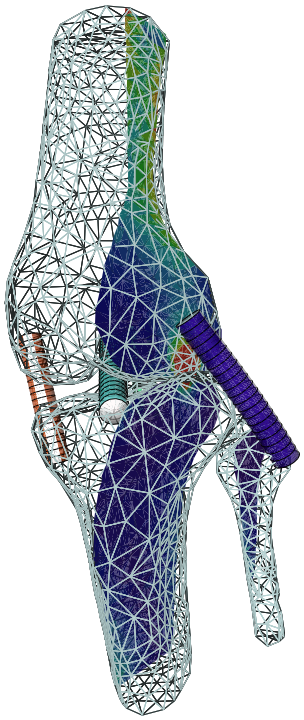
Inaugural-Dissertation
zur Erlangung der Doktorwürde
des Fachbereichs Mathematik und Informatik
der Freien Universität Berlin

Multidimensional Coupling in a Human Knee Model



vorgelegt von
Oliver Sander
am 3. 6. 2008

Gutachter:
Prof. Dr. Ralf Kornhuber (Berlin)
Prof. Dr. Peter Bastian (Stuttgart)

Datum der Disputation: 25. 9. 2008

Selbstständigkeitserklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbstständig angefertigt habe. Ich versichere, dass ich ausschließlich die angegebenen Quellen und Hilfen in Anspruch genommen habe.

Oliver Sander

Berlin, 3.6.2008

Contents

List of Symbols

Miscellaneous Symbols

$:$	Matrix inner product	27
$\langle \cdot, \cdot \rangle$	Scalar product in $\mathbb{R}^d$	31, 63, 127
$\lvert \cdot \rvert$	Euclidean norm in $\mathbb{R}^d$	31, 127
$\lvert \cdot \rvert$	Area of a surface	11, 70, 92
$\lVert \cdot \rVert_{\infty, T\mathrm{SE}(3)^n}$	Infinity norm on the tangent bundle of $\mathrm{SE}(3)^n$	83
$[\cdot]_\Phi$	Relative displacement with respect to the contact mapping Φ	31
$\times$	Vector product in $\mathbb{R}^3$	92
$\times$	Product of spaces	64
$\circ$	Map composition	31, 63
$'$ (*prime*)	Derivation with respect to the rod arclength s	68, 127
$\cdot^+$ (*superscript +*)	Moore–Penrose pseudo inverse	130
$\oplus$	Direct sum of vector spaces	53, 64
$\hat{\cdot}$ (*hat accent*)	Hat map $\mathbb{R}^3 \to \mathbb{A}^3$	65, 127
∇	Gradient	79
$\boldsymbol{\nabla}$	Deformation gradient	24
∇^2	Hessian	79
∇_w, ∇_v	Gradient with respect to w, v	81

Greek Letters

Γ	Interface boundary for a multidimensional coupling	91
$\Gamma_{i,C}$	Contact boundary of domain i	30
Γ_i	Contact boundary of domain i (in Sec. 3.5)	40
Γ_I	Interface boundary for a homogeneous coupling	88
$\Gamma_{i,D}$, $\Gamma_{i,N}$	Dirichlet/Neumann boundary of domain i	30
γ	Ellipticity constant	28
γ	Engineering shear strain	11
$\boldsymbol{\varepsilon}$	Linear strain tensor	25
η_1, η_2	Parameters controlling the trust-region radius	78
$\{\theta\}$	Dual mortar basis function	32
θ	Dirichlet–Neumann damping parameter	96
$\boldsymbol{\nu}$	Normal vector	23
$\hat{\boldsymbol{\nu}}$	Linearly interpolated normal vector	28
$\boldsymbol{\nu}_0, \boldsymbol{\nu}_l$	Rod normals at $s = 0$ and $s = l$, respectively	91
ν	Poisson ratio	12, 27, 70
Ξ	Piecewise linear homeomophism between two triangulated surfaces	41
π	Boundary parametrization function	46
ρ	Trust-region radius	77

$\boldsymbol{\sigma}$	Second Piola-Kirchhoff stress tensor	25
$\hat{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\sigma}}$	Constitutive relations	26
$\boldsymbol{\sigma_\nu}, \boldsymbol{\sigma}_T$	Normal and tangential components of the normal stress, respectively	31
Υ	Map from a total force/moment pair to a corresponding Neumann value field	100
Φ	Contact mapping	31
$\boldsymbol{\varphi}$	Configuration of a nonlinear elastic body	23
φ	Rod configuration	67, 91
$\hat{\varphi}$	Stress-free rod configuration	69
$\{\psi\}$, $\{\boldsymbol{\psi}\}$	Scalar/vector-valued nodal basis	29
$\{\tilde{\psi}\}$	Mortar-transformed basis function	37
Ω	Domain	23

Roman Letters

A_i	Shear and tension stiffnesses of a linear diagonal rod material	70
$\mathcal{A}$	Rod cross-section	67
$\mathbb{A}^n$	Space of all antisymmetric $n \times n$ matrices	65
Av	Interface averaging operator	93
$a(\cdot,\cdot)$	Bilinear form of linear elasticity	27
$a(\cdot,\cdot)$	Form of nonlinear elasticity	88
$\mathbf{B}_k$	Infinitesimal rotation about the director $\mathbf{d}_k$	68
$C_\mathbf{n}$, $C_\mathbf{m}$	Sets of vector-valued functions on Γ with prescribed force/moment	97
$\mathbf{C}$	Hooke tensor	27
$\mathscr{C}$	Generic coarse grid correction operator	36
D	Matrix coupling nonmortar displacements to Lagrange multipliers	35
D	Derivative of a mapping between two manifolds	64
$\mathscr{DN}$	Dirichlet–Neumann operator	101
$\mathscr{DN}_\theta$	Damped Dirichlet–Neumann operator	101
DtN	Dirichlet-to-Neumann operator of the rod problem	102
d	Differential	79
d^2	Second-order differential	79
$\tilde{d}^P$	Projected coarse-grid correction	39
$\mathrm{dist}(\cdot,\cdot)$	Geodesic distance	64
$\mathbf{d}_i$	Director vectors	67
$\mathbf{E}$	Green–St. Venant strain tensor	24, 87
E	Young's modulus (elastic modulus)	12, 27, 70
$\mathcal{E}$	Pointwise interpolation operator	72
$\mathbf{e}_i$	Canonical basis vectors of $\mathbb{R}^d$	24
$\mathbf{f}$	Body force density	25, 69
$\mathcal{F}$	Averaged interface deformation	92
$\mathcal{G}$	Graph	40
G	Grid	28, 71
$\mathscr{GS}$	Projected Gauß–Seidel operator	36
g	Initial gap function	31
$g(\cdot,\cdot)$	Riemannian metric	63

g	Weak initial gap function	34
$\mathbb{H}$	Quaternion algebra	66
$\mathbb{H}_{\lvert 1\rvert}$	Unit quaternion algebra	66
H^1, $\mathbf{H}^1$	Scalar/vector-valued first order Sobolev space	27
H^1_0, $\mathbf{H}^1_0$	Scalar/vector-valued first order Sobolev space with zero trace on Γ_D	27
h	Grid size	28, 71
Id	Identity	25
$\mathrm{inj}(p)$, $\mathrm{inj}(M)$	Injectivity radius at $p \in M$, injectivity radius of M	73
J	Energy functional of linear elasticity	27
J_1, J_2	Second moments of area of a rod cross-section	70
J_3	Polar moment of inertia of a rod cross-section	70
$\hat{J}_{x_\nu}$	Lifted energy functional on $T_{x_\nu}M$	79
j	Rod energy functional	70
$\hat{j}_\nu$	Lifted rod energy functional on $T_{(r_\nu,q_\nu)}\mathrm{SE}(3)^n$	81
K_i	Bending and torsion stiffnesses of a linear diagonal rod material	70
$\mathcal{K}$	Strong admissible set	31
$\mathcal{K}^w$	Weak admissible set	32
$\mathcal{K}_h$	Mortar-discretized admissible set	33
$\mathcal{K}_{\mathrm{alg}}$	Algebraic admissible set	35
K_k^{tr}	Trust-region at the k-th iteration	77
$\widetilde{\mathcal{K}}_{\mathcal{Q}}$	Set of admissible corrections in $\mathbf{V}_h^2$	51
$\widetilde{\mathcal{K}}_{\mathrm{alg}}^{\nu+\frac{1}{2}}$	Defect obstacle after the ν-th presmoothing step	38
L^2, $\mathbf{L}^2$	Space of scalar/vector-valued square-integrable functions	27
l	Rest length of a rod	67
$\mathbf{l}$	External volume moments	69
$l(\cdot)$	Linear form of external forces	27
M	Matrix coupling mortar displacements to Lagrange multipliers	35
M^+	Cone of positive Lagrange multipliers	32
M_h^+	Discretized cone of positive Lagrange multipliers	33
$\mathbf{m}, \mathsf{m}$	Resultant moment across a rod cross-section in canonical (local) coordinates	68
$\mathbf{n}, \mathsf{n}$	Resultant force across a rod cross-section in canonical (local) coordinates	68
n	Number of non-Dirichlet grid vertices	29, 71
O	Rotation matrix from canonical to normal/tangent coordinates	36
O_Γ	Average orientation of a boundary patch	93
q	Geodesic interpolation	75
$q\mathbb{A}^3$	Space of all matrices m of the form $m = qA$, where $q \in \mathrm{SO}(3)$ is fixed and $A \in \mathbb{A}^3$ is antisymmetric	65
$\mathcal{R}$	Continuous extension from $\mathbf{H}^{1/2}(\Gamma)$ to $\mathbf{H}^1(\Omega)$	89
R	Retraction mapping	78
$\mathbb{R}^{n\times n}_+$	Space of all real $n \times n$ matrices with positive determinant	24
$\mathbb{S}^d$	Space of symmetric second-order tensors in $\mathbb{R}^d$	25
S^{yz}, S^{xz}	Reflections about the yz and xz planes, respectively	102
$\mathrm{SE}(3)$	Group of rigid-body motions in $\mathbb{R}^3$ (special Euclidean group)	67

$\widetilde{\mathrm{SE}}(3)$	Special Euclidean group restricted to parallel translations in z-direction	102
$\mathrm{SO}(d)$	Group of proper rotations in a $\mathbb{R}^d$ (special orthogonal group)	65
$\mathfrak{so}(d)$	Lie algebra of $\mathrm{SO}(d)$	65
S^d	Unit sphere in $\mathbb{R}^{d+1}$	28, 63
T	Truncation matrix	38
T	Constancy functional	99
$\mathbf{T}^{\varphi}$	Stress tensor in the deformed configuration (Cauchy stress)	25
$\mathbf{t}$	Surface force density	25, 88
$\mathbf{u}$	Displacement vector field	24
$\tilde{u}^{\nu+\frac{1}{2}}$	Iterate after nonlinear smoothing with respect to $\{\tilde{\psi}\}$	37
$\mathbf{u}, \mathsf{u}$	Rod bending and torsional strain in canonical (local) coordinates	68
$\hat{\mathsf{u}}$	Rod bending and torsional strain in local coordinates in the stress free configuration	69
$\mathcal{V}, \mathcal{V}_i$	Vertices of a grid (of grid i)	28
$\mathcal{V}_{i,C}$	Grid vertices on the contact boundary i	32
$\mathcal{V}_T$	Linear truncated subspace	38
$V_h, \mathbf{V}_h$	Scalar/vector-valued first-order finite element space	28
$V_{h,0}, \mathbf{V}_{h,0}$	Scalar/vector-valued first-order finite element space with zero trace on Γ_D	28
$V^2_{h,0}, \mathbf{V}^2_{h,0}$	Scalar/vector-valued second-order finite element space with zero trace on Γ_D	50
V_h^M	Geodesic finite element space with image on M	72
$V_{h,D}^M$	Geodesic finite element space with image on M and Dirichlet boundary conditions	76
$\mathbf{v}, \mathsf{v}$	Rod extension and shear strain in canonical (local) coordinates	68
$\hat{\mathsf{v}}$	Rod extension and shear strain in local coordinates in the stress free configuration	69
W	Hyperelastic energy density	26, 69
W	Trace space on the contact boundary	32
W^+	Space of positive trace functions	32

