the mechanical system. The bilinear form $a(.,.)$ is symmetric, so we get the minimization problems:
Find $\mathbf{u} \in V$ (or $\mathbf{u} \in \mathcal{K}$) such that

$$\mathbf{u} = \operatorname*{argmin}_{\mathbf{v} \in V \text{ or } \mathcal{K}} \frac{1}{2} a(\mathbf{v}, \mathbf{v}) - f(\mathbf{v}) \,. \tag{2.23}$$

These problems are equivalent to the problems (2.18) and (2.19), see [23, Theorem 3.9] for a proof.

2.4. Boundary integral representation

Sometimes it is desirable to reduce the given problem to a problem on the boundary only. This is especially the case for vanishing volume forces $\mathbf{f}$: If there is no contribution from the interior of Ω, the problem can be reduced to a description on the boundary only, eliminating the need for a volume mesh. The number of degrees of freedom is drastically reduced in the equation system. Further, problems with unbounded domains can be treated effectively without introducing artificial boundaries.
The tradeoff of this method is that the involved integral operators produce dense Galerkin matrices. The emerging double integrals need special treatment, as the integrands may become singular. See Appendix A.2.2 for viable quadrature rules.

We will not give any theoretical results here. For more rigid derivations and more general situations, see e.g. the monographs by Sauter and Schwab [39] or Hsiao and Wendland [22].

Assume that the domain Ω is a Lipschitz domain. We will use a different test space here that is defined only on the boundary $\Gamma = \partial\Omega$:

$$H^{1/2}(\Gamma) \;:=\; \left\{ v \in L^2(\Gamma) \;:\; \exists v' \in H^1(\Omega)\,,\; \operatorname{tr} v' = v \right\}. \tag{2.24}$$

This is not the usual definition of the Sobolev space $H^{1/2}(\Gamma)$: Sauter and Schwab define $H^{1/2}(\Gamma)$ in [39, Section 2.4] via the norm

$$\left\|\phi\right\|^2_{1/2,\Gamma} \;:=\; \left\|\phi_0\right\|^2_{L^2(\Gamma)} \;+\; \int_\Gamma \int_\Gamma \frac{\left(\phi_0(\mathbf{x}) - \phi_0(\mathbf{y})\right)^2}{\|\mathbf{x}-\mathbf{y}\|^d}\, \mathrm{d}s_y\, \mathrm{d}s_x$$

with $d = 2, 3$ and

$$\phi_0(\mathbf{x}) \;:=\; \sum_{U \in \mathcal{U}} \phi_U(\mathbf{x})\,,$$

where $\mathcal{U}$ is a decomposition of Γ into a finite number of disjoint subsets such that a local coordinate system can be given on every $U \in \mathcal{U}$. Theorem 2.6.8 in [39] ensures that these definitions are equivalent for Lipschitz domains.

Additionally, we introduce Sobolev spaces on boundary parts $\Gamma_0 \subset \Gamma$:

$$\begin{aligned} \tilde{H}^{1/2}(\Gamma_0) \;&:=\; \left\{ v \in H^{1/2}(\Gamma) \;:\; \operatorname{supp} v \subset \overline{\Gamma_0} \right\} \\ H^{1/2}(\Gamma_0) \;&:=\; \left\{ \tilde{v} \;:\; \exists v \in H^{1/2}(\Gamma),\, \tilde{v} = v|_{\Gamma_0} \right\} \end{aligned} \tag{2.25}$$

The difference between these spaces is that $\tilde{H}^{1/2}(\Gamma_0)$ only contains functions that are zero on the boundary of Γ_0, while functions in $H^{1/2}(\Gamma_0)$ may be nonzero on $\partial\Gamma_0$.

We will also need the dual spaces with respect to the $L^2(\Gamma_0)$ scalar product

$$\tilde{H}^{-1/2}(\Gamma_0) \;:=\; \left(H^{1/2}(\Gamma_0)\right)'; \qquad\qquad H^{-1/2}(\Gamma_0) \;:=\; \left(\tilde{H}^{1/2}(\Gamma_0)\right)'. \tag{2.26}$$

Finally, we write a boldface $\mathbf{H}^{\pm 1/2}$ for the product space $\left[H^{\pm 1/2}\right]^3$.

The fundamental solution for the Lamé equation with isotropic, homogeneous material (λ, μ) reads

$$E(\mathbf{x},\mathbf{y}) \;:=\; \frac{1}{8\pi\mu\,(\lambda+2\mu)} \left(\frac{(\lambda+3\mu)\,\delta_{ij}}{\|x-y\|} \;+\; (\lambda+\mu) \frac{(x_i-y_i)(x_j-y_j)}{\|\mathbf{x}-\mathbf{y}\|^3} \right).$$

We can now substitute the volume equation, $-\operatorname{div}\boldsymbol{\sigma}(\mathbf{u}) = \mathbf{f}$, by the equivalent representation formula

$$\mathbf{u}(\mathbf{x}) = \int_\Gamma E(\mathbf{x},\mathbf{y}) \Big(P_y \mathbf{u}(\mathbf{y}) \Big)\, \mathrm{d}s_y \;- \int_\Gamma P_y E(\mathbf{x},\mathbf{y}) \mathbf{u}(\mathbf{y})\, \mathrm{d}s_y \;+ \int_\Omega E(\mathbf{x},\mathbf{y}) \mathbf{f}(\mathbf{y})\, \mathrm{d}\mathbf{y}\,,\; \mathbf{x} \in \Omega_{\text{int}} \tag{2.27}$$

with the traction operator $P_y : \mathbf{u} \mapsto \sigma(\mathbf{u}) \cdot \mathbf{n}_y$. This is the *Somigliana identity*, see e.g. [25, Chapter IV.3] for details. If the volume force $\mathbf{f}$ is zero, the last term vanishes, and we are left with equations on the boundary only.

We now introduce the boundary integral operators

$$\begin{aligned}
\big(V\boldsymbol{\psi}\big)(\mathbf{x}) &:= \int_\Gamma E(\mathbf{x},\mathbf{y})\boldsymbol{\psi}(\mathbf{y})\,\mathrm{d}s_y \\
\big(K\boldsymbol{\varphi}\big)(\mathbf{x}) &:= \int_\Gamma P_y E^\top(\mathbf{x},\mathbf{y})\boldsymbol{\varphi}(\mathbf{y})\,\mathrm{d}s_y \\
\big(K'\boldsymbol{\psi}\big)(\mathbf{x}) &:= P_x \int_\Gamma E(\mathbf{x},\mathbf{y})\boldsymbol{\psi}(\mathbf{y})\,\mathrm{d}s_y \\
\big(W\boldsymbol{\varphi}\big)(\mathbf{x}) &:= -P_x \int_\Gamma P_y E^\top(\mathbf{x},\mathbf{y})\boldsymbol{\varphi}(\mathbf{y})\,\mathrm{d}s_y\,,
\end{aligned} \tag{2.28}$$

which have the mapping properties

$$\begin{aligned}
V &: \mathbf{H}^{-1/2}(\Gamma) \to \mathbf{H}^{1/2}(\Gamma) \\
K &: \mathbf{H}^{1/2}(\Gamma) \to \mathbf{H}^{1/2}(\Gamma) \\
K' &: \mathbf{H}^{-1/2}(\Gamma) \to \mathbf{H}^{-1/2}(\Gamma) \\
W &: \mathbf{H}^{1/2}(\Gamma) \to \mathbf{H}^{-1/2}(\Gamma)\,.
\end{aligned}$$

Passing to the limit $\mathbf{x} \to \Gamma$ in (2.27), we retrieve the *Calderón operator*

$$\begin{pmatrix} \mathbf{u} \\ P_x\mathbf{u} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}I - K & V \\ W & \frac{1}{2}I + K' \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ P_x\mathbf{u} \end{pmatrix}, \tag{2.29}$$

which resolves to the *Steklov-Poincaré operator*

$$\big(P_x\mathbf{u}\big)(\mathbf{x}) \;=\; \big(W\mathbf{u} \;+\; (\tfrac{1}{2}I + K')\,V^{-1}\,(\tfrac{1}{2}I + K)\mathbf{u}\big)(\mathbf{x}) \;=:\; \big(S\mathbf{u}\big)(\mathbf{x})\,. \tag{2.30}$$

S maps to $\mathbf{H}^{-1/2}(\Gamma)$ and induces a symmetric bilinear form on $\mathbf{H}^{1/2}(\Gamma)$. Retreating to an open boundary Γ_0, the bilinear form $\langle S\mathbf{u}, \mathbf{v}\rangle_{\Gamma_0}$ is continuous and elliptic on $\tilde{\mathbf{H}}^{1/2}(\Gamma_0)$.

In our case, Γ is decomposed into a homogeneous Dirichlet boundary Γ_D and a Neumann boundary Γ_N with given data $\mathbf{t}$. The open boundary Γ_0 is then the Neumann boundary Γ_N. We can apply the Galerkin method by multiplication with a test function and integration over Γ_N. Substituting $P_x\mathbf{u}$ by $\mathbf{t}$ on Γ_N gives the variational equation:
Find $\mathbf{u} \in \tilde{\mathbf{H}}^{1/2}(\Gamma_N)$ such that

$$\int_{\Gamma_N} \big(S\mathbf{u}\big)(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})\,\mathrm{d}s_x \;=\; \int_{\Gamma_N} \mathbf{t}(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})\,\mathrm{d}s_x \quad \forall \mathbf{v} \in \tilde{\mathbf{H}}^{1/2}(\Gamma_N)\,. \tag{2.31}$$

3. Hemivariational inequalities

In this chapter, we will state the contact problem with adhesion in terms of hemivariational inequalities. We demonstrate two different discretizations using finite elements and boundary elements, then show how each problem can be rewritten as a nonsmooth minimization problem.

A minimal example demonstrates that a hemivariational inequality may admit multiple solutions. For some cases, we can give general conditions for uniqueness of a solution. These are reported in Section 3.5: Theorem 3.10 links the coercivity constant of the bilinear form and the Lipschitz constant of a surface law to a condition for strict convexity. Theorem 3.16 shows that even if the surface law has a jump, a solution is still unique under certain conditions.

Finally, we recapitulate the Bundle-Newton method, which can be applied to solve the emerging nonsmooth minimization problem.

Remark 3.1: *We only cover nonsmooth boundary conditions here. As can be seen in the work by Haslinger et al. [19], nonsmooth contributions inside the domain could be modeled as well if the mapping* Π *in* (3.2)*, the* L^2 *inner product in* (3.8), (3.9) *and the space for the auxiliary function* ξ *are changed appropriately.*

3.1. Variational formulation

The problem of contact with adhesion was given in the PDE system (2.15). We are looking for solutions in the function space $V = \left[H^1_D(\Omega)\right]^3$ or a convex subset $\mathcal{K} \subset V$. Discarding the contact condition, we arrive at the following variational equation: Find $\mathbf{u} \in V$ such that

$$\int_\Omega \sigma(\mathbf{u}) : \mathbb{C} : \varepsilon(\mathbf{v}) \; - \; \int_{\Gamma_C} \sigma(\mathbf{u}) \cdot \mathbf{n} \cdot \mathbf{v} \, \mathrm{d}s_x \; = \; \int_\Omega \mathbf{f} \cdot \mathbf{v} \; + \; \int_{\Gamma_N} \mathbf{t} \cdot \mathbf{v} \, \mathrm{d}s_x \quad \forall \mathbf{v} \in V \, . \tag{3.1}$$

The second integral still needs to be converted. First, recall from (2.14) that the boundary traction can be linearly decomposed into a tangential and a normal part. As we assume that no friction occurs on the contact boundary, the tangential stress σ_T is zero. The integrand then contains only the normal part,

$$\sigma(\mathbf{u}(\mathbf{x})) \cdot \mathbf{n} \cdot \mathbf{v}(\mathbf{x}) \; = \; \Big(\sigma_T(\mathbf{u}(\mathbf{x})) \; + \; \sigma_N(\mathbf{u}(\mathbf{x}))\mathbf{n}\Big) \cdot \mathbf{v}(\mathbf{x}) \; = \; \sigma_N(\mathbf{u}(\mathbf{x}))\, \mathbf{n} \cdot \mathbf{v}(\mathbf{x}) \quad \text{on } \Gamma_C \, .$$

For the normal stress $\sigma_N(\mathbf{u})$, we introduce an auxiliary function $\xi \in L^2(\Gamma_C)$. If we demand that $\sigma_N(\mathbf{u}(\mathbf{x})) = \xi(\mathbf{x})$, we get from the classical problem that $\xi(\mathbf{x}) \in \hat{b}(-u_N(\mathbf{x}))$.

Additionally, we introduce a mapping Π. In our problem, we only need to select the negative third coefficient:

$$\begin{aligned} \Pi : \mathbb{R}^3 &\to \mathbb{R} \\ (x_1, x_2, x_3) &\mapsto -x_3 \,. \end{aligned} \tag{3.2}$$

Note that the conditions on Π in [19] would admit far general mappings, e.g. for non-local surface laws.

Using the usual shorthand notation $a(\cdot,\cdot)$ for the bilinear form and $L(\cdot)$ for the right hand side linear form, (3.1) is now given as the following hemivariational equation:
Find $(\mathbf{u}, \xi) \in V \times L^2(\Gamma_C)$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) \; + \; \int_{\Gamma_C} \xi(\mathbf{x}) \, (\Pi \mathbf{v})(\mathbf{x}) \, \mathrm{d}s_x \; &= L(\mathbf{v}) && \forall \mathbf{v} \in V \\ \xi(\mathbf{x}) \; &\in \hat{b}\Big((\Pi \mathbf{u})(\mathbf{x})\Big) \quad \text{a.e. } \mathbf{x} \in \Gamma_C \,, \end{aligned} \tag{3.3}$$

using the fact that on Γ_C holds $(\Pi \mathbf{v})(\mathbf{x}) = -v_3(\mathbf{x}) = \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}$.

If the contact condition is present, we can derive a variational inequality formulation on $\mathcal{K}$ by multiplying with a test function and integrating by parts. Like in the unconstrained case, we retrieve an integral over Γ_C with $\boldsymbol{\sigma} \cdot \mathbf{n}$ in the integrand. Again, we introduce the function ξ and get the hemivariational inequality:
Find $(\mathbf{u}, \xi) \in \mathcal{K} \times L^2(\Gamma_C)$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v} - \mathbf{u}) \; + \; \int_{\Gamma_C} \xi(\mathbf{x}) \, (\Pi \mathbf{v} - \Pi \mathbf{u})(\mathbf{x}) \, \mathrm{d}s_x \; &\geq L(\mathbf{v} - \mathbf{u}) && \forall \mathbf{v} \in \mathcal{K} \\ \xi(\mathbf{x}) \; &\in \hat{b}\Big((\Pi \mathbf{u})(\mathbf{x})\Big) \quad \text{a.e. } \mathbf{x} \in \Gamma_C \,. \end{aligned} \tag{3.4}$$

Alternatively, we can use the Steklov-Poincaré operator (2.30). The boundary is now decomposed $\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_C$, and we write Γ_0 for the free boundary $\Gamma_N \cup \Gamma_C$. The problem (2.31) is modified to:
Find $(\mathbf{u}, \xi) \in \tilde{\mathbf{H}}^{1/2}(\Gamma_0) \times L^2(\Gamma_C)$ such that

$$\int_{\Gamma_0} (S\mathbf{u}) \cdot \mathbf{v} \, \mathrm{d}s_x \; - \; \int_{\Gamma_C} (P_x \mathbf{u}) \cdot \mathbf{v} \, \mathrm{d}s_x \; = \; \int_{\Gamma_N} \mathbf{t} \cdot \mathbf{v} \, \mathrm{d}s_x \quad \forall \mathbf{v} \in \tilde{\mathbf{H}}^{1/2}(\Gamma_0) \,. \tag{3.5}$$

Again, we decompose the normal stress $P_x \mathbf{u} = \boldsymbol{\sigma} \cdot \mathbf{n}$ into a tangential and a normal part. Substituting $\Pi P_x \mathbf{u}$ by an auxiliary function ξ again, we get the hemivariational

equation:
Find $(\mathbf{u}, \xi) \in \tilde{\mathbf{H}}^{1/2}(\Gamma_0) \times L^2(\Gamma_C)$ such that

$$\begin{aligned} \int_{\Gamma_0} (S\mathbf{u}) \cdot \mathbf{v} \, \mathrm{d}s_x \; + \; \int_{\Gamma_C} \xi(\mathbf{x}) \, (\Pi\mathbf{v})(\mathbf{x}) \, \mathrm{d}s_x \; &= \; \int_{\Gamma_N} \mathbf{t} \cdot \mathbf{v} \, \mathrm{d}s_x \quad \forall \mathbf{v} \in \tilde{\mathbf{H}}^{1/2}(\Gamma_0) \\ \xi(\mathbf{x}) \; &\in \hat{b}\Big((\Pi\mathbf{u})(\mathbf{x})\Big) \qquad \text{a.e. } \mathbf{x} \in \Gamma_C \, . \end{aligned} \tag{3.6}$$

Finally, we can also state a hemivariational inequality in the case of contact if we choose an appropriate convex set $\mathcal{K}$ incorporating the contact condition.
Find $(\mathbf{u}, \xi) \in \mathcal{K} \times L^2(\Gamma_C)$ such that

$$\begin{aligned} \int_{\Gamma_0} (S\mathbf{u}) \cdot (\mathbf{v} - \mathbf{u}) \, \mathrm{d}s_x \; + \; \int_{\Gamma_C} \xi(\mathbf{x}) \, (\Pi\mathbf{v} - \Pi\mathbf{u})(\mathbf{x}) \, \mathrm{d}s_x \; &\geq \; \int_{\Gamma_N} \mathbf{t} \cdot (\mathbf{v} - \mathbf{u}) \, \mathrm{d}s_x \qquad \forall \mathbf{v} \in \mathcal{K} \\ \xi(\mathbf{x}) \; &\in \hat{b}\Big((\Pi\mathbf{u})(\mathbf{x})\Big) \qquad \text{a.e. } \mathbf{x} \in \Gamma_C \, . \end{aligned} \tag{3.7}$$

The hemivariational equations (3.3), (3.6) and the hemivariational inequalities (3.4), (3.7) are of the same structure: We have a symmetric, coercive bilinear form over a Sobolev space and a continuous linear form on the right side. The L^2 inner product on the left side and the inclusion condition $\xi \in \hat{b}(\Pi\mathbf{u})$ are actually the same.

In general, we can select a Sobolev space V, a symmetric, coercive bilinear form $a : V \times V \to \mathbb{R}$ and a linear form $L : V \to \mathbb{R}$. Then a generic hemivariational equation reads:
Find $(u, \xi) \in V \times L^2(\Gamma_C)$ such that

$$\begin{aligned} a(u, v) \; + \; \langle \xi, \Pi v \rangle_{0,\Gamma_C} \; &= L(v) \qquad \forall v \in V \\ \xi(\mathbf{x}) \; &\in \hat{b}\Big((\Pi u)(\mathbf{x})\Big) \quad \text{a.e. } \mathbf{x} \in \Gamma_C \, . \end{aligned} \tag{3.8}$$

A generic hemivariational inequality is given by:
Find $(u, \xi) \in \mathcal{K} \times L^2(\Gamma_C)$ such that

$$\begin{aligned} a(u, v - u) \; + \; \langle \xi, \Pi(v - u) \rangle_{0,\Gamma_C} \; &\geq L(v - u) \qquad \forall v \in \mathcal{K} \\ \xi(\mathbf{x}) \; &\in \hat{b}\Big((\Pi u)(\mathbf{x})\Big) \quad \text{a.e. } \mathbf{x} \in \Gamma_C \, . \end{aligned} \tag{3.9}$$

If $b(\cdot)$ is bounded, Haslinger et al. [19] proved that (3.8) and (3.9) have at least one solution. The proofs rely on finite-dimensional approximations, see Remark 3.2.

3.2. Approximation with Finite Elements and Boundary Elements

We need finite-dimensional subspaces of V and $\tilde{\mathbf{H}}^{1/2}(\Gamma_0)$ to compute numerical solutions of (3.4) and (3.7). Further we assume that Ω is a polyhedral domain.

The first step in the construction of finite-dimensional subspaces is the decomposition of the domain of interest with a triangulation. For the full domain $\Omega \subset \mathbb{R}^{2;3}$, the triangulation $\mathcal{T}_h(\Omega)$ will consist of triangles or tetrahedra. If only parts of the boundary are discretized, $\mathcal{T}_h(\Gamma)$ will consist of edges or faces.

The space $L^2(\Gamma_C)$ is used in both formulations, so we will approximate it first. Note that we do not need an explicit implementation later: The auxiliary function ξ will be eliminated in the minimization formulation.
The simplest space to use here is the space of piecewise constant functions,

$$\Xi_h := \left\{ \xi \in L^2(\Gamma_C) \,:\, \xi\big|_e \in \mathbb{P}_0 \;\forall e \in \mathcal{T}_h(\Gamma_C) \right\}.$$

3.2.1. Finite Element spaces

With a conforming triangulation $\mathcal{T}_h$ into triangles or tetrahedra T given, we can introduce the standard node-based, piecewise linear and continuous functions. We impose the contact conditions in the nodes of Γ_C only. The Finite Element space V_h and the convex set $\mathcal{K}_h$ of admissible displacements are defined as

$$\begin{aligned} V_h &:= \left\{ \mathbf{v} \in \mathbf{H}^1_D(\Omega) \,:\, \mathbf{v}|_T \in \left[\mathbb{P}_1\right]^3 \;\forall T \in \mathcal{T}_h(\Omega) \right\} \\ \mathcal{K}_h &:= \left\{ \mathbf{v} \in V_h \,:\, v_N(P_i) \;\leq\; \psi(P_i) \;\text{ for all mesh nodes } P_i \in \Gamma_C \right\}. \end{aligned}$$

We have $V_h \subset V$, but not necessarily $\mathcal{K}_h \subset \mathcal{K}$: The obstacle ψ may penetrate a deformed configuration in a part of a boundary edge or face. This problem does not occur in our benchmark configuration, where ψ represents a flat obstacle and full contact is assumed in the undeformed configuration.

The stiffness matrix A can be computed exactly from the basis functions $\boldsymbol{\psi}_i$ of V_h. The load vector $\mathbf{f}$ can also be computed exactly, or at least approximated up to a given precision with numerical quadrature.

3.2.2. Boundary Element spaces

As stressed before, we only need a mesh on the boundary Γ. The elements $E \in \mathcal{T}_h$ of this mesh are edges in the 2D case, or triangles in the 3D case. We will use the space of node-based, piecewise linear and continuous functions on $\Gamma_0 = \Gamma_C \cup \Gamma_N$. The contact conditions are only imposed in the nodes of Γ_C.

$$\begin{aligned} V_h &:= \left\{ \mathbf{v} \in \tilde{\mathbf{H}}^{1/2}_D(\Gamma_0) \,:\, \mathbf{v}|_E \in \left[\mathbb{P}_1\right]^3 \;\forall E \in \mathcal{T}_h \right\} \\ \mathcal{K}_h &:= \left\{ \mathbf{v} \in V_h \,:\, v_N(P_i) \;\leq\; \psi(P_i) \;\text{ for all mesh nodes } P_i \in \Gamma_C \right\}. \end{aligned}$$

In contrast to the previous section, the bilinear form $\langle S\cdot,\cdot\rangle$ can only be computed approximately. This originates from the definition

$$S \;:=\; W + \left(\tfrac{1}{2}I + K'\right) V^{-1} \left(\tfrac{1}{2}I + K\right),$$

which contains the inverse single layer operator $V^{-1} : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$. We can not compute the exact inverse, but only an approximation:
First, set up a space of piecewise constant functions on the full boundary Γ,

$$Y_h \;:=\; \left\{ \mathbf{y} \in \left[L^2(\Gamma)\right]^3 \;:\; \mathbf{y}\big|_E \in \left[\mathbb{P}_0\right]^3 \;\forall E \in \mathcal{T}_h \right\}.$$

We can now compute the individual Galerkin matrices W_h, I_h, K_h, K_h' and V_h for the spaces V_h and Y_h. The matrix V_h is symmetric and positive definite, so we can perform a Cholesky decomposition to invert it:

$$S_h \;:=\; W_h \;+\; \left(\tfrac{1}{2}I_h + K_h'\right) V_h^{-1} \left(\tfrac{1}{2}I_h + K_h\right). \tag{3.10}$$

The matrix S_h is an approximation of the Galerkin matrix originating from the Steklov-Poincaré operator S. Maischak and Stephan show in [31, Lemma 15] that the approximation error converges with optimal rate; see also Chernov [8, Section 1.4]. This implies that condition (3.52) in [19],

$$\begin{aligned} & V_h \ni y_h \rightharpoonup y\,,\; V_h \ni z_h \to z \;\text{ in } V \\ \Rightarrow\quad & a_h(y_h, z_h) \;\to\; a(y,z) \;\text{ and }\; a_h(z_h, y_h) \;\to\; a(z,y)\,, \end{aligned}$$

is fulfilled for $a(\mathbf{u},\mathbf{v}) = \langle S\mathbf{u},\mathbf{v}\rangle$ and $a_h(\mathbf{x},\mathbf{y}) = \mathbf{x}\cdot S_h\mathbf{y}$. Then we can apply Theorem 3.4 and Theorem 3.9 in [19], which claim that the approximated solutions converge to solutions of (3.8) and (3.9).

3.2.3. Algebraic formulation

An important part in the formulation is the matrix Λ, which takes the role of the projection Π in the discrete formulation. Λ maps from a basis of V_h to a basis of Ξ_h. In our case, it descriptively selects the degrees of freedom in V_h on Γ_C that point in normal direction and evaluates them in the edge midpoints. If we have a displacement field $\mathbf{u}_h \in V_h$ given by the coefficient vector $\mathbf{x}$, the vector $\Lambda\mathbf{x}$ will consist of the normal displacements of $\mathbf{u}_h$ in the midpoints $\mathbf{m}_i$ of $E_i \subset \Gamma_C$: With

$$\Xi_h \;=\; \operatorname{span}\left\{\psi_i(\mathbf{x})\right\}, \quad \psi_i(E_j) \;=\; \delta_{ij}\,, \quad \text{and} \quad V_h \;=\; \operatorname{span}\left\{\boldsymbol{\varphi}_j(\mathbf{x})\right\},$$

we get

$$\Lambda_{ij} \;=\; \boldsymbol{\varphi}_j(\mathbf{m}_i)\,. \tag{3.11}$$

Let the volume of a face E_i on Γ_C be given by $|E_i|$. For convenience, define the matrix E by setting the diagonal elements to $|E_i|$. Then the integral over Γ_C is plainly given by

$$\int_{\Gamma_C} \xi_h(\mathbf{x})\,(\Pi \mathbf{u}_h)(\mathbf{x})\,\mathrm{d}s_x \;=\; \sum_{E_i \subset \Gamma_C} |E_i| (\boldsymbol{\xi})_i \Big(\Lambda \mathbf{x}\Big)_i \;=\; \boldsymbol{\xi}^\top E \Lambda \mathbf{x}\,.$$

We obtain a fully discrete version of (3.9) by associating V_h with $\mathbb{R}^n$, the convex set $\mathcal{K}_h$ with $K_h \subset \mathbb{R}^n$, and Ξ_h with $\mathbb{R}^m$:
Find $(\mathbf{x}, \boldsymbol{\xi}) \in K_h \times \mathbb{R}^m$ such that

$$\begin{aligned} \mathbf{x}^\top A(\mathbf{x}-\mathbf{v}) \;+\; \boldsymbol{\xi}^\top E\Lambda(\mathbf{x}-\mathbf{v}) \;&\geq \mathbf{f}^\top(\mathbf{x}-\mathbf{v}) \qquad\qquad \forall \mathbf{v} \in K_h \\ \boldsymbol{\xi}_i \;&\in \hat{b}\Big(\Big(\Lambda \mathbf{x}\Big)_i\Big) \qquad \forall i:\; E_i \subset \Gamma_C\,. \end{aligned} \tag{3.12}$$

The fully discrete version of (3.8) is:
Find $(\mathbf{x}, \boldsymbol{\xi}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$\begin{aligned} A\mathbf{x} \;+\; (E\Lambda)^\top \boldsymbol{\xi} \;&= \mathbf{f} \\ \boldsymbol{\xi}_i \;&\in \hat{b}\Big(\Big(\Lambda \mathbf{x}\Big)_i\Big) \quad \forall i:\; E_i \subset \Gamma_C\,. \end{aligned} \tag{3.13}$$

Remark 3.2: *The hemivariational equation* (3.8) *and the hemivariational inequality* (3.9) *have at least one solution.*

For the equation case, we can employ [19, Theorem 3.4]. Conditions (3.24)-(3.26), (3.52) and (3.53) on the bilinear form $a(\cdot,\cdot)$ and on the linear form $L(\cdot)$ (there denoted as f) are fulfilled: For the finite element setting, $a(\cdot,\cdot)$ can be computed exactly; for the boundary element setting, the approximated Poincaré-Steklov operator S_h and the induced bilinear form $a_h(\cdot,\cdot)$ satisfy (3.52). The linear form $L(\cdot)$ can be computed as exactly as necessary with appropriate quadrature rules. Conditions (3.35), (3.48) and (3.49) on the approximation property of V_h are satisfied e.g. if a triangulation of maximal size h is used for the construction of V_h, choosing the parameters $q = q' = s = 2$. As we assume in our case that b does not depend on $\mathbf{x}$, condition (3.47) is satisfied if b is bounded. This is especially true in the given delamination laws.
For the inequality case, [19, Theorem 3.9] can be used. Again, we choose $q = q' = s = 2$ and see that the given conditions (i)–(vi) are fulfilled, if additionally the closed convex cone $\mathcal{K}$ is approximated well enough by the convex sets $\{\mathcal{K}_h\}$.

3.3. The minimization formulation

One possibility to find a numerical solution to the hemivariational inequality (3.9) is to transform it to an equivalent minimization formulation.
Haslinger et al. introduced a so-called *superpotential* $\mathcal{L}$ in [19]. This potential function is given by

$$\mathcal{L}(\mathbf{v}) \;:=\; \frac{1}{2} a(\mathbf{v},\,\mathbf{v}) \;-\; L(\mathbf{v}) \;+\; \int_{\Gamma_C} \int_0^{(\Pi \mathbf{v})(\mathbf{x})} b(t)\,\mathrm{d}t\,\mathrm{d}s_x\,. \tag{3.14}$$

In the following, we will see that $\mathcal{L}$ is closely connected to the hemivariational inequality.

Remark 3.3: *The potential function $\mathcal{L}$ is not restricted to linear elasticity. Carstensen et al. show in [2] that an elastoplastic problem can be written as a nonsmooth minimization problem. The same technique may be used to set up an elastoplastic contact problem with adhesion. Due to the adhesion term, the function $\mathcal{L}$ is nonsmooth anyway.*

3.3.1. Discrete approximation

As the integral over the boundary Γ_C can, in general, not be computed exactly, an approximation needs to be made.
The discrete formulation imposes the inclusion condition $\xi \in \hat{b}(\cdot)$ only in specified points, namely the element midpoints. If we employ the midpoint quadrature rule for the integral over Γ_C, the approximation is

$$\int_{\Gamma_C} \int_0^{(\Pi \mathbf{v})(\mathbf{x})} b(t)\,\mathrm{d}t \;\approx\; \sum_{E_i \subset \Gamma_C} \big|E_i\big| \int_0^{(\Pi \mathbf{v})(\mathbf{m}_i)} b(t)\,\mathrm{d}t \;=:\; \Psi_h(\mathbf{v})\,, \tag{3.15}$$

where $|E_i|$ is the volume (length or area) of surface element E_i, and $\mathbf{m}_i$ is its barycenter.

Taking the Galerkin matrix A and the load vector $\mathbf{f}$ from the discrete system (3.4) or (3.7), the approximated potential takes the form

$$\mathcal{L}(\mathbf{x}) \;:=\; \frac{1}{2}\mathbf{x}^\top A\mathbf{x} \;-\; \mathbf{f}^\top\mathbf{x} \;+\; \sum_i \big|E_i\big|\, B\big(\big(\Lambda\mathbf{x}\big)_i\big) \tag{3.16}$$

with $B(t) := \int_0^t b(\tau)\,\mathrm{d}\tau$.
Recall from (3.11) that the matrix Λ maps $\mathbf{x}$ to the contact faces $E_i \subset \Gamma_C$: The i-th coefficient of $\Lambda\mathbf{x}$ is the normal displacement in the midpoint of E_i. We can directly compute the gradient of the first two parts as $A\mathbf{x} - \mathbf{f}$, and as these contributions are smooth, Corollary 1 of Proposition 2.3.3 in [11] yields the Clarke derivative

$$\bar{\partial}\mathcal{L}(\mathbf{x}) \;=\; \partial\Big(\frac{1}{2}\mathbf{x}^\top A\mathbf{x}\Big) \;-\; \partial(\mathbf{f}^\top\mathbf{x}) \;+\; \bar{\partial}\Big[\sum_i \big|E_i\big|\, B\big(\big(\Lambda\mathbf{x}\big)_i\big)\Big]\,, \tag{3.17}$$

or in coordinate form

$$\frac{\bar{\partial}}{\partial x_i}\mathcal{L}(\mathbf{x}) \;=\; \Big(A\mathbf{x} - \mathbf{f}\Big)_i \;+\; \Psi_h^0(\mathbf{x},\, \mathbf{e}_i) \tag{3.18}$$

(recall that Ψ_h^0 is the generalized directional derivative from (2.3).)
The individual summands now have the partial Clarke derivative

$$\frac{\bar{\partial}}{\partial x_i}\big|E_j\big|\, B\big(\big(\Lambda\mathbf{x}\big)_j\big) \;=\; \big|E_j\big|\, \frac{\bar{\partial}}{\partial x_i} B\big(\big(\Lambda\mathbf{x}\big)_j\big) \;=\; \big|E_i\big|\, \Lambda_{ij}\, \hat{b}\big(\big(\Lambda\mathbf{x}\big)_j\big)\,,$$

the last equality holds because Λ has full rank ([11, Theorem 2.3.10]).

If at most one summand is nonsmooth in a point $\mathbf{x}$, we can write

$$\frac{\bar{\partial}}{\partial x_i}\mathcal{L}(\mathbf{x}) \;=\; \Big(A\mathbf{x}-\mathbf{f}\Big)_i \;+\; \sum_j \big|E_j\big|\,\Lambda_{ij}\,\hat{b}\Big(\big(\Lambda\mathbf{x}\big)_j\Big)\,;$$

$$\text{otherwise,}\quad \frac{\bar{\partial}}{\partial x_i}\mathcal{L}(\mathbf{x}) \;\subset\; \Big(A\mathbf{x}-\mathbf{f}\Big)_i \;+\; \sum_j \big|E_j\big|\,\Lambda_{ij}\,\hat{b}\Big(\big(\Lambda\mathbf{x}\big)_j\Big)\,.$$

3.3.2. Equivalence for discretized problems

If we assume that Λ is surjective and that one-sided limits $b(\tau^{\pm})$ exist for all $\tau \in \mathbb{R}$, it can be shown that finding a solution $(\mathbf{u}, \boldsymbol{\xi})$ of (3.8) or (3.9) is equivalent to finding a substationary point $\mathbf{u}$ in $\mathbb{R}^N$ (or K_h).

Theorem 3.4: (Haslinger, 1999)
Let $\mathbf{u} \in \mathbb{R}^n$ be a substationary point of $\mathcal{L}$ (i.e. $0 \in \bar{\partial}\mathcal{L}(\mathbf{u})$). Then there exists a $\boldsymbol{\xi} \in \mathbb{R}^m$ such that $\Lambda^\top\boldsymbol{\xi} \in \bar{\partial}\Psi(\mathbf{u})$, and $(\mathbf{u}, \boldsymbol{\xi})$ solves the discrete hemivariational equation (3.3) or (3.6), respectively.
Conversely, let $(\mathbf{u}, \boldsymbol{\xi})$ be a solution of (3.3) or (3.6). Let further Λ be surjective, and assume that one-sided limits $b(\tau^{\pm})$ exist for all $\tau \in \mathbb{R}$. Then $\mathbf{u}$ is a substationary point of $\mathcal{L}$, and $\Lambda^\top\boldsymbol{\xi} \in \bar{\partial}\Psi(\mathbf{u})$.

Proof. See [19], Theorem 3.5 (p.135 ff) and Theorem 3.6 (p.137). As the proofs do not depend on the actual construction of A, they can also be applied to the boundary element formulation.
The original proof of Theorem 3.6 only demands that P is surjective, where $\Lambda = P\Pi$; however, this condition is only used to show that Λ is also onto. ⌟

Theorem 3.5: (Haslinger, 1999)
Let $\mathbf{u} \in \mathbb{R}^n$ be a substationary point of $\mathcal{L}$ on K_h (i.e. $0 \in \bar{\partial}\mathcal{L}(\mathbf{u}) + N_{K_h}(\mathbf{u})$). Then there exists a $\boldsymbol{\xi} \in \mathbb{R}^m$ such that $\Lambda^\top\boldsymbol{\xi} \in \bar{\partial}\Psi(\mathbf{u})$, and $(\mathbf{u}, \boldsymbol{\xi})$ solves the discrete hemivariational inequality (3.4) or (3.7), respectively.
Conversely, let $(\mathbf{u}, \boldsymbol{\xi})$ be a solution of (3.4) or (3.7). Let further Λ be surjective, and assume that one-sided limits $b(\tau^{\pm})$ exist for all $\tau \in \mathbb{R}$. Then $\mathbf{u}$ is a substationary point of $\mathcal{L}$, and $\Lambda^\top\boldsymbol{\xi} \in \bar{\partial}\Psi(\mathbf{u})$.