Acronyms

ACL	Anterior cruciate ligament	7, 119
LCL	Lateral collateral ligament	7, 119
MCL	Medial collateral ligament	7, 119
MMG	Monotone Multigrid	35, 56, 83
PCL	Posterior cruciate ligament	7, 119
QLV	Quasilinear viscoelasticity	16
TNNMG	Truncated Nonsmooth Newton Multigrid	35

1 Introduction

Today's surgery performs a wide range of interventions which alter the mechanical behavior of the musculoskeletal system. Examples are total hip or knee replacements and osteotomies (cutting a bone and resetting it at an angle to remove a limb deformation). These methods can be improved considerably if detailed, patient-specific knowledge about the joint stresses is available. Total joint replacements still tend to loosen after about 15 years, and this seems to be due to the bone remodelling around the implant in response to the altered mechanical stimulus. A better understanding of the mechanics of joints could help to improve prosthetics design. Osteotomies frequently show unfavorable long-term side effects. The changed loading in the joints can lead to overly increased cartilage wear which may result in premature arthrosis. Here, a detailed knowledge about the stresses can help to develop improved surgery procedures.

The healing of bone fractures appears to be influenced by the mechanical loading conditions. Knowledge of the mechanics of the musculoskeletal system can help to improve therapies for accelerated tissue regeneration and design better fixation devices. Modern competitive sport puts an ever-increasing demand on the athlete as a mechanical system, and high performance combined with a low risk of injury can only be achieved if athlete-specific biomechanics are taken into account.

Biomechanists have used a wide range of methods to gain insight into the stresses in bones and ligaments, some analytical and some experimental. Structural analysis of beams, for example, works fairly well on long bones of reasonably uniform cross-section. However, whole bones have complex shapes at their ends and are acted on by complex loads. Hence beam analysis breaks down near the bone ends [49]. Analytical methods are almost useless in complex structures such as vertebrae or the skull [29]. The presence of material inhomogeneities, in particular cancellous bone, is another complicating factor.

One experimental method to obtain strains and stresses of bones and ligaments is to attach a number of strain gauges to a bone or a model of it [49]. This allows to measure strains directly and to infer stresses from assumed material parameters. On the other hand it is very time-consuming and restricted to surface strains. Photoelasticity gives good qualitative insights for planar models. However, it is not straightforward to extend this to a three-dimensional analysis. Also, material inhomogeneities cannot properly be taken into account.

These experimental methods measure stresses in models of bone or dead tissue specimen. An important step forward was the work of Bergmann et al. [13]. They constructed hip, knee, and other implants containing a pressure sensor. It was thus possible to measure dynamic joint stresses in vivo during a wide range of everyday activities. The downside of this method is that only accumulated stress values for the entire joint are obtained. No information about the spatial distribution of the stresses within the joint

becomes available.

Numerical simulation can provide new insights into this difficult problem. In fact, it is establishing itself as a third fundamental way of gaining knowledge besides theory and experiments. The term 'in silico experiment' is becoming popular as a synonym for computer simulations [2].

In engineering, the finite element method has proven extremely useful for the simulation of mechanical structures of all kinds. It can, in principle, provide stress distributions of arbitrarily high resolution for all kinds of loading situations. Also, it allows the easy study of the dependence of the system on parameters, e.g., material anisotropy [54]. The catch is that the cost in computer power may be so high as to effectively prohibit many simulations even on the most powerful hardware available. This cost is determined by the size of the problem, the equations used to model the system, and the algorithms used to solve these equations. More challenging equations can be used to gain precision if fast algorithms are available for their solution. This make the design of efficient algorithms a key point.

Besides being efficient these algorithms have to be reliable. While a numerical analyst may be fully aware of the shortcomings of a given algorithm this cannot be expected from end users which are specialists in other fields. A "great tendency to take the results [of commercial finite element codes] as they come, without independent checks" (Currey [29], p. 243) can still be observed. This problem is particularly relevant in biomechanics as it is very difficult to validate simulation results experimentally. While software users should constantly be reminded to keep a certain scepticism with regard to the output of the programs, the makers of such programs should strive to design their algorithms to be as reliable as possible.

Simulating the mechanical behavior of a human knee using finite elements is an extremely challenging task. Some of the difficulties are the coexistence of different materials, the large strains undergone by everything but the bones, the contact between the various objects, the multiscale nature of the materials and their resulting inhomogeneous and anisotropic material properties. Any resulting discrete system must be nonlinear, nondifferentiable, and very large. From the more practical side, at the current state of technology it is very difficult to get reliable material parameters (see Sec. 2.5), and to validate simulation results numerically.

To cope with such difficulties researchers have proposed a wide scale of simplified models. These range from simple linked chains of rigid bodies to complex finite element models. Each is suited to a particular purpose. This thesis presents a new model for the human knee which aims at the top end of this scale. The model incorporates femur, tibia, and fibula, and the four major ligaments. The bones are modeled using three-dimensional linear elasticity on patient-specific simplicial grids. Therefore, the model can provide detailed information about the stress distributions in the bones. The contact between individual bones is taken into account. For the ligaments Cosserat rods are used. These are a standard tool in structural engineering used to model long slender objects undergoing large deformations. While they cannot provide spatially resolved stresses within cross-sections they do not cause meshing problems the way three-dimensional finite strain models do. In this thesis we assume the rod cross-sections to be circular

and of constant radius. However the model also allows for cross-sections which are, for example, ellipsoidal, and which may vary with arc length.

The model is by no means complete and has to be considered as a basis for further development. Articular cartilage, the patella bone, and the menisci are all crucial for reliable stress predictions. Also ligaments wrap around bone in some places and therefore the inclusion of bone–ligament contact is important. Some results concerning dynamic problems have already been obtained [58]. All this is subject of future work. Nevertheless the model as it is today can already provide useful information.

Mathematically, the knee model leads to a heterogeneous nonsmooth domain decomposition problem. The contact problem between the bones is described by a linear equation with convex constraints, while the Cosserat rod problem is formulated as a minimization problem on a nonlinear Riemannian manifold. For both problems new solvers are proposed which are both efficient and reliable. Coupling the two subproblems in a single model also breaks new ground. We propose transmission conditions for the coupling of a three-dimensional linear elastic body and a one-dimensional nonlinear elastic Cosserat rod in the continuous and the discrete setting. We also describe a nonoverlapping Schwarz method which solves this coupled problem.

We present two sets of results for biomechanical problems. In Chapter 3 we treat the pure contact problem between femur and tibia using the Visible Human data set [3]. In Chapter 6 we give results for the complete knee model additionally incorporating the proximal fibula and four major ligaments. These simulation runs are purely prototypical. Their only purpose is to demonstrate the performance of the solution algorithms. The stress distributions computed do not have quantitative relevance.

Cornerstone of this thesis is the efficient and robust implementation of all algorithms that are presented. These implementations are based on the DUNE library, which has been developed over the last years in a joint effort by groups in Stuttgart, Freiburg, and, as part of this thesis, in Berlin. It is a new framework for high-performance numerical computations, and is based on three design ideas.

- **Abstraction:** Since a single grid manager can never satisfy the needs of all application writers, DUNE defines an abstract grid interface and provides an extendible set of implementations for this interface. Hence, applications can be made to run with different grid managers and switching from one grid implementation to another only requires minimal changes.

- **Efficiency:** All interface code is written in C++ using methods from generic programming. As a result, compilers are able to optimize away most interface code and DUNE applications run with an efficiency comparable to applications which are directly attached to a single grid implementation.

- **Code reuse:** DUNE supports the reuse of existing grid managers as legacy code hidden behind the abstract grid interface. For example, the grid managers of the finite element codes UG [10] and ALBERTA [81] are available within DUNE as `UGGrid` and `AlbertaGrid`, respectively.

Detailed information about DUNE can be found in [11, 12], and on the project homepage [1].

This thesis consists of five main chapters, which we will now summarize briefly. We begin with a short introduction to the mechanics of the human knee joint in Chapter 2. There we concentrate on the physiology and material behavior of bones and ligaments, motivating our modeling decisions.

The next chapter deals with multi-body contact problems in linear elasticity. The aim is the efficient and robust solution of large problems on domains with curvilinear boundaries. The first three sections are devoted to the formal statement of the continuous and discrete problems. We then introduce the Truncated Nonsmooth Newton Multigrid algorithm (TNNMG) as a globally convergent, efficient, and reliable method for such problems. Next we describe how the contact mapping can be implemented efficiently. We also briefly present a way to automatically construct parametrized boundaries (previously published in [61, 79]), and a hierarchical error estimation schemes which proves to be very effective. Numerical results are given which show that the TNNMG solves contact problems on curvilinear domains with the same convergence rate as linear multigrid methods on corresponding linear problems.

In Chapter 4 we present Cosserat rods as our model for ligaments. Being one-dimensional, Cosserat rods avoid problems with the meshing of long slender structures and with grid quality for large deformation elasticity problems. The configuration space of Cosserat rods has a nonlinear Lie group structure. We introduce geodesic finite element spaces as a systematic way to discretize function spaces which map onto nonlinear Riemannian manifolds. Standard optimization algorithms cannot be used to find minimizers of the hyperelastic energy functional due to this nonlinearity. Using ideas from [83] and [4, 5] we present a Riemannian trust-region solver for nonlinear Cosserat rod problems. The solver converges globally, and locally superlinear. Hence it is both reliable and efficient. Using the ∞-norm for the trust-region allows to use a monotone multigrid method as the inner solver. We close the chapter with a numerical example.

Chapter 5 covers the coupling of the bone and ligament models. The two main difficulties are the difference in dimension between the two models and the fact that the rod model is nonlinear and its solutions are generally not unique. Based on the transmission conditions for the coupling of two three-dimensional objects (which are derived rigorously), we present coupling conditions for the 1d-3d problem. A Dirichlet–Neumann scheme is then presented which solves the multidimensional coupled problem. In order to show that the coupling conditions are in fact reasonable we give an existence result for solutions to the coupled problem in the special case of problems showing certain symmetries. Uniqueness of solutions cannot be expected as the Cosserat rod problem by itself does not have unique solutions. In the last section we present a few numerical results for the Dirichlet–Neumann solver on test problems.

The last chapter then combines all previous results. Simulation results for a knee model are presented. The model includes the distal femur and the proximal tibia and fibula; as well as the anterior and posterior cruciate ligaments, and the medial and lateral collateral ligaments. A special section is dedicated to the DUNE libraries, which have been codeveloped as part of this thesis and which allow an unseen degree of flexibility

in the choice of grid managers.

An appendix contains the somewhat technical derivation of the gradients of the rod strains on tangent spaces of the rod configuration space. This is used in Chap. 4.

Many people have helped me during the preparation of this thesis. I would like to thank Prof. Ralf Kornhuber for his constant support and valuable advice; Prof. Dr. Dr. h.c. Peter Deuflhard for his continued interest in my work; Prof. Rolf Krause for getting me started with contact problems; the DUNE team, in particular Christian Engwer and Prof. Peter Bastian for bringing the right tools just when I needed them; Dr. Stefan Lang for his help with UG; Prof. John Maddocks for introducing me to Cosserat rods; Dr. Markus Heller and Dr. Bill Taylor for their input on biomechanics; Prof. Klaus Ecker for his help on differential geometry; Dr. Heiko Berninger, Carsten Gräser, and Dr. Carsten Hartmann for proof-reading and many fruitful discussions; last but not least, Christian Salzmann for his computer help.

2 Biomechanics of the Human Knee

This chapter gives a brief overview over the biomechanics of the human knee, particularly concentrating on the bones and ligaments. We also discuss the attachments of ligaments to bone and comment on the experimental difficulties encountered when trying to obtain reliable material parameters. Most of the information on bones in this chapter has been taken from the book by Currey [29], while the information on ligaments is mainly based on the review article by Weiss and Gardiner [93].

2.1 Structure and Function of the Human Knee Joint

The human knee is the joint of the lower extremity which connects the femur and the tibia. It involves four different bones: femur, tibia, fibula, and patella (kneecap). Strictly speaking, the human knee consists of two joints. The first is the femoro-tibial joint, which links the femur with the tibia. Its purpose is to transfer the body weight from femur to tibia and to provide for the freedom of movement necessary for walking. The second joint is the femoro-patellar joint, which connects the patella and the patellar groove on the front of the femur. The femoro-tibial joint is a so-called synovial joint, which means that it is contained in a synovial membrane, or joint capsule, and bathed in synovial fluid.

The articulating surfaces of the bones are covered with layers of cartilage. These layers, with a thickness varying locally between 0.5 mm and 3 mm, have a two-fold purpose. They act as lubrication and allow for nearly effortless motion along preferred anatomical directions. Also, the viscoelastic cartilage acts as a damping.

The different bones of the knee are held together by the skeletal ligaments, which are short bands of tough fibrous connective tissue. They guide normal joint motion and provide stability by restricting abnormal joint movement. Assisting in this are the congruent geometry of the articulating surfaces and the musculotendinous forces. Even in everyday activities ligaments can be subjected to very high tensile stresses. When overloaded they can disrupt completely, which is a common sports injury.

The four major ligaments of the knee are the anterior cruciate ligament (ACL), the posterior cruciate ligament (PCL), the lateral collateral ligament (LCL), and the medial collateral ligament (MCL) (see Fig. 2.1). Cutting experiments using cadaveric knees have shown that each ligament has specific primary and secondary functions. The ACL, for example, acts as a primary restraint against anterior tibial displacements and as a secondary restraint to tibial axial rotation. The PCL provides the primary resistance to posterior tibial displacements and also acts as a secondary restraint to external tibial rotation. The MCL is the primary structure resisting valgus rotations of the knee and internal tibial displacement. The LCL provides a primary restraint to varus rotation

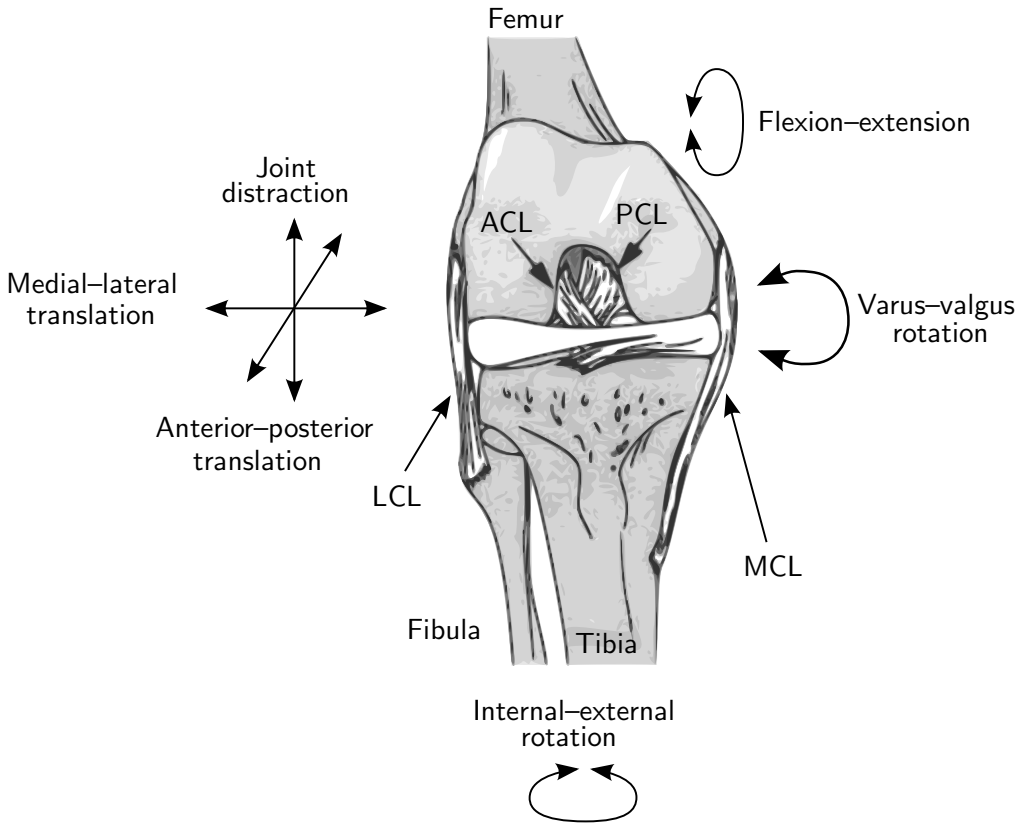


Figure 2.1: Anterior view of the right human knee joint. The femur, tibia, and fibula are shown together with the anterior cruciate ligament (ACL), posterior cruciate ligament (PCL), medial collateral ligament (MCL), and lateral collateral ligament (LCL). The patella is not shown. (Reprinted from Weiss and Gardiner [93].)

and external tibial rotation, and a secondary restraint to anterior and posterior tibial displacement [93]. Only knee flexions and extensions are unrestricted within a certain range.

Further joint lubrication and damping is provided by the two menisci. The menisci have a semilunar shape and are situated between the femoral condyles and the tibial plateau of the knee. They consist of fibrocartilage which contains thick layers of large collagen fibres. These give the menisci a relatively rough and fibrous appearance. The mechanical behavior is similar to articular cartilage, but with more pronounced anisotropies [70].

Various muscles attach to the bones in close proximity to the knee joint. Muscles taper off into tendons which then insert into the bone. Tendons differ from ligaments in their function, but have a very similar structure and mechanical behavior. Contractions of the muscles provide the forces to move the knee. In static loading situations such as an upright standing position they provide stability by active stabilization.

2.2 Bones

Bone is a multi-scale material and its composition and properties have to be described at several length scales. At the lowest level, bone can be considered as a composite material consisting of a fibrous protein, *collagen*, stiffened by an extremely dense filling of calcium phosphate crystals. There are other constituents, notably water, some proteins and polysaccharides, and, in many types of bone, living cells and blood vessels. The amount of water present in the bone is an important determinant of its mechanical behavior.

Collagen is a structural protein found in probably all metazoan animal phyla. It makes up more than half the protein in the human body. The collagen in bone is called *type 1* collagen. The protein molecule *tropocollagen* aggregates to form microfibrils, which become stabilized by intermolecular cross-links. Microfibrils in turn aggregate to form *fibrils*.

Impregnating and surrounding the collagen is the bone mineral, which is a variety of calcium phosphate. The precise nature of both the chemistry and the morphology of it is still a matter of some dispute. The reason is that mineral in bone comes in very small crystals with a very high surface-area-to-volume ratio. The size of the crystals is such that in one dimension it is only about 10 atomic layers thick [66]. This makes it reactive, and so most preparative techniques used for investigating it may cause alterations from the living state. Some of the bone mineral is the version of calcium phosphate called *hydroxyapatite* whose unit cell contains $Ca_{10}(PO_4)_6(OH)_2$. Reports of the visualization of the crystals directly support the view that the crystals are platelet-shaped [29].

Bone is permeated by various kinds of specialized cells. *Osteoblasts* are responsible for the formation of bone. They initially lay down the collagenous matrix, in which mineral is later deposited. Conversely, *osteoclasts* are bone-destroying cells. *Osteocytes* are living cells in the body of the bone. They derive from osteoblasts and their density varies from about 90,000 mm^{-3} (rats) to about 30,000 mm^{-3} (cows). They are trapped in the hard bone tissue and connect with neighboring osteocytes by means of processes

that are housed in little channels called *canaliculi*, of about 0.2–0.3 μm diameter [27].

Above the level of the collagen fibril and its associated mineral, mammalian bone exists in two usually fairly distinct forms: woven bone and lamellar bone.

Woven bone is usually laid down very quickly, more than 4 μm a day, most characteristically in the fetus and in the callus that is produced during fracture repair. The collagen in woven bone is variable, the fibrils being between 0.1–3 μm in diameter and oriented almost randomly, so that it is difficult to make out any preferred direction over distances greater than about a millimeter [20].

Lamellar bone is more precisely arranged, and is laid down much more slowly than woven bone (less than 1 μm a day [20]). The collagen fibrils and their associated mineral are arranged in sheets (lamellae), which often appear to alternate in thickness. The final degree of mineralization is less than that of woven bone.

On the length scale of millimeters there are four main types of bone that can be found in mammals. Two of them are woven and lamellar bone which can extend uniformly for several millimeters in all directions. But lamellar bone also exists in the quite separate form of *Haversian systems*. These form when many osteoclasts move forward in a concerted attack on the bone tissue. This forms a *cutting cone* of about 200 μm in diameter. As the cutting cone advances it leaves a cylindrical cavity behind. Almost as soon as the cavity forms, it begins to fill in. The walls of the cavity are made smooth, and bone is deposited on the internal surface in concentric lamellae. In humans, the whole process takes about two to four month. The Haversian system is the classic result of the process of remodeling, where the bone material is constantly renewed. The fourth type, *fibrolamellar bone*, tries to combine the fast deposition speed of woven bone with the stability of lamellar bone. It is found in sites where bones have to grow in diameter rather quickly. Essentially, an insubstantial scaffolding of woven bone or parallel-fibered bone is laid down quickly to be filled in more leisure with lamellar bone.

At the highest level of the hierarchy there is the mechanically important distinction between compact (cortical) and *cancellous* bone. The difference is visible to the naked eye. The former is solid, with spaces in it only for osteocytes, canaliculi, blood vessels and erosion cavities. In contrast, in cancellous bone there are large open spaces. The simplest kind of cancellous bone consists of cylindrical struts, about 0.1 mm in diameter. Each extends for about 1 mm before making a connection with one or more other struts, mostly roughly at right angles. In variations, the struts are partially replaced by little plates. Next to the preferred collagen fibril directions in lamellar bone, cancellous bone is another source of anisotropy, the struts frequently oriented mainly along preferred directions. The pores of cancellous bone are filled with *marrow fat*, which, at room temperature, is a viscous fluid. It seems unclear whether the bone marrow has any significant mechanical function. See the book by Currey [29], Sec. 7.9, for a discussion.