Proof. See [19], Theorem 3.10 (p.147). ⌟

3.4. A 1*d* example

In this section, we will use the Heaviside function Θ and its envelope $\hat{\Theta}$:

$$\Theta(t) := \begin{cases} 0, & t < 0 \\ 1, & t > 0 \end{cases} \qquad \hat{\Theta}(t) := \begin{cases} \{0\}, & t < 0 \\ [0,1], & t = 0 \\ \{1\}, & t > 0 \end{cases}$$

Consider the following ordinary differential equation with a nonsmooth boundary condition:
Find $u \in C^2(0,1)$ such that

$$\left[\begin{array}{rll} -u''(x) &= f(x), & x \in (0,1) \\ u(0) &= 0 & \\ u'(1) &\in -c_J \hat{\Theta}(u(1)) & \end{array}\right. \tag{P^{1D}}$$

with a constant $c_J \in \mathbb{R}$ setting the jump size in the nonsmooth boundary law. The associated hemivariational equation reads:
Find $u \in H^1_{(0)}(0,1)$ such that

$$\left[\begin{array}{rl} \displaystyle\int_0^1 u'(x)v'(x)\,\mathrm{d}x + \xi \cdot v(1) &= \displaystyle\int_0^1 f(x)v(x)\,\mathrm{d}x \qquad \forall v \in H^1_{(0)}(0,1) \\ & \xi \in c_J \hat{\Theta}(u(1)) \end{array}\right.$$

with $H^1_{(0)}(0,1) := \{v \in H^1(0,1) : v(0) = 0\}$. Note that Γ_C here only consists of the point $x = 1$.
We can now compute the minimization potential:

$$\mathcal{L}(v) := \frac{1}{2}\int_0^1 \left(v'(x)\right)^2 \mathrm{d}x - \int_0^1 f(x)v(x)\,\mathrm{d}x + \int_0^{v(1)} c_J\,\Theta(t)\,\mathrm{d}t\,.$$

If we partition the interval $(0,1)$ uniformly with stepwidth h and create piecewise linear basis functions on this partition, we get the finite element space $V_h \subset H^1_{(0)}$ with basis functions $\phi_1(x), \ldots, \phi_N(x)$. Then the stiffness matrix A, the load vector $\mathbf{f}$ and the matrix Π_h which selects the function value in $x = 1$ can be computed explicitly. (The matrix entry a_{NN} equals $1/h$, not $2/h$, because the boundary at $x = 1$ is not fixed, as opposed to the homogeneous boundary $x = 0$.)

$$A := \frac{1}{h}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{N\times N}, \qquad \mathbf{f} := \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \\ \vdots \\ f_N \end{pmatrix} \in \mathbb{R}^N,$$

$$\Pi_h := \begin{pmatrix} 0 & \ldots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{1\times N}.$$

The discrete minimization problem then reads:

$$\text{Find } \mathbf{u} \in \mathbb{R}^N \text{ s.t.} \quad \mathbf{u} = \arg\min_{\mathbf{v}\in\mathbb{R}^N} \mathcal{L}(\mathbf{v})$$

$$\text{with} \quad \mathcal{L}(\mathbf{v}) := \frac{1}{2}\mathbf{v}^\top A\mathbf{v} - \mathbf{f}^\top\mathbf{v} + c_J\Big(\Pi_h\mathbf{v}\Big)_+ .$$

As there is only one nonsmooth summand, the subdifferential of $\mathcal{L}$ can be explicitly given as

$$\bar\partial\mathcal{L}(\mathbf{v}) = A\mathbf{v} - \mathbf{f} + c_J\,\hat\Theta(v_N)\,\mathbf{e}_N\,, \tag{3.19}$$

where $\mathbf{e}_N$ is the N-th unit vector.

Now, we can characterize the stationary points. First recall that the entry $(A^{-1})_{N,N}$ is positive: The matrix A is positive definite, so its inverse A^{-1} is also positive definite. But then

$$(A^{-1})_{N,N} = \mathbf{e}_N^\top A^{-1}\mathbf{e}_N > 0 \quad\Rightarrow\quad \frac{1}{(A^{-1})_{N,N}} > 0\,.$$

We now impose the necessary condition that the first $N-1$ coefficients of the subgradient $\bar\partial\mathcal{L}(\mathbf{u})$ in (3.19) must be zero for a stationary point $\mathbf{u}$. This leaves a condition only on the last coefficient. With this, each stationary point can be parametrized as

$$\mathbf{u} = A^{-1}\Big(\mathbf{f} + t\mathbf{e}_N\Big), \quad t\in\mathbb{R}\,.$$

Most cases admit only one solution:

- If $c_J = 0$, we get the single solution $\mathbf{u} = A^{-1}\mathbf{f}$.
- If $c_J > 0$, we can only reach one stationary point. The solution is

$$\mathbf{u} = \begin{cases} A^{-1}\mathbf{f} & \text{for } \dfrac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}} < 0\,; \\ A^{-1}\Big(\mathbf{f} - \dfrac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}}\mathbf{e}_N\Big) & \text{for } \dfrac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}} \in [0, c_J]\,; \\ A^{-1}\Big(\mathbf{f} - c_J\mathbf{e}_N\Big) & \text{for } \dfrac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}} > c_J\,. \end{cases}$$

- If $c_J < 0$ and $\frac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}} \notin [c_J, 0]$, we have only one stationary point:

$$\mathbf{u} = \begin{cases} A^{-1}\mathbf{f} & \text{for } \dfrac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}} < c_J\,; \\ A^{-1}\Big(\mathbf{f} - c_J\mathbf{e}_N\Big) & \text{for } \dfrac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}} > 0\,. \end{cases}$$

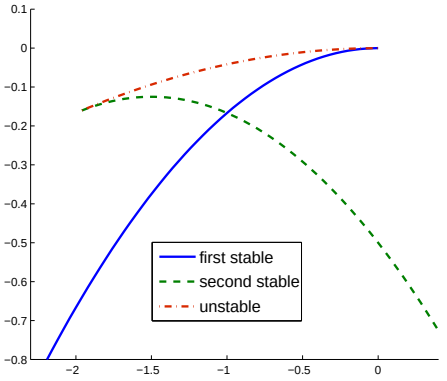


Figure 3.1.: Objective function values for stable and unstable solutions of (P^{1D}) for varying load $\mathbf{f} \in (-2.5, 0.5)$

Only in the case $c_J < 0$ and $\frac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}} \in (c_J, 0)$, we have three distinct stationary points:

$$\begin{aligned}
\mathbf{u}_{\min,1} &= A^{-1}\mathbf{f} \quad \text{is a local minimum;} \\
\mathbf{u}_{\min,2} &= A^{-1}\Big(\mathbf{f} - c_J\mathbf{e}_N\Big) \quad \text{is a local minimum;} \\
\mathbf{u}_{\max} &= A^{-1}\Big(\mathbf{f} - \frac{(A^{-1}\mathbf{f})_N}{(A^{-1})_{N,N}}\mathbf{e}_N\Big) \quad \text{is a local maximum.}
\end{aligned}$$

Corollary 3.6:
For A positive definite and $c_J < 0$, a potential function $\mathcal{L}$ of the form (3.16) may allow multiple distinct solutions. Then $\mathcal{L}$ can not be convex.

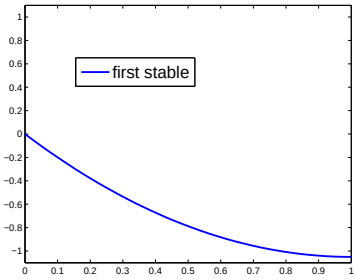
In the stationary points, the potential function evaluates as

$$\begin{aligned}
\mathcal{L}(\mathbf{u}_{\min,1}) &= -\frac{1}{2}\mathbf{f}^\top A^{-1}\mathbf{f}, \\
\mathcal{L}(\mathbf{u}_{\min,2}) &= -\frac{1}{2}\mathbf{f}^\top A^{-1}\mathbf{f} - \frac{1}{2}c_J^2(A^{-1})_{N,N} + c_J(A^{-1}\mathbf{f})_N \\
\mathcal{L}(\mathbf{u}_{\max}) &= -\frac{1}{2}\mathbf{f}^\top A^{-1}\mathbf{f} + \frac{1}{2}\frac{(A^{-1}\mathbf{f})_N{}^2}{(A^{-1})_{N,N}}.
\end{aligned}$$

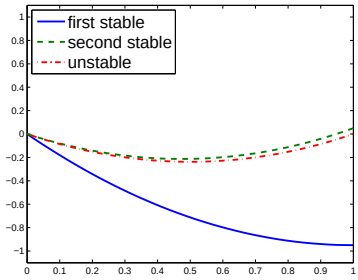
Now $\mathbf{u}_{\max}$ is a convex combination of $\mathbf{u}_{\min,1}$ and $\mathbf{u}_{\min,2}$, but $\mathcal{L}(\mathbf{u}_{\max}) > \mathcal{L}(\mathbf{u}_{\min,1})$ and $\mathcal{L}(\mathbf{u}_{\max}) > \mathcal{L}(\mathbf{u}_{\min,2})$. This means that $\mathcal{L}$ is not even quasiconvex here.

We performed computations for the case with multiple solutions, i.e. $c_J = -1$, for various constant loads $\mathbf{f}$.

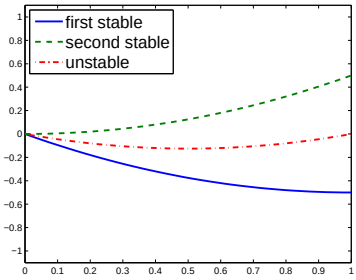
Figure 3.1 shows the extremal values of $\mathcal{L}$ with varying $\mathbf{f}$. We have a first stable solution $\mathbf{u}_{\min,1}$ for $\mathbf{f} \in (-\infty, 0]$ and a second stable solution $\mathbf{u}_{\min,2}$ for $\mathbf{f} \in [-2, \infty)$. The unstable solution $\mathbf{u}_{\max}$ is attained for $\mathbf{f} \in [-2, 0]$.
We show the solutions during the progress of increasing $\mathbf{f}$ in Figure 3.2. For $\mathbf{f} \leq -2$, there exists only one solution. This first stable solution fulfills the boundary condition $u'_{\min,1}(1) = 0$.
For $\mathbf{f} \in (-2, 0)$, we have three distinct stationary points of $\mathcal{L}$: Additionally to the first stable solution, we get a second stable solution that fulfills $u'_{\min,2}(1) = 1 = -c_J$. The unstable solution, which corresponds to a local maximum, attains a derivative between 0 and $-c_J$ in $x = 1$. This is allowed by the original problem, as $u_{\max}(1) = 0$ for $\mathbf{f} \in (-2, 0)$: For that case, we only demanded that $u'(1) \in [0, 1]$ through the Heaviside envelope function $\hat{\Theta}(t)$.
For $\mathbf{f} \geq 0$, only the second stable solution remains a stationary point of $\mathcal{L}$.



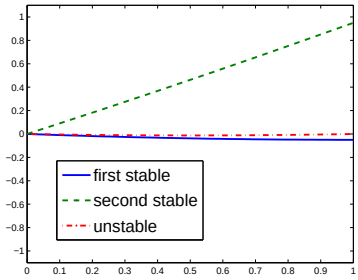
(a) **f** = −2.1



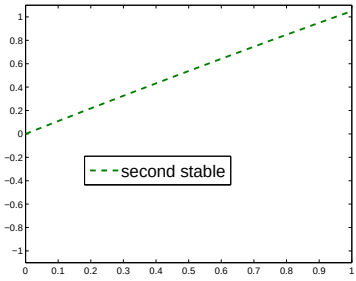
(b) **f** = −1.9



(c) **f** = −1



(d) **f** = −0.1



(e) **f** = +0.1

Figure 3.2.: Solutions of (P^{1D}) for different right hand sides with $c_J = -1$

3.5. Uniqueness results

This section is dedicated to conditions on the boundary law b such that the hemivariational inequality has a unique solution. Its main results are Theorem 3.10 and Theorem 3.16, which give conditions for the uniqueness of a solution. If b includes no jumps or only positive jumps, the conditions are global; for negative jumps, uniqueness can only be validated *a posteriori* when the solution is known.

Unique solvability of a problem is of fundamental importance for the numerical treatment: Only if a solution $\mathbf{u}$ can be uniquely determined, error measures of an approximate solution $\mathbf{u}^h$ can reasonably be given in terms of some norm. On the other hand, the original problem was re-cast into minimizing a potential in Section 3.3. An algorithm will look for stationary points, and uniqueness will guarantee that the algorithm will not converge to a local maximum or to a minimum that is not global.
Note however that due to the non-monotone contribution from b, we must in general expect multiple solutions of the original problem and its approximation.

3.5.1. Convexity

A first attempt will be to impose conditions on the problem such that the resulting potential function is strictly convex.
A strictly convex function with an additional growth condition has a unique minimizer:

Lemma 3.7:
Let X be a normed $\mathbb{R}$-vector space. Let $f : X \to \mathbb{R} \cup \{+\infty\}$ be strictly convex and lower semicontinuous, with $f(x) \to \infty$ as $\|x\| \to \infty$, $x \in X$.
Then f has a unique minimizer.

Proof. For a proof of existence, see Niculescu and Persson [34], Theorem C.1.1.

Assume that $x \in X$, $y \in X$, $x \neq y$, are two separate minimizers of f.
Define the midpoint of x and y, $m := \frac{1}{2}(x + y)$. We have $f(x) \leq f(m)$ and $f(y) \leq f(m)$. But then

$$\frac{1}{2}f(x) + \frac{1}{2}f(y) \;\leq\; \frac{1}{2}f(m) + \frac{1}{2}f(m) \;=\; f(m) \;<\; \frac{1}{2}f(x) + \frac{1}{2}f(y) \quad \text{by strict convexity.}$$

This is a contradiction, so any minimizer must be unique. □

Remark 3.8: *The growth condition, $f(x) \to \infty$ as $\|x\| \to \infty$, is necessary:*
For example, the exponential function is strictly convex, as

$$\left(e^x\right)'' \;=\; e^x \;>\; 0 \quad \forall x \in \mathbb{R}\,,$$

but it does not attain a minimum.

Lemma 3.9:
Let $f : X \to \mathbb{R}$ have a second-order upper Dini-directional derivative in every $z \in X$.
Let further $f^{DD}(z; d) < 0$ for all $0 \neq d \in X$.
Then f is strictly convex.

Proof. Theorem 2.1 of [48] states that

$$\forall x, y \in X\,, x \neq y \quad \exists z \in (x, y) \;:\quad f(y) \;\geq\; f(x) \;+\; \left\langle Df(x), y - x \right\rangle \;-\; \frac{1}{2} f^{DD}(z;\, y - x)\,,$$

where $z \in (x, y)$ means that $z = x + t(y - x)$ for some $t \in (0, 1)$.
As $f^{DD}(z;\, y - x) < 0$ holds, we get

$$\forall x, y \in X\,, x \neq y \;:\quad f(y) \;-\; f(x) \;>\; \left\langle Df(x), y - x \right\rangle .$$

This is just the first-order condition for strict convexity. □

We can now apply this generalized second-order condition to our potential function:

Theorem 3.10:
Assume that $b \in \mathrm{Lip}(\mathbb{R})$ with the Lipschitz constant c_L.
If this constant satisfies

$$c_L \;<\; \inf_{\substack{\mathbf{d} \in V\,: \\ \|\Pi \mathbf{d}\|_{0,\Gamma_C} = 1}} a(\mathbf{d}, \mathbf{d})\,, \tag{3.20}$$

the hemivariational inequality (3.9) has a unique solution.

Proof. First, there exists a positive constant ε such that

$$c_L + \varepsilon \;\leq\; \inf_{\substack{\mathbf{d} \in V\,: \\ \|\Pi \mathbf{d}\|_{0,\Gamma_C} = 1}} a(\mathbf{d}, \mathbf{d})\,.$$

As Π is a linear operator, the right-hand side can now be rewritten as a generalized Rayleigh quotient to have a condition on the full space V:

$$\begin{aligned} (c_L + \varepsilon) \quad &\leq \quad \inf_{\substack{\mathbf{d} \in V\,: \\ \mathbf{d} \notin \ker \Pi}} \frac{a(\mathbf{d}, \mathbf{d})}{\langle \Pi \mathbf{d}, \Pi \mathbf{d} \rangle_{\Gamma_C}} \\ \Rightarrow \quad a(\mathbf{d}, \mathbf{d}) \quad &\geq \quad (c_L + \varepsilon) \langle \Pi \mathbf{d}, \Pi \mathbf{d} \rangle_{\Gamma_C} \quad \forall \mathbf{d} \in V\,. \end{aligned}$$

The second inequality holds trivially for $\mathbf{d} \in \ker \Pi$, as the left side is non-negative and the right side is zero.

The following argument needs the Gâteaux derivative of $\mathcal{L}$. Writing $B(t) := \int_0^t b(\tau)\,\mathrm{d}\tau$, this is by definition

$$\begin{aligned}
&\Big(D\mathcal{L}(\mathbf{u}), \mathbf{d}\Big) \\
&= \lim_{t\downarrow 0} \frac{1}{t}\left(\frac{1}{2}a(\mathbf{u}+t\mathbf{d}, \mathbf{u}+t\mathbf{d}) - L(\mathbf{u}+t\mathbf{d}) + \int_{\Gamma_C} B(\Pi\mathbf{u} + t\Pi\mathbf{d})\,\mathrm{d}s_x - \mathcal{L}(\mathbf{u})\right) \\
&= a(\mathbf{u}, \mathbf{d}) - L(\mathbf{d}) + \int_{\Gamma_C} \lim_{t\downarrow 0} \frac{1}{t}\Big[B\big(\big(\Pi\mathbf{u}\big)(\mathbf{x}) + t\,\big(\Pi\mathbf{d}\big)(\mathbf{x})\big) - B\big(\big(\Pi\mathbf{u}\big)(\mathbf{x})\big)\Big]\,\mathrm{d}s_x\,.
\end{aligned}$$

Because $B(t)$ is continuously differentiable and its derivative is $b(t)$, we can write the integrand as

$$\lim_{t\downarrow 0} \frac{1}{t}\Big(B\big(\big(\Pi\mathbf{u}\big)(\mathbf{x}) + t\,\big(\Pi\mathbf{d}\big)(\mathbf{x})\big) - B\big(\big(\Pi\mathbf{u}\big)(\mathbf{x})\big)\Big) = \big(\Pi\mathbf{d}\big)(\mathbf{x})\; b\big(\big(\Pi\mathbf{u}\big)(\mathbf{x})\big)\,,$$

so the Gâteaux derivative of $\mathcal{L}$ in $\mathbf{d}$-direction is

$$\Big(D\mathcal{L}(\mathbf{u}), \mathbf{d}\Big) = a(\mathbf{u}, \mathbf{d}) - L(\mathbf{d}) + \Big\langle b(\Pi\mathbf{u}),\, \Pi\mathbf{d}\Big\rangle_{\Gamma_C}\,.$$

By assumption, there holds

$$a(\mathbf{d},\, \mathbf{d}) \geq c_L\Big\langle \Pi\mathbf{d}, \Pi\mathbf{d}\Big\rangle_{\Gamma_C} + \varepsilon\|\Pi\mathbf{d}\|^2_{0,\Gamma_C} = \int_{\Gamma_C} c_L\Big((\Pi\mathbf{d})(\mathbf{x})\Big)^2\,\mathrm{d}s_x + \varepsilon\|\Pi\mathbf{d}\|^2_{0,\Gamma_C}\,.$$

We can estimate the integrand as follows: We introduce an arbitrary element $\mathbf{u} \in V$ and $t > 0$. Write $(\Pi\mathbf{d})(\mathbf{x}) =: d_x$, $\delta := -td_x$ and $(\Pi\mathbf{u})(\mathbf{x}) =: u_x$. The Lipschitz continuity of b states that

$$c_L\delta^2 = c_L\,|\delta|\,|\delta| \geq |\delta|\,|b(u_x+\delta) - b(u_x)|\,.$$

If the arguments of both factors on the right side have the same sign (for example $\delta < 0$ and $b(u_x+\delta) - b(u_x) < 0$), we need not take the absolute values. On the other hand, if the arguments have different signs, the right side evaluates to a non-positive expression, and we retrieve

$$c_L\delta^2 \geq 0 \geq \delta\Big(b(u_x+\delta) - b(u_x)\Big)\,.$$

In both cases, there holds

$$c_L\delta^2 \geq \delta\Big(b(u_x+\delta) - b(u_x)\Big)\,.$$

Re-substituting δ by $-td_x$ and multiplying by t^{-2} on both sides, there holds for the integrand

$$c_L d_x^2 \geq \frac{d_x}{t}\Big(b(u_x - td_x) - b(u_x)\Big)\,,$$

and we get

$$a(\mathbf{d},\,\mathbf{d}) \;\geq \left\langle \Pi\mathbf{d},\; \frac{1}{t}\Big(b(\Pi\mathbf{u} - t\Pi\mathbf{d}) - b(\Pi\mathbf{u})\Big)\right\rangle_{\Gamma_C} \;+\; \varepsilon\|\Pi\mathbf{d}\|^2_{\Gamma_C}\,.$$

But now, we have

$$\begin{aligned}
0 \;&\geq\; \frac{1}{t}\left(-t\,a(\mathbf{d},\mathbf{d}) \;+\; \Big\langle b(\Pi\mathbf{u} - t\Pi\mathbf{d}) \;-\; b(\Pi\mathbf{u}),\, \Pi\mathbf{d}\Big\rangle_{\Gamma_C}\right) \;+\; \varepsilon\|\Pi\mathbf{d}\|^2_{0,\Gamma_C} \\
&=\; \frac{1}{t}\left(\; a(\mathbf{u} - t\mathbf{d},\mathbf{d}) \;+\; \Big\langle b(\Pi(\mathbf{u} - t\mathbf{d})), \Pi\mathbf{d}\Big\rangle_{\Gamma_C} \;-\; L(\mathbf{d})\right) \\
&\quad + \frac{1}{t}\left(- a(\mathbf{u},\mathbf{d}) \;-\; \Big\langle b(\Pi\mathbf{u}),\, \Pi\mathbf{d}\Big\rangle_{\Gamma_C} \;+\; L(\mathbf{d})\right) \\
&\quad + \varepsilon\|\Pi\mathbf{d}\|^2_{0,\Gamma_C} \\
&=\; \frac{1}{t}\Big((D\mathcal{L}(\mathbf{u} - t\mathbf{d}),\mathbf{d})_V \;-\; (D\mathcal{L}(\mathbf{u}),\mathbf{d})_V\Big) \;+\; \varepsilon\|\Pi\mathbf{d}\|^2_{0,\Gamma_C}\,.
\end{aligned}$$

Taking the limit $t \to 0$ on both sides, we arrive at

$$0 \;\geq\; \limsup_{t\downarrow 0} \frac{(D\mathcal{L}(\mathbf{u} - t\mathbf{d}),\mathbf{d})_V \;-\; (D\mathcal{L}(\mathbf{u}),\mathbf{d})_V}{t} \;+\; \varepsilon\|\Pi\mathbf{d}\|^2_{0,\Gamma_C}\,.$$

For strict convexity, we need to consider two cases:

- $\mathbf{d} \in \ker\Pi$: We have $\langle \mathbf{v}, \Pi\mathbf{d}\rangle_{\Gamma_C} = 0$ for an arbitrary $\mathbf{v}$.
  But then
  $$\limsup_{t\downarrow 0} \frac{(D\mathcal{L}(\mathbf{u} - t\mathbf{d}),\mathbf{d})_V \;-\; (D\mathcal{L}(\mathbf{u}),\mathbf{d})_V}{t} \;=\; -a(\mathbf{d},\mathbf{d}) \;<\; 0 \;\text{ for }\; \mathbf{d} \neq 0\,.$$

- $\mathbf{d} \notin \ker\Pi$: But then, we have
  $$\begin{aligned}
  &\limsup_{t\downarrow 0} \frac{(D\mathcal{L}(\mathbf{u} - t\mathbf{d}),\mathbf{d})_V \;-\; (D\mathcal{L}(\mathbf{u}),\mathbf{d})_V}{t} \\
  <\;& \limsup_{t\downarrow 0} \frac{(D\mathcal{L}(\mathbf{u} - t\mathbf{d}),\mathbf{d})_V \;-\; (D\mathcal{L}(\mathbf{u}),\mathbf{d})_V}{t} \;+\; \varepsilon\|\Pi\mathbf{d}\|^2_{0,\Gamma_C} \;\leq\; 0\,.
  \end{aligned}$$

In both cases, the second-order Dini derivative is negative, so $\mathcal{L}$ is strictly convex. Finally applying Lemma 3.7, we get the existence of a unique minimizer of $\mathcal{L}$. □

Remark 3.11: *If we assume that b is Lipschitz continuous, the discrete approximated potential function $\mathcal{L}$ in* (3.16) *is globally differentiable and has locally Lipschitz continuous derivatives: The quadratic part of $\mathcal{L}$ is smooth, so we only need to consider the last term in* (3.17)*. Because b is Lipschitz continuous, its anti-derivative B is globally differentiable.*

The space $C^{1,1}(X,\mathbb{R})$ of differentiable functions with locally Lipschitz gradients has been studied extensively, and the alternative second-order conditions in [49] may give additional uniqueness results. Further assumptions on b would lead to semismooth functions, for which specialized Newton methods exist [35].

3.5.2. Uniqueness for a C^0 potential

The second-order condition in Theorem 3.10 required that the function b is Lipschitz continuous. This confines the potential function $\mathcal{L}$ to $C^1(X, \mathbb{R})$, but more general surface laws would generate only $\mathcal{L} \in C^0(X, \mathbb{R})$.
Our next step is to introduce a discontinuity into b. We shall see that a positive jump will remain with the same condition for uniqueness as before, but a negative jump might lead to multiple solutions.

Note that integrals of the Heaviside function Θ can be expressed by

$$\int_a^b \Theta(t)\,\mathrm{d}t \;=\; \Theta(b)\,b \;-\; \Theta(a)\,a \;=\; (b)_+ \;-\; (a)_+ \quad . \tag{3.21}$$

Aim of the following auxiliary results is to extend the Lipschitz function $b(t)$ by adding $c_J\,\Theta(t - t^*)$, introducing a jump of size c_J in the point t^*.

Lemma 3.12:
Let $b(t) := b_{\mathrm{Lip}}(t) + c_J\,\Theta(t - t^*)$, where $t^* \in \mathbb{R}$ is fixed, $c_J \in \mathbb{R}$, and $b_{\mathrm{Lip}} \in \mathrm{Lip}(\mathbb{R})$ with Lipschitz constant c_L. Let $B(t) := \int_0^t b(\tau)\,\mathrm{d}\tau$.
Then

$$B(t+\delta) - B(t) - b(t)\,\delta \quad\leq\quad \frac{c_L}{2}\delta^2 \;+\; c_J\,(\Theta(t - t^* + \delta) - \Theta(t - t^*))\,(t - t^* + \delta)$$

and

$$B(t+\delta) - B(t) - b(t)\,\delta \quad\geq\quad -\frac{c_L}{2}\delta^2 \;+\; c_J\,(\Theta(t - t^* + \delta) - \Theta(t - t^*))\,(t - t^* + \delta)$$

for all $\delta \in \mathbb{R}$.

Proof. The first inequality follows from

$$\begin{aligned}
&B(t+\delta) - B(t) - b(t)\delta \\
&= \int_t^{t+\delta} \Big(b_{\mathrm{Lip}}(\tau) + c_J\,\Theta(\tau - t^*)\Big)\,\mathrm{d}\tau \;-\; b_{\mathrm{Lip}}(t)\int_t^{t+\delta}\mathrm{d}\tau \;-\; c_J\,\Theta(t - t^*)\delta \\
&= \int_t^{t+\delta} \Big(b_{\mathrm{Lip}}(\tau) - b_{\mathrm{Lip}}(t)\Big)\,\mathrm{d}\tau \;+\; c_J\int_t^{t+\delta}\Theta(\tau - t^*)\,\mathrm{d}\tau \;-\; c_J\Theta(t - t^*)\delta \\
&\leq \int_t^{t+\delta} c_L|\tau - t|\,\mathrm{d}\tau \\
&\quad + c_J\Big(\Theta(t - t^* + \delta)\,(t - t^* + \delta) \;-\; \Theta(t - t^*)\,(t - t^*)\Big) \;-\; c_J\Theta(t - t^*)\delta \\
&= c_L\int_0^{\delta} |\tau|\,\mathrm{d}\tau \;+\; c_J\,(\Theta(t - t^* + \delta) - \Theta(t - t^*))\,(t - t^* + \delta) \\
&\leq \frac{c_L}{2}\delta^2 \;+\; c_J\,(\Theta(t - t^* + \delta) - \Theta(t - t^*))\,(t - t^* + \delta)\,.
\end{aligned}$$

Because the variable δ may be negative, equality will not hold for all δ in the last step.
The second inequality can be proved analogously using

$$b_{\text{Lip}}(\tau) - b_{\text{Lip}}(t) \;\geq\; -c_L|\tau - t| \,. \qquad \square$$

The results of Lemma 3.12 and Theorem 3.16 can be extended to include a finite number of jumps:

Lemma 3.13:
Let $b(t) := b_{\text{Lip}}(t) + \sum_i c_i \Theta(t - t_i^*)$ be a piecewise Lipschitz function with a set of jumps, each in t_i^* with size c_i.
Then

$$B(t+\delta) - B(t) - b(t)\,\delta \;\leq\; \frac{c_L}{2}\delta^2 \;+\; \sum_i c_i \Big(\Theta(t - t_i^* + \delta) - \Theta(t - t_i^*)\Big)(t - t_i^* + \delta) \,.$$

Proof. This follows directly from Lemma 3.12, replacing the single jump c_J with a sum of jumps c_i. $\square$

Next, we show that the antiderivative B of a jumping function is locally Lipschitz continuous, but not differentiable in t^*.

Lemma 3.14:
Let b be defined as in Lemma 3.12.
Then the antiderivative $B(t) = \int_0^t b(\tau)\,\mathrm{d}\tau$ is locally Lipschitz. If $c_J \neq 0$, B is not differentiable in t^*.

Proof. For B being locally Lipschitz, we demand that

$$\forall t \in \mathbb{R} \quad \exists \varepsilon_0 > 0\,,\; c_L > 0 : \quad \forall |\varepsilon| \leq \varepsilon_0 \quad |B(t+\varepsilon) - B(t)| \;\leq\; c_L|\varepsilon| \,.$$

Choose an arbitrary, but fixed $t \in \mathbb{R}$, and an arbitrary constant $\varepsilon_0(t) > 0$. Selecting $c_L := \max\Big\{|b(t+\varepsilon)| \;:\; |\varepsilon| \leq \varepsilon_0(t)\Big\}$, we get for $\varepsilon \geq 0$

$$\begin{aligned} c_L\,|\varepsilon| \;&=\; \varepsilon\, c_L \;=\; \varepsilon \max_{|\varepsilon|\leq\varepsilon_0} |b(t+\varepsilon)| \;=\; \int_t^{t+\varepsilon} \max_{|\varepsilon|\leq\varepsilon_0} |b(t+\varepsilon)|\,\mathrm{d}\tau \\ \geq\; &\left|\int_t^{t+\varepsilon} b(\tau)\,\mathrm{d}\tau\right| \;=\; |B(t+\varepsilon) - B(t)| \,, \end{aligned}$$

and for $\varepsilon < 0$

$$\begin{aligned} c_L\,|\varepsilon| \;&=\; (-\varepsilon) \max_{|\varepsilon|\leq\varepsilon_0} |b(t+\varepsilon)| \;=\; \int_{t+\varepsilon}^{t} \max_{|\varepsilon|\leq\varepsilon_0} |b(t+\varepsilon)|\,\mathrm{d}\tau \\ \geq\; &\left|\int_{t+\varepsilon}^{t} b(\tau)\,\mathrm{d}\tau\right| \;=\; |B(t+\varepsilon) - B(t)| \,. \end{aligned}$$

If $c_J \neq 0$, we have the left- and right-side derivatives

$$\begin{aligned} \lim_{t \to t_+^*} B'(t) &= \lim_{t \to t_+^*} b(t) = b_{\mathrm{Lip}}(t^*) + c_J \\ \text{and} \quad \lim_{t \to t_-^*} B'(t) &= \lim_{t \to t_-^*} b(t) = b_{\mathrm{Lip}}(t^*)\,, \end{aligned}$$

thus B is not differentiable in t^*. □

In Theorem 3.16, we have to compare cut-off function terms and quadratic terms. This can be done by the following lemma.

Lemma 3.15:
For $\frac{1}{4} \leq a \in \mathbb{R}$, there holds $(t-a)_+ \leq t^2$.
For $\frac{1}{4} > a \in \mathbb{R}$, there holds $(t-a)_+ \leq t^2$ iff

$$t \in \mathbb{R} \setminus \left(\frac{1}{2} - \sqrt{\frac{1}{4} - a}\,,\ \frac{1}{2} + \sqrt{\frac{1}{4} - a} \right).$$

Proof. Let $\frac{1}{4} \leq a$. Then clearly, $(t-a)_+ = 0 \leq t^2$ if $t \leq a$. But for $t > a$, we have

$$t^2 - (t-a)_+ = t^2 - (t-a) \geq t^2 - t + \frac{1}{4} = \left(t - \frac{1}{2}\right)^2 \geq 0\,.$$

Let now $\frac{1}{4} > a$. Again, we only need to consider the case that $(t-a)_+ > 0$. But then, the functions t^2 and $t-a$ intersect in two points:

$$t^2 = t - a \quad \Leftrightarrow \quad t_{1,2} = \frac{1}{2} \pm \sqrt{\frac{1}{4} - a}\,.$$

It is clear that $t^2 > t - a$ is violated between these two points, but valid outside the given interval. □

We can finally give conditions for a solution $\mathbf{u}$ to be the unique minimizer of $\mathcal{L}$.

Theorem 3.16:
Let $\mathbf{u}$ be a stationary point of $\mathcal{L}$. Further, let the interface law b satisfy the conditions in Lemma 3.12.
Then $\mathbf{u}$ is the unique minimizer of $\mathcal{L}$ if one of the following conditions holds:

- The jump c_J is non-negative, and c_L satisfies

$$c_L < \inf_{\mathbf{d} \in V \setminus \ker \Pi} \frac{a(\mathbf{d}, \mathbf{d})}{\|\Gamma \Pi \mathbf{d}\|_{0,\Gamma_C}^2}\,. \tag{3.22}$$

- The jump c_J is negative,

$$\big(\Pi\mathbf{u}\big)(\mathbf{x}) \le t^* - \frac{1}{4}\,, \quad \text{and} \quad c_L - 2c_J \;<\; \inf_{\mathbf{d}\in V\setminus\ker\Pi} \frac{a(\mathbf{d},\mathbf{d})}{\|\Pi\mathbf{d}\|^2_{0,\Gamma_C}}\,. \tag{3.23}$$

Proof. If $\mathbf{u}$ is a stationary point of $\mathcal{L}$, we have that $0 \in \partial\mathcal{L}(\mathbf{u})$. Especially, we have a partial solution $\xi \in L^2(\Gamma_C)$ of (3.9) such that

$$a(\mathbf{u},\mathbf{v}-\mathbf{u}) \;-\; L(\mathbf{v}-\mathbf{u}) \;+\; \int_{\Gamma_C} \xi\left(\Pi\mathbf{v}-\Pi\mathbf{u}\right)\mathrm{d}s_x \;\ge\; 0\,.$$

Introduce the notations $u_{\mathbf{x}} := (\Pi\mathbf{u})(\mathbf{x})$ and $v_{\mathbf{x}} := (\Pi\mathbf{v})(\mathbf{x})$ where appropriate. Consider the following expression for an arbitrary $\mathbf{v}$:

$$\begin{aligned}
\mathcal{L}(\mathbf{v}) - \mathcal{L}(\mathbf{u}) \;=\;& \frac{1}{2}a(\mathbf{v},\mathbf{v}) \;-\; \frac{1}{2}a(\mathbf{u},\mathbf{u}) \;-\; L(\mathbf{v}) \;+\; L(\mathbf{u}) \\
&+ \int_{\Gamma_C} B(\big(\Pi\mathbf{v}\big)(\mathbf{x}))\,\mathrm{d}s_x \;-\; \int_{\Gamma_C} B(\big(\Pi\mathbf{u}\big)(\mathbf{x}))\,\mathrm{d}s_x \\
\ge\;& \frac{1}{2}a(\mathbf{v},\mathbf{v}) \;-\; \frac{1}{2}a(\mathbf{u},\mathbf{u}) \;-\; L(\mathbf{v}-\mathbf{u}) \;+\; \int_{\Gamma_C} \left(B(v_{\mathbf{x}}) \;-\; B(u_{\mathbf{x}})\right)\mathrm{d}s_x \\
&- \; a(\mathbf{u},\mathbf{v}-\mathbf{u}) \;+\; L(\mathbf{v}-\mathbf{u}) \;-\; \int_{\Gamma_C} \xi\,(v_{\mathbf{x}} - u_{\mathbf{x}})\,\mathrm{d}s_x \\
=\;& \frac{1}{2}a(\mathbf{v}-\mathbf{u},\mathbf{v}-\mathbf{u}) \;+\; \int_{\Gamma_C} \left(B(v_{\mathbf{x}}) \;-\; B(u_{\mathbf{x}}) \;-\; \xi\,(v_{\mathbf{x}} - u_{\mathbf{x}})\right)\mathrm{d}s_x \\
\ge\;& \frac{1}{2}a(\mathbf{d},\mathbf{d}) \;-\; \frac{1}{2}c_L \int_{\Gamma_C} (d_{\mathbf{x}})^2\,\mathrm{d}s_x \\
&+ \; c_J \int_{\Gamma_C} \Big(\Theta(u_{\mathbf{x}} - t^* + d_{\mathbf{x}}) \;-\; \Theta(u_{\mathbf{x}} - t^*)\Big)\Big(u_{\mathbf{x}} \;-\; t^* \;+\; d_{\mathbf{x}}\Big)\,\mathrm{d}s_x
\end{aligned} \tag{3.24}$$

employing Lemma 3.12 and writing $\mathbf{d} := \mathbf{v} - \mathbf{u}$. If this expression is positive for $\mathbf{d} \neq 0$, the stationary point $\mathbf{u}$ is a global minimizer.