Mechanical Behavior of Bone Material

As a multiscale material, the appropriate mathematical descriptions of the mechanical behavior of bone differ on the different length scales. Being mainly interested in modeling entire joints, we restrict this overview to macroscopic specimen. There it is reasonable

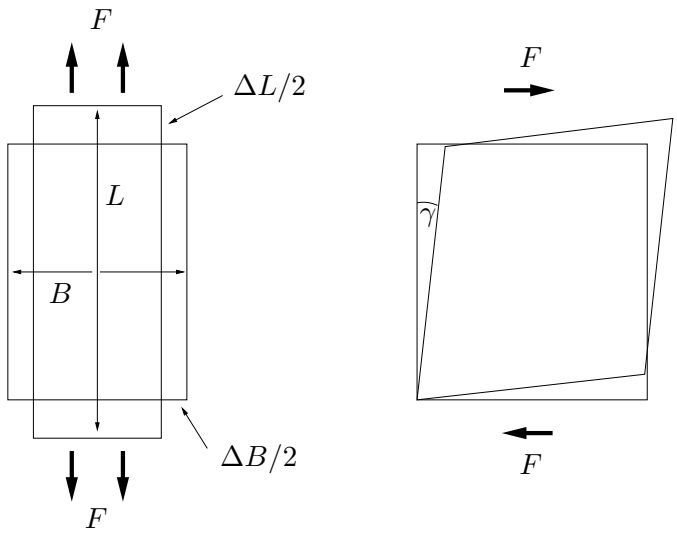


Figure 2.2: Normal strains (left) and shear strains (right).

to model bone as a continuum (of not necessarily homogeneous and isotropic properties) and use the well-established techniques of continuum mechanics for its description.

In the engineering and biomechanics literature, introductions to the elastic behavior of materials usually start with considering small test objects under loading.[1] Consider a bar of length L and width B acted on by a force F tending to stretch it. The bar will undergo an increase in length ΔL and a decrease in width ΔB (Fig. 2.2). The proportional changes in length, $\Delta L/L$ and width $\Delta B/B$, are called *normal strains*. If the force F is instead acting in parallel to a side of the bar, then the resulting deformation is called a *shear strain*. It is measured as the change in angle γ undergone by two lines originally at right angles, measured in radians. In bone we are usually dealing with small strains, $\Delta L/L$ less than 0.005 and γ less than 0.1. This justifies the choice of the *linear* strain tensor in Sec. 3.1.

Stresses are defined as the intensity of force acting across a plane. Take an area A in some plane in a body which is small enough for the forces acting across it to be essentially uniform. For convenience, choose an orthonormal coordinate system such that the plane is normal to the z-axis. The vector F of forces across the plane can be decomposed into a normal force F_{zz} and two shear forces F_{zx} and F_{zy}. If we scale the forces with the area $|A|$ we obtain the normal stress $\sigma_{zz} = F_{zz}/|A|$ and the two shear stresses $\tau_{zx} = F_{zx}/|A|$ and $\tau_{zy} = F_{zy}/|A|$. It can be shown mathematically that the forces across three nonparallel planes are enough to describe the complete state of stress at a point. In other words, it is always possible to rotate the imaginary test object in such a way that there are no shear stresses, and there remains just one normal stress on each face. These remaining normal stresses are called the *principal stresses*.

The relationship between stresses and strains describes the properties of the material

[1] A more rigorous description of linear elasticity is given in Sec. 3.1.

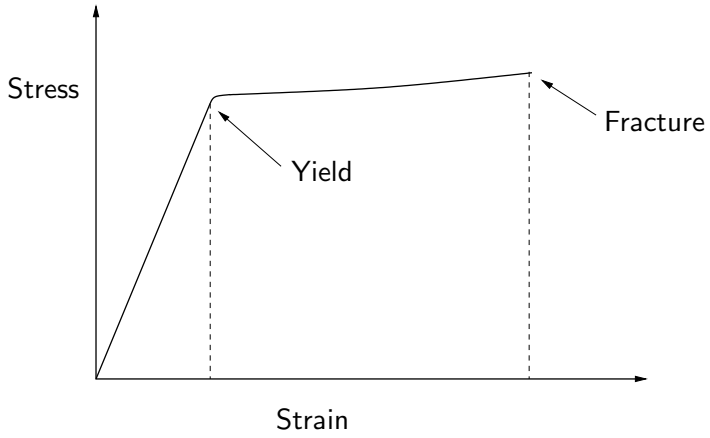


Figure 2.3: A load-deformation curve of a bone specimen loaded in tension.

E_1	E_2	E_3	G_{12}	G_{13}	G_{23}	ν_{12}	ν_{13}	ν_{23}	ν_{21}	ν_{31}	ν_{32}
12.8	15.6	20.1	4.7	5.7	6.7	0.28	0.29	0.26	0.37	0.45	0.34

Table 2.1: Elastic moduli for human Haversian bone (in GPa) measured by ultrasound. The data was determined by Ashman et al. [8] and is presented here as cited by Currey [29]. Cancellous bone material seems to be slightly softer, but exactly how much is unclear, since it is very difficult to perform reliable experiments.

under consideration, and is termed *constitutive law*. For all biomechanical materials the determination of suitable constitutive laws has been subject of much research. Suppose we load a small specimen of compact bone in tension until it breaks. While doing so we monitor the load and the deformation it causes. Scaled with the dimensions of the specimen this yields a stress–strain curve as in Fig. 2.3. Starting from the origin, there is a part where the stress varies linearly with the strain. At the *yield point* the curve flattens considerable, and now increasing strain involves little extra stress. Eventually the specimen breaks. Bones seem to be designed so that the loads placed on them in life usually do not lead beyond the yield point. The slope E of the linear part of the curve is termed *Young's modulus*, and determines the stiffness of the material. It is expressed in Pascal ([Pa] = Newton per square meter) and in bone it has a value of roughly 10–20 GPa. If a similar experiment is performed for a shear deformation γ, the corresponding slope is the *shear modulus* $G = \tau/\gamma$. Another quantity frequently given is the *Poisson ratio* $\nu = E/2G - 1$.

The stress–strain curve in Fig. 2.3 really only gives part of the truth. Bone is an *anisotropic* material, i.e., it behaves differently in different directions. On the length scale of lamellar bone this is due to the bone tissue being arranged in preferred directions, the name-giving lamellae. On the macroscopic scale, direction dependence also enters in form of the trabeculae in cancellous bone. To describe the elastic properties of anisotropic

materials, more parameters than just E and G are necessary. In the most general case, the linear approximation of the stress–strain relationship contains 21 free parameters. Such a number is difficult to handle, and, far more importantly, extremely difficult to measure. If the material under consideration is assumed to contain at least partial symmetries, then the number of parameters can be reduced. Table 2.1 lists some moduli of compact bone that have been determined experimentally by Ashman et al. [8]. Among the material parameters found in the literature there is quite a lot of variation. Currey [29], p. 54, conjectures that a large part of this variation is real and not caused by sloppy experimental work.

The distinction between cortical and cancellous bone is relevant for the modeling of macroscopic bone specimen. The cortical bone, which forms the outer shell of long bones, has fairly homogeneous properties. Cancellous bone, due to its sponge-like structure, shows anisotropic and heterogeneous behavior. Some of these heterogeneities can be extracted from certain gauged CT scans (see, e.g., Yosibash et al. [97]).

Bone is slightly viscoelastic, and its stiffness is to some extent strain-rate dependent. Carter and Caler [23] suggested that Young's modulus is proportional to $(\text{strain rate})^{0.05}$. That would mean that a thousandfold increase in strain rate will result in an apparent increase of Young's modulus of 40%. Strain-rate dependence can therefore be neglected in all but certain impact scenarios.

There have been many attempts at explaining the mechanical behavior of bone tissue on a given length scale by using information from lower scales. This multiscale modeling is a field on its own; an overview can be found in [29, Sec. 3.7]. Multiscale modeling can lead to justifications for well-known ad hoc material models as well as to proposals for new ones. Effective macroscopical material laws can be derived by simulation of the bone microstructure on a small representative volume (homogenization; see, e.g., [47]).

In conclusion, there is quite a number of choices among the family of continuum mechanics material laws. The appropriate choice depends on the problem at hand, and how much effort one is willing to put into the retrieval of material parameters. In this thesis, we will model macroscopic bones using a simple isotropic, homogeneous, linear elastic material law. Disregarding anisotropic and heterogeneous effects in bone is not truly justified when the main interest are the detailed stresses in bone. We choose the simple model in order to simplify the exposition of the contact handling and the coupling of the bones and ligaments. There are no conceptual difficulties in using more involved laws such as the one in [97]. The actual values for Young's modulus E and the Poisson ratio ν were taken from the literature. We used $E = 17$ GPa and $\nu = 0.3$ computed as an average of the values in Table 2.1 throughout.

2.3 Ligaments

Like bone material, ligament tissue is a multi-scale material. It is a composite which consists of a ground substance matrix reinforced by collagen fibres and elastin. The ground substance matrix is composed of proteoglycans, glycolipids, and fibroblasts, and holds large amounts of water. Indeed, about two thirds of the weight of normal ligament

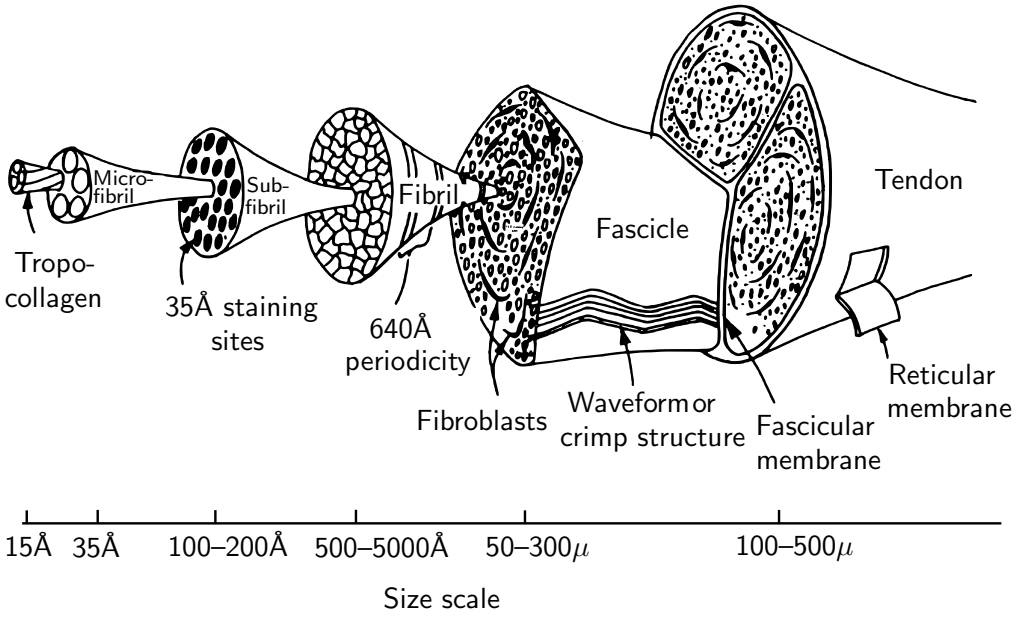


Figure 2.4: The structural hierarchy of ligament and tendon. (Reprinted from Weiss and Gardiner [93], who adapted it from Kastelic et al. [50].)

is made up of water. 70 to 80% of the remaining dry weight is made up by collagen, which is responsible for the great resistance to tensile stresses of ligaments.