The second integrand is always non-negative:

- If $u_{\mathbf{x}} - t^* + d_{\mathbf{x}} < 0$, the first Heaviside function evaluates to zero; but the second Heaviside function only evaluates to a non-negative number, leaving a non-positive first factor.
- If $u_{\mathbf{x}} - t^* + d_{\mathbf{x}} > 0$, the first factor in the second integral of (3.24) can not be negative by the same argument.

For a nonnegative jump $c_J \ge 0$, we can then discard the second integral. If we use assumption (3.22),

$$a(\mathbf{d},\mathbf{d}) \;>\; c_L\|\Pi\mathbf{d}\|^2_{0,\Gamma_C} \quad \forall 0 \neq \mathbf{d} \in V\,,$$

we can see that condition (3.24) results in $\mathcal{L}(\mathbf{v}) > \mathcal{L}(\mathbf{u})$ if $\mathbf{v} \neq \mathbf{u}$.
Note that this would just be the convexity condition on $\mathcal{L}$ if b would not contain any jumps.

For a negative jump $c_J < 0$, the assumption (3.23) states that

$$a(\mathbf{d}, \mathbf{d}) \;>\; (c_L - 2c_J)\|\Pi\mathbf{d}\|^2_{0,\Gamma_C} \quad \forall \mathbf{d} \in V\,. \tag{3.25}$$

The term $-2c_J\|\Pi\mathbf{d}\|^2_{0,\Gamma_C}$ can be bounded from below, using the assumption that $t^* - u_\mathbf{x} \geq \frac{1}{4}$ holds on Γ_C:

$$\begin{aligned}
\int_{\Gamma_C} d_\mathbf{x}^2 \,\mathrm{d}s_x \;&=\; \int_{\Gamma_C} d_\mathbf{x}^2 \,\mathrm{d}s_x \;-\; \int_{\Gamma_C} (u_\mathbf{x} - t^*)_+ \,\mathrm{d}s_x \;-\; \int_{\Gamma_C} \Theta(u_\mathbf{x} - t^*) d_\mathbf{x} \,\mathrm{d}s_x \\
&=\; \int_{\Gamma_C} d_\mathbf{x}^2 \,\mathrm{d}s_x \;-\; \int_{\Gamma_C} \Theta(u_\mathbf{x} - t^*) \left(u_\mathbf{x} - t^* + d_\mathbf{x}\right) \mathrm{d}s_x \\
\text{with Lemma 3.15:}\quad &\geq\; \int_{\Gamma_C} (d_\mathbf{x} - t^* + u_\mathbf{x})_+ \,\mathrm{d}s_x \;-\; \int_{\Gamma_C} \Theta(u_\mathbf{x} - t^*) \left(u_\mathbf{x} - t^* + d_\mathbf{x}\right) \mathrm{d}s_x \\
&=\; \int_{\Gamma_C} \Big(\Theta(u_\mathbf{x} - t^* + d_\mathbf{x}) - \Theta(u_\mathbf{x} - t^*)\Big) \left(u_\mathbf{x} - t^* + d_\mathbf{x}\right) \mathrm{d}s_x\,.
\end{aligned}$$

As $-2c_J > 0$, equations (3.25) and (3.24) return again that $\mathcal{L}(\mathbf{v}) > \mathcal{L}(\mathbf{u})$ for all $\mathbf{v} \neq \mathbf{u}$. □

If the inequality in (3.20) is relaxed to a not-strict inequality, Theorem 3.10 asserts only convexity, not strict convexity. Multiple solutions may occur. As the level sets of a convex function are convex, the set of minimizers of the objective function is simply connected and convex.

We can also relax the strict inequalities in (3.22) and (3.23) to "normal" inequalities. Then, we only retrieve $\mathcal{L}(\mathbf{v}) \geq \mathcal{L}(\mathbf{u})$ for all $\mathbf{v} \in V$. The set of solutions is also convex:

Corollary 3.17:
Under either of the two weakened assumptions

$$c_J \geq 0\,, \quad c_L \;\leq\; \inf_{\mathbf{d} \in V \setminus \ker\Pi} \frac{a(\mathbf{d}, \mathbf{d})}{\|\Pi\mathbf{d}\|^2_{0,\Gamma_C}}\,, \tag{3.26}$$

$$\text{or} \quad c_J < 0\,, \quad c_L - 2c_J \;\leq\; \inf_{\mathbf{d} \in V \setminus \ker\Pi} \frac{a(\mathbf{d}, \mathbf{d})}{\|\Pi\mathbf{d}\|^2_{0,\Gamma_C}}\,, \text{ and } \left(\Pi\mathbf{u}_i\right)(\mathbf{x}) \leq t^* - \frac{1}{4} \quad \forall i\,, \tag{3.27}$$

all stationary points $\mathbf{u}_i$ of $\mathcal{L}$ have the same (minimal) value $\mathcal{L}(\mathbf{u}_i) =: \mathcal{L}_{\min}$.
Moreover, the set $\{\mathbf{u}_i\} \subset \mathcal{K}$ of solutions is convex.

Proof. Under the weakened assumptions, the same steps as in the proof of Theorem 3.16 result in

$$\mathcal{L}(\mathbf{v}) \;\geq\; \mathcal{L}(\mathbf{u}_i) \quad \forall \mathbf{v} \in V$$

for each stationary point $\mathbf{u}_i$. Then it is immediately clear that all stationary points $\mathbf{u}_i$ have the same value $\mathcal{L}_{\min}$ in $\mathcal{L}$.

Taking two stationary points $\mathbf{u}_1$ and $\mathbf{u}_2$, we get from [11, Theorem 2.3.7, p.41] that

$$\mathcal{L}(\mathbf{u}_2) - \mathcal{L}(\mathbf{u}_1) \;=\; 0 \;\in\; \left\langle \overline{\partial}\mathcal{L}(\mathbf{u}_0),\, \mathbf{u}_2 - \mathbf{u}_1 \right\rangle \tag{3.28}$$

for some convex combination $\mathbf{u}_0 \;=\; \mathbf{u}_1 + \vartheta(\mathbf{u}_2 - \mathbf{u}_1)$ with $\vartheta \in (0,1)$. Then $\mathbf{u}_0$ is also a stationary point, and we get $\mathcal{L}(\mathbf{u}_0) = \mathcal{L}_{\min}$.

It remains to prove that every convex combination of two stationary points $\mathbf{u}_1$, $\mathbf{u}_2$ is again a stationary point. Select an arbitrary convex combination $\mathbf{w} = \mathbf{u}_1 + \vartheta(\mathbf{u}_2 - \mathbf{u}_1)$ with $\vartheta \in (0,1)$. We can now use a bisection algorithm. Set the boundaries to $\mathbf{u}_L^{(0)} := \mathbf{u}_1$

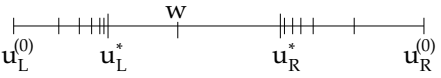


Figure 3.3.: Bisection algorithm towards $\mathbf{w}$

and $\mathbf{u}_R^{(0)} := \mathbf{u}_2$. In every step n, we can find a new stationary point $\mathbf{u}_0^{(n)} \in (\mathbf{u}_L^{(n)}, \mathbf{u}_R^{(n)})$, which has positive distance from $\mathbf{u}_L^{(n)}$ and $\mathbf{u}_R^{(n)}$. This $\mathbf{u}_0^{(n)}$ sets a new boundary for step $(n+1)$ such that the new interval $\left[\mathbf{u}_L^{(n+1)}, \mathbf{u}_R^{(n+1)}\right]$ contains $\mathbf{w}$. Moreover, $\mathbf{u}_0^{(n)}$ attains the value $\mathcal{L}(\mathbf{u}_0^{(n)}) = \mathcal{L}_{\min}$.

If the interval size of $\left[\mathbf{u}_L^{(n+1)}, \mathbf{u}_R^{(n+1)}\right]$ decreases to zero for $n \to \infty$, there holds

$$\lim_{n\to\infty} \mathcal{L}(\mathbf{u}_L^{(n)}) \;=\; \lim_{n\to\infty} \mathcal{L}(\mathbf{u}_R^{(n)}) \;=\; \mathcal{L}_{\min} \;=\; \mathcal{L}(\mathbf{w})$$

due to the continuity of $\mathcal{L}$.

If the interval boundaries converge such that they are bounded away from $\mathbf{w}$,

$$\mathbf{u}_L^* \;:=\; \lim_{n\to\infty} \mathbf{u}_L^{(n)} \;\neq\; \mathbf{u}_R^* \;:=\; \lim_{n\to\infty} \mathbf{u}_R^{(n)}\,,$$

the continuity of $\mathcal{L}$ yields that $\mathbf{u}_L^*$ and $\mathbf{u}_R^*$ are stationary points of $\mathcal{L}$. But then, we can restart the bisection with these elements as new boundaries. The bisection can be repeated until $\mathbf{u}_L^*$ and $\mathbf{u}_R^*$ converge to $\mathbf{w}$.

We now know that every convex combination $\mathbf{w}$ of two minimizers $\mathbf{u}_1$ and $\mathbf{u}_2$ is again a minimizer, so the set of solutions is convex. □

3.6. The Bundle-Newton method

According to Lemma 3.14, the potential $\mathcal{L}$ is locally Lipschitz, but not necessarily differentiable. It is also not convex in general, as shown in Corollary 3.6. The minimization method needs to regard this. Second order derivatives are not available in all points, and first order derivatives might only be given as elements of the Clarke subdifferential. Further, these differentials may vary rapidly, raising the need for an adapted line search strategy.

In this section, we present the Bundle-Newton method [29] and demonstrate the application to our model problem. We finally suggest a modification of the solver for the search direction, which may exploit the sparse structure of the system matrix better.

Before going into details, we state the central convergence statements:

Theorem 3.18:
Let the sequence $\{\mathbf{x}^{(k)}\}$ of iterates in the Bundle-Newton algorithm be bounded. Further, let the sequence $A^{(k)}$ of iteration matrices be bounded.
Then every accumulation point of $\{\mathbf{x}^{(k)}\}$ is a stationary point of the objective function, i.e. the algorithm converges.

Proof. See [29, Theorem 3.8]. The boundedness of the inverse Schur complement H_k of $A^{(k)}$ is a consequence of the boundedness of $A^{(k)}$. Note that Lukšan and Vlček claim that the sequence $\{H_k\}$ can be forced into boundedness by a modification of the matrix $G_p^{(k)}$. ⌟

Theorem 3.19:
Let the Bundle-Newton algorithm produce an infinite number of serious steps $\mathbf{x}^{(k)}$ that converge to $\mathbf{x}^*$. Further let the objective function f be strongly convex with continuous second-order derivatives in a neighborhood of $\mathbf{x}^*$. Let $A^{(k)}$ be bounded.
Then, after a sufficient number of steps, the algorithm will generate Newton iterates, resulting in superlinear convergence.

Proof. See [29, Theorem 4.4]; its assumptions have been collapsed into the given form. ⌟

Bundle methods for nonsmooth optimization regard that differentials may vary rapidly and that a second order derivative may not be available in all points: Usually, only a function evaluation and an arbitrary subdifferential element are needed in each iteration point. This class of methods goes back to the concepts of cutting-plane and ε-subgradient algorithms.
The key idea is to approximate the objective function $f(x)$ by a bundle of functions

$f_i^\#(x)$ which are easier to treat, i.e. linear or quadratic functions. In every iteration step k, f is then approximated by the maximum of all $f_i^\#(x)$ in the bundle I_k:

$$f(x) \approx f_k^\square(x) := \max_{i \in I_k} \left\{ f_i^\#(x) \right\}.$$

Note that minimizing $f_k^\square(x)$ is equivalent to minimizing a simple linear function with nonlinear constraints:

$$\begin{aligned} &\min_{x \in \mathbb{R}^n,\, t \in \mathbb{R}} t \\ &\text{s.t. } f_i^\#(x) \le t \quad \forall i \in I_k \end{aligned}$$

Bundle methods provide approximation strategies for the functions $f_i^\#(x)$, as well as updating algorithms for the bundle I_k.

The Bundle-Newton method was introduced by Lukšan and and Vlček [29]. The approximating functions $f_i^\#(x)$ are chosen to be quadratic: If

$$f_i^\#(x) = f(y^{(i)}) + (g^{(i)})^\top (x - y^{(i)}) + \frac{1}{2} \rho^i (x - y^{(i)})^\top G^{(i)} (x - y^{(i)})$$

for sample points $y^{(i)}$, subgradient elements $g^{(i)} \in \partial f(y^{(i)})$ and second order derivatives $G(\tilde{y}^{(i)})$ "close enough" ($\tilde{y}^{(i)} \approx y^{(i)}$), superlinear convergence of the method can be proven for a large class of problems [29, Theorem 4.4].

Bundle updates are performed by categorizing the iteration steps as *short steps*, *null steps* and *serious steps*, which imply improvement of the iteration point $x^{(k)}$, the bundle I_k, or both.

3.6.1. Preprocessing of the problem

It is advisable to reduce the number of degrees of freedom as far as possible; this is especially true in the 3D case, where millions of unknowns might appear in a discretization.

We are looking for a substationary point of $\mathcal{L}$. Assume that the degrees of freedom are ordered such that the first n basis functions of V expose no normal displacement on the boundary part Γ_C. These functions will be filtered out by the matrix Λ and thus will only produce a quadratic contribution. Then, the Galerkin matrix and vector will have the block structure

$$A = \begin{pmatrix} \bar{A} & B \\ B^\top & C \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} \mathbf{f}^1 \\ \mathbf{f}^2 \end{pmatrix}.$$

A solution vector will have the structure $(\mathbf{x}^1, \mathbf{x}^2)$. By a Schur complement, we can now express $\mathbf{x}^1$ in terms of $\mathbf{x}^2$:

$$\mathbf{x}^1 = \bar{A}^{-1} \left(\mathbf{f}^1 - B\mathbf{x}^2 \right), \tag{3.29}$$

which we can again insert into the definition of $\mathcal{L}$ from (3.16),

$$\begin{aligned}\mathcal{L}(\mathbf{x}) &= \frac{1}{2}\begin{pmatrix}\mathbf{x}^1\\ \mathbf{x}^2\end{pmatrix}^\top \begin{pmatrix}\bar{A} & B\\ B^\top & C\end{pmatrix}\begin{pmatrix}\mathbf{x}^1\\ \mathbf{x}^2\end{pmatrix} - \begin{pmatrix}\mathbf{f}^1\\ \mathbf{f}^2\end{pmatrix}^\top \begin{pmatrix}\mathbf{x}^1\\ \mathbf{x}^2\end{pmatrix} + \Psi_h(\mathbf{x}^2)\\ &= \frac{1}{2}\mathbf{x}^1\bar{A}\mathbf{x}^1 + \mathbf{x}^1 B\mathbf{x}^2 + \frac{1}{2}\mathbf{x}^2 C\mathbf{x}^2 - \mathbf{f}^1\mathbf{x}^1 - \mathbf{f}^2\mathbf{x}^2 + \Psi_h(\mathbf{x}^2)\\ &= \frac{1}{2}x^2\big(C - B^\top\bar{A}^{-1}B\big)x^2 - \big(\mathbf{f}^2 - \mathbf{f}^1\bar{A}^{-1}B\big)x^2 - \frac{1}{2}\mathbf{f}^1\bar{A}^{-1}\mathbf{f}^1 + \Psi_h(\mathbf{x}^2)\,.\end{aligned} \tag{3.30}$$

As we only want to minimize $\mathcal{L}$, we can drop the constant term $\mathbf{f}^1\bar{A}^{-1}\mathbf{f}^1$. Writing

$$\underline{C} := C - B^\top\bar{A}^{-1}B\,, \quad \underline{\mathbf{f}} := \mathbf{f}^2 - \mathbf{f}^1\bar{A}^{-1}B\,,$$

the new objective function results again in a quadratic problem, augmented by a nonsmooth part:

$$\underline{\mathcal{L}}(\mathbf{x}^2) := \frac{1}{2}\mathbf{x}^2\underline{C}\mathbf{x}^2 - \underline{\mathbf{f}}\mathbf{x}^2 + \Psi_h(\mathbf{x}^2)\,. \tag{3.31}$$

In a postprocessing step, we can retrieve $\mathbf{x}^1$ with (3.29).

3.6.2. Algorithm description

This description follows [29]. Some parameters were culled, as their values are only needed for theoretical convergence (i_r, i_m, C_G sufficiently large) or need to be fixed for higher order convergence ($\omega = 1$).

Each iteration of the algorithm is composed of three steps: The determination of a search direction $\mathbf{d}^{(k)}$, a line search returning the step length t, and a bundle update managing the set of functions $f_i^\#(\mathbf{x})$. A stopping criterion can be applied already after the direction-search step.

A list of parameters can be passed to the algorithm:

- Counters and indices:
  M: maximal bundle size
  i_m: matrix selection parameter
- control values:
  $\varepsilon \geq 0$: final tolerance
  $\gamma > 0$: distance measure parameter
  $t_0 \in (0, 1)$ to distinguish small and serious steps

Further, we have persistent auxiliary variables to transfer information in consecutive iteration steps:

$s_p^{(k)} \in \mathbb{R}$, correction term for the approximation error
$f_p^{(k)} \in \mathbb{R}$, approximated value of f
$\mathbf{g}_p^{(k)} \in \mathbb{R}^n$, approximated gradient
$G_p^{(k)} \in \mathbb{R}^{n\times n}$, approximated Hessian
$i_n \in \mathbb{N}$, the number of consecutive null steps
$i_s \in \mathbb{N}$, counting the number of serious steps since the last reset

Each element j of the bundle will consist of the following parts:

$\mathbf{y}_j^{(k)} \in \mathbb{R}^n$: a trial point *(needs not be stored explicitly)*
$f_j^{(k)} \in \mathbb{R}$: the function evaluation in $\mathbf{y}_j^{(k)}$
$s_j^{(k)} \in \mathbb{R}$: the approximation error in $\mathbf{x}^{(k)}$
$\mathbf{g}_j^{(k)} \in \mathbb{R}^n$: an element of the subgradient $\bar{\partial} f(\mathbf{y}_j^{(k)})$
$G_j^{(k)} \in \mathbb{R}^{n\times n}$: Hessian in $\mathbf{y}_j^{(k)}$ or a nearby point
$\rho_j \in \{0, 1\}$ selects linear or quadratic approximation

Finally, let $\mathbf{x}^{(k)}$ be the sequence converging to a minimal argument $\mathbf{x}$, $\mathbf{g}^{(k)}$ an element of the subgradient of f in $\mathbf{x}^{(k)}$, and $G^{(k)}$ an approximated Hessian in $\mathbf{x}^{(k)}$.

$k = 1, 2, \ldots$ is the iteration variable.

Initialization

Select a first approximation $\mathbf{x}^{(1)}$; set the counters $k := 1$, $i_s := 0$, $i_n := 0$.
The auxiliary variables can be initialized:

$s_p^{(1)} := 0$
$f_p^{(1)} := f(\mathbf{x}^{(1)})$
$\mathbf{g}_p^{(1)} :\in \bar{\partial} f(\mathbf{x}^{(1)})$
$G_p^{(1)} := \nabla^2 f(\mathbf{x}^{(1)})$, or in a point close to $\mathbf{x}^{(1)}$ where a Hessian of f is available

Set $\mathbf{g}^{(1)} := \mathbf{g}_p^{(1)}$ and $G^{(1)} := G_p^{(1)}$. The first bundle element is

$$f_1^{(1)} = f(\mathbf{x}^{(1)}), \qquad s_1^{(1)} = 0, \qquad \mathbf{g}_1^{(1)} = \mathbf{g}^{(1)},$$
$$G_1^{(1)} = G^{(1)}, \qquad \rho_1 = 1.$$

Search direction

To find a search direction $\mathbf{d}^{(k)}$, a quadratic minimization problem with linear inequality constraints needs to be solved. This introduces the vector of Lagrange multipliers, $\boldsymbol{\lambda}^{(k)}$.

Define an auxiliary matrix $A = A^{(k)}$:

- If $i_n > i_m$, take the matrix from the last step, $A^{(k-1)}$.
- Otherwise:
  - If the last two steps were classified as *serious* and $\lambda_{k-1}^{(k-1)} = 1$, take $G^{(k)}$;
  - else, take the persistent Hessian $G_p^{(k)}$.

In the second case, the matrix A may be adjusted to be positive definite; this can be done by adding a multiple of the identity matrix.

The original problem is stated in Section 3.6.3. Its objective function is augmented by the Lagrange multipliers $\boldsymbol{\lambda}^{(k)}$ and $\lambda_p^{(k)}$, finally its number of unknowns can be reduced to the bundle size $\#B$.
Setting $\alpha_j^{(k)} := \max\left\{\left|f_j^{(k)} - f(\mathbf{x}^{(k)})\right|,\ \gamma s_j^{(k)}\right\}$ and $\alpha_p^{(k)} := \max\left\{\left|f_p^{(k)} - f(\mathbf{x}^{(k)})\right|,\ \gamma s_p^{(k)}\right\}$, the augmented problem states:

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^{\#B},\,\lambda_p\in\mathbb{R}} \frac{1}{2}\Big(\sum_{j=1}^{\#B}\lambda_j\mathbf{g}_j^{(k)} + \lambda_p\mathbf{g}_p^{(k)}\Big)^\top A^{-1}\Big(\sum_{j=1}^{\#B}\lambda_j\mathbf{g}_j^{(k)} + \lambda_p\mathbf{g}_p^{(k)}\Big) + \sum_{j=1}^{\#B}\lambda_j\alpha_j^{(k)} + \lambda_p\alpha_p^{(k)}$$
$$\text{s.t. } \lambda_j \geq 0 \quad \forall j; \qquad \lambda_p \geq 0; \qquad \sum_{j=1}^{\#B}\lambda_j + \lambda_p = 1\,. \tag{3.32}$$

The new search direction can be determined as

$$\mathbf{d}^{(k+1)} := -\sum_{j=1}^{\#B}\lambda_j^{(k)}A^{-1}\mathbf{g}_j^{(k)} - \lambda_p^{(k)}A^{-1}\mathbf{g}_p^{(k)}\,. \tag{3.33}$$

Some auxiliary variables can now be computed:

$$\hat{v}^{(k)} := -\mathbf{d}^{(k+1)\top}A\mathbf{d}^{(k)} - \sum_{j=1}^{\#B}\lambda_j^{(k)} - \lambda_p^{(k)}\alpha_p^{(k)}\,,$$

$$\tilde{\mathbf{g}}_p^{(k)} := \sum_{j=1}^{\#B}\lambda_j^{(k)}\mathbf{g}_j^{(k)} + \lambda_p^{(k)}\mathbf{g}_p^{(k)}\,, \qquad \tilde{f}_p^{(k)} := \sum_{j=1}^{\#B}\lambda_j^{(k)}f_j^{(k)} + \lambda_p^{(k)}f_p^{(k)}\,,$$

$$\tilde{s}_p^{(k)} := \sum_{j=1}^{\#B}\lambda_j^{(k)}s_j^{(k)} + \lambda_p^{(k)}s_p^{(k)}\,, \qquad \tilde{\alpha}_p^{(k)} := \max\left\{\left|\tilde{f}_p^{(k)} - f(\mathbf{x}^{(k)})\right|,\ \gamma\tilde{s}_p^{(k)}\right\}\,,$$

$$v^{(k)} := -\frac{1}{2}\tilde{\mathbf{g}}_p^{(k)}A^{-1}\tilde{\mathbf{g}}_p^{(k)} - \tilde{\alpha}_p^{(k)}\,, \qquad w^{(k)} := -\frac{1}{2}v^{(k)} + \frac{1}{2}\tilde{\alpha}_p^{(k)}\,,$$

and the matrix $G_p^{(k)}$ is updated to

$$G_p^{(k+1)} := \sum_{j=1}^{\#B}\lambda_j^{(k)}\rho_j G_j^{(k)} + \lambda_p^{(k)}G_p^{(k)}\,.$$

The stopping criterion is $w^{(k)} \leq \varepsilon$; it can be evaluated before the matrix update.

Line search

The step lengths t_L, t_R are determined through a line search algorithm by Kiwiel [24]. This algorithm depends on some further parameters:

$m_L \in (0, \frac{1}{2})$ and $m_R \in (m_L, 1)$: conditions for serious and short steps
$0 < C_S \in \mathbb{R}$: maximal distance $\|x^{(k)} - y^{(k)}\|$
$\zeta \in (0, \frac{1}{2})$ and $1 \leq \theta \in \mathbb{R}$: interpolation parameters for $t \in (t_L, t_U)$

Contrary to the original description, the matrix A may not be recalculated in every line search step to reduce the computational complexity.

First, the bounds are initialized by $t_L = 0$, $t_U = 1$, and $t = 1$.

The following steps are repeated until one of the exit conditions is met:

- Compute $f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)})$. If this step is good, i.e. $f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)}) \leq f(\mathbf{x}^{(k)}) + m_L t v^{(k)}$, raise the lower bound ($t_L \leftarrow t$), otherwise drop the upper bound ($t_U \leftarrow t$).
- If $t_L \geq t_0$ (serious step),
  set $t_R := t_L$ and **exit**.
- Compute new line search data:

$$\mathbf{g} :\in \bar{\partial} f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)}) \,, \qquad \rho := \begin{cases} 1, & i_n \leq 3, \\ 0, & \text{else} \end{cases},$$

$$f_v := f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)}) \;+\; (t_L - t)\mathbf{g}^\top \mathbf{d}^{(k)} \;+\; \frac{1}{2}\rho(t_L - t)^2 (\mathbf{d}^{(k)})^\top A \mathbf{d}^{(k)} \,,$$

$$\beta := \max\left\{ |f_v - f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)})| \,,\; \gamma |t_L - t| \|\mathbf{d}^{(k)}\| \right\}$$

- If $(t - t_L)\|\mathbf{d}^{(k)}\| \leq C_S$ and

$$-\beta + (\mathbf{d}^{(k)})^\top \left(\mathbf{g} + \rho(t_L - t) A \mathbf{d}^{(k)} \right) \;\geq\; m_R v^{(k)} \,,$$

  set $t_R := t_L$ and **exit**.
- Select a new $t \in [t_L + \zeta(t_U - t_L)^\theta,\; t_U - \zeta(t_U - t_L)^\theta]$ by interpolation.

Bundle and variables update

Update the iteration variables for the next step with values from the line search:

$$\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + t_L \mathbf{d}^{(k)} \,, \qquad \mathbf{y}^{(k+1)} := \mathbf{x}^{(k)} + t\mathbf{d}^{(k)} \,,$$
$$f^{(k+1)} := f(\mathbf{y}^{(k+1)}) \,, \qquad \mathbf{g}^{(k+1)} :\in \bar{\partial} f(\mathbf{y}^{(k+1)}) \,,$$
$$G^{(k+1)} := \nabla^2 f(\tilde{\mathbf{y}}) \,,$$

where $\tilde{\mathbf{y}} \approx \mathbf{y}^{(k+1)}$ is a point where f is smooth enough.
For the persistent variables, the new values are

$$s_p^{(k+1)} := \tilde{s}_p^{(k)} + t_L \|\mathbf{d}^{(k)}\|, \qquad f_p^{(k+1)} := \tilde{f}_p^{(k)} + t_L (\mathbf{d}^{(k)})^\top \tilde{\mathbf{g}}_j^{(k)} + \frac{1}{2} t_L^2 (\mathbf{d}^{(k)})^\top G_p^{(k+1)} \mathbf{d}^{(k)},$$
$$\mathbf{g}_p^{(k+1)} := \tilde{\mathbf{g}}_p^{(k)} + t_L G_p^{(k+1)} \mathbf{d}^{(k)}.$$

The counters are also updated: If $t_L < t_0$ (short step), increase i_n by 1; otherwise, set i_n to 0 and increase i_s by 1.

Now the entries of each bundle element j need to be adjusted:

$$s_j^{(k+1)} := s_j^{(k)} + t_L \|\mathbf{d}^{(k)}\|, \qquad f_j^{(k+1)} := f_j^{(k)} + t_L (\mathbf{d}^{(k)})^\top \mathbf{g}_j^{(k)} + \frac{1}{2} \rho_j t_L^2 (\mathbf{d}^{(k)})^\top G^{(j)} \mathbf{d}^{(k)},$$
$$\mathbf{g}_j^{(k+1)} := \mathbf{g}_j^{(k)} + \rho_j t_L G^{(j)} \mathbf{d}^{(k)}.$$

A new element is added to the bundle:

$$f_{(k+1)}^{(k+1)} := f^{(k+1)} + (t_L - t)(\mathbf{d}^{(k)})^\top \mathbf{g}^{(k+1)} + \frac{1}{2} \rho_{(k+1)} (t_L - t)^2 (\mathbf{d}^{(k)})^\top G^{(k+1)} \mathbf{d}^{(k)},$$
$$s_{(k+1)}^{(k+1)} := t \|\mathbf{d}^{(k)}\|, \qquad \mathbf{g}_{(k+1)}^{(k+1)} := \mathbf{g}^{(k+1)} + \rho_{(k+1)} (t_L - t) G^{(k+1)} \mathbf{d}^{(k+1)},$$
$$G_{(k+1)}^{(k+1)} := G^{(k+1)}, \qquad \rho_{(k+1)} := \begin{cases} 1, & i_n \le 3, \\ 0, & \text{else} \end{cases}.$$

Finally, the new bundle is set up by any bundle elements, provided that the oldest element is eliminated, and the new element is included.

3.6.3. Alternative determination of $\mathbf{d}^{(k)}$

When computing the search direction $\mathbf{d}^{(k)}$, a constrained quadratic minimization problem needs to be solved. The algorithm later makes use of the Lagrange multipliers $\boldsymbol{\lambda}^{(k)}$ and $\lambda_p^{(k)}$, so an augmented objective function is minimized under simple bounds instead. This results in the minimization formulation (3.32).
The dimension of that problem equals the size of the bundle, which can be bounded to a small number. However, linear equation systems with n unknowns need to be solved first to compute the vectors $A^{-1}\mathbf{g}_j^{(k)}$ and $A^{-1}\mathbf{g}_p^{(k)}$, which are necessary to set up the actual $(\#B + 1) \times (\#B + 1)$ system matrix. The expensive computations are:

- solving $\#B + 1$ large linear equation systems with a symmetric, positive definite matrix,
- solving a small-scale minimization problem with $\#B + 1$ unknowns and simple inequality constraints.

Instead, one can also regard the original minimization problem:

$$\begin{aligned} \min_{\mathbf{d}\in\mathbb{R}^n,\, v\in\mathbb{R}} \; & v \;+\; \frac{1}{2}\mathbf{d}^\top A^{(k)}\mathbf{d} \\ \text{s.t.} \quad & -\alpha_j^{(k)} \;+\; \mathbf{d}^\top \mathbf{g}_j^{(k)} \;\leq\; v \\ & -\alpha_p^{(k)} \;+\; \mathbf{d}^\top \mathbf{g}_p^{(k)} \;\leq\; v\,. \end{aligned} \tag{3.34}$$

This problem has the full dimension of $n + 1$ variables; the number of inequality constraints is $\#B + 1$. Such problems can be solved efficiently using active-set or interior point methods.

When a minimizer $(\mathbf{d}, v)$ is found, the multipliers $\boldsymbol{\lambda}^{(k)}$ and $\lambda_p^{(k)}$ can be computed by solving the over-determined equation system

$$\begin{pmatrix} \vdots & \vdots \\ \mathbf{g}_j^{(k)} & \mathbf{g}_p^{(k)} \\ \vdots & \vdots \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}^{(k)} \\ \lambda_p^{(k)} \end{pmatrix} \;=\; -A^{(k)}\mathbf{d}^{(k)}\,.$$

The expensive computations are:

- solving a sequence of large linear equation systems with a symmetric, indefinite matrix,
- solving one more equation system to compute $A^{-1}\tilde{\mathbf{g}}_p^{(k)}$ for the auxiliary variable $v^{(k)}$.

If the matrix A is dense, solving several linear equation systems in the original computation can be done by first computing a Cholesky decomposition of A (recall that A was, if necessary, modified to be positive definite). The number of systems to be solved itself is then not critical. On the other hand, the number of solved systems is critical for a sparse matrix A, as the inverse of A is not given explicitly. If the constrained problem (3.34) can be solved in less than $\#B+1$ steps, this ansatz is more efficient.

4. A primal-dual active set method

The minimization of a nondifferentiable objective function is computationally expensive. Section 3.6.1 showed the block structure of the minimization problem and a possible Schur complement strategy. In every iteration step for the Bundle-Newton method, we had to compute a dense approximate Hessian matrix

$$\nabla^2 \underline{\mathcal{L}}(\mathbf{x}^{(k)}) \;\approx\; \underline{C} \;+\; M(\hat{b}, \Lambda \mathbf{x}^{(k)}) \,.$$

Without this preprocessing, each iteration step would have needed the Hessian matrix

$$\nabla^2 \mathcal{L}(\mathbf{x}^{(k)}) \;\approx\; A + \begin{pmatrix} 0 & 0 \\ 0 & M(\hat{b}, \Lambda \mathbf{x}^{(k)}) \end{pmatrix} ,$$

which has large sparse blocks, as A is sparse. Both options leave us with a dense Hessian, effectively constraining the number of unknowns on the contact boundary to several thousands in present-day computers.

However, if the function $b(t)$ has the structure

$$b(t) \;=\; -c_J \Theta(\delta - t) \,,$$

and if we may impose certain further assumptions, we can provide another method to solve the hemivariational inequality. This method will also include the contact condition directly.
The use of a primal-dual active set strategy for a specific hemivariational inequality was first proposed by Hintermüller et al. [20] for a membrane problem. In this chapter, we use this algorithm as a subproblem solver in an iterative scheme for elastic problems. The final algorithm is given in Section 4.4.

Remark 4.1: *We can not directly apply the results from [20] to the linear elastic problem, as the proof of convergence in Lemma 4.5 can only be transferred to other differential operators than $-\Delta$ that provide a maximum principle. However, a general maximum principle has not yet been found for the differential operator $-\operatorname{div} \boldsymbol{\sigma}(\cdot)$. Special configurations may expose a maximum principle, e.g. by exploiting symmetry. There are counter-examples for a maximum principle in terms of the principal stresses, for an overview see Wheeler [47]. A scaled version based on the modulus of displacements was given by Agmon [1] and Fichera [14], stating*

$$\sup_{\Omega} |\mathbf{u}| \;\leq\; H \sup_{\partial\Omega} |\mathbf{u}| \,;$$

but in our case, a maximum principle would be needed that depends on the displacement in normal direction, $\mathbf{u} \cdot \mathbf{n}$.

The general strategy of our solution method is to decompose the contact problem with adhesion (2.17) in three subproblems:
First, we cut off a layer of thickness h from the contact boundary Γ_C. In the remaining domain Ω^1, we simply have to solve a linear-elastic problem with mixed boundary conditions.
Second, we decompose the solution in the h-layer into a part $(u_1, u_2, 0)$ and a part $(0, 0, u_3)$. For the $(u_1, u_2, 0)$-part, we again have to solve a linear-elastic problem.
Finally, the u_3-part can be reformulated into a membrane problem, which can be solved with the primal-dual active set method.
Coupling all three subproblems results in an iterative method to solve the full problem.

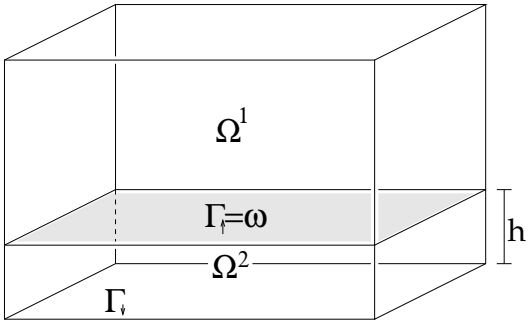


Figure 4.1.: Decomposition into two subdomains: Ω^1 and bottom layer Ω^2

Remark 4.2: *In our strategy, we need to decompose the problem in the bottom layer Ω^2 into two subproblems. It is also possible to reduce the deformed bottom layer to the plate equation $\Delta^2 u = f$. The problems $(u_1, u_2, 0)$ and $(0, 0, u_3)$ then completely decouple, see Ciarlet [10, Section 1.4]. The use of conforming Finite Elements would then lead to quadratic test functions, and the theory in [19] would need to be adapted.*

The assumptions we impose are as follows:

- The contact surface Γ_C has no curvature. Here, we choose a configuration such that Γ_C is inside the $x_3 = 0$ plane.
- The x_3 derivative of u_3 can be neglected close to the contact surface.
  In the same region, the derivatives $u_{1,13}$ and $u_{2,23}$ are continuous.
- Close to the contact surface, the third components of the given boundary tractions t_3 and the volume force f_3 are smooth.