Collagen is formed from a structural hierarchy. The exact levels of organization still seem to be subject of some discussion, and there is evidence that they are tissue-specific [93]. Fig. 2.4 shows the hierarchy proposed by Kastelic et al. [50] for the type 1 collagen in rat tail tendon. At the beginning of the process there are certain modifications of linear polypeptide chains, which coil up to form *tropocollagen* molecules. Five such molecules wind up together to yield a collagen *microfibril*. Similarly to the type 1 collagen found in bone (p. 9), these in turn form subfibrils, which then form fibrils (Fig. 2.4). The fibrils are packed together to form fiber bundles. With a diameter in the range of 1 to 12 μm, these can be observed with a light microscope. Under polarized light the unloaded fibres display a clear banding. The collagen has a longitudinal waveform referred to as the *crimp pattern*. This pattern disappears when the ligament is loaded.

The collagen is surrounded by a connective tissue which is known as the ground substance matrix. Proteoglycans, which are the main constituent, aggregate with hyaluronic acid to form hydrophilic molecules. These associate with water to form the gel-like extracellular matrix. This interaction is in part responsible for holding the large amount of water in ligament. Although these hydrophilic molecules constitute less than 1% of the ligament dry weight, their role is very important, as they are partly responsible for holding the collagen together.

The water and its interaction with the ground substance matrix is also responsible for some of the viscoelastic properties of ligament tissue. Movement of water molecules is

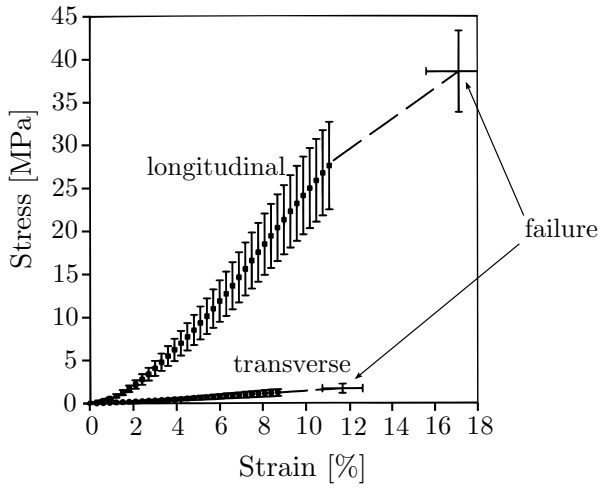


Figure 2.5: Stress–strain curves for human MCL, tested parallel (9 specimen) and transverse (7 specimen) to the collagen fiber direction. Error bars indicate the standard error. (Reprinted from Weiss and Gardiner [93], originally in Quapp and Weiss [76].)

inhibited by charged proteoglycan molecules. Because of the large fraction of water many constitutive equations assume ligaments to be incompressible. Unlike other collagenous soft tissue, such as skin, which retains most of its water even under high pressure, ligament tissue has been shown to lose some water under cyclic loading. However, the exact quantity is unknown. The mechanical importance of this change in hydration, and hence in volume, is an open question.

The third component of ligaments is *elastin*, which makes up for less than 1% of the dry weight. Elastin is an insoluble protein which takes on a complex coiled arrangement when unstressed. This arrangement stretches out when the elastin is stressed. This behavior of elastin accounts for a small part of a ligament's resistance to tension and its elastic properties. Ligaments with a high elastin content are less stiff and undergo larger strains before failure [93].

Constitutive Models for Ligaments

When ligament tissue is loaded under tension, the resulting stress–strain plot shows two distinct regions. While the material behavior is linear beyond a certain strain, for small strains the curve shows a markedly nonlinear, upwarded-curved behavior (Fig. 2.5). This initial nonlinear section of the stress-strain curve is frequently called the *toe region*. A widely accepted theory sees the collagen crimp as the reason for this nonlinear material response. In the unstrained state, the collagen fibres are coiled up and do not contribute noticeably to the macroscopical stiffness. As the ligament is stretched the collagen fibres straighten. At the beginning of the linear section they are all recruited and the linear behavior is effectively the behavior of the collagen fibres [93]. Various researchers have

found values in the range of 330 MPa for this linear elastic modulus [93, Table 1].

This recruiting theory has lead many researchers to propose microstructural models for the observed behavior. Viidik [89] and Frisen et al. [39] represented the elastic response by many individual linearly elastic components. Each of these components represented a collagen fibril of different initial length in its unloaded and crimped form. Fibres were recruited as the ligament was loaded, and at high loads the linear behavior of the fibrils became visible. This simple model was thus able to reproduce the uniaxial response of ligaments.

To properly capture the spatially varying, anisotropic character of ligament tissue, a wide variety of three-dimensional constitutive equations has been proposed. Simple ones describe ligaments as a nonlinear, homogeneous, isotropic material. More advanced models assume transverse isotropy, with the principal material direction following the fiber direction of the tissue. Some try to capture the composite structure by explicitly modeling collagen fibres in addition to the material continuum. The article by Weiss and Gardiner [93] gives a good overview over existing theories and many further references.

The elastic properties of ligaments are relatively insensitive to the strain rate [40, 93]. Viscoelasticity does play a role though in situations with very high strain rates such as impact scenarios. There is also a long-term viscoelastic effect which leads to the ligaments being slightly softer after a number of load cycles.[2] A viscoelastic theory for ligaments has to take both the short-term and the long-term viscoelastic effects into account.

Among the many viscoelastic models proposed for ligament tissue the quasi-linear viscoelasticity (QLV) of Fung [40] has been the most successful. The underlying idea is to take the stress at a given time as the convolution of the elastic response with a relaxation function. Fung assumed this relaxation function to be scalar, i.e., isotropic, and proposed a specific form. The expression contains three material parameters τ_1, τ_2, and c, which can be measured using stress–relaxation experiments. QLV has been used successfully to model many types of biological soft tissue [93].

This has been just a very small selection of the material models proposed in the literature. Both elastic and viscoelastic models come in many different flavors. Also, there has been a growing interest to include poroelastic effects. The interested reader is again referred to Weiss and Gardiner [93]. In this thesis, we have chosen to only model the linear part of the ligament stress–strain relationship, again to not clutter the exposition with to many details. The extension to nonlinear elastic material models is straightforward.

Geometric Modelling of Ligaments

The diversity of constitutive models for ligaments is paralleled by the diversity of the different approaches concerning the geometric modeling. If the interest is the accurate description of spatially resolved ligament stresses and interactions with the surrounding soft tissue and bone, a full three-dimensional continuum representation cannot be

[2] This effect is well-known to athletes.

avoided. Few such models have been used, though, because of their inherent difficulties. Constitutive models for three-dimensional fiber-reinforced continua which describe the nonlinear anisotropic response are difficult to construct, and the parameters they contain are hard to determine. Also, the large deformations undergone by ligaments can lead to severe mesh problems. The survey article of Weiss and Gardiner [93] provides an overview over a few three-dimensional models.

When the interest is more on the overall joint kinematics than on detailed stresses in the ligaments, one-dimensional models can be used. These can be much easier to handle, because the constitutive equations for one-dimensional models are simpler and there are no problems with the mesh quality. The easiest models consist of a single linear or nonlinear spring. Others use several springs for each ligament, with individual springs representing fiber bundles within a ligament. Blankevoort and Huiskes [15] used line elements and allowed them to follow the curved edge of a contacting bone, hence including contact in their model.

All these one-dimensional ligament models can only represent tensile stresses, and possibly contact stresses in some models. Yet in vivo, ligaments also experience some shear and transverse loading. In particular at insertion sites complex loading patterns are not uncommon [68]. A geometric approach which can support more complex loading situations are Cosserat rods. In addition to tension, which is also supported by line elements, they can express large deformation shear, bending, and torsion, while retaining the advantages of a one-dimensional model. We will use Cosserat rods for the simulation of ligaments, and they are introduced formally in Sec. 4.2. To our knowledge, Cosserat rods have not been used previously to model ligament mechanics.

As a compromise between the simplicity of one-dimensional models and the power of expression of three-dimensional models, two-dimensional shell models have been used several times [93].

2.4 The Attachment of Ligaments to Bone

The junction of a ligament or a tendon to bone is called an *insertion site*, or *enthesis*. It is quite complex and can vary greatly from ligament to ligament as well as between the two ends of a single ligament. When attaching two materials of widely different Young's modulus, large stress discontinuities can occur over the interface, which makes adhesion difficult. In engineering technology, the solution is often a complex knot or other fastening. In joints the problem has been solved by fusing ligament fibres into the bone [29].

Broadly, insertion sites have been categorized into two classes, *direct* and *indirect*. Direct insertions occur, for example, at the femoral attachment of the medial collateral ligament and anterior cruciate ligament of the knee. They are usually well-defined areas with a sharp boundary between the bone and the attaching ligament (Fig. 2.6, left). Most ligament fibrils at direct insertion sites are deep fibrils that meet the bone at approximately right angles.

Typical direct entheses can be divided in four regions. Going from ligament to bone,

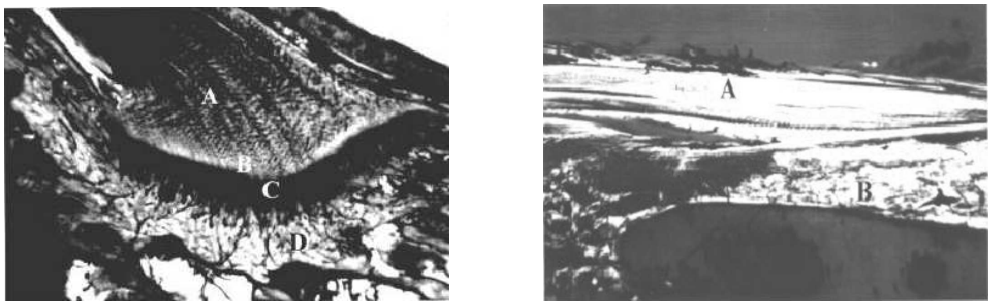


Figure 2.6: Left: Direct insertion site at the femoral insertion of a rabbit MCL. Four distinct zones can be clearly observed: ligament (A), uncalcified fibrocartilage (B), calcified fibrocartilage (C), and bone (D). Right: Indirect insertion site at the tibial insertion of a rabbit MCL. The fiber (A) enters the bone (B) at an oblique angle. (Reprinted from Weiss and Gardiner [93].)

these are

1. Ordinary ligament.
2. A fibrocartilage region, about 300 μm wide. In this region cartilage cells appear, lying in rows in the extracellular matrix of the ligament. The cross-sectional area of the tendon increases slightly to accommodate the cells.
3. A mineralized fibrocartilage region, about 200 μm wide. There is a sharp boundary between this region and the last. The mineralization does not start gradually, but appears as a clear tideline.
4. The mineralized fibrocartilage, containing mineralized tendinous fibers, merges imperceptibly into the rest of the bone, with no clear point where the fibers stop and the bone begins.

At indirect insertion sites, the collagen fibres meet the bone at an oblique angle (Fig. 2.6, right). The result is a more gradual transition between soft and hard tissue and a larger area of attachment. The superficial fibres dominate at indirect insertion sites and they attach to the bone mainly by blending with the periosteum. No fibrocartilagenous transitional zone can be observed in indirect insertions. Indirect insertions can be found at the tibial attachment of the MCL.