- Saint-Venant's principle allows us to replace surface normal tractions by a volume force close to the contact surface. (See e.g. [42] for an overview of this principle.)

4.1. Domain decomposition

Our strategy is to split the original problem into three subproblems. We resort again to the classical formulation, as we need to perform some further processing in the second domain Ω^2. The original equation (2.17) is split into one part for each domain.
Additionally, we need transmission conditions on the interface $\Gamma_\uparrow$ between Ω^1 and Ω^2. Let $\mathbf{u}^{(1)}$ be the solution in Ω^1 and $\mathbf{u}^{(2)}$ the solution in Ω^2. The first condition demands continuity of the displacement field:

$$\lim_{\Omega^1 \ni \mathbf{x} \to \mathbf{x}^*} \mathbf{u}^{(1)}(\mathbf{x}) \stackrel{!}{=} \lim_{\Omega^2 \ni \mathbf{x} \to \mathbf{x}^*} \mathbf{u}^{(2)}(\mathbf{x}), \quad \mathbf{x}^* \in \Gamma_\uparrow \tag{4.1}$$

This condition is immediately clear: If it is violated, the body exposes a crack along the interface.

The second condition demands continuity of the stress vector along $\Gamma_\uparrow$:

$$\lim_{\Omega^1 \ni \mathbf{x} \to \mathbf{x}^*} \sigma(\mathbf{u}^{(1)})(\mathbf{x}) \cdot \mathbf{n}^1 \stackrel{!}{=} \lim_{\Omega^2 \ni \mathbf{x} \to \mathbf{x}^*} \sigma(\mathbf{u}^{(2)})(\mathbf{x}) \cdot \mathbf{n}^2, \quad \mathbf{x}^* \in \Gamma_\uparrow \tag{4.2}$$

Descriptively, this condition enforces a balance of forces on $\Gamma_\uparrow$. If it is violated, the body can not be in mechanical equilibrium. The mathematical justification of this is given when deriving the variational formulation in both subdomains: If a test function $\mathbf{v}$ is $\mathbf{v}^{(1)}$ in Ω^1 and $\mathbf{v}^{(2)}$ in Ω^2, we get an additional integral

$$\int_{\Gamma_\uparrow} \left(\sigma(\mathbf{u}^{(1)}) \cdot \mathbf{n}^1 \; + \; \sigma(\mathbf{u}^{(2)}) \cdot \mathbf{n}^2 \right) \cdot \mathbf{v} \, \mathrm{d}s_x \,,$$

which only vanishes if the transmission condition is fulfilled. (Note that $\mathbf{n}^1 = -\mathbf{n}^2$.)

If no nonlinear contribution is present, the domain decomposition method iterates the following steps, starting with some displacement function λ^k on $\Gamma_\uparrow$:

- Solve a problem with inhomogeneous Dirichlet boundary conditions in Ω^1:

$$\begin{aligned} -\operatorname{div} \sigma(\mathbf{u}^{(1);k}) &= \mathbf{f} \quad \text{in } \Omega^1 \\ \mathbf{u}^{(1);k} &= 0 \quad \text{on } \partial\Omega^1 \cap \Gamma_D \\ \sigma(\mathbf{u}^{(1);k}) \cdot \mathbf{n} &= \mathbf{t} \quad \text{on } \partial\Omega^1 \cap \Gamma_N \\ \mathbf{u}^{(1);k} &= \lambda^k \quad \text{on } \Gamma_\uparrow \,. \end{aligned}$$

- Retrieve the stress vector field $\sigma(\mathbf{u}^{(1);k}) \cdot \mathbf{n}$ on $\Gamma_\uparrow$.

- Solve a problem in Ω^2:

$$\begin{aligned} -\operatorname{div} \sigma(\mathbf{u}^{(2);k}) &= \mathbf{f} \quad \text{in } \Omega^2 \\ \mathbf{u}^{(2);k} &= 0 \quad \text{on } \partial\Omega^2 \cap \Gamma_D \\ \sigma(\mathbf{u}^{(2);k}) \cdot \mathbf{n} &= \mathbf{t} \quad \text{on } \partial\Omega^2 \cap \Gamma_N \\ \sigma(\mathbf{u}^{(2);k}) \cdot \mathbf{n}^2 &= \sigma(\mathbf{u}^{(1);k}) \cdot \mathbf{n}^1 \quad \text{on } \Gamma_\uparrow \,. \end{aligned}$$

- Update λ:

$$\lambda^{k+1} := \vartheta \, \mathbf{u}^{(2);k}\big|_{\Gamma_\uparrow} + (1-\vartheta)\lambda^k$$

If we choose the damping parameter $\vartheta \in (0, \vartheta_{\max})$ small enough, this Richardson iteration converges to a solution, see e.g. Quarteroni and Valli [36, Theorem 4.2.2, p.118ff] for a general proof. The specialization to elastic problems can be found in the same book in Section 5.2.

The problem in Ω^1 now contains the most degrees of freedom. But these unknowns now only appear in a linear-elastic problem with mixed boundary conditions. The transmission of the interface data is not difficult, as we have matching meshes for λ^k and $\mathbf{u}^{(1);k}$ on $\Gamma_\uparrow$.

Denote by an index 1 the restriction of a, V and L to Ω^1. Extending λ^k into Ω^1 by a function $\mathcal{E}\lambda^k \in V_1$, we get a linear variational equation with a homogeneous Dirichlet boundary condition on $\Gamma_\uparrow$:
Find $\mathbf{u}_*^{(1)} \in V_{1\,(0)}^1$ such that

$$a_1(\mathbf{u}_*^{(1)}, \mathbf{v}) \;=\; L(\mathbf{v}) \;-\; a_1(\mathcal{E}\lambda^k, \mathbf{v}) \quad \forall \mathbf{v} \in V_{1\,(0)}^1 \,.$$

The solution is then $\mathbf{u}^{(1);k} = \mathbf{u}_*^{(1)} + \mathcal{E}\lambda^k$.

The same can be applied to the finite-dimensional approximation. With piecewise linear functions for $\mathbf{u}$ and λ, a valid extension is simply done by matching pointwise displacements on $\Gamma_\uparrow$ and setting all other coefficients in Ω^1 to zero.

4.2. Active set method for the membrane

We will split the subproblem in Ω^2 even further into two problems in Section 4.3. The nonlinear behavior is then reduced to a membrane problem with contact and adhesion.

In this section, we consider a domain $\Omega \subset \mathbb{R}^2$.

The primal-dual active set method that we use was proposed by Hintermüller, Kovtunenko and Kunisch [20]. The problem under consideration is there:

Given $\Omega \subset \mathbb{R}^2$ and a material parameter D, find a scalar function $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -D\Delta u &= f + \xi \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \\ u &\geq \psi \quad \text{in } \Omega \end{aligned} \tag{4.3}$$

$$\begin{aligned} \text{with} \quad \xi(\mathbf{x}) &= 0 \quad \text{if } u(\mathbf{x}) > \psi(\mathbf{x}) + \delta\,, \\ \xi(\mathbf{x}) &= -c_J \quad \text{if } u(\mathbf{x}) \in \Big(\psi(\mathbf{x}),\ \psi(\mathbf{x}) + \delta\Big]\,, \\ \xi(\mathbf{x}) &\geq -c_J \quad \text{if } u(\mathbf{x}) = \psi(\mathbf{x})\,. \end{aligned}$$

Here, a reaction force of size c_J will occur if the gap size falls below δ.

This problem is then transferred into a hemivariational inequality. Selecting a convex cone

$$\mathcal{K} \;:=\; \Big\{u \in H_0^1(\Omega) \;:\; u(\mathbf{x}) \geq \psi(\mathbf{x})\ \forall \mathbf{x} \in \Omega\Big\}\,,$$

the problem now states:
Find $u \in \mathcal{K}$ such that

$$D\int_\Omega \nabla u \cdot \nabla(v-u)\,\mathrm{d}x \;+\; c_J \int_\Omega \Theta\Big(\delta - (\psi - u)\Big)\,\mathrm{d}x \;\geq\; \int_\Omega f\,(v-u)\,\mathrm{d}x \quad \forall v \in \mathcal{K}\,. \tag{4.4}$$

Hintermüller et al. then introduce a first Lagrange multiplier $\lambda \in M_+$ for the non-penetration condition, where

$$M_+ \;:=\; \Big\{\lambda \in L^2(\Omega) \;:\; \lambda \geq 0 \ \text{ a.e. in } \Omega\Big\}\,.$$

This leads to the (nonlinear) variational equation

$$D\int_\Omega \nabla u \cdot \nabla v\,\mathrm{d}x \;+\; c_J \int_\Omega \Theta\Big(\delta - (\psi - u)\Big)v\,\mathrm{d}x \;-\; \int_\Omega \lambda v\,\mathrm{d}x \;=\; \int_\Omega f v\,\mathrm{d}x\,. \tag{4.5}$$

A second Lagrange multiplier p is introduced to model the reaction force from the adhesion term, leading to the following result:

Lemma 4.3:
There exists a pair $(u, \lambda) \in \Big(\mathcal{K} \cap H^2(\Omega)\Big) \times M_+$ such that equation (4.5) and the complementarity system

$$\lambda \geq 0\,, \quad u \geq \psi\,, \quad \int_\Omega \lambda(u-\psi)\,\mathrm{d}x = 0$$

are satisfied. u satisfies the hemivariational inequality (4.4). Define the multiplier p by

$$p \;:=\; c_J\Theta(\delta - (\psi - u)) \;\in\; M_+\,,$$

and let $\xi := \lambda - p$, then (u, ξ) solves the original problem (4.3).

Proof. see [20]. ⌟

Next, we introduce the active sets for the continuous problem. For this, we first choose an arbitrary, but fixed constant $c > 0$. The active and inactive set for the contact condition are then

$$\mathcal{A}_c := \left\{x \in \Omega : \lambda(x) - c\Big(u(x) - \psi(x)\Big) > 0\right\}, \tag{4.6}$$

$$\mathcal{I}_c := \left\{x \in \Omega : \lambda(x) - c\Big(u(x) - \psi(x)\Big) \leq 0\right\}. \tag{4.7}$$

The active and inactive set for the adhesion force are

$$\mathcal{A}_p := \{x \in \Omega : u(x) \leq \psi(x) + \delta\}, \tag{4.8}$$

$$\mathcal{I}_p := \{x \in \Omega : u(x) > \psi(x) + \delta\}. \tag{4.9}$$

With these sets defined, we can now state the problem we will actually use in Section 4.2.1. If Algorithm 4.4 in Section 4.2.1 converges, the solution of this problem is a stationary solution of that algorithm. Lemma 5 and Theorem 1 in [20] confirm that this solution is a solution of the original hemivariational inequality.

The reformulated problem is now:
Find $v \in H_0^1(\Omega)$ such that

$$\begin{aligned} D\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x + \int_\Omega pv \,\mathrm{d}x - \int_\Omega \lambda v \,\mathrm{d}x &= \int_\Omega fv \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega) \\ u = \psi \text{ on } \mathcal{A}_c \qquad \lambda &= 0 \text{ on } \mathcal{I}_c \\ p = c_J \text{ on } \mathcal{A}_p \qquad p &= 0 \text{ on } \mathcal{I}_p \end{aligned} \tag{4.10}$$

4.2.1. Active set algorithm for the continuous problem

We first state the algorithm for the continuous problem. While there still exists no proof that this algorithm will converge to a solution of (4.4), its discrete version (4.8) will stop after a finite number of steps.

Algorithm 4.4:
Initialize $\mathcal{A}_p^{(0)} := \Omega$ and $\mathcal{A}_c^{(0)}$ arbitrarily. Then iterate the following steps:

1. Solve the linear subproblem

$$D\int_\Omega \nabla u^{(n)} \cdot \nabla v \,\mathrm{d}x - \int_\Omega \lambda^{(n)} v \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x - \int_\Omega p^{(n)} v \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega)$$

with the equality conditions

$$u^{(n)} = \psi \text{ on } \mathcal{A}_c^{(n-1)} \quad \text{and} \quad \lambda^{(n)} = 0 \text{ on } \mathcal{I}_c^{(n-1)} = \Omega \setminus \mathcal{A}_c^{(n-1)}$$

and the reaction force $p^{(n)}$ defined as

$$p^{(n)}(x) := \begin{cases} c_J\,, & x \in \mathcal{A}_p^{(n-1)} \\ 0\,, & x \notin \mathcal{A}_p^{(n-1)} \end{cases}.$$

2. Update the active sets:

$$\begin{aligned} \mathcal{A}_c^{(n)} &:= \left\{ x \in \Omega \,:\, \lambda^{(n)}(x) - c\Big(u^{(n)}(x) - \psi(x)\Big) \,>\, 0 \right\}, \\ \mathcal{I}_c^{(n)} &:= \Omega \setminus \mathcal{A}_c^{(n)}, \\ \mathcal{A}_p^{(n)} &:= \left\{ x \in \Omega \,:\, u^{(n)}(x) \,\leq\, \psi(x) + \delta \right\}, \\ \mathcal{I}_p^{(n)} &:= \Omega \setminus \mathcal{A}_p^{(n)}. \end{aligned}$$

3. Check if the active sets were changed, i.e. $\mathcal{A}_c^{(n)} \subsetneq \mathcal{A}_c^{(n-1)}$ or $\mathcal{A}_p^{(n)} \subsetneq \mathcal{A}_p^{(n-1)}$. Otherwise, stop.

♦

The iteration elements $\mathcal{A}_c^{(n)}$, $\mathcal{A}_p^{(n)}$, $u^{(n)}$ and $p^{(n)}$ are subject to an invariant of the algorithm, as stated in the following lemma. Note that this lemma does not imply convergence to a solution.

Lemma 4.5:
If the boundary of $I_c^{(n)}$ is C^2-regular for all n, we have the following monotonicity relations:

$$\psi \,\leq\, u^{(2)} \,\leq\, \cdots \,\leq\, u^{(n)} \,\leq\, u^{(n+1)} \,\leq\, \cdots \tag{4.11}$$

$$\Omega \,\supseteq\, \mathcal{A}_c^{(1)} \,\supseteq\, \cdots \,\supseteq\, \mathcal{A}_c^{(n)} \,\supseteq\, \mathcal{A}_c^{(n+1)} \,\supseteq\, \cdots \tag{4.12}$$

$$c_J = p^{(1)} \,\geq\, p^{(2)} \,\geq\, \cdots \,\geq\, p^{(n)} \,\geq\, p^{(n+1)} \,\geq\, \cdots \tag{4.13}$$

$$\Omega \,=\, \mathcal{A}_p^{(0)} \,\supseteq\, \mathcal{A}_p^{(1)} \,\supseteq\, \cdots \supseteq \mathcal{A}_p^{(n)} \,\supseteq\, \mathcal{A}_p^{(n+1)} \,\supseteq\, \cdots \tag{4.14}$$

Proof. see [20].
The proof is analogous to the proof of Lemma 4.9 for the discrete version. It relies on the maximum principle for the Laplacian, which is replaced in the discrete version by the M-matrix property given in Section 4.2.2 for the stiffness matrix. ⛶

4.2.2. M-matrices

An M-matrix $A \in \mathbb{R}^{n\times n}$ is a matrix of the type

$$A \,=\, sI - B\,, \quad s \geq \rho(B)\,, \quad B \geq 0\,,$$

where I is the $n \times n$ identity matrix, and $\mathbb{R}^{n\times n} \ni B \geq 0$ means that B has only positive elements. One can immediately see here that all off-diagonal entries of A are negative, raising need to define the set of Z-matrices by

$$Z^{n\times n} := \left\{A = (a_{ij}) \in \mathbb{R}^{n\times n} : a_{ij} \leq 0,\, i \neq j\right\},$$

which is then a superset of M-matrices.

Let a decomposition $\mathcal{T}_h$ of Ω into triangles be given. Then the following lemma holds:

Lemma 4.6:
Let the interior angles of T be acute for all $T \in \mathcal{T}_h$. Let A be the stiffness matrix resulting from the Laplace operator, using the standard piecewise linear, conforming FE basis over $\mathcal{T}_h$.
Then A is an M-matrix. Moreover, if we have a homogeneous Dirichlet boundary part, A is a nonsingular M-matrix.

Proof. Assembling the global stiffness matrix A, we compute a local stiffness matrix A_T on each element. This is then added to the global matrix in the according coordinates. Note that diagonal entries of A_T will be added only to diagonal entries of A, and off-diagonal entries of A_T will be added only to off-diagonal entries of A.

Recall the definition (A.3) of local basis functions and their gradients (in local coordinates) on the reference element:

$$\begin{aligned}
\bar{\varphi}_1(\xi,\eta) &= 1-\xi-\eta\,, & \nabla_\xi \bar{\varphi}_1 &= \begin{pmatrix}-1\\-1\end{pmatrix};\\
\bar{\varphi}_2(\xi,\eta) &= \xi\,; & \nabla_\xi \bar{\varphi}_2 &= \begin{pmatrix}1\\0\end{pmatrix};\\
\bar{\varphi}_3(\xi,\eta) &= \eta\,; & \nabla_\xi \bar{\varphi}_3 &= \begin{pmatrix}0\\1\end{pmatrix}.
\end{aligned}$$

From (A.2), we know that $\nabla_x = H^{-\top}\nabla_\xi$. The local stiffness matrix entries are then

$$\left(A_T\right)_{ij} = \int_T \nabla_x\varphi_i \cdot \nabla_x\varphi_j \,\mathrm{d}\mathbf{x} = \nabla_\xi\bar{\varphi}_i H^{-1}H^{-\top}\nabla_\xi\bar{\varphi}_j\,|T|\,.$$

First, we show that the diagonal entries of A_T are positive. Taking $i = j$, the gradient term reduces to

$$\nabla_\xi\bar{\varphi}_i H^{-1}H^{-\top}\nabla_\xi\bar{\varphi}_i = \|H^{-\top}\nabla_\xi\bar{\varphi}_i\|^2\,,$$

and as $H^{-\top}$ is invertible, this expression is positive. Further, $|T|$ is positive, so $(A_T)_{ii} > 0$ (no summation over i).

Second, we show that the off-diagonal entries of A_T are negative. A_T is symmetric, so we only need to check three cases. Before that, we transfer the condition on the angles to a more convenient form.

Let $\measuredangle(\overline{P_0P_1}, \overline{P_0P_2}) \in (-\frac{\pi}{2}, \frac{\pi}{2})$. This results in

$$\begin{aligned} (0,1] \ni \cos\measuredangle(\overline{P_0P_1}, \overline{P_0P_2}) &= \frac{(P_1-P_0)\cdot(P_2-P_0)}{\|P_1-P_0\|\,\|P_2-P_0\|} \\ \Rightarrow \quad (P_1-P_0)\cdot(P_2-P_0) &= \begin{pmatrix} h_{11} \\ h_{21} \end{pmatrix} \cdot \begin{pmatrix} h_{12} \\ h_{22} \end{pmatrix} = h_{11}h_{12} + h_{21}h_{22} > 0\,. \end{aligned}$$

Similarly, let $\measuredangle(\overline{P_1P_2}, \overline{P_1P_0}) \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Then

$$\begin{aligned} (0,1] \ni \cos\measuredangle(\overline{P_1P_2}, \overline{P_1P_0}) &= \frac{(P_2-P_0+P_0-P_1)\cdot(P_0-P_1)}{\|P_2-P_1\|\,\|P_0-P_1\|} \\ \Rightarrow \quad (P_2-P_0)\cdot(P_0-P_1) + (P_0-P_1)\cdot(P_0-P_1) & \\ = \begin{pmatrix} h_{12} \\ h_{22} \end{pmatrix} \cdot \begin{pmatrix} -h_{11} \\ -h_{21} \end{pmatrix} + \|P_1-P_0\|^2 &= -h_{11}h_{12} - h_{21}h_{22} + h_{11}^2 + h_{21}^2 > 0\,. \end{aligned}$$

Finally, let $\measuredangle(\overline{P_2P_0}, \overline{P_2P_1}) \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Then

$$\begin{aligned} (0,1] \ni \cos\measuredangle(\overline{P_2P_0}, \overline{P_2P_1}) &= \frac{(P_0-P_2)\cdot(P_1-P_0+P_0-P_2)}{\|P_0-P_2\|\,\|P_1-P_2\|} \\ \Rightarrow \quad (P_0-P_2)\cdot(P_1-P_0) + (P_0-P_2)\cdot(P_0-P_2) & \\ = \begin{pmatrix} -h_{12} \\ -h_{22} \end{pmatrix} \cdot \begin{pmatrix} h_{11} \\ h_{21} \end{pmatrix} + \|P_2-P_0\|^2 &= -h_{11}h_{12} - h_{21}h_{22} + h_{12}^2 + h_{22}^2 > 0\,. \end{aligned}$$

Further, recall from (A.4) that

$$H^{-\top} = \frac{1}{\det H} \begin{pmatrix} h_{22} & -h_{21} \\ -h_{12} & h_{11} \end{pmatrix}.$$

We can now check the three combinations of gradients: All we need to show now is that $\nabla_\xi \phi_i H^{-1} H^{-\top} \nabla_\xi \phi_j < 0$ for $i \neq j$.

$$\begin{aligned} \begin{pmatrix} 1 \\ 0 \end{pmatrix} H^{-1}H^{-\top} \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= \underbrace{\left(\frac{1}{\det H}\right)^2}_{>0} \begin{pmatrix} h_{22} \\ -h_{12} \end{pmatrix} \cdot \begin{pmatrix} -h_{21} \\ h_{11} \end{pmatrix} \\ &= (\det H)^{-2}\left(-h_{11}h_{12} - h_{21}h_{22}\right) < 0\,; \\ \begin{pmatrix} -1 \\ -1 \end{pmatrix} H^{-1}H^{-\top} \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= (\det H)^{-2} \begin{pmatrix} -h_{22}+h_{21} \\ h_{12}-h_{11} \end{pmatrix} \cdot \begin{pmatrix} -h_{21} \\ h_{11} \end{pmatrix} \\ &= (\det H)^{-2}\left(-h_{11}^2 - h_{21}^2 + h_{11}h_{12} + h_{21}h_{22}\right) < 0\,; \\ \begin{pmatrix} -1 \\ -1 \end{pmatrix} H^{-1}H^{-\top} \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= (\det H)^{-2} \begin{pmatrix} -h_{22}+h_{21} \\ h_{12}-h_{11} \end{pmatrix} \cdot \begin{pmatrix} h_{22} \\ -h_{12} \end{pmatrix} \\ &= (\det H)^{-2}\left(-h_{12}^2 - h_{22}^2 + h_{11}h_{12} + h_{21}h_{22}\right) < 0\,. \end{aligned}$$

Thus the off-diagonal entries of A_T are negative.

If we assemble the global matrix A, there will only be positive contributions to the diagonal entries, and only negative contributions to the off-diagonal entries. Thus A is a Z-matrix. Theorem 6.4.6 from [5, p.149], statement (C_8), claims that A is an M-matrix if every real eigenvalue of A is nonnegative.
By construction, A is symmetric, and all eigenvalues are real. Moreover, as $a(.,.)$ is positive semi-definite on the FE subspace over $\mathcal{T}_h$, all eigenvalues of A are nonnegative.

Let further a Dirichlet boundary part be given. Then the bilinear form $a(.,.)$ is even elliptic on the FE subspace over $\mathcal{T}_h$ due to Poincaré's inequality, so the eigenvalues of A are real and positive. □

A further result will be needed later:

Lemma 4.7:
Let A be given as in Lemma 4.6.
Then every reordering $PAP^\top$ with a permutation matrix P is again an M-matrix. Further, every block-diagonal submatrix A_{bl} of $PAP^\top$ is an M-matrix. If it is invertible, there holds

$$\left(A_{\text{bl}}\right)^{-1}_{ij} \geq 0 \quad \forall i, j\,.$$

Proof. For the first proposition, note that P and $P^\top$ will swap the same rows and columns. Then the diagonal entries of A are mapped to the diagonal of $PAP^\top$; off-diagonal entries will be mapped to off-diagonal entries again.

For the second proposition, note that a block-diagonal submatrix of $PAP^\top$ is a reordering of A with some rows and columns removed. Removing row and column i from the stiffness matrix is equivalent to taking ϕ_i out of the basis of the FE space V_h. Lemma 4.6 can be applied again with a smaller FE space.
From [5, p.134], condition (N_{38}) in Theorem 6.2.3, we know that A_{bl} is inverse-positive, i.e. its inverse has only nonnegative entries. □

4.2.3. Active set algorithm for the discrete problem

We can now state a discretized problem. For our purposes, we need to substitute the Dirichlet boundary condition by a Neumann boundary condition on a part of $\partial\Omega$. Decompose $\partial\Omega$ in two disjoint parts, $\partial\Omega = \overline{\Gamma_D \cup \Gamma_N}$, where $\operatorname{vol}\Gamma_D > 0$. (It is cleared in Remark 4.11 why we need a Dirichlet boundary.) The problem is then:

Given $\Omega \subset \mathbb{R}^2$ and $g \in L^2(\Gamma_N)$, find a scalar function $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -D\Delta u &= f + \xi \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma_D \\ \frac{\partial u}{\partial \mathbf{n}} &= g \quad \text{on } \Gamma_N \\ u &\geq \psi \quad \text{in } \Omega \end{aligned} \tag{4.15}$$

$$\begin{aligned} \text{with} \quad \xi(\mathbf{x}) &= 0\,, \quad u(\mathbf{x}) > \psi(\mathbf{x}) + \delta \\ \xi(\mathbf{x}) &= -c_J\,, \quad u(\mathbf{x}) \in \Big(\psi(\mathbf{x}),\ \psi(\mathbf{x}) + \delta\Big] \\ \xi(\mathbf{x}) &\geq -c_J\,, \quad u(\mathbf{x}) = \psi(\mathbf{x})\,. \end{aligned}$$

Selecting a finite-dimensional subspace $V_h \subset H^1_D(\Omega)$ is straightforward. We introduce a partition $\mathcal{T}_h$ of Ω into acute triangles and define V_h by piecewise linear functions on $\mathcal{T}_h$. We can create a basis ϕ^i of V_h associated with the inner nodes of $\mathcal{T}_h$:

$$\begin{aligned} V_h &:= \Big\{ v \in C^0(\Omega) : \ v|_T \in \mathbb{P}_1(T)\ \forall T \in \mathcal{T}_h\,,\ v|_{\partial\Omega} = 0 \Big\} \\ V_h &= \operatorname{span}\Big(\phi^1(x), \ldots, \phi^n(x)\Big) \quad \text{with } \phi^i(P^j_{\text{inner}}) = \delta_{ij} \end{aligned}$$

We also need a subset $M_h^+ \subset M_+$ for the Lagrange multiplier p; this can be done by setting up a finite-dimensional subspace $M_h \subset L^2(\Omega)$. M_h is piecewise constant on $\mathcal{T}_h$, and a basis ψ^i of M_h can be associated with the elements of $\mathcal{T}_h$:

$$\begin{aligned} M_h &:= \Big\{ m \in L^2(\Omega) : \ m|_T \in \mathbb{P}_0(T)\ \forall T \in \mathcal{T}_h \Big\} \\ M_h &= \operatorname{span}\Big(\psi^1(x), \ldots, \psi^k(x)\Big) \quad \text{with } \ \psi^i\big|_{T_j} = \delta_{ij} \end{aligned}$$

According to [20], we need to define a discretized active set $\mathcal{A}_p$ for the adhesion, first by an approximation $\mathcal{A}_{p,h}$ as union of elements, later by a discretization through nodal points.

We use a slight digression here. The proof of convergence for the active-set algorithm 4.9 remains true if, for our discretization, condition (44) in [20] holds. That is,

$$\vec{p}(\mathcal{A}) \geq \vec{p}(\mathcal{B}) \text{ if and only if } \mathcal{A} \supset \mathcal{B} \tag{4.16}$$

holds componentwise, where the "load vector" from adhesion is

$$\Big(\vec{p}(\mathcal{A})\Big)_i = c_J \int_\Omega \phi^i(\mathbf{x})\, \chi_{\mathcal{A}}(\mathbf{x})\, \mathrm{d}\mathbf{x}\,.$$

For this, the active set $\mathcal{A}_p$ is approximated by a union of elements $\mathcal{A}_{p,h} \subset \mathcal{A}_p$: As we have only piecewise linear functions in V_h, we may create an indicator vector $\vec{c_p}$ (of the same dimension as M_h) as follows,

- compute the values of u_h in all three corners of T_j
- if the adhesion condition is active in all points, set $\left(\vec{c}_p\right)_j$ to 1
- otherwise, set $\left(\vec{c}_p\right)_j$ to 0.

This fulfills the condition

$$T_j \subset \mathcal{A}_{p,h} \;\Leftrightarrow\; T_j \subset \mathcal{A}_p \,.$$

By refining the mesh, we have

$$\mathcal{A}_{p,h} \overset{h\to 0}{\to} \mathcal{A}_p \,.$$

If we now compute a rectangular Galerkin matrix by testing piecewise constant functions ψ against piecewise linear functions ϕ,

$$\left(M_{\text{mix}}\right)_{ij} \;=\; \int_\Omega \phi^i(\mathbf{x})\, \psi^j(\mathbf{x})\, \mathrm{d}\mathbf{x} \,,$$

we can express $\vec{p}(\mathcal{A}_p)$ by

$$\vec{p} \;=\; c_J\, M_{\text{mix}} \vec{c}_p \,.$$

The non-penetration condition $u \geq \psi$ in (4.3) is imposed in the mesh points only. If we assume ψ to be constant (i.e. the contact surface is flat), this is equivalent to the non-penetration condition $u_h(\mathbf{x}) \geq \psi$, $u_h \in V_h$, as we only use piecewise linear functions. Create an "obstacle vector" from the values of ψ in the inner points of $\mathcal{T}_h$,

$$(\vec{\psi})_i \;=\; \psi(P^i_{\text{inner}}) \,.$$

Let the coefficient vector of u_h be $\vec{u}_h$, i.e.

$$u_h(x) \;=\; \sum_i (\vec{u}_h)_i \phi^i(x) \,,$$

then the discrete non-penetration condition states

$$\vec{u}_h \;\geq\; \vec{\psi}$$

in each component.

The stiffness matrix A is defined by

$$A_{ij} \;:=\; D \int_\Omega \nabla\phi^i \cdot \nabla\phi^j \,\mathrm{d}x \,,$$

and the load vector f is

$$f_i \;:=\; \int_\Omega f\phi^i \,\mathrm{d}x \;+\; \int_{\Gamma_N} g\phi^i \,\mathrm{d}s_x \,.$$

Now we introduce a discrete Lagrange multiplier $\vec{\lambda}$ for the pointwise non-penetration condition. Then the discretization of (4.10) reads

$$\begin{aligned} A\vec{u}_h \; - \; \vec{\lambda} \; &= \vec{f} \; - \; \vec{p}(\mathcal{A}_p) \\ (\vec{u}_h)_i \; &= (\vec{\psi})_i \quad \text{for } P^i_{\text{inner}} \in \mathcal{A}_c \\ (\vec{\lambda})_i \; &= 0 \quad \text{for } P^i_{\text{inner}} \notin \mathcal{A}_c \; , \end{aligned} \tag{4.17}$$

or, rewritten as a linear subproblem with transient active sets $\mathcal{A}_c^{(n)}$ and $\mathcal{A}_p^{(n)}$,

$$\begin{aligned} A\vec{u}_h^{(n)} \; - \; \vec{\lambda}^{(n)} \; &= \vec{f} \; - \; \vec{p}(\mathcal{A}_p^{(n-1)}) \\ \left(\vec{u}_h^{(n)}\right)_i \; &= (\vec{\psi})_i \quad \text{for } P^i_{\text{inner}} \in \mathcal{A}_c^{(n-1)} \\ \left(\vec{\lambda}^{(n)}\right)_i \; &= 0 \quad \text{for } P^i_{\text{inner}} \notin \mathcal{A}_c^{(n-1)} \; . \end{aligned} \tag{4.18}$$

The active set algorithm for the discrete problem is then:

Algorithm 4.8:
Initialize the set $\mathcal{A}_{p,h}^{(0)} := \mathcal{T}_h$ and $\mathcal{A}_{c,h}^{(0)}$ arbitrarily. Then iterate the following steps:

1. Solve the linear subproblem (4.18) with $\mathcal{A}_{p,h}^{(n-1)}$ and $\mathcal{A}_{c,h}^{(n-1)}$.
2. Update $\mathcal{A}_{p,h}^{(n)}$ and $\mathcal{A}_{c,h}^{(n)}$ using the solution vectors $\vec{\lambda}^{(n)}$ and $\vec{u}_h^{(n)}$.
3. If one of the active sets changed in this step, continue; otherwise, stop.

♦

Lemma 4.9:
Let the stiffness matrix A be an M-matrix.
Then the preceding algorithm converges to a solution $(\vec{u}_h^*, \vec{\lambda}^*, \vec{p}^*)$ of (4.17) in a finite number of steps. Moreover, the algorithm has the following monotonicity invariants:

$$\vec{\psi} \; \leq \; \vec{u}_h^{(2)} \; \leq \; \cdots \; \leq \; \vec{u}_h^{(n)} \; \leq \; \cdots \; \leq \; \vec{u}_h^* \tag{4.19}$$

$$\mathcal{T}_h \; \supseteq \; \mathcal{A}_{c,h}^{(1)} \; \supseteq \; \cdots \; \supseteq \; \mathcal{A}_{c,h}^{(n)} \; \supseteq \; \cdots \; \supseteq \; \mathcal{A}_{c,h}^* \tag{4.20}$$

$$c_J = \vec{p}^{(1)} \; \geq \; \vec{p}^{(2)} \; \geq \; \cdots \; \geq \; \vec{p}^{(n)} \; \geq \; \cdots \geq \vec{p}^* \tag{4.21}$$

$$\Omega \; = \; \mathcal{A}_{p,h}^{(0)} \; \supseteq \; \mathcal{A}_{p,h}^{(1)} \; \supseteq \; \cdots \; \supseteq \mathcal{A}_{p,h}^{(n)} \; \supseteq \; \cdots \supseteq \mathcal{A}_{p,h}^* \tag{4.22}$$

Proof. This proof follows the proof of Theorem 2 in [20]. For readability, we drop the index h here.
Define the three vectors

$$\vec{\delta}_u^{(n-1)} \; := \; \vec{u}^{(n)} - \vec{u}^{(n-1)} \, , \quad \vec{\delta}_\lambda^{(n-1)} \; := \; \vec{\lambda}^{(n)} - \vec{\lambda}^{(n-1)} \, , \quad \vec{\delta}_p^{(n-1)} \; := \; M_{\text{mix}}(\vec{c}_p^{(n)} - \vec{c}_p^{(n-1)})$$

For a fixed n, we can deduce the following:
If $\mathcal{A}_p^{(n)} \subseteq \mathcal{A}_p^{(n-1)}$, we know that $\vec{\delta}_p^{(n)} \leq 0$ due to (4.16).

As $\vec{f}$ stays constant during the iteration, we get from (4.18) that

$$A\delta_u^{(n)} \;=\; \delta_\lambda^{(n)} - \delta_p^{(n)} \,. \tag{4.23}$$

Now let two selection matrices for the active contact set $\mathcal{A}_c^{(n)}$ be given, $P_+^{(n)}$ and $P_-^{(n)}$. These are defined as follows:
The matrix $P_+^{(n)} : \mathbb{R}^n \to \mathbb{R}^{n_A}$ (where n_A is the number of active nodes in contact) selects only degrees of freedom where the constraint from A_c is active. Its entries are only 0 or 1, each line contains exactly one 1, thus the selection matrix is surjective.
The matrix $P_-^{(n)} : \mathbb{R}^n \to \mathbb{R}^{n-n_A}$ selects all degrees of freedom where the constraint is inactive. This matrix also contains exactly one 1 per line, has otherwise 0 as entry, and is surjective.
Attaching both matrices together results in a permutation matrix $P_{(n)} : \mathbb{R}^n \to \mathbb{R}^n$:

$$\begin{pmatrix} P_+^{(n)} \\ P_-^{(n)} \end{pmatrix} \;=\; P_{(n)} \;=\; (P_{(n)}^\top)^{-1}$$

$P_{(n)}$ can now sort rows and columns in A:

$$P_{(n)} A P_{(n)}^\top \;=\; \begin{pmatrix} A_{++} & A_{+-} \\ A_{-+} & A_{--} \end{pmatrix}$$

As the variables are reordered, we now have decompositions of $\vec{\delta}_u^{(n)}$, $\vec{\delta}_\lambda^{(n)}$ and $\vec{\delta}_p^{(n)}$ into two subvectors each:

$$P_{(n)}\vec{\delta}_u^{(n)} \;=\; \begin{pmatrix} \delta_+^u \\ \delta_-^u \end{pmatrix} \qquad P_{(n)}\vec{\delta}_\lambda^{(n)} \;=\; \begin{pmatrix} \delta_+^\lambda \\ \delta_-^\lambda \end{pmatrix} \qquad P_{(n)}\vec{\delta}_p^{(n)} \;=\; \begin{pmatrix} \delta_+^p \\ \delta_-^p \end{pmatrix} ,$$

omitting the vector arrow and the index $\cdot^{(n)}$ for readability.