Attaching ligaments to bone by blending the former into the latter is certainly mechanically effective. The low-modulus material penetrates the high-modulus one and becomes of high modulus itself within a millimeter or so. There are no real problems, therefore, in bonding the ligament to the bone. The critically important point is the continuity in the collagen fibrils right from true ligament into the heart of the bone. There is no line of weakness. Insertion sites are hence, mechanically, rigid junctions, and can be modeled as such. There is no torsion or bending possible of the ligament

end relative to the bone, other than the torsion and bending of the ligament itself away from the insertion. This motivates the coupling conditions adopted in Chap. 5 for the bone-ligament junctions.

2.5 The Problem of Getting Material Parameters

Any numerical simulation of a human joint which is expected to have more than just qualitative relevance is faced with the problem of how to get material parameters. A great many experiments for material properties for bones and ligaments have been published, and the results are varied. As mentioned in Sec. 2.2, Currey [29], p. 54, conjectures that a large part of this variation is real and not caused by sloppy experiments. Material values seem to differ from joint to joint, from test person to test person, and are even nonuniform in each single specimen. Details of the testing procedure can have a noticeable influence on the results. Important aspects are, for example, the orientation of the specimen in relation to the bone from which it came, whether the specimen is wet or dry (dry bone is stiffer and much more brittle), and the strain rate [29].

There are several ways to measure the elastic properties of bone. The most straightforward is to apply a load to a specimen and calculate the elastic properties from the resulting deformation (or vice versa). Using this methods it is relatively simple to determine Young's modulus in a variety of directions. Cancellous bone and bone full of cavities can be tested without there being too many worries about what precisely is being measured. Also, the effect of the strain rate can be investigated.

A second possibility is to measure the velocity of sound waves in bone. The velocity of sound in a medium is $\sqrt{E/\rho}$, E Young's modulus, and ρ the density. So in theory Young's modulus can be calculated from a knowledge of the sound velocity and the density. In reality this simple formula holds only for isotropic materials and is more complicated for anisotropic ones. Also, there are some methodological problems when it is applied to cancellous bone. However, ultrasonic testing can make possible the derivation of all the stiffness coefficients and, with difficulty, their determination all from the same specimen. Also, it can be applied to complexly shaped specimens, and in some circumstances, it can be used in vivo. Both the ultrasonically and the mechanically determined values for the properties show considerable variation as a function of orientation, with the stiffness measured along the length of the bone being about 1.6 to 2.4 times as great as that measured at right angles to it.

The ideas about stiffness and strength applicable to compact bone have to be severely modified for cancellous bone. The arrangement of trabeculae in space means that cancellous bone behaves like a structure as well as like a material. The loads in cancellous bone can be transferred from place to place by bending moments, and compressive loads may cause individual trabeculae to buckle. End effects are far more important in cancellous than in compact bone, and that makes mechanical testing very difficult. Tensile tests are made difficult by the problem of gripping the sponge-like structure which is cancellous bone. It may appear easier to perform compressive tests. However, the ends of cancellous specimen deform much more under compression than the middle, because

they consist of many isolated struts which are not supported sideways. A calculation of Young's modulus, assuming a uniform deformation of the entire specimen will systematically underestimate the stiffness of the bone specimen [29].

Most of these experimental methods cannot supply in vivo data, yet patient-specific measurements appear to be important, considering the natural variation in material parameters. In recent years there has been progress in using quantitative computer tomograph (QCT) scanning to obtain patient-specific in vivo material parameters. QCTs report the radiodensity of the material, which is measured using the Hounsfield scale. The radiodensity can be correlated with the material density, which in turn allows to estimate Young's modulus. See Yosibash et al. [97] for a comparison of numerical simulations using this spatially resolved Young's modulus with experimental data.

Testing the properties of ligament material is equally difficult. The testing procedure can influence the material properties considerably. Depending on precisely what kind of information is desired it is possible to perform a range of tests such as uniaxial tension, strip biaxial tension, and shear. All these tests can be performed at a fixed strain rate, or under creep and relaxation conditions in order to investigate the viscoelastic effects. Uniaxial tensile tests are commonly used to determine data on the one-dimensional, tensile properties of ligaments. Strip biaxial tests are used for samples with a low length-to-width ratio to minimize the tissue's lateral stretch. Shear tests provide a means to obtain information about the contributions of the collagen fibers and the ground substance matrix, by testing with various orientations of the shear deformation relative to the fiber directions. Further experiments try to quantify tissue permeability and assess the role of fluid flow in the mechanics of ligament materials.

There are numerous technical difficulties associated with tensile testing of ligament material. Ligament specimens are frequently small and are prone to slip from the clamps of the testing machine. Alternatively, the tissue may fail due to the stress induced by the clamping. Some investigator have tried to freeze the tissue at the clamped area but care has to be taken not to appreciably change the overall material properties. Also, test specimens may have irregular geometries which lead to very inhomogeneous stress distributions. In order to obtain a homogeneous distribution together with a strong grip, specimens are frequently cut in a dog-bone shape. For tests of bone–ligament complexes, special bone fixtures are used to ensure a strong grip and a positioning ensuring longitudinal loading of the ligament.

The calculation of stress in uniaxial tension experiments requires the accurate measurement of the applied load and the tissue cross-sectional area. While the former is fairly straightforward, the latter can be quite difficult due to the irregular shapes of ligament cross-sections. Calipers have been used extensively. Their advantage is the ease of use; however, they require the assumption of a regular cross-sectional shape. Also, they may introduce errors by deforming the tissue. In recent years, noncontact methods, such as laser micrometers have gained acceptance. Their accuracy is not impaired by measurement-induced deformations or shape assumptions.

In order to measure strain in soft tissues, many investigators divide the crosshead displacement of the testing machine by the initial length of the specimen. This may introduce large errors due to slack in the system, slippage in the clamps, or inhomo-

geneities in the strain field. To avoid some of these problems, pins have been attached to the insertion sites and various other testing points and the distances between the pins have been measured. Different devices have been used for the distance measurement, e.g., calipers, but also liquid metal strain gauges or Hall effect strain transducers. Noncontact methods include the video dimension analyzer (VDA) which utilizes a video image of the test specimen. In its simplest form, two or more reference lines are drawn on the ligament surface perpendicular to the loading axis. The VDA system tracks the distance between the lines and converts the distance into a voltage. More advanced video systems use special image processing software to track point markers instead of lines. To a certain extend, this allows the measuring of two-dimensional strain as well as quantifying strain inhomogeneities.

Since experimental determination of material parameters is so involved, few groups can afford to do their own measurements. Most works in biomechanical simulation take their parameters from the literature. So do we, and all numerical results in this thesis will use the values $E = 17$ GPa and $\nu = 0.3$ for bone and $E = 330$ MPa, $\nu = 0.3$ for ligament. The bone values are an average of the results in Table 2.1, while the values for ligaments were reported in [93, Table 1].

3 Two-Body Contact Problems on Domains with Curved Boundaries

The simulation of contact between femur and tibia is a central difficulty in the reliable simulation of the mechanics of the human knee. The main problems are the correct formulation of two-body contact for objects on domains with curved boundaries and the efficient and robust solution of the resulting systems. This chapter presents an algorithm for the fast and reliable solution of large two-body contact problems on free-form geometries. In subsequent chapters this algorithm will be used as part of a solver for a more complex knee model.

Contact problems have been treated extensively both in the mathematical and engineering literature. Comprehensive treatments can be found in the monographs of Kikuchi and Oden [52] and Laursen [64]. It is impossible to provide a complete list of articles on the discretization and efficient solution of two-body contact problems. Our work is mainly based on results by Kornhuber and Krause [57] and Wohlmuth and Krause [95]. The basic idea of the new solver stems from Gräser and Kornhuber [42].

3.1 Linear Elasticity

We begin by presenting linear elasticity as a mathematical model for the mechanical behavior of human bone. This choice has been justified in Sec. 2.2, where a few more complex material models for bone were also discussed.

Let Ω be a bounded, open, connected subset of $\mathbb{R}^d$. The dimension d can be two or three; however we will focus on the case $d = 3$. We shall think of the closure $\overline{\Omega}$ of the set Ω as representing the volume occupied by a body 'before it is deformed'. For this reason, the set $\overline{\Omega}$ is called the *reference configuration*. In the absence of external forces we assume the body to be in equilibrium there. We denote the domain boundary by $\partial\Omega$ and suppose that it is piecewise Lipschitz continuous. Then, by Rademacher's theorem, there is an outward unit normal vector defined almost everywhere on $\partial\Omega$ [25, pp. 32–35], which we denote by $\boldsymbol{\nu}$. The boundary is further supposed to consist of two disjoint subsets Γ_D and Γ_N, where Γ_D is expected to have nonempty $(d-1)$-dimensional measure and $\overline{\Gamma}_D \cup \overline{\Gamma}_N = \partial\Omega$. We assume that the body is clamped at Γ_D and that surface force (Neumann) boundary conditions are prescribed on Γ_N.

If the body is subjected to surface and volume forces it will deform and take on a new equilibrium configuration (see Fig. 3.1). This new configuration is described by a function

$$\boldsymbol{\varphi} : \Omega \to \mathbb{R}^d,$$

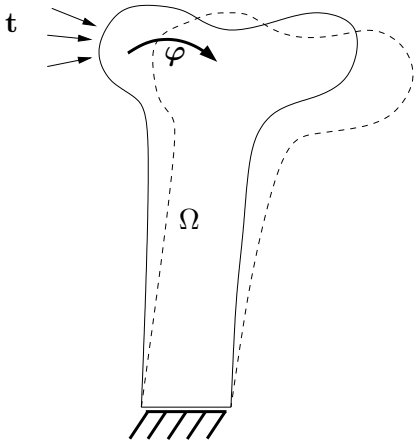


Figure 3.1: Reference configuration Ω (solid line) and deformed configuration (dashed).

which assigns to each point in the reference configuration its position in the deformed configuration. In order to make sense physically, the mapping $\boldsymbol{\varphi}$ must be orientation-preserving in $\overline{\Omega}$ and injective on Ω. In linear elasticity, the function

$$\mathbf{u} : \Omega \to \mathbb{R}^d, \qquad \mathbf{u}(x) = \boldsymbol{\varphi}(x) - x, \tag{3.1}$$

is more commonly used. We follow the convention of the continuum mechanics community to use bold-face symbols for vector-valued quantities.

Let $\mathbf{e}_i$, $0 \leq i < d$ be the canonical basis vectors of $\mathbb{R}^d$. For a given deformation function

$$\boldsymbol{\varphi} = \sum_{i=0}^{d-1} \varphi_i \mathbf{e}_i$$

we define at each $x \in \Omega$ the *deformation gradient*

$$\boldsymbol{\nabla}\boldsymbol{\varphi} = \begin{pmatrix} \frac{\partial \varphi_0}{\partial x_0} & \cdots & \frac{\partial \varphi_0}{\partial x_{d-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \varphi_{d-1}}{\partial x_0} & \cdots & \frac{\partial \varphi_{d-1}}{\partial x_{d-1}} \end{pmatrix}. \tag{3.2}$$

Since, by assumption, a deformation is orientation-preserving, we have

$$\det \boldsymbol{\nabla}\boldsymbol{\varphi}(x) > 0$$

for all $x \in \Omega$. In particular, for all $x \in \Omega$, $\boldsymbol{\nabla}\boldsymbol{\varphi}(x)$ is invertible. We denote the set of all real $d \times d$ matrices with positive determinant by $\mathbb{R}^{d\times d}_+$.