As P is invertible, we can multiply the linear system (4.23) by P from the left to get

$$PAP^\top P\delta_u^{(n)} \;=\; P\delta_\lambda^{(n)} - P\delta_p^{(n)} \,,$$

or in expansion

$$\begin{pmatrix} A_{++} & A_{+-} \\ A_{-+} & A_{--} \end{pmatrix} \begin{pmatrix} \delta_+^u \\ \delta_-^u \end{pmatrix} \;=\; \begin{pmatrix} \delta_+^\lambda \\ \delta_-^\lambda \end{pmatrix} - \begin{pmatrix} \delta_+^p \\ \delta_-^p \end{pmatrix} . \tag{4.24}$$

The second line can be re-ordered to state

$$\begin{aligned} A_{++}\delta_-^u \;&=\; -A_{-+}\delta_+^u \;+\; \delta_+^\lambda \;-\; \delta_-^p \\ \Leftrightarrow \qquad \delta_-^u \;&=\; -A_{++}^{-1}A_{-+}\delta_+^u \;+\; A_{++}^{-1}\Big(\delta_+^\lambda \;-\; \delta_-^p\Big) \end{aligned} \tag{4.25}$$

We already know that $\delta_+^u \geq 0$, because $\vec{u}^{(n)}$ equals $\vec{\psi}$ where the contact is active. Similarly, $\delta_-^\lambda \geq 0$ because $\vec{\lambda}^{(n)}$ equals zero where the contact is inactive.
The matrix A_{-+} is an off-diagonal block submatrix of A, which is an M-matrix. This means that all entries of A_{-+} are nonpositive. Using Lemma 4.7, we further notice that A_{++}^{-1} has only nonnegative entries. We get

$$\begin{aligned} & \delta_p^{(n)} \leq 0 \\ \Rightarrow \quad \delta_-^u \;=\; - \underbrace{A_{++}^{-1}}_{\geq 0} \underbrace{A_{-+}}_{\leq 0} \underbrace{\delta_+^u}_{\geq 0} \;+\; \underbrace{A_{++}^{-1}}_{\geq 0} \Big(\underbrace{\delta_+^\lambda}_{\geq 0} \;-\; \underbrace{\delta_-^p}_{\geq 0} \Big) \;\geq\; 0\,. \end{aligned} \tag{4.26}$$

It remains to show that $\delta_p^{(n)} \leq 0$.

We can start an induction at $n = 1$: We know that $\mathcal{A}_p^{(0)} = \mathcal{T}_h$, so $\mathcal{A}_p^{(1)}$ can not be larger. But then (4.16) again gives $\vec{p}^{(1)} \geq \vec{p}^{(2)}$.
Assume that in induction step n, there holds $\mathcal{A}_p^{(n)} \subseteq \mathcal{A}_p^{(n-1)}$. This results in $\vec{\delta}_p^{(n)} \leq 0$. From (4.26) follows that $\vec{\delta}_u^{(n)} \geq 0$, which immediately implies $\mathcal{A}_p^{(n+1)} \subseteq \mathcal{A}_p^{(n)}$, or $\vec{\delta}_p^{(n+1)} \leq 0$.

Finally, the algorithm converges in a finite number of steps: Both active sets $\mathcal{A}_c$ and $\mathcal{A}_p$ can only grow smaller in each step. As soon as both sets were not changed, the algorithm is converged. As we can only remove $\#\mathcal{T}_h$ elements from $\mathcal{A}_p$ and $\#\{P_{\text{free}}\}$ nodes from $\mathcal{A}_c$ before both sets are empty, the algorithm will terminate after at most $\#\mathcal{T}_h + \#\{P_{\text{free}}\} + 1$ steps. □

Remark 4.10: *Because the M-matrix property still holds if a Neumann boundary is present, the proof of Lemma 4.9 does not need to be modified. This is different in Lemma 4.5, where the maximum principle for δ_-^u is applied in the original proof with $\delta_-^u = 0$ on $\partial\Omega \cap \partial\mathcal{I}_c^{(n-1)}$. It would remain to be proven that $\delta_-^u \geq 0$ on $\Gamma_N \cap \partial\mathcal{I}_c^{(n-1)}$.*

The linear subproblem (4.18) in Algorithm 4.8 can be solved as follows:
Using the matrix decomposition of A with P_+ and P_- from the proof of Lemma 4.9, we can multiply the linear equation system

$$A\vec{u}_h - \vec{\lambda} \;=\; \vec{f} - \vec{p}$$

with P from the left to get the equivalent system

$$\Leftrightarrow \quad PAP^\top P\vec{u}_h - P\vec{\lambda} \;=\; P\vec{f} - P\vec{p}\,.$$

Problem (4.18) fixes some coefficients of $\vec{u}_h$ and $\vec{\lambda}$. This can be expressed by introducing the following decompositions:

$$P\vec{u}_h \;=\; \begin{pmatrix} P_+ \\ P_- \end{pmatrix} \vec{u}_h \;=\; \begin{pmatrix} P_+\vec{\psi} \\ \vec{u}_- \end{pmatrix}; \qquad\qquad P\vec{\lambda} \;=\; \begin{pmatrix} P_+ \\ P_- \end{pmatrix} \vec{\lambda} \;=\; \begin{pmatrix} \vec{\lambda}_+ \\ 0 \end{pmatrix}.$$

Using the representation of $PAP^\top$ in the proof of Lemma 4.9, we get the equivalent system

$$\Leftrightarrow \quad \begin{pmatrix} A_{++} & A_{+-} \\ A_{-+} & A_{--} \end{pmatrix} \begin{pmatrix} P_+\vec{\psi} \\ \vec{u}_- \end{pmatrix} - \begin{pmatrix} \vec{\lambda}_+ \\ 0 \end{pmatrix} = \begin{pmatrix} P_+ \\ P_- \end{pmatrix} \vec{f} - \begin{pmatrix} P_+ \\ P_- \end{pmatrix} \vec{p}\,, \tag{4.27}$$

which can now be decomposed into

$$A_{--}\vec{u}_- = P_-\vec{f} - P_-\vec{p} - A_{-+}P_+\vec{\psi} \tag{4.28}$$

$$\vec{\lambda}_+ = A_{+-}\vec{u}_- - P_+\vec{f} + P_+\vec{p} + A_{++}P_+\vec{\psi}\,. \tag{4.29}$$

We can now solve first for $\vec{u}_-$ and insert this solution into the second equation. Finally, the solution vectors $\vec{u}_h$ and $\vec{\lambda}$ can be restored by

$$\vec{u}_h = P^\top \begin{pmatrix} P_+\vec{\psi} \\ \vec{u}_- \end{pmatrix} = P_+^\top P_+\vec{\psi} + P_-^\top \vec{u}_-\,; \quad \vec{\lambda}_h = P^\top \begin{pmatrix} \vec{\lambda}_+ \\ 0 \end{pmatrix} = P_+^\top \vec{\lambda}_+\,.$$

Remark 4.11: *The linear subproblem has a unique solution: The matrix A_{--} is a diagonal sub-block of $PAP^\top$, i.e. it is the Galerkin matrix on a subspace $\tilde{V}_h \subset V_h$ with the basis $\{\phi^{(1)}, \ldots, \phi^{(n)}\} \subset \{\phi^1, \ldots, \phi^n\}$. As $a(.,.)$ is positive definite on V_h, it is also positive definite on $\tilde{V}_h$.*

We may even relax the condition $\mathrm{vol}\,\Gamma_D > 0$. If we assume that A_{--} is a strict sub-block of $PAP^\top$, it is positive definite:
The classical formulation for the pure Neumann problem reads

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ \frac{\partial u}{\partial \mathbf{n}} &= g \quad \text{on } \partial\Omega\,. \end{aligned}$$

The bilinear form $a(.,.)$ is elliptic on $H^1(\Omega)/\mathbb{R}$. But then, it is also elliptic on a finite-dimensional subspace $V_h/\mathbb{R} \subset H^1(\Omega)/\mathbb{R}$, and we get the unique solution $u^* \in V_h/\mathbb{R}$, respectively the affine solution set $\{u^* + c, c \in \mathbb{R}\}$.
The FE solution states that $u_h(P^i) = (\vec{u}_h)_i$ for all mesh points P^i and the solution vector $\vec{u}_h$ from the linear equation system. If we fix one degree of freedom, this means we remove one function ϕ^k from the basis of V_h. We get

$$V_h = \tilde{V}_h \oplus \mathit{span}\,\{\phi^k\} = \tilde{V}_h \oplus \mathbb{R}\,.$$

Then $a(.,.)$ is elliptic on $\tilde{V}_h$. Removing more degrees of freedom corresponds to taking subspaces $\subset \tilde{V}_h$, where $a(.,.)$ remains elliptic.

The condition $\mathrm{vol}\,\Gamma_D > 0$ can then be relaxed to the condition

$$\mathcal{A}_c^{(n)} \neq \emptyset \quad \forall n\,,$$

the membrane must be in contact in at least one point in every iteration step n.

4.3. Subproblems in Ω^2

First, change some of the notation to improve readability: Let ω be the (x_1, x_2) shape of Ω^2 such that

$$\Omega^2 \;=\; \omega \times [0, h]\,.$$

Further, write Γ_D instead of $\Gamma_D \cap \partial\Omega^2$ and Γ_N instead of $\Gamma_N \cap \partial\Omega^2$. The 2D equivalents of these boundaries are then

$$\gamma_D \;:\; \gamma_D \times [0, h] \;=\; \Gamma_D\,; \qquad \gamma_N \;:\; \gamma_N \times [0, h] \;=\; \Gamma_N\,.$$

Denote the upper boundary by $\omega \times \{h\} := \Gamma_\uparrow$ and the lower boundary by $\omega \times \{0\} := \Gamma_\downarrow$. The prescribed traction from the problem in Ω^1 on the transmission boundary $\Gamma_\uparrow$ is $\left(\sigma(\mathbf{u}) \cdot \mathbf{n}\right)^1$.

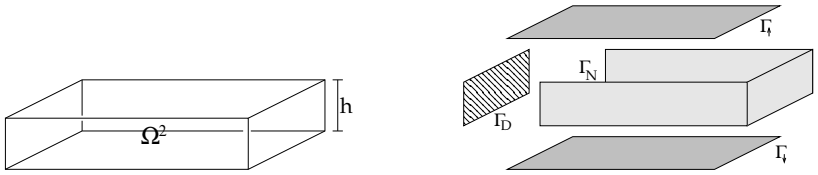


Figure 4.2.: Domain Ω^2 of the lower subproblems; new boundary markers

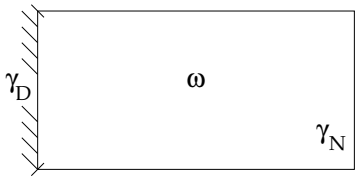


Figure 4.3.: Membrane ω with boundary markers

The original subproblem on Ω^2 can then be written in the following form:
Find $\mathbf{u} = (u_1, u_2, u_3)$ such that

$$\begin{aligned}
-\operatorname{div}\sigma(\mathbf{u}) &= \mathbf{f} \quad \text{in } \Omega^2 \\
\mathbf{u} &= 0 \quad \text{on } \Gamma_D \\
\sigma(\mathbf{u}) \cdot \mathbf{n} &= \mathbf{t} \quad \text{on } \Gamma_N \\
\sigma(\mathbf{u}) \cdot \mathbf{n} &= \left(\sigma(\mathbf{u}) \cdot \mathbf{n}\right)^1 \quad \text{on } \Gamma_\uparrow \\
\sigma_N(\mathbf{u}) &= b(u_N)\,, \\
\sigma_t(\mathbf{u}) &= 0 \quad \text{on } \Gamma_\downarrow
\end{aligned} \tag{4.30}$$

When Ω^2 is thin (h is small), one can argue through the Saint-Venant principle that the normal component of exterior tension on $\Gamma_\uparrow$ and $\Gamma_\downarrow$ may be approximated by a volume force. This is done by moving boundary stresses on test volumes into the interior of Ω^2; taking $h \to 0$ will improve the quality of this approximation.
We then can set the exterior tensions on $\Gamma_\uparrow$ and $\Gamma_\downarrow$ to zero in the x_3 component. Note that replacing the tension term $\sigma_N(\mathbf{u}) = b(u_N)$ by a volume force, we get $\sigma_N(\mathbf{u}) = 0$ on $\Gamma_\downarrow$. Combined with $\boldsymbol{\sigma}_t(\mathbf{u}) = 0$, this gives $\boldsymbol{\sigma} \cdot \mathbf{n} = 0$ on $\Gamma_\downarrow$.

Decompose the traction from Ω^1 into two vectors:

$$\Big(\sigma(\mathbf{u}) \cdot \mathbf{n}\Big)^1 = \begin{pmatrix} s_1 \\ s_2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ s_3 \end{pmatrix} = \mathbf{s}_{12} + s_3 \mathbf{e}_3$$

The new problem now reads as follows:
Find $\mathbf{u} = (u_1, u_2, u_3)$ such that

$$\begin{aligned} -\operatorname{div} \sigma(\mathbf{u}) &= \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} + \frac{1}{h} \begin{pmatrix} 0 \\ 0 \\ s_3^\uparrow - b(u_N) \end{pmatrix} \quad \text{in } \Omega^2 \\ \sigma(\mathbf{u}) \cdot \mathbf{n} &= \mathbf{t} \quad \text{on } \Gamma_N \\ \mathbf{u} &= 0 \quad \text{on } \Gamma_D \\ \sigma(\mathbf{u}) \cdot \mathbf{n} &= \mathbf{s}_{12} \quad \text{on } \Gamma_\uparrow \\ \sigma(\mathbf{u}) \cdot \mathbf{n} &= 0 \quad \text{on } \Gamma_\downarrow . \end{aligned} \tag{4.31}$$

Here, $s_3^\uparrow(x_1, x_2, x_3) := s_3(x_1, x_2, h)$ is the continuation from $\Gamma_\uparrow$ into Ω^2. The normal vector on $\Gamma_{\uparrow;\downarrow}$ is $\pm\mathbf{e}_3$, which changes the sign of $b(u_N)$ in the volume force term.

As $u_{3,3} = 0$ in Ω^2, the stress tensor becomes

$$\sigma(\mathbf{u}) = \begin{pmatrix} (\lambda + 2\mu)u_{1,1} + \lambda u_{2,2} & \mu(u_{1,2} + u_{2,1}) & \mu(u_{1,3} + u_{3,1}) \\ \mu(u_{1,2} + u_{2,1}) & (\lambda + 2\mu)u_{2,2} + \lambda u_{1,1} & \mu(u_{2,3} + u_{3,2}) \\ \mu(u_{1,3} + u_{3,1}) & \mu(u_{2,3} + u_{3,2}) & \lambda(u_{1,1} + u_{2,2}) \end{pmatrix}, \tag{4.32}$$

and we get the PDE system

$$-\Big((\lambda + \mu)(u_{1,11} + u_{2,12}) + \mu\Delta u_1\Big) = f_1 \tag{4.33}$$

$$-\Big((\lambda + \mu)(u_{1,12} + u_{2,22}) + \mu\Delta u_2\Big) = f_2 \tag{4.34}$$

$$-\Big((\lambda + \mu)(u_{1,13} + u_{2,23}) + \mu\Delta u_3\Big) = f_3 + \frac{1}{h}(s_3^\uparrow - b(u_N)) \tag{4.35}$$

for the volume term in (4.31). Note that the Laplacian in the third equation, Δu_3, is actually only the 2D Laplacian, as $u_{3,33} = 0$.

This system is now decoupled into a first part with the governing equations (4.33) and (4.34) and a second part under equation (4.35). Both problems will be used in a

staggered iteration, assuming the other function to be given respectively. For this, it is useful to write down partial strain and stress tensors:
For the $(u_1, u_2, 0)$ function, we get

$$\nabla \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix} = \begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ u_{2,1} & u_{2,2} & u_{2,3} \\ 0 & 0 & 0 \end{pmatrix} \quad \Rightarrow \quad \boldsymbol{\varepsilon} \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix} = \begin{pmatrix} u_{1,1} & \frac{1}{2}(u_{2,1} + u_{1,2}) & \frac{1}{2}u_{1,3} \\ \frac{1}{2}(u_{2,1} + u_{1,2}) & u_{2,2} & \frac{1}{2}u_{2,3} \\ \frac{1}{2}u_{1,3} & \frac{1}{2}u_{2,3} & 0 \end{pmatrix} \tag{4.36}$$

$$\boldsymbol{\sigma} \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda(u_{1,1} + u_{2,2}) + 2\mu u_{1,1} & \mu(u_{2,1} + u_{1,2}) & \mu u_{1,3} \\ \mu(u_{2,1} + u_{1,2}) & \lambda(u_{1,1} + u_{2,2}) + 2\mu u_{2,2} & \mu u_{2,3} \\ \mu u_{1,3} & \mu u_{2,3} & \lambda(u_{1,1} + u_{2,2}) \end{pmatrix} \tag{4.37}$$

$$-\operatorname{div} \boldsymbol{\sigma} \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix} = \begin{pmatrix} -(\lambda + \mu)(u_{1,11} + u_{2,12}) - \mu \Delta u_1 \\ -(\lambda + \mu)(u_{1,12} + u_{2,22}) - \mu \Delta u_2 \\ -(\lambda + \mu)(u_{1,13} + u_{2,23}) \end{pmatrix}. \tag{4.38}$$

For the $(0, 0, u_3)$ function, we get

$$\nabla \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ u_{3,1} & u_{3,2} & 0 \end{pmatrix} \quad \Rightarrow \quad \boldsymbol{\varepsilon} \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \frac{1}{2}u_{3,1} \\ 0 & 0 & \frac{1}{2}u_{3,2} \\ \frac{1}{2}u_{3,1} & \frac{1}{2}u_{3,2} & 0 \end{pmatrix} \tag{4.39}$$

$$\boldsymbol{\sigma} \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \mu u_{3,1} \\ 0 & 0 & \mu u_{3,2} \\ \mu u_{3,1} & \mu u_{3,2} & 0 \end{pmatrix} \tag{4.40}$$

$$-\operatorname{div} \boldsymbol{\sigma} \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\mu \Delta u_3 \end{pmatrix}, \tag{4.41}$$

where again Δ reduces to the 2D Laplacian.

The first two governing equations, (4.33) and (4.34), then provide

$$-\operatorname{div} \boldsymbol{\sigma} \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ -(\lambda + \mu)(u_{1,13} + u_{2,23}) \end{pmatrix}, \tag{4.42}$$

where the third component is only repeated on the right side. Likewise, the third governing equation (4.35) then provides

$$-\operatorname{div} \boldsymbol{\sigma} \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\mu \Delta u_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ f_3 + \frac{1}{h}(s_3^{\uparrow} - b(u_N)) + (\lambda + \mu)(u_{1,13} + u_{2,23}) \end{pmatrix}, \tag{4.43}$$

where the first two components simply state $0 = 0$ here.

Moving known functions to the right side, we get the PDE for the first subproblem,

$$-\operatorname{div} \boldsymbol{\sigma} \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 + \frac{1}{h}(s_3^{\uparrow} - b(u_N)) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \mu \Delta u_3 \end{pmatrix} \tag{4.44}$$

and the PDE for the second subproblem,

$$-\mu\Delta u_3 \;= f_3 + \frac{1}{h}(s_3^{\uparrow} - b(u_N)) + (\lambda+\mu)(u_{1,13} + u_{2,23})\,. \tag{4.45}$$

If we otherwise add (4.42) and (4.43), we arrive again at (4.31).

4.3.1. First subproblem: $(u_1, u_2, 0)$

Let us first assume that u_3 is given. As we are now looking for a solution $(u_1, u_2, 0)$ satisfying (4.44), the appropriate solution space is $V := \left[H^1_D(\Omega^2)\right]^2 \times \{0\}$. We get the variational formulation:
Find $(u_1, u_2, 0) \in V$ such that

$$\begin{aligned}\int_{\Omega^2} \boldsymbol{\sigma}\begin{pmatrix}u_1\\u_2\\0\end{pmatrix} : \boldsymbol{\varepsilon}(\mathbf{v})\,\mathrm{d}x \;&=\; \int_{\Omega^2} \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}x \;+\; \frac{1}{h}\int_{\Omega^2}\begin{pmatrix}0\\0\\(s_3^{\uparrow} - b(u_N))\end{pmatrix}\cdot\mathbf{v}\,\mathrm{d}x \\ &\quad+\; \mu\int_{\Omega^2}\begin{pmatrix}0\\0\\\Delta u_3\end{pmatrix}\cdot\mathbf{v}\,\mathrm{d}x \;+\; \int_{\partial\Omega^2}\boldsymbol{\sigma}\begin{pmatrix}u_1\\u_2\\0\end{pmatrix}\cdot\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}s_x \qquad \forall \mathbf{v}\in V\,.\end{aligned} \tag{4.46}$$

The second rhs integral is zero by construction of V, and the volume force term f_3 will also be ignored in the first integral. These terms are taken care of in the second subproblem.

The term Δu_3 is ill-defined if we only demand $u_3 \in H^1_D(\Omega^2)$. It will nevertheless vanish if we apply the divergence theorem: Use (4.39) to write

$$\begin{pmatrix}0\\0\\\Delta u_3\end{pmatrix} \;=\; \operatorname{div}\begin{pmatrix}0&0&0\\0&0&0\\u_{3,1}&u_{3,2}&0\end{pmatrix} \;=\; \operatorname{div}\nabla\begin{pmatrix}0\\0\\u_3\end{pmatrix}\,.$$

The product rule of differentiation can be extended to products of tensors, see e.g. [3]: If A is a tensor of order 2 and $\mathbf{b}$ is a tensor of order 1, we get

$$\operatorname{div}(A^{\top}\mathbf{b}) \;=\; \mathbf{b}\cdot\operatorname{div}A \;+\; A : (\nabla\mathbf{b})\,.$$

Together with the divergence theorem, we get

$$\begin{aligned}&\int_{\Omega^2}\left(\operatorname{div}\nabla\begin{pmatrix}0\\0\\u_3\end{pmatrix}\right)\cdot\mathbf{v}\,\mathrm{d}x \;=\; \int_{\Omega^2}\operatorname{div}\left(\left(\nabla\begin{pmatrix}0\\0\\u_3\end{pmatrix}\right)^{\top}\cdot\mathbf{v}\right)\mathrm{d}x \;-\; \int_{\Omega^2}\left(\nabla\begin{pmatrix}0\\0\\u_3\end{pmatrix}\right) : \nabla\mathbf{v}\,\mathrm{d}s_x \\ &= \int_{\partial\Omega^2}\mathbf{n}\cdot\begin{pmatrix}0&0&u_{3,1}\\0&0&u_{3,2}\\0&0&0\end{pmatrix}\cdot\begin{pmatrix}v_1\\v_2\\0\end{pmatrix}\mathrm{d}s_x \;+\; \int_{\Omega^2}\begin{pmatrix}0&0&0\\0&0&0\\u_{3,1}&u_{3,2}&0\end{pmatrix} : \begin{pmatrix}v_{1,1}&v_{1,2}&v_{1,3}\\v_{2,1}&v_{2,2}&v_{2,3}\\0&0&0\end{pmatrix}\mathrm{d}x \\ &= 0\,.\end{aligned}$$

The last integral in the variational formulation (4.46) still has a $\sigma(u_1, u_2, 0) \cdot \mathbf{n}$ part, but we want $\sigma(u_1, u_2, u_3) \cdot \mathbf{n}$, which equals $\mathbf{s}_{12}$ on $\Gamma_\uparrow$, $\mathbf{t}$ on Γ_N and 0 on $\Gamma_\downarrow$. We add and subtract a correction part $\sigma(0, 0, u_3) \cdot \mathbf{n}$ and note that

$$\sigma\begin{pmatrix}0\\0\\u_3\end{pmatrix}\cdot\mathbf{v}\cdot\mathbf{n} \;=\; \mathbf{n}\cdot\begin{pmatrix}0 & 0 & \mu u_{3,1}\\ 0 & 0 & \mu u_{3,2}\\ \mu u_{3,1} & \mu u_{3,2} & 0\end{pmatrix}\cdot\mathbf{v} \;=\; \mathbf{n}\cdot\begin{pmatrix}0\\0\\ \mu u_{3,1}v_1 \,+\, \mu u_{3,2}v_2\end{pmatrix}.$$

This contribution is only nonzero on $\Gamma_\uparrow \cup \Gamma_\downarrow$, where the normal vector $\mathbf{n}$ has a nozero third component.

We get the final variational formulation:
Find $(u_1, u_2, 0) \in V$ such that

$$\begin{aligned}\int_{\Omega^2} \sigma\begin{pmatrix}u_1\\u_2\\0\end{pmatrix} : \varepsilon(\mathbf{v})\,\mathrm{d}x \;=\; &\int_{\Omega^2} \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}x \;+\; \int_{\Gamma_N} \mathbf{t}\cdot\mathbf{v}\,\mathrm{d}s_x \;+\; \int_{\Gamma_\uparrow} \mathbf{s}_{12}\cdot\mathbf{v}\,\mathrm{d}s_x \\ &- \mu\int_{\Gamma_\uparrow}\left(u_{3,1}v_1 + u_{3,2}v_2\right)\mathrm{d}s_x \;+\; \mu\int_{\Gamma_\downarrow}\left(u_{3,1}v_1 + u_{3,2}v_2\right)\mathrm{d}s_x \;\forall \mathbf{v}\in V\,.\end{aligned} \tag{4.47}$$

4.3.2. Second subproblem: $(0, 0, u_3)$

Now assume that u_1 and u_2 are given.

In equation (4.45), u_3 only depends on x_1 and x_2. This also holds for $s_3^\uparrow$ and $b(u_N)$, which are just extensions from the surfaces $\Gamma_\uparrow$ and $\Gamma_\downarrow$. As we assumed that f_3, $u_{1,13}$ and $u_{2,23}$ are continuous in Ω^2, they may be approximated up to order h by taking $f_3(x_1, x_2, h)$ or $f_3(x_1, x_2, 0)$ (evaluating them on $\Gamma_\uparrow$ or $\Gamma_\downarrow$). But then, (4.45) is effectively reduced to a 2D problem in ω.

As we are looking for a solution $u_3 \in H^1_D(\omega)$, we may multiply by a test function $v \in H^1_D(\omega)$ and integrate by parts to get

$$\begin{aligned}\mu\int_\omega \nabla u_3\cdot\nabla v\,\mathrm{d}x \;=\; &\frac{1}{h}\int_\omega s_3 v\,\mathrm{d}x \;+\; \int_\omega f_3 v\,\mathrm{d}x \;-\; \frac{1}{h}\int_\omega b(u_N)v\,\mathrm{d}x \;+\; \mu\int_{\gamma_N}\frac{\partial u_3}{\partial \mathbf{n}}v\,\mathrm{d}s_x \\ &+\; (\lambda+\mu)\int_\omega (u_{1,13}+u_{2,23})v\,\mathrm{d}x \qquad \forall v\in H^1_D(\omega)\,.\end{aligned} \tag{4.48}$$

Now

$$u_{1,13} + u_{2,23} \;=\; \operatorname{div}\begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix}.$$

We can apply the divergence theorem on the last integral:

$$\begin{aligned}\int_\omega \operatorname{div}\begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix} v\,\mathrm{d}x &= \int_\omega \operatorname{div}\Big(\begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix} v\Big)\,\mathrm{d}x - \int_\omega \begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix}\cdot\nabla v\,\mathrm{d}x\\ &= \int_{\gamma_N}\begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix}\cdot\mathbf{n}\,v\,\mathrm{d}s_x - \int_\omega \begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix}\cdot\nabla v\,\mathrm{d}x\,. \end{aligned} \tag{4.49}$$

We still need to adapt the Neumann condition, i.e. $\frac{\partial u_3}{\partial \mathbf{n}}$. The normal vector $\mathbf{n}$ will be in the $\mathbf{x}_3 = 0$ plane, so we need

$$\boldsymbol{\sigma}\cdot\mathbf{n} = n_1\big(\sigma_{i1}\big)_i + n_2\big(\sigma_{i2}\big)_i \overset{!}{=} \mathbf{t} \qquad \textit{(no summation)}$$

on Γ_N in (4.31).
Derivatives of u_3 appear only in the last row of the stress tensor (4.32). Comparing the last component in terms of u_i, we get

$$t_3 \overset{!}{=} \mu\big(n_1(u_{1,3}+u_{3,1}) + n_2(u_{2,3}+u_{3,2})\big)$$

or equivalently

$$\nabla u_3\cdot\mathbf{n} = \frac{1}{\mu}t_3 - \begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix}\cdot\mathbf{n}\,.$$

As we assumed that t_3 is smooth on Γ_N, we can follow that $t_3(x_1,x_2,x_3) = t_3(x_1,x_2,0) + O(x_3)$. Substituting

$$\tilde{t}_3(x_1,x_2) := t_3(x_1,x_2,h)\,,$$

the boundary condition on Γ_N is independent of x_3, so we can restrict it to γ_N.

The final variational formulation is then:
Find $u_3 \in H^1_D(\omega)$ such that

$$\begin{aligned}\mu\int_\omega \nabla u_3\cdot\nabla v\,\mathrm{d}x &= \frac{1}{h}\int_\omega s_3 v\,\mathrm{d}x + \int_\omega f_3 v\,\mathrm{d}x - \frac{1}{h}\int_\omega b(u_N)v\,\mathrm{d}x\\ &\quad + \int_{\gamma_N}\tilde{t}_3 v\,\mathrm{d}s_x + \lambda\int_{\gamma_N}\begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix}\mathbf{n}\,v\,\mathrm{d}s_x \\ &\quad - (\lambda+\mu)\int_\omega\begin{pmatrix}u_{1,3}\\u_{2,3}\end{pmatrix}\cdot\nabla v\,\mathrm{d}x \qquad \forall v\in H^1_D(\omega)\,.\end{aligned} \tag{4.50}$$

(Note that the Neumann boundary condition cancelled out the μ part of the divergence contribution in (4.49).)

Remark 4.12: *Problem* (4.47) *depends on the derivatives of* u_3*. Likewise, problem* (4.50) *depends on derivatives of* u_1 *and* u_2*. We can take these values from an iterate solution in* Ω^1*, or from the respectively other problem in* Ω^2*.*

If we would make the further assumption

$$u_{1,3} = u_{2,3} = 0 \quad \textit{in } \Omega^2\,,$$

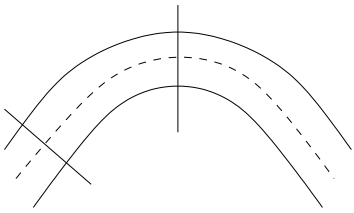


Figure 4.4.: Mindlin-Reissner hypothesis: Rotation of the normal vector for a bending plate

the last two integrals in (4.50) *would be zero. We then could solve* (4.50) *first. The other subproblem would reduce to a problem in* ω *as well, and the last integral would also vanish. This assumption would, however, violate the Mindlin-Reissner hypothesis [6, VI.6], where* u_1 *and* u_2 *explicitly depend on* x_3*. Figure 4.4 demonstrates this: The thickness* h *of a plate under load remains constant, but* u_1 *and* u_2 *may vary along the thickness. The domain* ω *under deformation is represented by the dashed middle line. The left cut exposes that the transversal displacements depend on* x_3 *when extended along the middle surface's normal.*

4.4. Solution algorithm

Using the decomposition into three subproblems, the solution algorithm now is as follows:

Algorithm 4.13:
Select the damping parameters ϑ_0 for the problem in Ω^1, ϑ_1 for the membrane subproblem, and ϑ_2 for the $(u_1, u_2, 0)$ subproblem in Ω^2.
Denote the vectors of unknowns in Ω^1 by $\vec{x}_\Omega^{(n)}$, the unknowns for the membrane problem by $\vec{x}_m^{(n)}$, and the unknowns for the $(u_1, u_2, 0)$ subproblem by $\vec{x}_{12}^{(n)}$, where n is the iteration index.

Select starting vectors $\vec{x}_\Omega^{(0)}$, $\vec{x}_m^{(0)}$ and $\vec{x}_{12}^{(0)}$. Then iterate the following steps until convergence is achieved:

- Solve the linear elastic problem in Ω^1 with inhomogeneous Dirichlet boundary conditions: The displacements on $\Gamma_\uparrow$ are given by $\vec{x}_m^{(n)}$ and $\vec{x}_{12}^{(n)}$. Denote the solution by $\tilde{x}_\Omega$.
- Compute the normal traction $\boldsymbol{\sigma} \cdot \mathbf{n}$ for the old vector $\vec{x}_\Omega^{(n)}$ and decompose them into $\mathbf{s}_{12}^{(n)}$ and $s_3^{(n)}$.
- Solve the membrane problem (4.50) with the load $s_3^{(n)}$, using the primal-dual active set algorithm 4.8. Also solve the other problem in Ω^2 with the load $\mathbf{s}_{12}^{(n)}$,

which is a linear elastic problem.
Denote the solutions by $\tilde{x}_m$ and $\tilde{x}_{12}$.

- Update the vectors with damping parameters:

$$\begin{aligned}\vec{x}_\Omega^{(n+1)} &:= (1-\vartheta_0)\vec{x}_\Omega^{(n)} + \tilde{x}_\Omega \\ \vec{x}_m^{(n+1)} &:= (1-\vartheta_1)\vec{x}_m^{(n)} + \tilde{x}_m \\ \vec{x}_{12}^{(n+1)} &:= (1-\vartheta_2)\vec{x}_{12}^{(n)} + \tilde{x}_{12}\ .\end{aligned}$$

♦

4.5. Implementation issues

All functions are represented by linear combinations of basis functions in the discrete version. In particular, all used FE spaces are:

- $\mathbf{V}^1$ is the space of continuous, piecewise linear functions $\Omega^1 \to \mathbb{R}^3$
- $\mathbf{V}_{12}^2$ is the space of continuous, piecewise linear functions $\Omega^2 \to \mathbb{R}^3$; in the construction of this space, no degrees of freedom are created for the third component.
- $\mathbf{W}^2$ is the space of possibly discontinuous, piecewise constant functions $\Omega^2 \to \mathbb{R}^3$.
- V^ω is the space of continuous, piecewise linear functions $\omega \to \mathbb{R}$.
- W^ω is the space of possibly discontinuous, piecewise constant functions $\omega \to \mathbb{R}$.
- $\mathbf{W}^\omega$ is the space of possibly discontinuous, piecewise constant functions $\omega \to \mathbb{R}^3$.

These spaces may have some appropriate boundary conditions on Γ_D or γ_D. Each discrete space is spanned by a collection of basis functions with likewise notation, e.g.

$$V^\omega \;=\; \operatorname{span}\left\{\varphi_i^\omega\right\}, \quad i = 1,\dots,N\ .$$

Here, functions φ are continuous functions (with space symbol V) and ψ are piecewise constant functions (with space symbol W). A boldface symbol $\mathbf{V}$ or $\boldsymbol{\varphi}$ indicates a mapping to $\mathbb{R}^3$.

In this section, we will only deal with functions in these spaces; only the given data $\mathbf{f}$ and $\mathbf{t}$ are an exception here and need to be approximated. For the sake of readability, the index h is omitted here.

4.5.1. Left side bilinear forms

On the left side, we have the bilinear forms

$$\int_{\Omega^2} \sigma(\left(\boldsymbol{\varphi}^2_{12}\right)_i) : \varepsilon(\left(\boldsymbol{\varphi}^2_{12}\right)_j)\,\mathrm{d}\mathbf{x} \tag{4.51}$$

and

$$\int_{\omega} \nabla\varphi^{\omega}_i \cdot \nabla\varphi^{\omega}_j\,\mathrm{d}x \tag{4.52}$$

for the equations (4.47) and (4.50). Applying these forms to $\mathbf{V}^2_{12} \times \mathbf{V}^2_{12}$ and $V^{\omega} \times V^{\omega}$, respectively, we will get the standard stiffness matrices A_{Ω_2} and A_{ω}.

4.5.2. Given forces $\mathbf{f}$ and tractions $\mathbf{t}$

The volume force $\mathbf{f} \in \left[H^{-1}(\Omega)\right]^3$ can be interpolated by piecewise constant functions, so we choose an approximation $\mathbf{f}_h \in \mathbf{W}^2$.
The interpolation vector $\vec{f}$ of $\mathbf{f}$ is defined by the coefficients f_i,

$$\mathbf{f}(\mathbf{x}) \;\approx\; \mathbf{f}_h(\mathbf{x}) \;=\; \sum_i f_i\,\boldsymbol{\psi}^2_i(\mathbf{x})\,.$$

We also interpolate the given boundary traction $\mathbf{t} \in \left[H^{-1/2}(\Gamma_N)\right]^3$: For implementation reasons, it is easier to express this interpolation in the full space $\mathbf{W}^2$. As the associated bilinear form (4.54) will only act on the boundary Γ_N, the exact nature of the extension does not matter: We only need to take care that the interpolation is performed on the boundary.
The interpolation vector $\vec{t}$ of $\mathbf{t}$ is defined by the coefficients t_i,

$$\mathbf{t}(\mathbf{x}) \;\approx\; \mathbf{t}_h(\mathbf{x}) \;=\; \sum_i t_i\,\boldsymbol{\psi}^2_i(\mathbf{x})\Big|_{\Gamma_N}\,.$$

If we define the mass bilinear form

$$\begin{aligned} a^2_{\mathrm{mass}} \;:\; \left[L^2(\Omega^2)\right]^3 \times \left[L^2(\Omega^2)\right]^3 \;&\to\; \mathbb{R} \\ (\boldsymbol{\psi}^i, \boldsymbol{\psi}^j) \;&\mapsto\; \int_{\Omega^2} \boldsymbol{\psi}^i(\mathbf{x}) \cdot \boldsymbol{\psi}^j(\mathbf{x})\,\mathrm{d}\mathbf{x}\,, \end{aligned} \tag{4.53}$$

we can compute the Galerkin matrix

$$M_{\Omega} \;:=\; M(\mathbf{V}^2_{12};\, \mathbf{W}^2) \;:=\; \left(a^2_{\mathrm{mass}}(\left(\boldsymbol{\varphi}^2_{12}\right)_i, \boldsymbol{\psi}^2_j)\right)_{ij}\,.$$

Likewise, we can define the mass boundary bilinear form

$$\begin{aligned} a^2_{\text{mass};\Gamma_N} &: \left[L^2(\Omega^2)\right]^3 \times \left[L^2(\Omega^2)\right]^3 \;\to\; \mathbb{R} \\ &(\boldsymbol{\psi}^i, \boldsymbol{\psi}^j) \;\mapsto\; \int_{\Gamma_N} \boldsymbol{\psi}^i(\mathbf{x}) \cdot \boldsymbol{\psi}^j(\mathbf{x}) \,\mathrm{d}s_x \end{aligned} \tag{4.54}$$

and the Galerkin matrix

$$M_{\Gamma_N} \;:=\; M_{\Gamma_N}(\mathbf{V}^2_{12};\, \mathbf{W}^2) \;:=\; \left(a^2_{\text{mass};\Gamma_N}(\left(\boldsymbol{\varphi}^2_{12}\right)_i, \boldsymbol{\psi}^2_j)\right)_{ij}.$$

Now we can compute the volume and boundary load vectors as matrix-vector products:

$$\begin{aligned} \int_{\Omega^2} \mathbf{f}_h(\mathbf{x}) \cdot \left(\boldsymbol{\varphi}^2_{12}\right)_j(\mathbf{x}) \,\mathrm{d}\mathbf{x} \;&=\; \left(M(\mathbf{V}^2_{12};\, \mathbf{W}^2) \cdot \vec{f}\right)_j, \\ \int_{\Gamma_N} \mathbf{t}_h(\mathbf{x}) \cdot \left(\boldsymbol{\varphi}^2_{12}\right)_j(\mathbf{x}) \,\mathrm{d}s_x \;&=\; \left(M_{\Gamma_N}(\mathbf{V}^2_{12};\, \mathbf{W}^2) \cdot \vec{t}\right)_j. \end{aligned}$$

Further, the boundary stress in the u_3 equation (4.50), given as $\tilde{t}_3$, can be interpolated to a W^ω function, represented by $\vec{t}_3$. Then we use the mass boundary bilinear form

$$\begin{aligned} a^\omega_{\text{mass};\gamma_N} &: L^2(\omega) \times L^2(\omega) \;\to\; \mathbb{R} \\ &(\psi^i, \psi^j) \;\mapsto\; \int_{\gamma_N} \psi^i(\mathbf{x}) \cdot \psi^j(\mathbf{x}) \,\mathrm{d}s_x \end{aligned} \tag{4.55}$$

and the Galerkin matrix

$$M_{\gamma_N} \;:=\; M_{\gamma_N}(V^\omega;\, W^\omega) \;:=\; \left(a^\omega_{\text{mass};\gamma_N}(\varphi^\omega_i, \psi^\omega_j)\right)_{ij}$$

to get the boundary load vector on ω:

$$\int_{\gamma_N} \tilde{t}_3 \varphi^\omega_j \,\mathrm{d}s_x \;\approx\; \left(M_{\gamma_N}(V^\omega;\, W^\omega) \cdot \vec{t}_3\right)_j.$$

If we have an interpolation $\vec{f_3}$ of the volume force's third component, f_3, we can use the mass bilinear form

$$\begin{aligned} a^\omega_{\text{mass}} &: L^2(\omega) \times L^2(\omega) \;\to\; \mathbb{R} \\ &(\psi^i, \psi^j) \;\mapsto\; \int_\omega \psi^i(\mathbf{x}) \psi^j(\mathbf{x}) \,\mathrm{d}s_x \end{aligned} \tag{4.56}$$

and the resulting Galerkin matrix

$$M_\omega \;:=\; M_\omega(V^\omega;\, W^\omega) \;:=\; \left(a^\omega_{\text{mass}}(\phi^\omega_i, \psi^\omega_j)\right)_{ij}$$

to get the vector representation

$$\int_\omega f_3 \phi_j \,\mathrm{d}x \;\approx\; \left(M_\omega(V^\omega; W^\omega) \cdot \vec{f_3}\right)_j.$$

4.5.3. Stress transmission from Ω^1

The stress tensor of a function $\mathbf{u}^1 \in \mathbf{V}^1$ in Ω^1 can be computed by a postprocessing: The functions in $\mathbf{V}^1$ are piecewise linear on the given triangulation, so their derivatives are piecewise constant. Then the stress tensor is constant on each element in Ω^1. We may restrict this tensor to the transmission boundary $\Gamma_\uparrow$ and multiply by the normal vector from Ω^1, which is $-\mathbf{e}_3$. Effectively, if we evaluate the stress tensor from $\mathbf{u}^1$ on each element adjacent to $\Gamma_\uparrow$ and take the negative of the third column, we get a representation of $\sigma(\mathbf{u}^1)\cdot\mathbf{n}$ which could immediately be expressed in the space $\mathbf{W}^\omega$, as ω and $\Gamma_\uparrow$ describe the same domain. But in our specific implementation, functions on volumes and functions on boundaries differ conceptually and are stored in separate program parts. As a consequence, we set up another representation of $\sigma(\mathbf{u}^1)\cdot\mathbf{n}$ in the space $\mathbf{W}^2$ by some arbitrary extension.

The first two components of the stress vector $\sigma(\mathbf{u}^1)\cdot\mathbf{n}$ are now represented by $\vec{s}_{12}$, i.e. $\mathbf{s}_{12} \in \mathbf{W}^2$. This leads to a bilinear form

$$\begin{aligned} a^2_{\text{mass};\Gamma_\uparrow} &: \left[L^2(\Omega^2)\right]^3 \times \left[L^2(\Omega^2)\right]^3 \;\to\; \mathbb{R} \\ &(\boldsymbol{\psi}^i, \boldsymbol{\psi}^j) \;\mapsto\; \int_{\Gamma_\uparrow} \boldsymbol{\psi}^i(\mathbf{x})\cdot\boldsymbol{\psi}^j(\mathbf{x})\,\mathrm{d}s_x\,, \end{aligned} \tag{4.57}$$

and the Galerkin matrix

$$M_{\Gamma_\uparrow} \;:=\; M_{\Gamma_\uparrow}(\mathbf{V}^2_{12};\mathbf{W}^2) \;:=\; \left(a^2_{\text{mass};\Gamma_\uparrow}(\left(\boldsymbol{\varphi}^2_{12}\right)_i, \boldsymbol{\psi}^2_j)\right)_{ij}$$

gives us the vector representation

$$\int_{\Gamma_\uparrow} \mathbf{s}_{12}\cdot\left(\boldsymbol{\varphi}^2_{12}\right)_j \mathrm{d}s_x \;=\; \left(M_{\Gamma_\uparrow}(\mathbf{V}^2_{12};\mathbf{W}^2)\cdot\vec{s}_{12}\right)_j$$

for the third integral on the right side of (4.47).

For equation (4.50), we use a representation of the third component $(\sigma(\mathbf{u}^1)\cdot\mathbf{n})_3 = -\sigma_N(\mathbf{u}^1)$ in W^ω by the vector $\vec{s}_3$. This allows us to re-use the bilinear form (4.56) and the matrix M_ω:

$$\int_\omega s_3\phi_j\,\mathrm{d}x \;=\; \left(M_\omega(V^\omega;W^\omega)\cdot\vec{s}_3\right)_j\,.$$

4.5.4. Transmission from u_3 to $(u_1, u_2, 0)$

The first transmission contribution stems from the term

$$\int_{\Gamma_\uparrow}\left(\nabla\begin{pmatrix}0\\0\\u_3\end{pmatrix}\right)^\top\cdot\boldsymbol{\varphi}^2_j\cdot\mathbf{n}\,\mathrm{d}s_x \;=\; \int_{\Gamma_\uparrow}\left(u_{3,1}(\boldsymbol{\varphi}^2_j)_1 + u_{3,2}(\boldsymbol{\varphi}^2_j)_2\right)n_3\,\mathrm{d}s_x\,.$$

The normal vector on $\Gamma_\uparrow$ is $\mathbf{e}_3$, so $n_3 = +1$.
Here, we can use the bilinear form $a_{\text{mass};\Gamma_\uparrow}$ from (4.57), and the Galerkin matrix $M_{\Gamma_\uparrow}$ is already known. Let the 2D gradient of u_3 be given in $\mathbf{W}^\omega$, i.e. as piecewise constant 3D vectors (setting the third component to zero), with the coefficient vector $\vec{u}_{3,\nabla}$:

$$\nabla u_3(x_1, x_2) \;=:\; \sum_i \left(\vec{u}_{3,\nabla}\right)_i \boldsymbol{\psi}_i(x_1, x_2)\,.$$

If we use an arbitrary extension of ∇u_3 into $\mathbf{W}^2$, e.g. by mapping the representation $\vec{u}_{3,\nabla}$ in $\mathbf{W}^\omega$ to $\vec{u}_{3,\nabla,\Omega}$ in $\mathbf{W}^2$, we can state

$$\int_{\Gamma_\uparrow} \left(u_{3,1}(\boldsymbol{\varphi}_j^2)_1 + u_{3,2}(\boldsymbol{\varphi}_j^2)_2\right) n_3 \,\mathrm{d}s_x \;=\; \left(M_{\Gamma_\uparrow}(\mathbf{V}_{12}^2;\, \mathbf{W}^2) \cdot \vec{u}_{3,\nabla,\Omega}\right)_j\,.$$

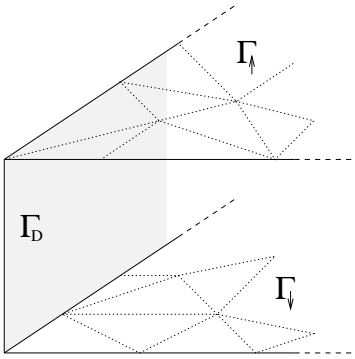


Figure 4.5.: Nonmatching surface meshes on $\Gamma_\uparrow$ and $\Gamma_\downarrow$

The other mixed contribution (on $\Gamma_\downarrow$) needs more attention. We may take the trace of $(u_1, u_2, 0)$ on the bottom, but u_3 is not defined there. As we assumed that $u_{3,3} = 0$ in Ω^2, we can extend

$$u_3(x_1, x_2, 0) = u_3(x_1, x_2, h)\,.$$

Here, a problem occurs when the problem is discretized. The surface mesh $\mathtt{T}^\uparrow$ on $\Gamma_\uparrow$ will generally not match the surface mesh $\mathtt{T}^\downarrow$ on $\Gamma_\downarrow$.
First, note that $n_3 = -1$ on $\Gamma_\downarrow$, so we get

$$\int_{\Gamma_\downarrow} \left(\nabla \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix}\right)^\top \cdot \boldsymbol{\varphi}_j^2 \cdot \mathbf{n}\,\mathrm{d}s_x \;=\; -\int_{\Gamma_\downarrow} \left((\boldsymbol{\varphi}_j^2)_1(\mathbf{x}_{123})\, u_{3,1}(\mathbf{x}_{12}) \;+\; (\boldsymbol{\varphi}_j^2)_2(\mathbf{x}_{123})\, u_{3,2}(\mathbf{x}_{12})\right) \mathrm{d}s_x\,.$$

The support of a function $\boldsymbol{\varphi}_j^2 \in \mathbf{V}_{12}^2$, restricted to $\Gamma_\downarrow$, is a set of triangles $T_n^\downarrow \in \mathtt{T}^\downarrow$. The support of a function $\boldsymbol{\psi}_i^\omega \in \mathbf{W}^\omega$ is a set of triangles $T_m^\uparrow \in \mathtt{T}^\uparrow$. (If we use piecewise constant functions $\boldsymbol{\psi}_i^\omega$, the support is just one triangle.)

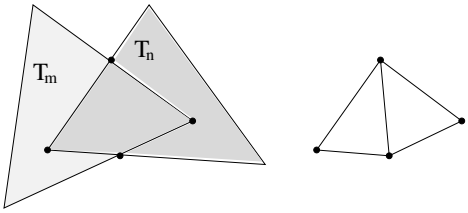


Figure 4.6.: Intersection of two triangles T_m and T_n with corner points; generic decomposition into triangles T_α

The support of the integrand is then the intersection of triangles T_m and T_n in the $x_3 = 0$ plane. It can be subdivided into subtriangles T_α (see Section 4.5.7).

We can now write down the associated bilinear form and its implementation:

$$\begin{aligned} a^2_{\Gamma_\downarrow} \;&: \left[H^1(\Omega^2)\right]^3 \times \left[L^2(\omega)\right]^3 \;\to\; \mathbb{R} \\ &(\boldsymbol{\varphi}^i, \boldsymbol{\psi}^j) \;\mapsto\; -\int_{\Gamma_\downarrow} \left((\boldsymbol{\varphi}^2_j)_1 \, u_{3,1}(x_1, x_2) + (\boldsymbol{\varphi}^2_j)_2 \, u_{3,2}(x_1, x_2) \right) \mathrm{d}\mathbf{x} \end{aligned} \tag{4.58}$$

Let the support elements be given as

$$\operatorname{supp}\left(\boldsymbol{\varphi}^i \big|_{\Gamma_\downarrow} \right) \;=\; \bigcup_m T_m \,, \quad \operatorname{supp} \boldsymbol{\psi}^j \;=\; \bigcup_n T_n \,.$$

Decomposing all intersections $T_m \cap T_n$ into triangles $\mathtt{T}_\alpha$, we get

$$a^2_{\Gamma_\downarrow}(\boldsymbol{\varphi}^i, \boldsymbol{\psi}^j) \;=\; -\sum_m \sum_n \sum_\alpha \int_{\mathtt{T}_\alpha} \left((\boldsymbol{\varphi}^2_j)_1 \, u_{3,1}(x_1, x_2) + (\boldsymbol{\varphi}^2_j)_2 \, u_{3,2}(x_1, x_2) \right) \mathrm{d}\mathbf{x} \,.$$

We need the Galerkin matrix

$$M_{\Gamma_\downarrow} \;:=\; M_{\Gamma_\downarrow}(\mathbf{V}^2_{12}; \mathbf{W}^{(\omega)}) \;:=\; \left(a^2_{\Gamma_\downarrow}\left(\left(\boldsymbol{\varphi}^2_{12} \right)_i, \boldsymbol{\psi}^{(\omega)}_j \right) \right)_{ij}$$

to compute the vector entries

$$\int_{\Gamma_\downarrow} \left(\nabla \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix} \right)^\top \cdot \boldsymbol{\varphi}^2_j \cdot \mathbf{n} \, \mathrm{d}s_x \;=\; \left(M_{\Gamma_\downarrow}(\mathbf{V}^2_{12}; \mathbf{W}^{(\omega)}) \cdot \vec{u}_{3,\nabla} \right)_j \,.$$

4.5.5. Transmission from $(u_1, u_2, 0)$ to u_3

This transmission does not need extensions like the transmission from u_3 to $(u_1, u_2, 0)$, so no polygonal intersections need to be made. Let some function $(u_1, u_2, 0)$ be given, either from the problem in Ω^1 or from a solution of the first subproblem in

Ω^2. Then we can compute the gradient of $(u_1, u_2, 0)$, which is piecewise constant on the elements of Ω^1 (or Ω^2). The vector $(u_{1,3}, u_{2,3}, 0)$ can now be stored as the third row of this gradient, restricted to $\Gamma_\uparrow$. The appropriate space for this is $\mathbf{W}^\omega$, and the coefficient vector is denoted by $\vec{u}_{\partial,3}$.

We introduce the bilinear form

$$\begin{aligned} a^\omega_{\mathrm{mix},\gamma} &: L^2(\omega) \times \left[L^2(\omega)\right]^3 \to \mathbb{R} \\ (\psi^i, \boldsymbol{\psi}^j) &\mapsto \int_{\gamma_N} \psi^i \begin{pmatrix} \psi^j_1 \\ \psi^j_2 \end{pmatrix} \cdot \mathbf{n} \, \mathrm{d}s_x \end{aligned} \tag{4.59}$$

and the Galerkin matrix

$$M_{\mathrm{mix},\gamma} := M_{\mathrm{mix},\gamma}(V^\omega; \mathbf{W}^\omega) := \left(a^\omega_{\mathrm{mix},\gamma}(\varphi^\omega_i, \boldsymbol{\psi}^\omega_j)\right)_{ij}$$

to rewrite the γ_N integral in equation (4.50) as follows:

$$\int_{\gamma_N} \begin{pmatrix} u_{1,3} \\ u_{2,3} \end{pmatrix} \cdot \mathbf{n} \, v \, \mathrm{d}s_x = \left(M_{\mathrm{mix},\gamma}(V^\omega; \mathbf{W}^\omega) \cdot \vec{u}_{\partial,3}\right)_j .$$

For the last integral, we use the bilinear form

$$\begin{aligned} a^\omega_{\mathrm{mix}} &: H^1(\omega) \times \left[L^2(\omega)\right]^3 \to \mathbb{R} \\ (\varphi^i, \boldsymbol{\psi}^j) &\mapsto \int_\omega \nabla\varphi^i \cdot \begin{pmatrix} \psi^j_1 \\ \psi^j_2 \end{pmatrix} \mathrm{d}\mathbf{x} \end{aligned} \tag{4.60}$$

which induces the Galerkin matrix

$$M_{\mathrm{mix},\omega} := M_{\mathrm{mix},\omega}(V^\omega; \mathbf{W}^\omega) := \left(a^\omega_{\mathrm{mix}}(\varphi^\omega_i, \boldsymbol{\psi}^\omega_j)\right)_{ij} .$$

The last integral then is

$$\int_\omega \begin{pmatrix} u_{1,3} \\ u_{2,3} \end{pmatrix} \cdot \nabla v \, \mathrm{d}\mathbf{x} = \left(M_{\mathrm{mix},\omega}(V^\omega; \mathbf{W}^\omega) \cdot \vec{u}_{\partial,3}\right)_j .$$

4.5.6. Implementation blocks

With the Galerkin matrices defined as in the previous section, we can set up the linear systems to solve the equations for $(u_1, u_2, 0)$ and u_3.
For the $(u_1, u_2, 0)$ equation, the data is given as follows:

- The volume force $\mathbf{f}$ is given in $\mathbf{W}^2$ by the coefficient vector $\vec{f}$.
- The traction $\mathbf{t}$ on Γ_N is given in $\mathbf{W}^2$ by $\vec{t}$.
- The stress vector in Ω^1 is transferred to $\mathbf{W}^2$ with the vector $\vec{s}_{12}$.

- The 3D gradient of a given u_3 has the coefficient vector $\vec{u}_{3,\nabla}$ in $\mathbf{W}^\omega$.
- The 3D gradient $\vec{u}_{3,\nabla}$ is extended into $\mathbf{W}^2$ by $\vec{u}_{3,\nabla,\Omega}$.

All matrices and the vectors $\vec{f}$ and $\vec{t}$ can be computed before any iteration. Only the vectors $\vec{s}_{12}$, $\vec{u}_{3,\nabla}$ and $\vec{u}_{3,\nabla,\Omega}$ vary in the iteration.

The first problem (4.47) is then, in matrix form,

$$A_{\Omega^2}\,\vec{x} = M_\Omega\,\vec{f} + M_{\Gamma_N}\,\vec{t} + M_{\Gamma_\uparrow}\,\vec{s}_{12} - \mu\,M_{\Gamma_\uparrow}\,\vec{u}_{3,\nabla,\Omega} - \mu\,M_{\Gamma_\downarrow}\,\vec{u}_{3,\nabla}\,. \tag{4.61}$$

For the u_3 equation, the data is given as follows:

- The normal stress in Ω^1 is transferred to W^ω by the vector $\vec{s}_3$.
- The third component of the volume force $\mathbf{f}$ is given in W^ω by the vector $\vec{f}_3$.
- An extension of the boundary stress part t_3 is given in W^ω by $\vec{t}_3$.
- The x_3 derivatives of given u_1 and u_2 are given as a 3D vector with zero third component in $\mathbf{W}^\omega$ by the coefficient vector $\vec{u}_{\partial,3}$.

Again, all matrices and the vectors $\vec{f}_3$ and $\vec{t}_3$ can be computed in advance. Only $\vec{s}_3$ and $\vec{u}_{\partial,3}$ vary in the iteration.

The second problem (4.50) in matrix form is then

$$\begin{aligned} A_\omega\,\vec{x} = {} & \frac{1}{h}M_\omega\,\vec{s}_3 + M_\omega\,\vec{f} - \frac{1}{h}\Big(\int_\omega b(-u_3)\phi_j\,\mathrm{d}x\Big)_j \\ & + M_{\gamma_N}\,\vec{t}_3 + \lambda\,M_{\mathrm{mix},\gamma}\,\vec{u}_{\partial,3} \\ & - (\lambda+\mu)\,M_{\mathrm{mix},\omega}\,\vec{u}_{\partial,3}\,. \end{aligned} \tag{4.62}$$

4.5.7. Intersection of triangles

The intersection of two triangles T_m and T_n is a convex set, as both triangles are convex. It may have up to six corner points. As several cases can occur in the intersection of two triangles, we stick with a generic algorithm here.

Algorithm 4.14:
First, compute the extremal points of $T_m \cap T_n$:

- Create an empty set P_C of corner points.
- Check all corner nodes $P^1_{1,2,3}$ of T_m: If $P^1_i \in T_n$, add it to P_C
- Check all corner nodes $P^2_{1,2,3}$ of T_n: If $P^2_i \in T_m$, add it to P_C

- Compute the intersections of all triangle edges of T_m and all triangle edges of T_n. If the intersection point P_e is inside edge e_i^m of T_m and inside edge e_j^n of T_n, add it to P_C.

Now the intersection polygon is $T_m \cap T_n = \operatorname{conv} P_C$, and P_C is a set of points $x_i \in \partial(\operatorname{conv} P_C)$.

Next, re-arrange these points:

- Compute the midpoint of the convex hull,

$$P_m := \frac{1}{N} \sum_{i=1}^{N} x_i \,.$$

- Compute the direction vectors $d_i = x_i - P_m$.
- With the angle $\vartheta_i = \measuredangle(d_i)$, sort the d_i counter-clockwise.

Now the convex hull can be decomposed into triangles intersecting only on edges,

$$\operatorname{conv} P_C = \bigcup_{i=1}^{N} \triangle\Big(P_m,\, P_m + d_i,\, P_m + d_{(i \bmod N)+1}\Big)\,. \qquad \blacklozenge$$

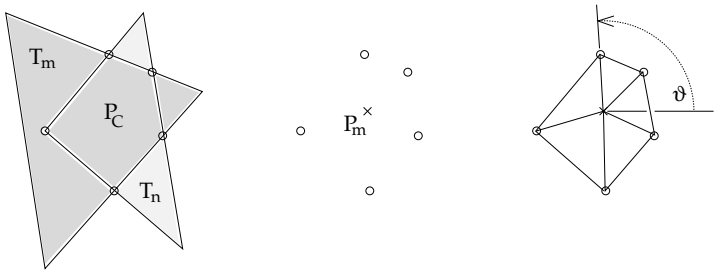


Figure 4.7.: $T_m \cap T_n$: Corner points, midpoint and decomposition into sorted triangles

5. Numerical experiments

In this chapter, we will demonstrate some numerical experiments on the benchmark test from the introduction. All computations were performed using piecewise linear functions on triangles or tetrahedra for the displacement. The programs were run on two desktop workstations and on the computing cluster at the Institute for Applied Mathematics.

The decomposition into triangles or tetrahedra was done by the external libraries `CGAL` [7] for 2D meshes, and `tetgen` by Si [41, 45] for 3D volume and surface meshes. The reference implementation `PNEW` by Lukšan and Vlček [28] was used for the minimization of the nonsmooth objective functions. All other software for computation and postprocessing was written along with this thesis by the author.

Remark 5.1: *As we had no exact solution for the benchmark problems, we computed solutions on a very fine mesh and assumed that these were close enough to a solution of the continuous problem.*
The benchmark problems employ a non-monotone delamination law, and we cannot assure that the problems admit only one solution. If a large error for a numerical solution is given, it may still be close to another solution of the continuous problem.

5.1. Adaptive refinement

A wide-spread method to keep the number of unknowns as low as possible is the use of error estimators and adaptive refinement. For our Finite Element computations, we used a heuristic residual error estimator. A general step of the refinement algorithm is as follows:

Algorithm 5.2:
After finding an approximate solution $\mathbf{u}_h$ in V_h, based on the mesh $\mathcal{T}_h$, go through the following steps:

- For every triangle or tetrahedron $T \in \mathcal{T}_h$, compute the error indicator η_T.
- Compute the maximal indicator $\eta_{\max}$.
- Mark all elements T with $\eta_T > \vartheta \eta_{\max}$ for refinement.
- Mark further elements such that no hanging nodes occur.
- Refine the mesh.

With the new, finer mesh, one can re-start the algorithm again. ♦

In this algorithm, $\vartheta \in (0,1)$ is a control parameter. More elements are marked in one refinement step if ϑ is small.

The local residual error indicator

$$\eta_T^2 := h_T^2 \left\| \operatorname{div} \boldsymbol{\sigma}(\mathbf{u}^h) + \pi_0 \mathbf{f} \right\|_{0;T}^2 + \sum_{E_{\text{int}}(T)} h_E \left\| \left[\boldsymbol{\sigma}(\mathbf{u}^h) \cdot \mathbf{n}_E \right] \right\|_{0;E}^2 + \sum_{E_{\text{ext},N}(T)} h_E \left\| \boldsymbol{\sigma}(\mathbf{u}^h) \cdot \mathbf{n}_E - \pi_0 \mathbf{t} \right\|_{0;E}^2 \tag{5.1}$$

can be given for each $T \in \mathcal{T}_h$. If the problem only has Dirichlet and Neumann boundaries, i.e. no contact or other nonlinear terms are present, one can prove that η_T is efficient and reliable: See [46] for a derivation, especially Section 3.6 for linear elasticity.

Some comments on each term are in order. First, note that h_T denotes the diameter of the element T, and h_E denotes the length of edge E.

The first term is the L^2 norm of the residual in the differential equation. We demand that $-\operatorname{div} \boldsymbol{\sigma} = \mathbf{f}$, so this term describes the approximation error of the volume term. For implementation reasons, the function $\mathbf{f}$ can be approximated by a constant value $\pi_0 \mathbf{f}$ on T. We used the value of $\mathbf{f}$ in the element midpoint here.
The second term is a summation over all interior edges or faces of T. As another element T' is attached to T along E, we can compute the stress vector on E in both elements, and its jump $[\boldsymbol{\sigma} \cdot \mathbf{n}]$ across this face. In the classical formulation, $\boldsymbol{\sigma}$ is continuous; the error to that is given by the second term.
The third term is a summation over all exterior edges or faces of T that are attached to the Neumann boundary. Here, we have a prescribed traction $\mathbf{t}$. The deviation of the solution's traction $\boldsymbol{\sigma}(\mathbf{u}^h) \cdot \mathbf{n}$ from $\mathbf{t}$ is the approximation error here. Again, we use an interpolation by a constant value $\pi_0 \mathbf{t}$ here for implementation reasons.

The hemivariational inequality has an additional boundary Γ_C. Here, the exact solution satisfies

$$\xi(\mathbf{x}) \in \hat{b}\Big((\Pi \mathbf{u})(\mathbf{x})\Big) \quad \text{a.e. } \mathbf{x} \in \Gamma_C \,,$$

as we demand that

$$\xi(\mathbf{x}) = \sigma_N(\mathbf{x}) \quad \text{a.e. } \mathbf{x} \in \Gamma_C \,.$$

In our case, Π extracts the negative normal displacement.

The function $b(t)$ represents a force in normal direction, so if we treat it like a given force $\mathbf{t}$, we get an additional contribution

$$\sum_{E_{\text{ext},C}(T)} h_E \left\| \sigma_N(\mathbf{u}^h) - b\Big((\Pi \mathbf{u}^h)(\mathbf{x})\Big) \right\|_{0;E}^2 \tag{5.2}$$

on the edges or faces on Γ_C.
Note that the function b is used here instead of $\hat{b}$, which would be set-valued. There

may appear the case that a normal displacement is attained such that $\hat{b}$ would return more than one element. This, however, did in practice not appear in our computations.
Again, we interpolate the given traction $b((\Pi\mathbf{u})(\mathbf{x}))\cdot\mathbf{n}$ by a constant function, e.g. by evaluation in the midpoint $\mathbf{m}$ of E. We can now state the residual error indicator for the hemivariational inequality:

$$\begin{aligned}\eta_T^2 \;:=\; & h_T^2 \left\|\operatorname{div}\sigma(\mathbf{u}^h)+\pi_0\mathbf{f}\right\|_{0;T}^2 \;+\; \sum_{E_{\mathrm{int}}(T)} h_E \left\|\left[\sigma(\mathbf{u}^h)\cdot\mathbf{n}_E\right]\right\|_{0;E}^2 \\ & + \sum_{E_{\mathrm{ext},N}(T)} h_E \left\|\sigma(\mathbf{u}^h)\cdot\mathbf{n}_E \;-\; \pi_0\mathbf{t}\right\|_{0;E}^2 \;+\; \sum_{E_{\mathrm{ext},C}(T)} h_E \left\|\sigma_N(\mathbf{u}^h) \;-\; b\Big((\Pi\mathbf{u}^h)(\mathbf{m})\Big)\right\|_{0;E}^2 .\end{aligned} \tag{5.3}$$

Note that this estimator is only heuristic, but it performed well in our benchmark computations.

5.2. 2D benchmark

The following 2D benchmark was computed with Finite Elements.

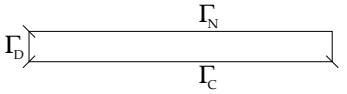


Figure 5.1.: Geometry of the 2D benchmark

The problem domain is a rectangle of height 1 and width 10:

$$\Omega \;=\; \Big\{\mathbf{x}\in\mathbb{R}^2 \;:\; x_1\in[-5,5]\,,\; x_2\in[0,1]\Big\}\,.$$

On the lower boundary, contact with adhesion may occur. This boundary Γ_C consists of all $\mathbf{x}\in\Omega$ with $x_2=0$, leaving the normal vector $\mathbf{n}=(0,-1)$. The Dirichlet boundary Γ_D consists of all $\mathbf{x}\in\Omega$ with $x_1=-5$. Finally, all other boundary parts constitute the Neumann boundary Γ_N. We prescribe a constant force on the upper right part only, where $x_1\in[4,5]$ and $x_2=1$. Here $\mathbf{t}=(0,f_N)$ with a given f_N; on the remaining parts of Γ_N, we set $\mathbf{t}=(0,0)$.

The chosen material parameters are $\lambda = 1.211\cdot10^{11}$ and $\mu = 8.077\cdot10^{10}$, which corresponds to the material parameters $E = 210\,\mathrm{GPa}$ and $\nu = 0.3$ for steel. The benchmarks in [19] and [4] use serrated reaction forces b with several jags. As the qualitative behavior does not change, we use an exemplary function b with two jags

only. Its envelope $\hat{b}$ with generic constants A_i and t_i is given as follows:

$$\hat{b}(t) := \begin{cases} (-\infty, 0], & t = 0 \\ \left\{\frac{A_1}{t_1} t\right\}, & t \in (0, t_1) \\ [A_2, A_1], & t = t_1 \\ \left\{\frac{A_3 - A_2}{t_2 - t_1} t + \frac{t_2 A_2 - t_1 A_3}{t_2 - t_1}\right\}, & \in (t_1, t_2) \\ [0, A_3], & t = t_2 \\ \{0\}, & t > t_2 \end{cases} . \tag{5.4}$$

The anti-derivative we use is then

$$B(t) = \int_0^t b(\tau)\,\mathrm{d}\tau = \begin{cases} \frac{A_1}{2t_1} t^2, & t \in [0, t_1] \\ \frac{A_3 - A_2}{2(t_2 - t_1)}(t^2 - t_1^2) + \frac{t_2 A_2 - t_1 A_3}{t_2 - t_1}(t - t_1) + \frac{A_1 t_1}{2}, & t \in (t_1, t_2] \\ \frac{A_3 - A_2}{2(t_2 - t_1)}(t_2^2 - t_1^2) + t_2 A_2 - t_1 A_3 + \frac{A_1 t_1}{2}, & t > t_2 \end{cases} . \tag{5.5}$$

We used $t_1 = -0.02$ and $t_2 = -0.1$ for the displacements before tear-off, with the forces $A_1 = 20 \cdot 10^6$, $A_2 = 8 \cdot 10^6$ and $A_3 = 10 \cdot 10^6$.

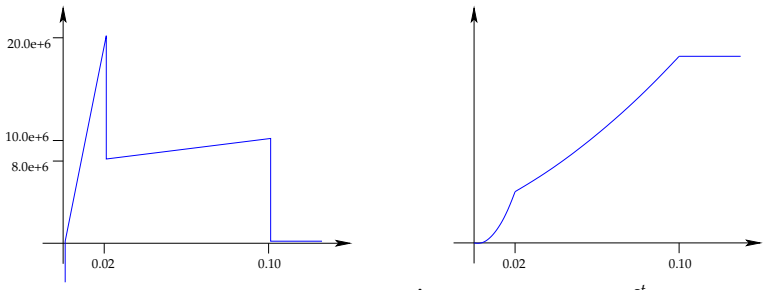


Figure 5.2.: Reaction force function $\hat{b}(t)$; anti-derivative $\int_0^t b(\tau)\,\mathrm{d}\tau$

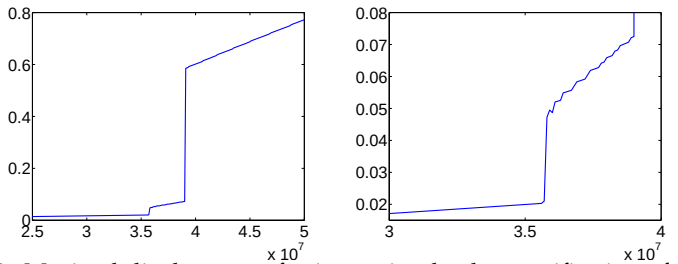


Figure 5.3.: Maximal displacement for increasing load; magnification of the lower left area

When forces between $25 \cdot 10^6$ and $50 \cdot 10^6$ are incrementally applied, three parts can be isolated. Figure 5.3 shows the norm of the maximal displacement for different

forces. Note that the maximal displacement always appeared at the upper-right corner of Ω.

In the first part (up to ca. $36 \cdot 10^6$), no delamination takes place. The displaced mesh and the reaction forces along the contact boundary are shown in Figure 5.4 for a representative force of $30 \cdot 10^6$. The first "crack" has not yet appeared here.

In the second part (up to ca. $39 \cdot 10^6$), partial delamination occurs. Figure 5.5 shows the displaced mesh and the reaction forces for a load of $37.5 \cdot 10^6$.

Finally, a section of the contact boundary delaminates completely in the third part. No reaction force is given for this section, as can be seen in Figure 5.6, which was computed for a load of $40 \cdot 10^6$. A stress singularity appears in the lower left corner. In the next figures, the color scale denotes the *von Mises stress* in MPa. As we computed a plane strain problem in 2D, the von Mises stress computes as

$$\sigma_Y \;=\; \sqrt{\left(\sigma_{11}^2 + \sigma_{22}^2\right)(\nu^2 - \nu + 1) \;+\; \sigma_{11}\sigma_{22}(2\nu^2 - 2\nu - 1) \;+\; 3\sigma_{12}^2}\,,$$

where the Poisson number ν is

$$\nu \;=\; \frac{\lambda}{2(\lambda + \mu)}\,.$$

The computations were performed with 128 degrees of freedom in normal direction on Γ_C, the full problem consisted of 4224 unknowns.