The deformation gradient completely specifies the local deformation to within first order [25]. However, it does not exhibit the invariance under rigid-body motions one would expect from a measure of deformation. To obtain this invariance, the *right Green–St. Venant strain tensor*

$$\begin{aligned} \mathbf{E} &: \Omega \to \mathbb{S}^d \\ \mathbf{E}(x) &= \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{\varphi}^T(x)\boldsymbol{\nabla}\boldsymbol{\varphi}(x) - \mathrm{Id}) \end{aligned}$$

is introduced, where $\mathbb{S}^d$ is the space of symmetric second-order tensors in $\mathbb{R}^d$ and Id denotes the identity. The tensor $\mathbf{E}$ is invariant under rigid-body translations and rotations [25, Thm. 1.8-1]. In terms of the displacements $\mathbf{u}$ we have

$$\mathbf{E} = \frac{1}{2}(\boldsymbol{\nabla}\mathbf{u}^T + \boldsymbol{\nabla}\mathbf{u} + \boldsymbol{\nabla}\mathbf{u}^T\boldsymbol{\nabla}\mathbf{u}). \tag{3.3}$$

If it is known a priori that strains and rigid-body motions remain small, then the quadratic term in (3.3) can be truncated to yield the linear strain tensor

$$\boldsymbol{\varepsilon} = \frac{1}{2}(\boldsymbol{\nabla}\mathbf{u}^T + \boldsymbol{\nabla}\mathbf{u}).$$

As bone only undergoes small strains when subjected to physiological loadings, we will use $\boldsymbol{\varepsilon}$ as the strain measure unless explicitly stated otherwise. The linear strain is invariant under translations, but not under rotations.

Deforming a body will lead to stresses, which, intuitively, are the forces that want to return the body to its stress-free configuration. The existence of stresses is stated by the following fundamental axiom of continuum mechanics. For a detailed treatment of the necessary smoothness assumptions see [25].

Axiom 3.1.1 (Stress principle of Euler and Cauchy)**.** *Consider a body occupying a deformed configuration $\overline{\Omega}^{\boldsymbol{\varphi}} = \boldsymbol{\varphi}(\overline{\Omega})$, and subjected to applied forces represented by densities $\mathbf{f}^{\boldsymbol{\varphi}} : \Omega^{\boldsymbol{\varphi}} \to \mathbb{R}^3$ and $\mathbf{t}^{\boldsymbol{\varphi}} : \Gamma_N^{\boldsymbol{\varphi}} = \boldsymbol{\varphi}(\Gamma_N) \to \mathbb{R}^d$. Then, if $\Omega^{\boldsymbol{\varphi}}$, $\mathbf{f}^{\boldsymbol{\varphi}}$, and $\mathbf{t}^{\boldsymbol{\varphi}}$ are sufficiently smooth, there exists a continuously differentiable field of symmetric tensors*

$$\mathbf{T}^{\boldsymbol{\varphi}} : \overline{\Omega}^{\boldsymbol{\varphi}} \to \mathbb{S}^d,$$

such that

$$-\operatorname{div}\mathbf{T}^{\boldsymbol{\varphi}}(x^{\boldsymbol{\varphi}}) = \mathbf{f}^{\boldsymbol{\varphi}}(x^{\boldsymbol{\varphi}}) \qquad \text{for all } x^{\boldsymbol{\varphi}} \in \overline{\Omega}^{\boldsymbol{\varphi}} \tag{3.4}$$

and

$$\mathbf{T}^{\boldsymbol{\varphi}}(x^{\boldsymbol{\varphi}})\boldsymbol{\nu}^{\boldsymbol{\varphi}} = \mathbf{t}^{\boldsymbol{\varphi}}(x^{\boldsymbol{\varphi}}) \qquad \text{for all } x^{\boldsymbol{\varphi}} \in \Gamma_N^{\boldsymbol{\varphi}}.$$

The divergence in (3.4) *is to be interpreted with respect to the spatial coordinates $x^{\boldsymbol{\varphi}} = \boldsymbol{\varphi}(x)$, and $\boldsymbol{\nu}^{\boldsymbol{\varphi}}$ is the unit outer normal vector along $\Gamma_N^{\boldsymbol{\varphi}}$.*

The equations (3.4) are the equilibrium equations on the deformed configuration. They need to be reformulated as equations on the undeformed configuration Ω. To this end, we introduce the *second Piola-Kirchhoff stress tensor* $\boldsymbol{\sigma} : \overline{\Omega} \to \mathbb{S}^d$ by setting

$$\boldsymbol{\sigma}(x) = \det(\boldsymbol{\nabla}\boldsymbol{\varphi}(x))\boldsymbol{\nabla}\boldsymbol{\varphi}(x)^{-1}\mathbf{T}^{\boldsymbol{\varphi}}(\boldsymbol{\varphi}(x))\boldsymbol{\nabla}\boldsymbol{\varphi}(x)^{-T}.$$

The tensor $\boldsymbol{\sigma}$ is again symmetric. In problems of linear elasticity, where we assume $\boldsymbol{\nabla}\boldsymbol{\varphi} \simeq \mathrm{Id}$, the difference between $\boldsymbol{\sigma}$ and $\mathbf{T}^{\boldsymbol{\varphi}}(\boldsymbol{\varphi}(x))$ can be disregarded. Also, the deformed quantities can be replaced with the undeformed ones. Eq. (3.4) then reads

$$\begin{aligned} -\operatorname{div}\boldsymbol{\sigma}(x) &= \mathbf{f}(x) \qquad \text{for all } x \in \Omega, \\ \boldsymbol{\sigma}(x)\boldsymbol{\nu} &= \mathbf{t}(x) \qquad \text{for all } x \in \Gamma_N. \end{aligned} \tag{3.5}$$

These are the equations of equilibrium on the reference configuration. To obtain a well-posed problem, the Dirichlet boundary conditions

$$\mathbf{u} = 0 \qquad \text{on } \Gamma_D$$

need to be prescribed.

So far our considerations have been independent of a specific material. Properties of the material in question enter the picture in form of *constitutive relations*, which link stresses to strains. A material is called *elastic*, if the stress $\boldsymbol{\sigma}$ at a point $x \in \Omega$ depends only on the deformation gradient at x at the same time, and possibly on x,

$$\boldsymbol{\sigma}(x) = \hat{\boldsymbol{\sigma}}(\boldsymbol{\nabla}\boldsymbol{\varphi}(x), x). \tag{3.6}$$

The response function $\hat{\boldsymbol{\sigma}} : \mathbb{R}^{d\times d}_+ \times \overline{\Omega} \to \mathbb{S}^d$ characterizes the material. By the principle of frame indifference, the dependence on $\boldsymbol{\nabla}\boldsymbol{\varphi}$ must be on the strain only, i.e., there is a second response function $\tilde{\boldsymbol{\sigma}} : \mathbb{S}^d \times \overline{\Omega} \to \mathbb{S}^d$ with

$$\hat{\boldsymbol{\sigma}}(\boldsymbol{\nabla}\boldsymbol{\varphi}(x), x) = \tilde{\boldsymbol{\sigma}}(\boldsymbol{\nabla}\boldsymbol{\varphi}(x)^T \boldsymbol{\nabla}\boldsymbol{\varphi}(x), x).$$

A material is called *homogeneous* if (3.6) does not depend on its second argument. It is called *isotropic* if it behaves the same 'in all directions', i.e.,

$$\hat{\boldsymbol{\sigma}}(\mathbf{F}\mathbf{Q}, x) = \hat{\boldsymbol{\sigma}}(\mathbf{F}, x) \qquad \text{for all } \mathbf{F} \in \mathbb{R}^{d\times d}_+ \text{ and } \mathbf{Q} \in \mathrm{SO}(d).$$

A material is *hyperelastic* if there exists a *stored energy function*

$$W : \mathbb{R}^{d\times d}_+ \times \overline{\Omega} \to \mathbb{R}$$

such that the stress is a derivative of this function

$$\boldsymbol{\sigma}(x) = \hat{\boldsymbol{\sigma}}(\mathbf{F}, x) = \frac{\partial W}{\partial \mathbf{F}}(\mathbf{F}, x), \qquad \text{for all } \mathbf{F} = \boldsymbol{\nabla}\boldsymbol{\varphi} \in \mathbb{R}^{d\times d}_+ \text{ and } x \in \overline{\Omega}.$$

Here $\partial W/\partial \mathbf{F}$ denotes the matrix of partial derivatives of W with respect to the components of $\mathbf{F}$. If, additionally, the applied forces are *conservative*, i.e., can be written as gradients of an energy functional, solving the boundary value problem (3.5) is formally equivalent to finding the stationary points of a total energy functional $\mathcal{J}$, subject to the constraints that $\det \boldsymbol{\nabla}\boldsymbol{\varphi} > 0$ and that $\boldsymbol{\varphi}$ assumes the Dirichlet boundary values on Γ_D [25]. If the applied loads are dead loads, i.e., independent of $\boldsymbol{\varphi}$, the total energy is given by

$$\mathcal{J}(\boldsymbol{\varphi}) = \int_\Omega W(\boldsymbol{\nabla}\boldsymbol{\varphi}(x), x)\, dx - \int_\Omega \mathbf{f}\boldsymbol{\varphi}\, dx - \int_{\Gamma_N} \mathbf{t}\boldsymbol{\varphi}\, ds. \tag{3.7}$$

Arguments from thermomechanics suggest that the assumption of hyperelasticity is physically reasonable, as an elastic material is hyperelastic if and only if the work done in closed processes is nonnegative [45].

For certain materials one may assume a linear relationship

$$\boldsymbol{\sigma} = \mathbf{C} : \boldsymbol{\varepsilon} \tag{3.8}$$

between the second Piola-Kirchhoff stress $\boldsymbol{\sigma}$ and the strain $\boldsymbol{\varepsilon}$. The proportionality factor $\mathbf{C}$ is a fourth-order tensor which is called Hooke tensor. The : symbol denotes tensor contraction and is defined by the component-wise relation

$$\boldsymbol{\sigma}_{ij} = \sum_{k,l=0}^{d-1} \mathbf{C}_{ijkl}\boldsymbol{\varepsilon}_{kl}, \qquad \text{for } i,j \in \{0,\dots,d-1\}.$$

The d^4 components of $\mathbf{C}$ are subject to various restrictions due to the symmetries of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$. If the material is isotropic is turns out that $\mathbf{C}$ depends only on two parameters E and ν, and that (3.8) reduces to

$$\boldsymbol{\sigma}(\boldsymbol{\varepsilon}) = \frac{E\nu}{(1+\nu)(1-2\nu)}(\operatorname{tr}\boldsymbol{\varepsilon})\mathrm{Id} + \frac{E}{1+\nu}\boldsymbol{\varepsilon}. \tag{3.9}$$

Materials which behave like (3.9) are called *St. Venant–Kirchhoff materials*. The parameters $E > 0$ and $0 < \nu < 1/2$ are called Young's modulus (or elastic modulus) and Poisson ratio, respectively.