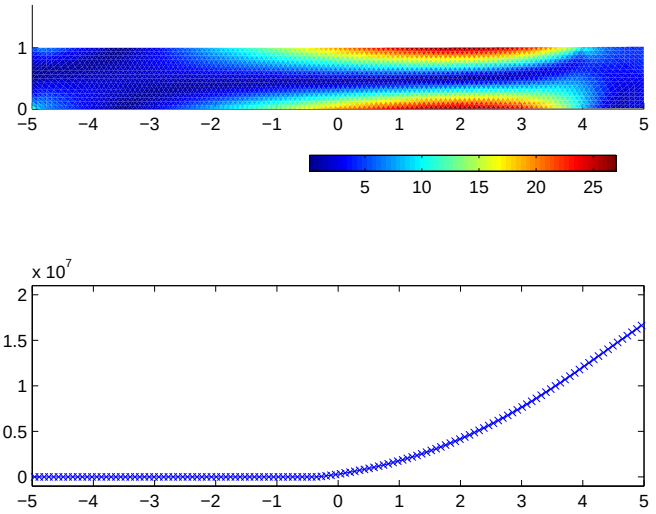


Figure 5.4.: Deformed configuration and reaction forces along Γ_C for a load of $30 \cdot 10^6$

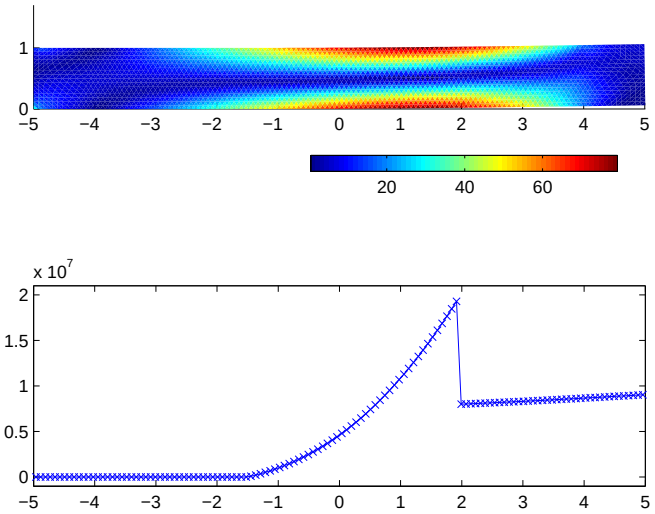


Figure 5.5.: Deformed configuration and reaction forces along Γ_C for a load of $37.5 \cdot 10^6$

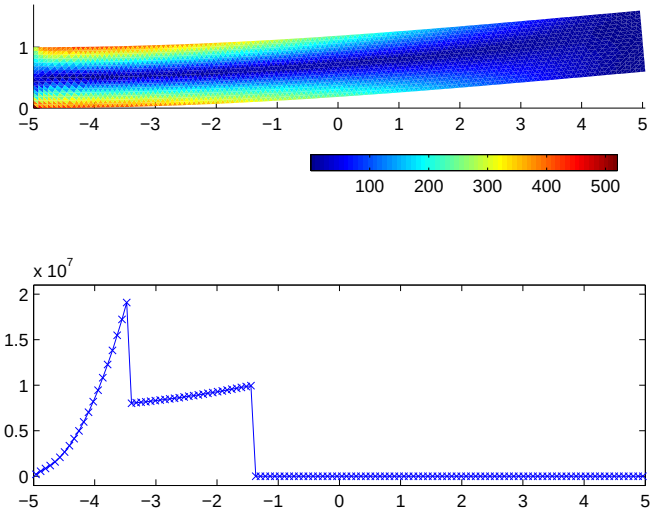


Figure 5.6.: Deformed configuration and reaction forces along Γ_C for a load of $40 \cdot 10^6$

The error induced by the method depends on the load: If no delamination occurs, the error is generally smaller than in the delamination case. Several reasons underline this: For a nonsmooth reaction force, we may have multiple solutions. Our computations only give the error to one specific solution for a very fine mesh. Further, small variations in the load move the point of delamination on the contact boundary, so the displacement is not a smooth function of the given load. Finally, there is no displacement in the lower left corner if no delamination is present: In Figure 5.4, this corner is in contact, and the boundary nodes close to it are not displaced. In contrast, Figure 5.6 shows that there is a displacement in this corner, which results in a stress peak.

N	$\left\|u^h - u^*\right\|_{L^2(\Omega)}$	EOC(N)
32	1.89e-3	—
94	1.07e-3	0.53
316	3.18e-4	1.00
1140	1.51e-4	0.58
4290	6.71e-5	0.61
N	$\left\|u^h - u^*\right\|_{L^2(\Omega)}$	EOC(N)
32	1.89e-3	—
76	1.08e-3	0.65
186	6.32e-4	0.60
604	1.29e-4	1.35
2112	6.80e-5	0.51
2270	4.46e-5	5.85
2800	2.47e-5	2.82
2932	1.76e-5	6.86 (!)
5088	2.22e-6	3.79

Table 5.1.: L^2 error for load $10 \cdot 10^6$, uniform and adaptive refinement

Figure 5.7 shows the convergence of the L^2 error if only a small force is applied, i.e. the reaction forces on Γ_C are smooth. Figure 5.8 shows the convergence of the L^2 error if a large force is applied. We used finer solutions as reference solution for the error computation, which were obtained by three further refinement steps from the uniform scheme ($N \approx 250.000$ for the first case, $N \approx 180.000$ for the second case).

The obtained errors and estimated orders of convergence are given in Tables 5.1 and 5.1. Note that some orders of convergence for the adaptive schemes are off the scale, which is probably due to a combination of bad previous steps and coincidence.

N	$\left\|u^h - u^*\right\|_{L^2(\Omega)}$	EOC(N)
238	1.85e-1	—
824	4.92e-2	1.07
3040	2.56e-2	0.50
11648	1.50e-3	2.11
N	$\left\|u^h - u^*\right\|_{L^2(\Omega)}$	EOC(N)
238	1.78e-1	—
252	1.32e-1	5.21
348	7.19e-2	1.88
472	4.70e-2	1.40
524	4.08e-2	1.34
814	1.64e-2	2.08
1090	4.06e-3	4.78
1324	2.51e-3	2.47

Table 5.2.: L^2 error for load $40 \cdot 10^6$, uniform and adaptive refinement

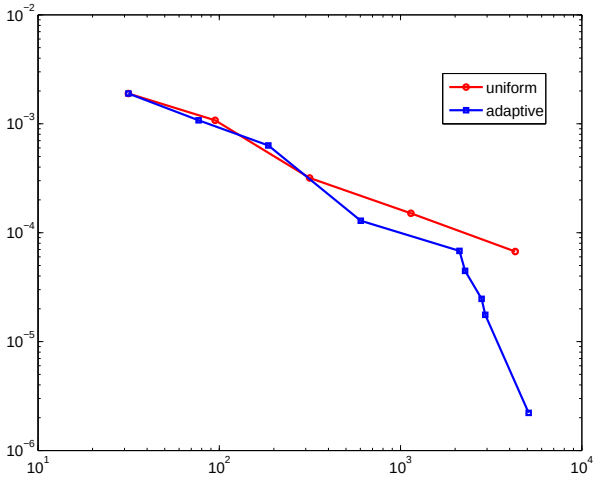


Figure 5.7.: L^2 error vs. number of unknowns for 2D uniform and adaptive refinement, load $10 \cdot 10^6$ (no delamination)

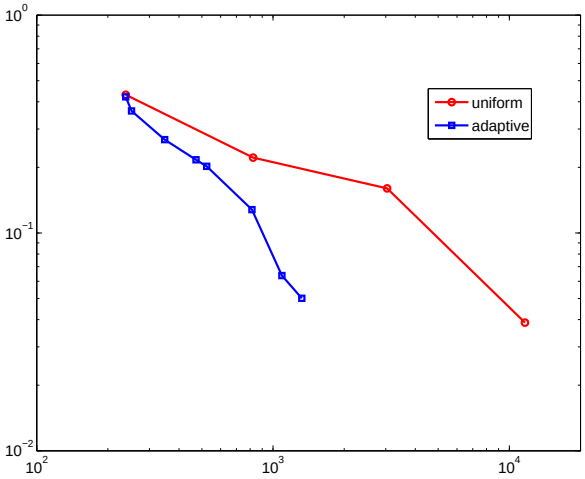


Figure 5.8.: L^2 error vs. number of unknowns for 2D uniform and adaptive refinement, load $40 \cdot 10^6$ (with delamination)

Parallelization

The main computational effort for the used problem sizes was the Schur complement matrix in Section 3.6.1. To get the system matrix

$$\underline{C} = C - B^\top \bar{A}^{-1} B \,,$$

one equation system $\bar{A}x = b$ had to be solved for each degree of freedom on Γ_C. These systems can be solved in parallel. This was implemented using the shared-memory model with the `OpenMP` programming interface. For 2, 8 and 16 virtual cores with hyperthreading (1, 2 and 8 physical cores), nearly linear scaling could be observed.

The Bundle-Newton solver took a larger fraction of computing time as the number of unknowns grew. Note that the reference implementation is not parallelized and uses a full LR decomposition of $\underline{C}$, so there is still some potential here.

Further, the computation of stiffness matrices and load vectors has been done in parallel. This part took only a small fraction of the total computation time, so the improvement was not too significant here.

The same parallelization was used for the FE and BE computations in 3D with the Bundle-Newton method.

5.3. 3D benchmarks

5.3.1. Finite Element computation

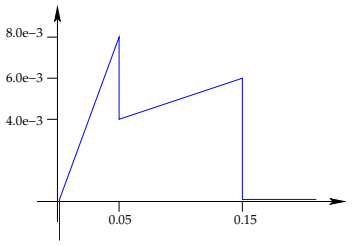


Figure 5.9.: Reaction force function $\hat{b}(t)$

The configuration is given in Figure 5.10. We could not identify matching parameters for convergence of the Bundle-Newton method with realistic material parameters and a large number of unknowns, so we used $\lambda = \mu = 1$ instead. The delamination law in Figure 5.9 is of the same type as in the 2D benchmark, Figure 5.2, but with different numbers.

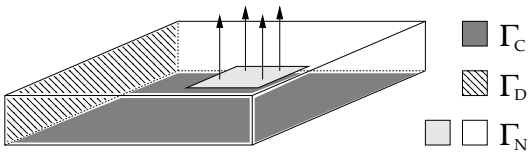


Figure 5.10.: Reference configuration for the 3D FE and BE benchmark

The geometry size is $7.5 \times 4 \times 0.4$. Γ_D is the block's back side ($x_1 = 0$). The load is applied on a 1×1 square on the upper left corner ($x_1 \geq 6.5$, $x_2 \geq 3$) in x_3 direction.

The deformed mesh for a load of 0.2 is shown in Figure 5.11. The lower picture shows the distribution of reaction forces, i.e. the function ξ^h on the contact boundary elements. Note that the front of red-colored elements denotes the maximal reaction force from the first tip of $\hat{b}(\cdot)$. The elements towards the lower left corner already undergo only the second, lower reaction force. This situation corresponds to the 2D situation in Figure 5.5.

The mesh was adaptively refined using the error indicator (5.3).

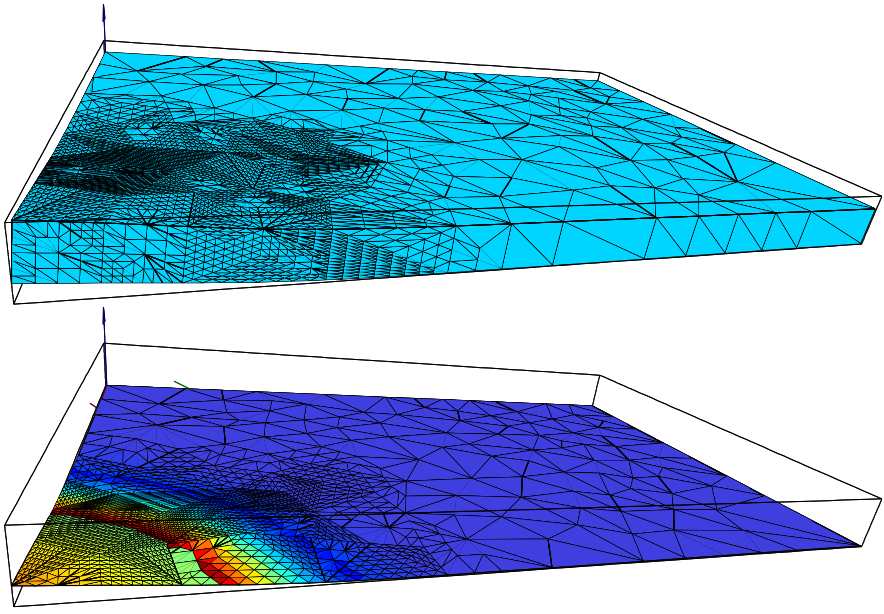


Figure 5.11.: Deformed mesh and reaction force distribution, FE computation

5.3.2. Boundary Element computation

The same configuration was used for the BE computation, but only uniform refinement was performed. The results are plotted in Figure 5.12.

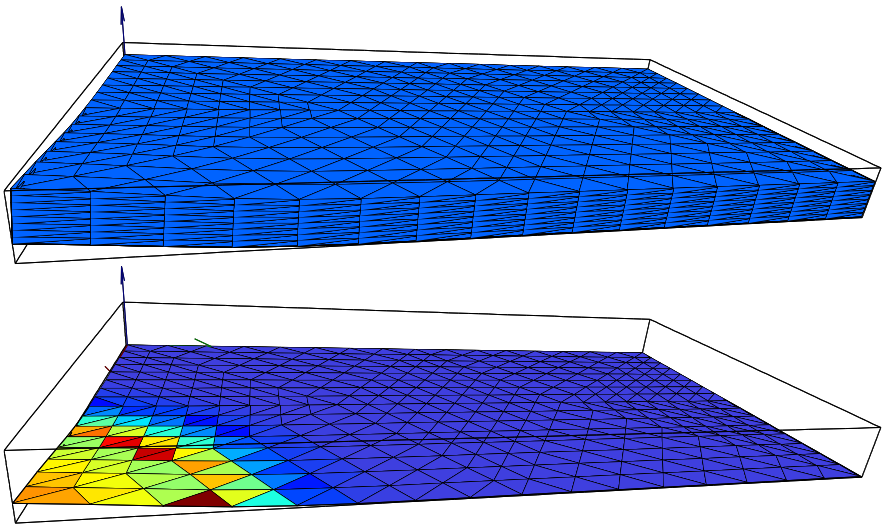


Figure 5.12.: Deformed mesh and reaction force distribution, BE computation

5.3.3. Finite Elements with PDAS

In this example, the domain Ω is a $5 \times 5 \times 2$ block, of which a bottom layer with thickness h is chipped off. We used $h = 0.15$ for our computation. Again, $x_1 = 0$ is the Dirichlet boundary Γ_D, and the load is applied on a 1×1 square on the top. This time, we push down in x_3 direction. We can apply the primal-dual active set algorithm as long as $\hat{b}(\cdot)$ is constant and positive up to some point, then zero. The plane $x_3 = -0.8$ is used as an obstacle; in a distance of maximally 0.2 to the obstacle, a reaction force of size 0.2 appears.

The convergence of Algorithm 4.13 depends strongly on the choice of the damping parameters ϑ_i. Here, we chose $\vartheta_0 = 1.0$ and $\vartheta_1 = \vartheta_2 = 0.01$. The choice of $\vartheta_0 = 1$ corresponds to the direct stress transmission in the Dirichlet-Neumann iteration in [36, Section 1.3]. The parameters ϑ_1, ϑ_2 needed to be chosen very small. Note that this also was the case for a purely linear reference benchmark that we implemented for the domain decomposition method. Small parameters do not break the algorithm's convergence, they only influence the number of steps up to some desired accuracy.

The convergence for the algorithm, applied to the adhesion benchmark, is documented in Figure 5.13: Two norms are displayed for two different mesh sizes. First, we computed the energy norm for the membrane update steps; second, we computed the energy norm for the update steps in the linear elastic block Ω^1. As the stiffness matrices are already known, we can simply compute

$$\begin{aligned}\left\|u_3^{(n+1)} - u_3^{(n)}\right\|^2_{H^1(\omega)} &= \left(x_3^{(n+1)} - x_3^{(n)}\right)^\top A_{\text{Lap}} \left(x_3^{(n+1)} - x_3^{(n)}\right);\\ \left\|\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}\right\|^2_{\mathbf{H}^1(\Omega^1)} &= \left(x_\Omega^{(n+1)} - x_\Omega^{(n)}\right)^\top A_{\text{Lamé}} \left(x_\Omega^{(n+1)} - x_\Omega^{(n)}\right).\end{aligned}$$

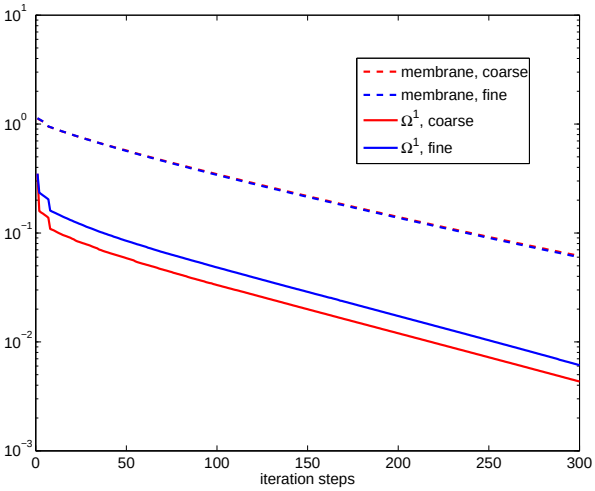


Figure 5.13.: Energy norm of update steps, with inner PDAS iteration

The "coarse" mesh was created by `tetgen` with a volume constraint of 0.001. This resulted in 42,354 degrees of freedom in Ω^1 and 1273 degrees of freedom for the membrane problem in $\Gamma_\uparrow$. The transmission boundary $\Gamma_\uparrow$ consisted of 2473 elements. The "fine" mesh was created with a volume constraint of 0.0001. Here, we got 413,460 degrees of freedom in Ω^1 and 6044 degrees of freedom for the membrane problem in $\Gamma_\uparrow$. The transmission boundary $\Gamma_\uparrow$ consisted of 11,931 elements.

Note that the convergence rates for the displayed subproblems are very similar. Moreover, the size of the update steps does not strongly depend on the meshwidth for the problem in Ω^1; for the membrane subproblem, there seems to be no dependence on the meshwidth. The same behavior was retrieved for other meshes with volume constraints of 0.003 and 0.0003, which have been left out of Figure 5.13 for the sake of readability.

step	0.003	0.001	0.0003	0.0001
50	0.039	0.058	0.069	0.084
100	0.023	0.032	0.039	0.048
150	0.014	0.020	0.024	0.028
200	0.0085	0.012	0.015	0.017
250	0.0052	0.0071	0.0088	0.010
300	0.0035	0.0043	0.0055	0.0061

step	0.003	0.001	0.0003	0.0001
50	0.57	0.57	0.56	0.56
100	0.34	0.35	0.34	0.34
150	0.22	0.22	0.21	0.21
200	0.14	0.14	0.14	0.14
250	0.091	0.095	0.091	0.089
300	0.061	0.064	0.062	0.060

Table 5.3.: Energy norm of update steps in Ω^1 (left) and ω (right) for meshes with $v_{\max} = 0.003, 0.001, 0.0003, 0.0001$

Figure 5.14 shows the deformed mesh. The vertical, red surface is the Dirichlet boundary. The membrane is shown in dark blue, it is in contact in the lower right corner.

The von Mises stress distribution inside the domain is given in Figure 5.15, which shows the deformed mesh from another perspective.

Parallelization and efficient computation

Again, the computation of stiffness matrices and load vectors was done in parallel with a shared-memory model, using `OpenMP`. This was important for the matrix $M_{\Gamma_\downarrow}$, where each element on $\Gamma_\uparrow$ had to be compared with each element on $\Gamma_\downarrow$ due to the naive ansatz we used. The matrix could be computed in seconds through parallelization.

The main computational cost arose from solving the problem in Ω^1. Our central intent was to show the method itself. Thus, we solved that problem with a plain CG method. The only option to parallelize the computation was the implementation of a parallel matrix-vector multiplication. This scaled only sublinearly, as the memory bus showed to be the bottleneck.
A large set of preconditioners exists that can be applied to large scale linear-elastic problems. This includes domain decomposition methods and block preconditioning, see e.g. Saad [38, Chapter 12] for an overview.

Although the PDAS method for the membrane problem needed several iterations in each step, the linear elastic problem in Ω^1 took more time to solve. This can be

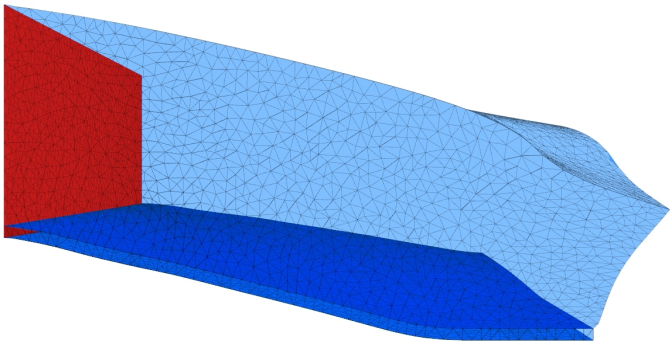


Figure 5.14.: Deformed mesh, PDAS computation

seen in Table 5.4 for four different meshes with the volume constraint as given in the first line. Recall that we had three subproblems: The linear elastic problem in Ω^1, the $(u_1, u_2, 0)$ problem in Ω^2, and the membrane problem in ω. We give the number of unknowns in each problem, together with the solution time. For the first and second problem, we give the average number of CG iterations; for the membrane problem, we give the average number of PDAS iterations (recall that each PDAS step contains one full CG run). Additionally, we state the condition number of each stiffness matrix.

The second subproblem took a surprisingly large fraction of the total computation time. This is due to the ill-conditioned system matrix, so more CG iterations are needed.

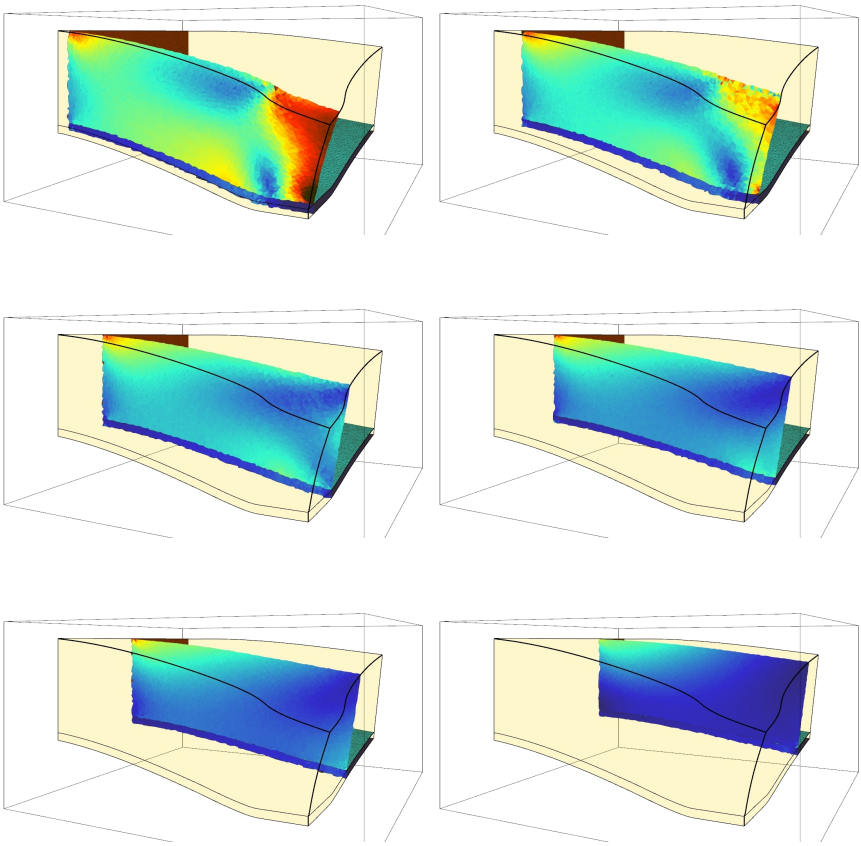


Figure 5.15.: Deformed mesh, PDAS computation: stress distribution

	$v_{\max}$	0.003	0.001	0.0003	0.0001
in Ω^1:	N	14364	42354	139077	413460
	time	39s	262s	1236s	5688s
	CG steps	316	454	653	947
	$\kappa(A_{\Omega^1})$	1.7e+3	3.7e+3	8.8e+3	1.6e+4
$(u_1, u_2, 0)$:	N	2588	4770	14108	33872
	time	17s	58s	333s	1055s
	CG steps	425	522	871	1045
	$\kappa(A_{\Omega^2})$	4.1e+4	5.4e+4	1.6e+5	2.5e+5
membrane:	N	660	1273	3349	6044
	time	11s	24s	73s	285s
	PDAS steps	36	45	64	81
	$\kappa(A_\omega)$	2.1e+3	4.2e+3	1.1e+4	1.9e+4

Table 5.4.: Computation times and average iteration steps for all subproblems

A. Implementation

A.1. Basis functions and reference elements

Although the following definitions are common, we state them again, as the proof of Lemma 4.6 and the quadrature rules in section A.2 rely on them.

A.1.1. 2D elements

Define the reference element by the simplex

$$\bar{T} := \left\{(\xi, \eta) \in \mathbb{R}^2 : \ \xi \in [0,1],\ \eta \in [0,1],\ \xi + \eta \leq 1\right\}. \tag{A.1}$$

The transformation from $\bar{T}$ to an actual element T_i is done by the affine mapping

$$\begin{aligned} h : \bar{T} \subset \mathbb{R}^2 \ &\to \ T_i \\ (\xi_1, \xi_2) \ &\mapsto \begin{pmatrix} d_1^{(1)} & d_1^{(2)} \\ d_2^{(1)} & d_2^{(2)} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + P^{(0)} \ = \ H\boldsymbol{\xi} + P^{(0)}\,, \end{aligned} \tag{A.2}$$

where $P^{(k)}$ are the element nodes and $d^{(k)} := P^{(k)} - P^{(0)}$. This mapping is bijective because H is invertible: The direction vectors $d^{(k)}$ are linearly independent.

The standard linear local basis functions on $\bar{T}$, which are used for 2D FEM and 3D BEM on surface meshes, are given by

$$\bar{\varphi}_1(\xi, \eta) \ = \ 1 - \xi - \eta\,; \quad \bar{\varphi}_2(\xi, \eta) \ = \ \xi\,; \quad \bar{\varphi}_3(\xi, \eta) \ = \ \eta\,. \tag{A.3}$$

The inverse of H can be computed explicitly:

$$H^{-1} \ = \ \frac{1}{\det H} \begin{pmatrix} H_{22} & -H_{12} \\ -H_{21} & H_{11} \end{pmatrix} =: \begin{pmatrix} h_{11}^- & h_{12}^- \\ h_{21}^- & h_{22}^- \end{pmatrix} \tag{A.4}$$

Writing the inverse of (A.2) as

$$\begin{aligned} \xi_1(x_1, x_2) \ &= h_{11}^-(x_1 - P_1^{(0)}) + h_{12}^-(x_2 - P_2^{(0)}) \\ \xi_2(x_1, x_2) \ &= h_{21}^-(x_1 - P_1^{(0)}) + h_{22}^-(x_2 - P_2^{(0)})\,, \end{aligned}$$

the total differentials of a function $\varphi(\mathbf{x})$ are

$$\begin{aligned}\frac{\partial\varphi}{\partial x_1} &= \frac{\partial\varphi}{\partial\xi_1}\frac{\partial\xi_1}{\partial x_1} + \frac{\partial\varphi}{\partial\xi_2}\frac{\partial\xi_2}{\partial x_1} = h_{11}^{-}\frac{\partial\varphi}{\partial\xi_1} + h_{21}^{-}\frac{\partial\varphi}{\partial\xi_2}\\ \frac{\partial\varphi}{\partial x_2} &= \frac{\partial\varphi}{\partial\xi_1}\frac{\partial\xi_1}{\partial x_2} + \frac{\partial\varphi}{\partial\xi_2}\frac{\partial\xi_2}{\partial x_2} = h_{12}^{-}\frac{\partial\varphi}{\partial\xi_1} + h_{22}^{-}\frac{\partial\varphi}{\partial\xi_2}.\end{aligned}$$

The gradient of a function φ in global coordinates (x_1, y_1) on T_i can then be expressed in local coordinates (ξ, η) on $\bar{T}$:

$$\nabla_x\varphi(\mathbf{x}) = H^{-\top}\nabla_\xi\varphi(h(\boldsymbol{\xi})). \tag{A.5}$$

A.1.2. 3D elements

The reference element is now

$$\bar{T} := \left\{(\xi,\eta,\zeta)\in\mathbb{R}^3 : \xi\in[0,1],\ \eta\in[0,1],\ \zeta\in[0,1],\ \xi+\eta+\zeta\le 1\right\}. \tag{A.6}$$

Then we get the transformation to an actual element T_i:

$$\begin{aligned}h : \bar{T}\subset\mathbb{R}^3 &\to T_i\\ (\xi_1,\xi_2,\xi_3) &\mapsto \begin{pmatrix} d_1^{(1)} & d_1^{(2)} & d_1^{(3)}\\ d_2^{(1)} & d_2^{(2)} & d_2^{(3)}\\ d_3^{(1)} & d_3^{(2)} & d_3^{(3)}\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\\ \xi_3\end{pmatrix} + P^{(0)}.\end{aligned} \tag{A.7}$$

Again, h and H are invertible because $d^{(k)}$ are linearly independent.

The standard linear local basis functions on $\bar{T}$, which are used for 2D FEM and 3D BEM on surface meshes, are given by

$$\begin{aligned}\bar{\varphi}_1(\xi,\eta,\zeta) &= 1-\xi-\eta-\zeta; & \bar{\varphi}_2(\xi,\eta,\zeta) &= \xi;\\ \bar{\varphi}_3(\xi,\eta,\zeta) &= \eta; & \bar{\varphi}_4(\xi,\eta,\zeta) &= \zeta.\end{aligned} \tag{A.8}$$

A.2. Quadrature rules

To evaluate the Galerkin matrix entries, integrals over single elements need to be computed. These elements are, for our benchmarks, triangles (for FE computations in $2d$) or tetrahedra (for FE computations in $3d$). For BE computations in $3d$, nested integrations over an inner and an outer element need to be performed.

The involved matrix H is constant, giving the integral transformation

$$\int_{T_i} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \left|\det H\right|\int_{\bar{T}} f(h(\boldsymbol{\xi}))\,\mathrm{d}\xi$$

The following sections describe the actual computation of matrix entries, as actually used in our benchmarks:

A.2.1. Regular quadrature

For the finite element method, the Galerkin matrix entries have the following form:

$$a_{ij} \;:=\; \sum_{T\subset(\operatorname{supp}\phi_i\cap\operatorname{supp}\phi_j)} \int_T \Big(D\boldsymbol{\varphi}_i(\mathbf{x})\Big)^{\top} \alpha(\mathbf{x}) \Big(D\boldsymbol{\varphi}_j(\mathbf{x})\Big)\,\mathrm{d}\mathbf{x}\,, \tag{A.9}$$

where $\boldsymbol{\varphi}_{i/j}$ are basis functions of V_h, D is a differential operator (e.g. the gradient) and $\alpha(\mathbf{x})$ is a given mapping preserving ellipticity of a. In the case of homogeneous linear elasticity, D was taken as the strain tensor ε_{ij}, and α was assumed to be the constant Hooke tensor $\mathbb{C}_{ijkl}$.
These integrals can be evaluated analytically. However, the computation of Galerkin matrices was not a time-critical step in our simulations, compared to the solution algorithms. Numerical quadrature was used to provide flexibility in the choice of local basis functions.

Here, we first use a tensor product rule to create quadrature nodes on the unit square or the unit cube. Next, the so called *Duffy transformation* is used to map these onto the reference triangle.
From a one-dimensional Gauss quadrature rule,

$$\int_{-1}^{1} f(\tau)\,\mathrm{d}\tau \;\approx\; \sum_{i=1}^{n} \tilde{\omega}_i\, f(\tilde{\tau}_i)\,,$$

we can set up a quadrature on $(0,1)$ by

$$\tau_i \;:=\; \frac{\tilde{\tau}_i+1}{2}\,; \quad \omega_i \;:=\; \frac{1}{2}\tilde{\omega}_i\,.$$

Nodes and weights for the reference square $(0,1)^2$ and cube $(0,1)^3$ are defined by the tensor product rule

$$\begin{aligned}
\boldsymbol{\tau}_i^{\square} &:= (\tau_k,\tau_l)\,, & \omega_i^{\square} &:= \omega_k\omega_l\,, & i &= n(k-1)+l\,;\\
\boldsymbol{\tau}_i^{\square\!\square\!\square} &:= (\tau_k,\tau_l,\tau_m)\,, & \omega_i^{\square\!\square\!\square} &:= \omega_k\omega_l\omega_m\,, & i &= n^2(k-1)+n(l-1)+m
\end{aligned}$$

for $k,l,m\in\{1,\dots,n\}$.

The Duffy transformation from the square to the reference triangle is given by

$$h_D \,:\, (\tau_1^{\square},\,\tau_2^{\square}) \;\mapsto\; (\tau_1^{\square},\,(1-\tau_1^{\square})\tau_2^{\square})\,, \tag{A.10}$$

$$\int_0^1 (1-\tau_1^{\square}) \int_0^1 f\Big(\tau_1^{\square},\,(1-\tau_1^{\square})\tau_2^{\square}\Big)\,\mathrm{d}\tau_2^{\square}\,\mathrm{d}\tau_1^{\square} \;=\; \int_0^1\int_0^{1-\xi} f(\xi,\eta)\,\mathrm{d}\eta\,\mathrm{d}\xi\,, \tag{A.11}$$

where (ξ,η) are local coordinates in the triangle. Thus, the quadrature nodes $\boldsymbol{\tau}_i^{\square}$ and weights $\omega_i^{\square}$ on the square are transformed to

$$\boldsymbol{\xi}_i \;=\; (\xi_i,\,\eta_i) \;:=\; (\tau_{i,1}^{\square},\,(1-\tau_{i,1}^{\square})\tau_{i,2}^{\square}) \;=\; \left(\frac{\tilde{\tau}_k+1}{2},\,\Big(1-\frac{\tilde{\tau}_k+1}{2}\Big)\frac{\tilde{\tau}_l+1}{2}\right) \tag{A.12}$$

$$\omega_i \;=\; (1-\tau_{i,1}^{\square})\,\omega_i^{\square} \;=\; \frac{1-\tilde{\tau}_k}{8}\tilde{\omega}_k\tilde{\omega}_l \tag{A.13}$$

for $i = n(k-1)+l$, where $k,l \in \{1,\dots,n\}$. The $2d$ quadrature rule on the reference triangle then has n^2 nodes and weights.

Similarly, the transformation from the cube to the reference tetrahedron is given by

$$h_D : (\tau_1^{\square}, \tau_2^{\square}, \tau_3^{\square}) \mapsto (\tau_1^{\square}, (1-\tau_1^{\square})\tau_2^{\square}, (1-\tau_1^{\square})(1-\tau_2^{\square})\tau_3^{\square}) , \tag{A.14}$$

$$\int_0^1 (1-\tau_1^{\square}) \int_0^1 (1-\tau_1^{\square})(1-\tau_2^{\square}) \int_0^1 f\Big(\tau_1^{\square}, (1-\tau_1^{\square})\tau_2^{\square}, (1-\tau_1^{\square})(1-\tau_2^{\square})\tau_3^{\square}\Big)\,\mathrm{d}\boldsymbol{\tau}$$
$$= \int_0^1 \int_0^{1-\xi} \int_0^{1-\xi-\eta} f(\xi,\eta,\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\eta\,\mathrm{d}\xi , \tag{A.15}$$

where the local coordinates are (ξ,η,ζ). The quadrature nodes $\boldsymbol{\tau}_i^{\square}$ and weights $\omega_i^{\square}$ are transformed to nodes and weights on the reference tetrahedron,

$$\boldsymbol{\xi}_i = (\xi_i, \eta_i, \zeta_i) := (\tau_{i,1}^{\square}, (1-\tau_{i,1}^{\square})\tau_{i,2}^{\square}, (1-\tau_{i,1}^{\square})(1-\tau_{i,2}^{\square})\tau_{i,3}^{\square}) \tag{A.16}$$
$$= \Big(\frac{\tilde\tau_k+1}{2}, \Big(1-\frac{\tilde\tau_k+1}{2}\Big)\frac{\tilde\tau_l+1}{2}, \Big(1-\frac{\tilde\tau_k+1}{2}\Big)\Big(1-\frac{\tilde\tau_l+1}{2}\Big)\frac{\tilde\tau_m+1}{2}\Big)$$

$$\omega_i = (1-\tau_{i,1}^{\square})^2(1-\tau_{i,2}^{\square})\,\omega_i^{\square} = \frac{(1-\tilde\tau_k)^2(1-\tilde\tau_l)}{8}\tilde\omega_k\tilde\omega_l\tilde\omega_m \tag{A.17}$$

for $i = n^2(k-1)+n(l-1)+m$, where $k,l,m \in \{1,\dots,n\}$. The $3d$ rule for the reference tetrahedron will have n^3 nodes and weights.

Finally, we retrieve the quadrature rule

$$\int_{\bar T} f(\boldsymbol{\xi})\,\mathrm{d}\xi \approx \sum_{i=1}^{(N)} \omega_i\, f(\boldsymbol{\xi}_i) . \tag{A.18}$$

Remark A.1: *The transformed quadrature rules are still exact to some polynomial orders $p_\xi, p_\eta(, p_\zeta)$. Note however that the original order of exactness is diminished due to the transformation factors, and due to the fact that nodes in the reference element are represented by products of quadrature nodes.*

Remark A.2: *The Duffy transformation was originally used to reduce orders of singularities fixed in one specified triangle (tetrahedron) corner by contracting an edge (face). This leads to a concentration of nodes towards one node, leaving these quadrature rules not symmetric. They are also not* minimal *in the sense that a rule for the same order of exactness may be given with a smaller number of nodes. Derivation of a general formula for minimal, symmetric quadrature nodes is still an open problem.*

A survey of quadrature rules on various element shapes was given by Stroud [44]; some minimal rules are also given explicitly there.

A.2.2. Adaptive quadrature

For the boundary element method, the Galerkin matrix is in general fully populated. Entries will have the following form:

$$a_{ij} := \left\langle P\varphi_i, \psi_j \right\rangle_\Sigma = \sum_{T_i \subset \operatorname{supp} \varphi_i} \sum_{T_j \subset \operatorname{supp} \psi_j} \int_{T_j} \int_{T_i} k(\mathbf{x}, \mathbf{y}) \varphi_i(\mathbf{y}) \, \mathrm{d}s_y \, \psi_j(\mathbf{x}) \, \mathrm{d}s_x \tag{A.19}$$

Here, P is an integral operator with the kernel $k(\cdot, \cdot)$, and φ_i, ψ_j are basis functions of the test and ansatz spaces. Note that for the double layer matrix K, these spaces are actually different. Also note that for the hypersingular operator, the surface gradients of φ_i and ψ_j are used instead.
Only the $3d$ case is considered here, where $\mathcal{T}$ is a decomposition of the boundary of interest (Γ or Σ) into triangles T_i.

An analytical representation of these integrals is only given for special cases. A quite general analysis, including analytical evaluation on rectangles for integral operators associated with the Laplace, Lamé and Helmholtz problems, was done by Maischak [30, p.151ff].
For the Lamé equation, no analytical evaluation on triangles is known yet, so a quadrature scheme has to be used again.

The kernel k creates singularities in the integrand. For this reason, increasing the number of quadrature points might not lead to an increase of accuracy: The approximation error is given in terms of higher order derivatives of the integrand, which are unbounded.

Example A.3: *For a $1d$ Gaussian quadrature rule, the approximation error can be expressed as*

$$\int_{-1}^{1} f(x) \, dx - \sum_{i=1}^{n} \omega_i f(x_i) = \frac{f^{(2n)}(\xi)}{(2n)!} \int_{-1}^{1} \left[\frac{n!}{(2n)!} \frac{d^n}{dx^n} \left(x^2 - 1 \right)^n \right]^2 dx$$

for some $\xi \in (-1, 1)$, see [43]. The function

$$f(x) = \frac{1}{\sqrt{x+1}}$$

is integrable on $(-1, 1)$ (the integral is $2\sqrt{2}$), but singular in -1. Its n-th derivative is

$$f^{(n)}(x) = \frac{(-1)^n \, (n)!!}{2^n \, (x+1)^{\frac{(2n+1)}{2}}} \, ,$$

which is unbounded for $x \to -1$, so the standard approximation may not ensure convergence here.

In our case, the kernel $k(\mathbf{x}, \mathbf{y})$ will have a singularity for $\mathbf{x} \to \mathbf{y}$. For the integration on an outer triangle T_i and an inner triangle T_j, there are several different cases to be considered. The transformations are performed according to [13] and [39, 5.2], but other quadrature rules could be used here that take singularities into account (e.g. the triangle quadrature rule in [40], which would need to be nested).
Note that an intermediate reference triangle $\tilde{T}$ is used here:

$$\tilde{T} = \left\{(\xi, \eta) : \xi \in [0,1],\ \eta \in [0, \xi]\right\}$$

The final quadrature rules (A.20), (A.22) are again given in local coordinates of our reference triangle $\bar{T}$.

Case 0: No intersection

Here, the triangles do not intersect at all, $T_i \cap T_j = \{\}$.
In this case, the points $\mathbf{y} \in T_i$ are bounded away from $\mathbf{x} \in T_j$, and no singularity will occur. We can use regular quadrature rules on both triangles, as in (A.12), (A.13).

Case 1: One common vertex

In this case, $T_i \cap T_j = \{p\}$, where p is a vertex of both T_i and T_j. The double integral is now expressed as a four-dimensional integral over $\tilde{T} \times \tilde{T}$ with the singularity at the origin. This domain is then decomposed into two domains,

$$\begin{aligned} D_1 &= \left\{\lambda_1 \in [0,1];\ \lambda_2 \in [0, \lambda_1];\ \lambda_3 \in [0, \lambda_1];\ \lambda_4 \in [0, \lambda_3]\right\} \\ D_2 &= \left\{\lambda_1 \in [0, \lambda_3];\ \lambda_2 \in [0, \lambda_1];\ \lambda_3 \in [0,1];\ \lambda_4 \in [0, \lambda_3]\right\}, \end{aligned}$$

which are both transformed to the unit hypercube $(0,1)^4$ to cope with the singularity. These transformations are

$$\begin{aligned} D_1: \quad (\lambda_1, \lambda_2) &\mapsto (\lambda_4, \lambda_4\lambda_1) \\ (\lambda_3, \lambda_4) &\mapsto (\lambda_4\lambda_2, \lambda_4\lambda_2\lambda_3) \\ D_2: \quad (\lambda_1, \lambda_2) &\mapsto (\lambda_4\lambda_2, \lambda_4\lambda_2\lambda_3) \\ (\lambda_3, \lambda_4) &\mapsto (\lambda_4, \lambda_4\lambda_1) \end{aligned}$$

The integral over $\tilde{T} \times \tilde{T}$ is then given by

$$\int_{(0,1)^4} \lambda_4^3\lambda_2 \left(f(\lambda_4, \lambda_4\lambda_1, \lambda_4\lambda_2, \lambda_4\lambda_2\lambda_3) + f(\lambda_4\lambda_2, \lambda_4\lambda_2\lambda_3, \lambda_4, \lambda_4\lambda_1)\right) \mathrm{d}\lambda_{1,2,3,4},$$

and finally we can apply the translations

$$\begin{aligned} (\tilde{\lambda}_1, \tilde{\lambda}_2) &\mapsto (\tilde{\lambda}_1 - \tilde{\lambda}_2, \tilde{\lambda}_2) \\ (\tilde{\lambda}_3, \tilde{\lambda}_4) &\mapsto (\tilde{\lambda}_3 - \tilde{\lambda}_4, \tilde{\lambda}_4) \end{aligned}$$

from $\tilde{T}$ to the original reference triangle $\bar{T}$:

$$\begin{aligned}
&\int_{\bar{T}}\int_{\bar{T}} f(\xi_1,\,\xi_2,\,\xi_3,\,\xi_4)\,\mathrm{d}\xi_{1,2,3,4} \\
&= \int_{(0,1)^4} \lambda_4^3\lambda_2\, f\Big(\lambda_4,\,\lambda_4(\lambda_1-\lambda_2),\,\lambda_4\lambda_2,\,\lambda_4\lambda_2(\lambda_3-\lambda_4)\Big)\,\mathrm{d}\lambda_{1,2,3,4} \\
&\quad + \int_{(0,1)^4} \lambda_4^3\lambda_2\, f\Big(\lambda_4\lambda_2,\,\lambda_4\lambda_2(\lambda_3-\lambda_4),\,\lambda_4,\,\lambda_4(\lambda_1-\lambda_2)\Big)\,\mathrm{d}\lambda_{1,2,3,4}\,.
\end{aligned}$$

As in the regular case, we take a $1d$ Gaussian quadrature $(\tilde{\tau}_\nu,\,\tilde{\omega}_\nu)_\nu$ as a generating rule for the integration over $(0,1)^4$ by applying the tensor product: The nodes and weights are again normalized to the interval $(0,1)$ by

$$\tau_\nu \;:=\; \frac{\tilde{\tau}_\nu+1}{2}\,,\quad \omega_\nu \;:=\; \frac{\tilde{\omega}_\nu}{2}\,.$$

This results in the quadrature nodes and weights

$$\begin{aligned}
\boldsymbol{\xi}_i^1 &= \Big(\tau_n-\tau_n\tau_l,\,\tau_n\tau_l\Big) & \boldsymbol{\eta}_i^1 &= \Big(\tau_k\tau_n-\tau_k\tau_m\tau_n,\,\tau_k\tau_m\tau_n\Big)\\
\boldsymbol{\xi}_i^2 &= \Big(\tau_k\tau_n-\tau_k\tau_m\tau_n,\,\tau_k\tau_m\tau_n\Big) & \boldsymbol{\eta}_i^2 &= \Big(\tau_n-\tau_n\tau_l,\,\tau_n\tau_l\Big)\\
\omega_i &= \tau_l\Big(\tau_n-\tau_n\tau_l\Big)^3\,\omega_k\omega_l\omega_m\omega_n\,.
\end{aligned} \tag{A.20}$$

The quadrature on the actual elements is then

$$\begin{aligned}
\int_{T_j\times T_i} k(\mathbf{x},\mathbf{y})\,\boldsymbol{\varphi}_i(\mathbf{y})\boldsymbol{\psi}_j(\mathbf{x}) \;\approx\;& |T_i|\,|T_j|\sum_{\nu=1}^{(N)}\omega_\nu\, k\Big(h_i(\boldsymbol{\xi}_\nu^1),h_j(\boldsymbol{\eta}_\nu^1)\Big)\,\boldsymbol{\varphi}_i\Big(h_j(\boldsymbol{\eta}_\nu^1)\Big)\,\boldsymbol{\psi}_j\Big(h_i(\boldsymbol{\xi}_\nu^1)\Big)\\
&+ |T_i|\,|T_j|\sum_{\nu=1}^{(N)}\omega_\nu\, k\Big(h_i(\boldsymbol{\xi}_\nu^2),h_j(\boldsymbol{\eta}_\nu^2)\Big)\,\boldsymbol{\varphi}_i\Big(h_j(\boldsymbol{\eta}_\nu^2)\Big)\,\boldsymbol{\psi}_j\Big(h_i(\boldsymbol{\xi}_\nu^2)\Big)\,.
\end{aligned} \tag{A.21}$$

Remark A.4: *The nodes $\boldsymbol{\xi}_\nu \mapsto h_i(\boldsymbol{\xi}_\nu)$ and $\boldsymbol{\eta}_\nu \mapsto h_j(\boldsymbol{\eta}_\nu)$ are chosen to be concentrated towards the corners $P_i^{(0)}$ and $P_j^{(0)}$ of T_i and T_j. If the triangles are connected in other points, there are two possibilities:*

- *Renumber the triangle nodes, such that h_i and h_j map the common node to $P^{(0)}$.*
- *Transform the quadrature nodes by*

$$\begin{aligned}
\tilde{\boldsymbol{\xi}}_i &= \begin{pmatrix}-1 & -1\\ 1 & 0\end{pmatrix}\boldsymbol{\xi}_i + \begin{pmatrix}1\\0\end{pmatrix}\\
\textit{and}\quad \tilde{\boldsymbol{\xi}}_i &= \begin{pmatrix}0 & 1\\ -1 & -1\end{pmatrix}\boldsymbol{\xi}_i + \begin{pmatrix}0\\1\end{pmatrix}
\end{aligned}$$

  and store the nodes for each pairing, leading to 9 stored transformed quadrature rules.

Case 2: One common edge (two common vertices)

If T_i and T_j share a full common edge e, the double integral over $\tilde{T}\times\tilde{T}$ is transformed to the integral

$$\int_0^1 \int_{-\lambda_4}^{1-\lambda_4} \int_0^{\lambda_1+\lambda_4} \int_0^{\lambda_4} f(\lambda_4,\, \lambda_3,\, \lambda_1+\lambda_4,\, \lambda_2)\, \mathrm{d}\lambda_{1,2,3,4}\,,$$

mapping the edge e to $(0,0,0,t)$, $t\in[0,1]$.
The domain is now decomposed into five subdomains,

$$\begin{aligned}
D_1 &= \left\{\lambda_1\in[-1,0]\,;\ \lambda_2\in[0,1+\lambda_1]\,;\ \lambda_3\in[0,\lambda_2-\lambda_1]\,;\ \lambda_4\in[\lambda_2-\lambda_1,1]\right\}\\
D_2 &= \left\{\lambda_1\in[-1,0]\,;\ \lambda_2\in[0,1+\lambda_1]\,;\ \lambda_3\in[\lambda_2-\lambda_1,1]\,;\ \lambda_4\in[\lambda_3,1]\right\}\\
D_3 &= \left\{\lambda_1\in[0,1]\,;\ \lambda_2\in[0,\lambda_1]\,;\ \lambda_3\in[0,1-\lambda_1]\,;\ \lambda_4\in[\lambda_3,1-\lambda_1]\right\}\\
D_4 &= \left\{\lambda_1\in[0,1]\,;\ \lambda_2\in[\lambda_1,1]\,;\ \lambda_3\in[0,\lambda_2-\lambda_1]\,;\ \lambda_4\in[\lambda_2-\lambda_1,1-\lambda_1]\right\}\\
D_5 &= \left\{\lambda_1\in[0,1]\,;\ \lambda_2\in[\lambda_1,1]\,;\ \lambda_3\in[\lambda_2-\lambda_1,1-\lambda_1]\,;\ \lambda_4\in[\lambda_3,1-\lambda_1]\right\},
\end{aligned}$$

which are again mapped from $(0,1)^4$ by

$$\begin{aligned}
D_1:\quad & (\lambda_1,\,\lambda_2) \mapsto (\lambda_4,\,-\lambda_4\lambda_1\lambda_2)\\
& (\lambda_3,\,\lambda_4) \mapsto (\lambda_4\lambda_1(1-\lambda_2),\,\lambda_4\lambda_1\lambda_3)\\
D_2:\quad & (\lambda_1,\,\lambda_2) \mapsto (\lambda_4,\,-\lambda_1\lambda_2\lambda_3\lambda_4)\\
& (\lambda_3,\,\lambda_4) \mapsto (\lambda_4\lambda_1\lambda_2(1-\lambda_3),\,\lambda_4\lambda_1)\\
D_3:\quad & (\lambda_1,\,\lambda_2) \mapsto (\lambda_4(1-\lambda_1\lambda_2),\,\lambda_4\lambda_1\lambda_2)\\
& (\lambda_3,\,\lambda_4) \mapsto (\lambda_1\lambda_2\lambda_3\lambda_4,\,\lambda_4\lambda_1(1-\lambda_2))\\
D_4:\quad & (\lambda_1,\,\lambda_2) \mapsto (\lambda_4(1-\lambda_1\lambda_2\lambda_3),\,\lambda_1\lambda_2\lambda_3\lambda_4)\\
& (\lambda_3,\,\lambda_4) \mapsto (\lambda_4\lambda_1,\,\lambda_4\lambda_1\lambda_2(1-\lambda_3))\\
D_5:\quad & (\lambda_1,\,\lambda_2) \mapsto (\lambda_4(1-\lambda_1\lambda_2\lambda_3),\,\lambda_1\lambda_2\lambda_3\lambda_4)\\
& (\lambda_3,\,\lambda_4) \mapsto (\lambda_4\lambda_1\lambda_2,\,\lambda_4\lambda_1(1-\lambda_2\lambda_3))\,.
\end{aligned}$$

The Jacobian is $\lambda_4^3\lambda_1^2$ for the D_1 transformation and $\lambda_4^3\lambda_1^2\lambda_2$ for the other transformations.
We can use the $(0,1)$-normalized rule (τ_ν,ω_ν) again and apply the final transforma-

tion to map to our reference element, returning the quadrature nodes and weights

$$\begin{aligned}
\boldsymbol{\xi}_i^1 &= \left(\tau_n - \tau_n\tau_k\tau_m,\ \tau_n\tau_k\tau_m\right) & \boldsymbol{\eta}_i^1 &= \left(\tau_n - \tau_k\tau_n,\ \tau_k\tau_n - \tau_k\tau_l\tau_n\right)\\
\omega_i^1 &= \tau_n^3\tau_k^2\,\omega_k\omega_l\omega_m\omega_n
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\xi}_i^2 &= \left(\tau_n - \tau_n\tau_k,\ \tau_n\tau_k\right) & \boldsymbol{\eta}_i^2 &= \left(\tau_n - \tau_n\tau_k\tau_l,\ \tau_n\tau_k\tau_l - \tau_k\tau_l\tau_m\tau_n\right)\\
\boldsymbol{\xi}_i^3 &= \boldsymbol{\eta}_i^1 & \boldsymbol{\eta}_i^3 &= \left(\tau_n - \tau_k\tau_l\tau_m\tau_n,\ \tau_k\tau_l\tau_m\tau_n\right)\\
\boldsymbol{\xi}_i^4 &= \boldsymbol{\eta}_i^2 & \boldsymbol{\eta}_i^4 &= \boldsymbol{\xi}_i^2\\
\boldsymbol{\xi}_i^5 &= \left(\tau_n - \tau_n\tau_k,\ \tau_n\tau_k - \tau_k\tau_l\tau_m\tau_n\right) & \boldsymbol{\eta}_i^5 &= \left(\tau_n - \tau_n\tau_k\tau_l,\ \tau_n\tau_k\tau_l\right)\\
\omega_i &= \tau_n^3\tau_k^2\tau_l\,\omega_k\omega_l\omega_m\omega_n\,. & & \qquad\qquad (A.22)
\end{aligned}$$

(The coordinates were permutated and combined here in order to minimize computational costs.)

Again, terms can be collected like in (A.21), and permutations of the edges lead to 9 transformed quadrature rules like in Remark A.4.

Case 3: Common face (three common vertices)

If $T_i = T_j$, the integration domain $\bar{T}\times\bar{T}$ is decomposed into six subdomains $D_1, \dots, D_6$. All of these are then mapped to $\bar{T}_3 \times (0,1)$, where $\bar{T}_3$ is the reference tetrahedron, using the transformations

$$\begin{aligned}
D_1:\quad & (\lambda_1,\lambda_2) \mapsto (\lambda_1+\lambda_2+\lambda_3,\ \lambda_1+\lambda_2)\\
& (\lambda_3,\lambda_4) \mapsto (\lambda_1(1-\lambda_4)+\lambda_2+\lambda_3,\ \lambda_2)\\
D_2:\quad & (\lambda_1,\lambda_2) \mapsto (\lambda_1(1-\lambda_4)+\lambda_2+\lambda_3,\ \lambda_1(1-\lambda_4)+\lambda_2)\\
& (\lambda_3,\lambda_4) \mapsto (\lambda_1+\lambda_2+\lambda_3,\ \lambda_2)\\
D_3:\quad & (\lambda_1,\lambda_2) \mapsto (\lambda_1+\lambda_2+\lambda_3,\ \lambda_1\lambda_4+\lambda_2)\\
& (\lambda_3,\lambda_4) \mapsto (\lambda_2+\lambda_3,\ \lambda_2)\\
D_4:\quad & (\lambda_1,\lambda_2) \mapsto (\lambda_1(1-\lambda_4)+\lambda_2+\lambda_3,\ \lambda_2)\\
& (\lambda_3,\lambda_4) \mapsto (\lambda_1+\lambda_2+\lambda_3,\ \lambda_1+\lambda_2)\\
D_5:\quad & (\lambda_1,\lambda_2) \mapsto (\lambda_1+\lambda_2+\lambda_3,\ \lambda_2)\\
& (\lambda_3,\lambda_4) \mapsto (\lambda_1(1-\lambda_4)+\lambda_2+\lambda_3,\ \lambda_1(1-\lambda_4)+\lambda_2)\\
D_6:\quad & (\lambda_1,\lambda_2) \mapsto (\lambda_2+\lambda_3,\ \lambda_2)\\
& (\lambda_3,\lambda_4) \mapsto (\lambda_1+\lambda_2+\lambda_3,\ \lambda_1\lambda_4+\lambda_2)\,;
\end{aligned}$$

here, $(\lambda_1,\lambda_2,\lambda_3) \in \bar{T}_3$ and $\lambda_4 \in (0,1)$. The integrand is now analytic.
We use regular quadrature rules $(\boldsymbol{\tau}_\nu^\triangle, \omega_\nu^\triangle)$ on the tetrahedron and (τ_k, ω_k) on the

interval. Finally, mapping the quadrature nodes to our reference triangle $\bar{T}$, we arrive at the following nodes and weights:

$$\begin{aligned}
\xi_i^1 &= \left(\tau_{\nu,3}^{\triangle},\ \tau_{\nu,1}^{\triangle} + \tau_{\nu,2}^{\triangle}\right) & \eta_i^1 &= \left(\tau_{\nu,1}^{\triangle}(1-\tau_k) + \tau_{\nu,3}^{\triangle},\ \tau_{\nu,2}^{\triangle}\right) \\
\xi_i^2 &= \left(\tau_{\nu,3}^{\triangle},\ \tau_{\nu,1}^{\triangle}(1-\tau_k) + \tau_{\nu,2}^{\triangle}\right) & \eta_i^2 &= \left(\tau_{\nu,1}^{\triangle} + \tau_{\nu,3}^{\triangle},\ \tau_{\nu,2}^{\triangle}\right) \\
\xi_i^3 &= \left(\tau_{\nu,1}^{\triangle} - \tau_{\nu,1}^{\triangle}\tau_k + \tau_{\nu,3}^{\triangle},\ \tau_{\nu,1}^{\triangle}\tau_k + \tau_{\nu,2}^{\triangle}\right) & \eta_i^3 &= \left(\tau_{\nu,3}^{\triangle},\ \tau_{\nu,2}^{\triangle}\right) \\
\xi_i^4 &= \left(\tau_{\nu,1}^{\triangle}(1-\tau_k) + \tau_{\nu,3}^{\triangle},\ \tau_{\nu,2}^{\triangle}\right) & \eta_i^4 &= \left(\tau_{\nu,3}^{\triangle},\ \tau_{\nu,1}^{\triangle} + \tau_{\nu,2}^{\triangle}\right) \\
\xi_i^5 &= \left(\tau_{\nu,1}^{\triangle} + \tau_{\nu,3}^{\triangle},\ \tau_{\nu,2}^{\triangle}\right) & \eta_i^5 &= \left(\tau_{\nu,3}^{\triangle},\ \tau_{\nu,1}^{\triangle}(1-\tau_k) + \tau_{\nu,2}^{\triangle}\right) \\
\xi_i^6 &= \left(\tau_{\nu,3}^{\triangle},\ \tau_{\nu,2}^{\triangle}\right) & \eta_i^6 &= \left(\tau_{\nu,1}^{\triangle} + \tau_{\nu,3}^{\triangle} - \tau_{\nu,1}^{\triangle}\tau_k,\ \tau_{\nu,1}^{\triangle}\tau_k + \tau_{\nu,2}^{\triangle}\right) \\
\omega_i &= \tau_{\nu,1}^{\triangle}\omega_k\omega_\nu^{\triangle} & & \qquad\qquad (A.23)
\end{aligned}$$

with $i = n(k-1)+\nu$, where n is the number of quadrature nodes in the tetrahedron, $\nu \in \{1,\dots,n\}$ and $k \in \{1,\dots,N\}$.

A.3. Conforming adaptive refinement

Section 5.1 introduces an error indicator. It is desirable to refine the given mesh only in places where the error is assumed to be large, as a coarser mesh will result in smaller equation systems.

If the mesh is refined uniformly, no hanging nodes are introduced. This is different for adaptive refinement techniques, where hanging nodes need to be taken care of. A refinement algorithm needs to eliminate hanging nodes and edges. A further requirement is that the quality of single elements must not decrease: For a 2D mesh, the angles of an element need to be bounded from below; for a 3D mesh, the tetrahedra must not degenerate to flat "slivers" (see [41, Section 5.1] for degenerate elements). For linear problems, this mesh quality has a direct impact on the error constants and the matrix condition. An example for this is given in Verfürth [46, Section 1.2].

In our implementation, we follow the strategies suggested in [46, Section 4.1]. For a 2D mesh, we use the following directions:

Algorithm A.5:

- Retrieve the maximal error indicator $\eta_{\max}$ and select some threshold constant $\vartheta \in (0,1)$.
- Mark all elements $T \in \mathcal{T}_h$ with $\eta_T > \vartheta\eta_{\max}$ `red`.

- For hanging nodes, mark the adjacent elements with a compatible `green` or `blue` marker. If the adjacent element has been refined before with `green` or `blue`, mark it `red` to retain mesh quality.
- Repeat the last step until all hanging nodes are removed. ♦

Here, a `red` refinement means that the triangle is decomposed into four similar subtriangles; a `green` refinement means that one edge is bisected; and a `blue` refinement means that two edges are marked for bisection. Here, we are free to choose a rotational direction $\texttt{blue}\cdot^{+}$ or $\texttt{blue}\cdot^{-}$ (see Figure A.1), and it is advisable to take the one with the largest minimal angle for mesh quality.

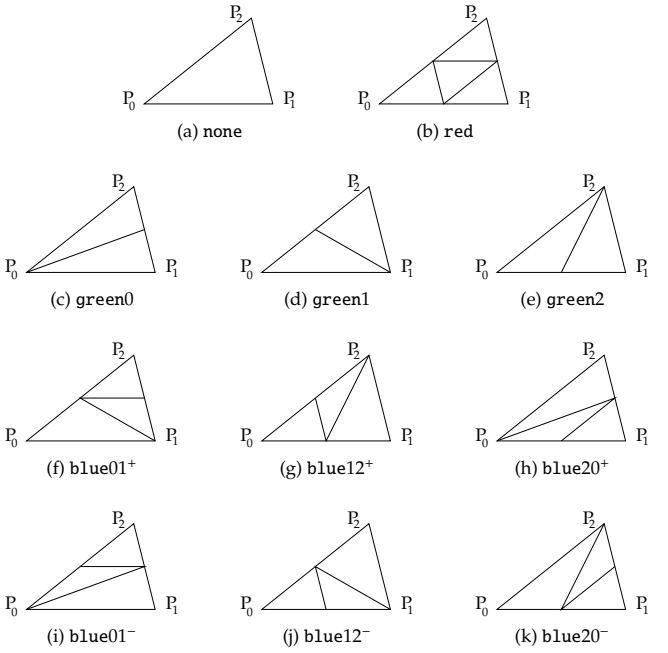


Figure A.1.: Possible refinements of a triangular element by edge bisection

The refinement of a tetrahedron is more involved, as there are more cases. In general, we proceed with the same refinement algorithm:

Algorithm A.6:

- Retrieve the maximal error indicator $\eta_{\max}$ and select some threshold constant $\vartheta \in (0, 1)$.

- Mark all elements $T \in \mathcal{T}_h$ with $\eta_T > \vartheta \eta_{\max}$ `red`.
- For hanging nodes or hanging edges, mark the adjacent elements with a compatible `green` or `blue` marker. This marker is chosen such that the refinement trace across the joint face matches: On a tetrahedron face, the refinement trace will be one of the possible 2D refinements given in Figure A.1. Here, the + and − versions of `blue` refinement are both needed, as the orientation of interface triangles will flip when matching traces.
  If the given refinements are not able to resolve all hanging nodes, there are three or more edges marked for refinement, and they are not coplanar. Mark this element `red`.
  If the adjacent element has been refined before with `green` or `blue`, mark it `red` to retain mesh quality.
- Repeat the last step until all hanging nodes and edges are removed. ♦

We get six general types of refinement, from which all cases are deduced by rotations and reflections:

1. `none`, the element is not refined.
2. `green2`, two faces are subdivided by a 2D `green` refinement. This effectively cuts the element in half along one edge.
3. `blue1green2`, one face is subdivided by `blue` and two faces are subdivided by `green` refinements. This cuts the element in three parts that all share one vertex.
4. `green4`, all faces are subdivided by `green` refinements. This first cuts the element into four parts along two planes.
5. `red1green3`, one face is subdivided by `red` and the other faces are subdivided by `green` refinements. This cuts the element in four parts that all share one vertex.
6. `red4`, all faces are subdivided by `red` refinements. This is similar to the 2D `red` refinement case. Note that an element will not be decomposed into similar sub-elements here: We can cut off the tips of a tetrahedron, giving four elements at the corners that are in fact scaled versions of the original element. The remaining part is an octahedron, where opposite faces are equal up to a half rotation. It can be decomposed into four further sub-elements.

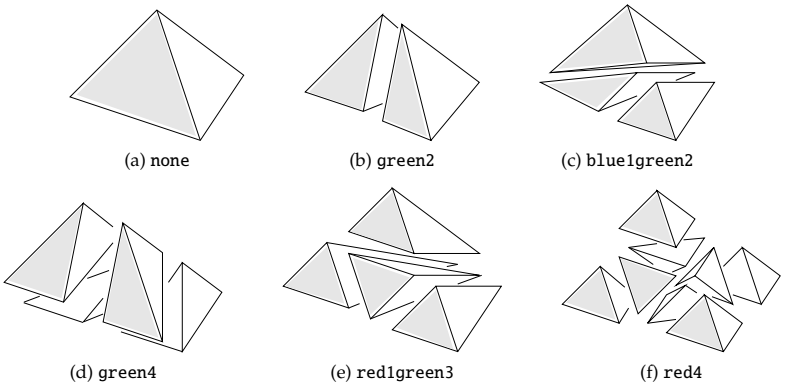


Figure A.2.: Possible refinements of a tetrahedral element by edge bisection

Bibliography

[1] S. Agmon. Maximum theorems for solutions of higher order elliptic equations. *Bull. Amer. Math. Soc.*, 66:77–80, 1960.

[2] J. Alberty, C. Carstensen, and D. Zarrabi. Adaptive numerical analysis in primal elastoplasticity with hardening. *Comput. Methods Appl. Mech. Engrg.*, 171(3-4):175–204, 1999.

[3] J. Altenbach and H. Altenbach. *Einführung in die Kontinuums-Mechanik*. Teubner Verlag, 1994.

[4] C. C. Baniotopoulos, J. Haslinger, and Z. Morávková. Mathematical modeling of delamination and nonmonotone friction problems by hemivariational inequalities. *Appl. Math.*, 50(1):1–25, 2005.

[5] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1979. Computer Science and Applied Mathematics.

[6] D. Braess. *Finite Elemente*. Springer-Verlag, 2007.

[7] CGAL. Computational Geometry Algorithms Library. http://www.cgal.org.

[8] A. Chernov. *Nonconforming boundary elements and finite elements for interface and contact problems with friction – hp-version for mortar, penalty and Nitsche's methods*. PhD dissertation, Universität Hannover, 2006.

[9] A. Chernov and E. P. Stephan. Adaptive BEM for contact problems with friction. In *IUTAM Symposium on Computational Methods in Contact Mechanics*, volume 3 of *IUTAM Bookser.*, pages 113–122. Springer, Dordrecht, 2007.

[10] P. G. Ciarlet. *Mathematical elasticity. Vol. II*, volume 27 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1997. Theory of plates.

[11] F. H. Clarke. *Optimization and nonsmooth analysis*, volume 5 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990.

[12] G. Duvaut and J.-L. Lions. *Inequalities in mechanics and physics*. Springer-Verlag, Berlin, 1976. Translated from the French by C. W. John, Grundlehren der Mathematischen Wissenschaften, 219.

[13] S. Erichsen and S. Sauter. Efficient automatic quadrature in 3-d Galerkin BEM. *Comput. Methods Appl. Mech. Engrg.*, 157(3-4):215–224, 1998. Seventh Conference on Numerical Methods and Computational Mechanics in Science and Engineering (NMCM 96) (Miskolc).

[14] G. Fichera. Il teorema del massimo modulo per l'equazione dell'elastostatica tridimensionale. *Arch. Rational Mech. Anal.*, 7:373–387, 1961.

[15] A. F. Filippov. *Differential equations with discontinuous righthand sides*, volume 18 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1988. Translated from the Russian.

[16] R. Glowinski, J.-L. Lions, and R. Trémolières. *Numerical analysis of variational inequalities*, volume 8 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1981. Translated from the French.

[17] C. Großmann and H.-G. Roos. *Numerical treatment of partial differential equations*. Universitext. Springer, Berlin, 2007. Translated and revised from the 3rd (2005) German edition by Martin Stynes.

[18] C. Hager and B. I. Wohlmuth. Nonlinear complementarity functions for plasticity problems with frictional contact. *Comput. Methods Appl. Mech. Engrg.*, 198(41-44):3411–3427, 2009.

[19] J. Haslinger, M. Miettinen, and P. D. Panagiotopoulos. *Finite Element Method for Hemivariational Inequalities*. Nonconvex Optimization and its Applications. Kluwer Academic Publishers, Dordrecht, 1999.

[20] M. Hintermüller, V. A. Kovtunenko, and K. Kunisch. Obstacle problems with cohesion: A hemi-variational inequality approach and its efficient numerical solution. Technical Report 2010-002, SpezialForschungsBereich F 32, Karl-Franzens Universität Graz, 2010.

[21] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. I*, volume 305 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993. Fundamentals.

[22] G. C. Hsiao and W. L. Wendland. *Boundary integral equations*, volume 164 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin, 2008.

[23] N. Kikuchi and J. T. Oden. *Contact problems in elasticity: a study of variational inequalities and finite element methods*, volume 8 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1988.

[24] K. C. Kiwiel. *Methods of descent for nondifferentiable optimization*, volume 1133 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1985.

[25] M. Kleiber. *Handbook of computational solid mechanics: survey and comparison of contemporary methods*. Springer-Verlag, Berlin, 1998.

[26] K. Kunisch and G. Stadler. Generalized Newton methods for the 2D-Signorini contact problem with friction in function space. *M2AN Math. Model. Numer. Anal.*, 39(4):827–854, 2005.

[27] L. D. Landau and E. M. Lifschitz. *Lehrbuch der theoretischen Physik. Band VII*. Akademie-Verlag, Berlin, sixth edition, 1989. Elastizitätstheorie.

[28] L. Lukšan and J. Vlček. Algorithm 811: NDA: algorithms for nondifferentiable optimization. *ACM Transactions on Mathematical Software*, 27(2):193–213, June 2001.

[29] L. Lukšan and J. Vlček. A bundle-newton method for nonsmooth unconstrained minimization. *Math. Progr.*, 83:373–391, 1998.

[30] M. Maischak. *hp-Methoden für Randintegralgleichungen bei* 3*D-Problemen, Theorie und Implementierung*. PhD dissertation, Universität Hannover, 1995.

[31] M. Maischak and E. P. Stephan. Adaptive *hp*-versions of BEM for Signorini problems. *Appl. Numer. Math.*, 54(3-4):425–449, 2005.

[32] Z. Naniewicz and P. D. Panagiotopoulos. *Mathematical theory of hemivariational inequalities and applications*, volume 188 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker Inc., New York, 1995.

[33] J. Nečas and I. Hlaváček. *Mathematical theory of elastic and elasto-plastic bodies: an introduction*, volume 3 of *Studies in Applied Mechanics*. Elsevier Scientific Publishing Co., Amsterdam, 1980.

[34] C. Niculescu and L.-E. Persson. *Convex Functions and Their Applications*. CMS Books in Mathematics. Canadian Mathematical Society, Halifax, 2006.

[35] L. Q. Qi and J. Sun. A nonsmooth version of Newton's method. *Math. Programming*, 58(3, Ser. A):353–367, 1993.

[36] A. Quarteroni and A. Valli. *Domain decomposition methods for partial differential equations*. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 1999. Oxford Science Publications.

[37] J. Rauch. Discontinuous semilinear differential equations and multiple valued maps. *Proc. Amer. Math. Soc.*, 64(2):277–282, 1977.

[38] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[39] S. Sauter and C. Schwab. *Randelementmethoden*. Vieweg + Teubner, 2004.

[40] C. Schwab. Variable order composite quadrature of singular and nearly singular integrals. *Computing*, 53(2):173–194, 1994.

[41] H. Si. Constrained Delaunay tetrahedral mesh generation and refinement. *Finite Elem. Anal. Des.*, 46(1-2):33–46, 2010.

[42] I. S. Sokolnikoff. *Mathematical theory of elasticity*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1956. 2d ed.

[43] J. Stoer. *Einführung in die numerische Mathematik. I*, volume 105 of *Heidelberger Taschenbücher [Heidelberg Paperbacks]*. Springer-Verlag, Berlin, third edition, 1979. Based on the lectures of F. L. Bauer.

[44] A. H. Stroud. *Approximate calculation of multiple integrals*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1971. Prentice-Hall Series in Automatic Computation.

[45] tetgen. A quality tetrahedral mesh generator. http://tetgen.berlios.de.

[46] R. Verfürth. *A Review of* a posteriori *Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley Teubner, 1996.

[47] L. T. Wheeler. Maximum principles in classical elasticity. In *Mathematical problems in elasticity*, volume 38 of *Ser. Adv. Math. Appl. Sci.*, pages 157–185. World Sci. Publ., River Edge, NJ, 1996.

[48] X. Q. Yang. Generalized second-order characterizations of convex functions. *J. Optim. Theory Appl.*, 82(1):173–180, 1994.

[49] X. Q. Yang and V. Jeyakumar. Generalized second-order directional derivatives and optimization with $C^{1,1}$ functions. *Optimization*, 26(3-4):165–185, 1992.

Curriculum Vitae

17. 12. 1981	born in Halle (Westf.)
06. 2001	Abitur at CJD-Gymnasium, Versmold
09. 2001 – 06. 2002	Community service, Diakoniestation Borgholzhausen
10. 2002 – 09. 2005	Studies at Universität Hannover *Mathematik, Studienrichtung Rechnergestützte Wissenschaften*
04. 2005	Vordiplom
09. 2005 – 09. 2006	Studies at Brunel University, Uxbridge, UK *Computational Mathematics with Modelling*
09. 2006	Master of Science
10. 2006 – 09. 2007	Studies at Universität Hannover *Mathematik, Studienrichtung Rechnergestützte Wissenschaften*
since 10. 2007	PhD student in the workgroup *Numerical Analysis*, IfAM, Leibniz University Hannover
10. 2007 – 09. 2010	scholarship holder, DFG-Graduiertenkolleg 615