Inserting the linear material law (3.8) into the equilibrium equations (3.5) we obtain the equations of linear elasticity in their strong form

$$\begin{aligned} -\operatorname{div}[\mathbf{C} : \boldsymbol{\varepsilon}(\mathbf{u})] &= \mathbf{f} && \text{in } \Omega, && \text{(3.10a)}\\ \mathbf{u} &= 0 && \text{on } \Gamma_D, && \text{(3.10b)}\\ \boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu} &= \mathbf{t} && \text{on } \Gamma_N. && \text{(3.10c)} \end{aligned}$$

This problem will now be reformulated in two useful ways. For further reference we introduce several function spaces. Let $L^2(\Omega)$ be the space of all square-integrable scalar functions on Ω and $H^1(\Omega)$ the Sobolev space of scalar functions on Ω which are weakly differentiable. Let $\mathbf{L}^2(\Omega) = (L^2(\Omega))^d$ and $\mathbf{H}^1(\Omega) = (H^1(\Omega))^d$ be their vector-valued counterparts. We denote by $H^1_0(\Omega)$ and $\mathbf{H}^1_0(\Omega)$ the subspaces of those functions in $H^1(\Omega)$ and $\mathbf{H}^1(\Omega)$, respectively, which are zero in the sense of traces on Γ_D. Finally, we introduce $\mathbf{H}^{1/2}(\Gamma_N)$ as the space of traces of functions in $\mathbf{H}^1(\Omega)$ on $\Gamma_N \subset \partial\Omega$.

Hooke's law corresponds to the quadratic energy density

$$W(\boldsymbol{\nabla}\boldsymbol{\varphi}(x), x) = \frac{1}{2}\boldsymbol{\varepsilon}(\mathbf{v}(x)) : \mathbf{C} : \boldsymbol{\varepsilon}(\mathbf{v}(x)).$$

With this choice, the hyperelastic energy functional (3.7) takes the form

$$J(\mathbf{v}) = \frac{1}{2}a(\mathbf{v},\mathbf{v}) - l(\mathbf{v}). \tag{3.11}$$

with the bilinear form

$$a(\mathbf{v},\mathbf{w}) = \int_\Omega \boldsymbol{\varepsilon}(\mathbf{v}(x)) : \mathbf{C} : \boldsymbol{\varepsilon}(\mathbf{w}(x))\,dx \qquad \mathbf{v},\mathbf{w} \in \mathbf{H}^1(\Omega), \tag{3.12}$$

and the linear form

$$l(\mathbf{v}) = \int_\Omega \mathbf{f}\mathbf{v}\,dx + \int_{\Gamma_N} \mathbf{t}\mathbf{v}\,ds, \qquad \mathbf{v} \in \mathbf{H}^1(\Omega). \tag{3.13}$$

Hence (3.10) corresponds to the minimization problem to find a $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ such that

$$J(\mathbf{u}) \leq J(\mathbf{v}), \qquad \text{for all } \mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{3.14}$$

By Korn's inequality, if Γ_D has positive $(d-1)$-dimensional measure, the bilinear form $a(\cdot,\cdot)$ is $\mathbf{H}_0^1$-elliptic, i.e., there exists a $\gamma > 0$ such that

$$a(\mathbf{v},\mathbf{v}) > \gamma \|\mathbf{v}\|^2 \qquad \text{for all } \mathbf{v} \in \mathbf{H}_0^1. \tag{3.15}$$

This is used in the following result.

Lemma 3.1.1. *Let Γ_D have positive $(d-1)$-dimensional measure. Then J is strictly convex and coercive on $\mathbf{H}_0^1$, and J has a unique minimum on $\mathbf{H}_0^1(\Omega)$.*

Proof. [35]. □

Alternatively, there is a corresponding variational formulation of the linear elasticity problem.

Lemma 3.1.2. *The minimization problem* (3.14) *is equivalent to finding $\mathbf{u} \in \mathbf{H}_0^1(\Omega)$ such that*

$$a(\mathbf{u},\mathbf{v}) = l(\mathbf{v}), \qquad \textit{for all } \mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{3.16}$$

Suppose that Ω is polygonal[1] and partitioned by a simplicial grid G. We assume that G resolves the partitioning of the boundary in Γ_D and Γ_N, and that it is quasiuniform, i.e., there exists a constant κ such that each element E contains a sphere of radius $\rho_E \geq h_E/\kappa$, where h_E is the diameter of E. We call $\mathcal{V}$ the set of grid vertices. The scalar quantity h denotes the length of the longest edge in G and is hence a measure of its resolution.

Since Ω is polygonal, the outward unit normal $\boldsymbol{\nu}$ is piecewise constant and discontinuous across edges of $\partial\Omega$. In order to obtain a well-defined normal at grid vertices we introduce the averaged vertex normal

$$\hat{\boldsymbol{\nu}}_p = \frac{\sum_{T \in \partial\Omega, p \in T} \boldsymbol{\nu}(T)}{\|\sum_{T \in \partial\Omega, p \in T} \boldsymbol{\nu}(T)\|} \tag{3.17}$$

for each vertex p of $\partial\Omega$. Here we have used T to denote the triangles of the grid boundary. The set of vertex normals can be extended by linear interpolation to yield a continuous normal field $\hat{\boldsymbol{\nu}} : \partial\Omega \to S^{d-1}$, where S^{d-1} is the $(d-1)$-dimensional unit sphere in $\mathbb{R}^d$.

Standard first order finite elements will be used to discretize (3.16). We call $\mathbf{V}_h(G) = \big(V_h(G)\big)^d$ the space of vector-valued, continuous functions which are linear on each simplex of G, and $\mathbf{V}_{h,0}$ the subset of those functions of $\mathbf{V}_h$ which are zero on Γ_D. Let $a(\cdot,\cdot)$ and $l(\cdot)$ be given by (3.12) and (3.13), respectively. Since $\mathbf{V}_{h,0} \subset \mathbf{H}_0^1$ these forms are well-defined on $\mathbf{V}_{h,0}$, and $a(\cdot,\cdot)$ is $\mathbf{V}_{h,0}$-elliptic. The discrete counterpart of the continuous minimization problem (3.14) is to find a $\mathbf{u}_h \in \mathbf{V}_{h,0}$ such that

$$J(\mathbf{u}_h) \leq J(\mathbf{v}_h), \qquad \text{for all } \mathbf{v}_h \in \mathbf{V}_h. \tag{3.18}$$

[1] We will show in Sec. 3.6 how nonpolygonal domains can be approximated arbitrarily well.

Note that since J is coercive and strictly convex on $\mathbf{H}_0^1$ it is so in particular on $\mathbf{V}_{h,0}$.

Making use of the ellipticity of $a(\cdot,\cdot)$ on $\mathbf{V}_{h,0}$ we get the following result (see, e.g., [21]). Denote by $\|\cdot\|_1$ the norm in $\mathbf{H}^1(\Omega)$ and by $\mathbf{H}^2(\Omega)$ the second-order Sobolev space with the norm $\|\cdot\|_2$.

Lemma 3.1.3. *The minimization problem* (3.18) *has a unique solution* $\mathbf{u}_h$*. If* $\mathbf{u}$ *is the unique solution of the continuous problem* (3.14) *and* $\mathbf{u} \in \mathbf{H}^2(\Omega)$ *then*

$$\|\mathbf{u}_h - \mathbf{u}\|_1 \leq Ch\|\mathbf{u}\|_2$$

holds with a constant C *independent of* h*.*

Just as in the continuous case there is a variational equality corresponding to the minimization formulation (3.18).

Lemma 3.1.4. *The function* $\mathbf{u}_h$ *is a solution of* (3.18) *if and only if* $\mathbf{u}_h \in \mathbf{V}_{h,0}$ *and*

$$a(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h) \qquad \text{for all } \mathbf{v}_h \in \mathbf{V}_{h,0}. \tag{3.19}$$

Let $\psi_p \in V_{h,0}$ be the scalar nodal basis function corresponding to the vertex p. Using the canonical basis vectors $\mathbf{e}_i$, $i = 0, \ldots, d-1$ of $\mathbb{R}^d$ we also define the vector-valued nodal basis functions $\boldsymbol{\psi}_{p,i} = \psi_p \mathbf{e}_i$, $p \in \mathcal{V} \setminus \Gamma_D$, $i = 0, \ldots, d-1$. Let n be the number of non-Dirichlet vertices in G. Writing finite element functions with respect to this basis we arrive at the algebraic energy functional

$$J(v) = \frac{1}{2} v^T A v - bv \qquad v \in \mathbb{R}^{dn}. \tag{3.20}$$

In an abuse of notation we have given it the same symbol as the corresponding functional on $\mathbf{V}_h$. The matrix $A \in \mathbb{R}^{dn \times dn}$ has a block structure of $d \times d$ submatrices. For $p, q \in \mathcal{V}$, the block entry A_{pq}, $p, q \in \mathcal{V}$ is given by

$$(A_{pq})_{ij} = \int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{\psi}_{p,i}(x)) : \mathbf{C} : \boldsymbol{\varepsilon}(\boldsymbol{\psi}_{q,j}(x))\, dx, \tag{3.21}$$

and the right hand side vector $b \in \mathbb{R}^{dn}$ is defined by

$$(b_p)_i = \int_\Omega \mathbf{f} \boldsymbol{\psi}_{p,i}\, ds + \int_{\Gamma_N} \mathbf{t} \boldsymbol{\psi}_{p,i}\, ds. \tag{3.22}$$

Here and in the following we use two indices to specify the components of a vector $b \in \mathbb{R}^{dn}$. By $(b_p)_i$ we mean the component corresponding to the vertex p and the i-the coordinate direction. The variational equality (3.16) reduces to the linear system

$$Au = b,$$

where $u \in \mathbb{R}^{dn}$ is the vector of coefficients of the solution function $\mathbf{u}_h$.

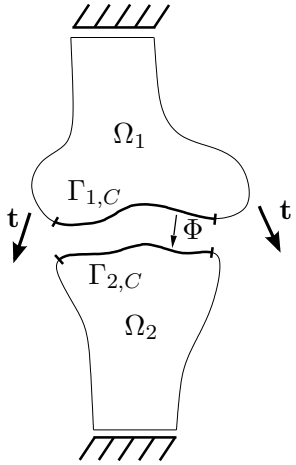


Figure 3.2: Two-body contact problem.

3.2 Two-Body Contact in Linear Elasticity

In this section we introduce the formulation of the contact conditions. Our exposition is mainly based on [95]. However, there it was implicitly assumed that the two contact boundaries coincide in their reference configurations. If this is not the case, a *contact mapping* is necessary which identifies the two contact boundaries. Formulations including contact mappings can be found, e.g., in [17, 34].

Let Ω_1, Ω_2 be two disjoint domains, each with a piecewise Lipschitz continuous boundary, and set $\Omega = \Omega_1 \cup \Omega_2$. The boundaries are each supposed to consist of three pairwise disjoint subsets $\Gamma_{i,D}$, $\Gamma_{i,N}$, $\Gamma_{i,C}$ with

$$\partial\Omega_i = \overline{\Gamma}_{i,D} \cup \overline{\Gamma}_{i,N} \cup \overline{\Gamma}_{i,C}, \qquad i \in \{1,2\}.$$

Both $\Gamma_{1,D}$ and $\Gamma_{2,D}$ are expected to have positive $(d-1)$-dimensional measure. The boundary parts $\Gamma_{1,C}$ and $\Gamma_{2,C}$ are the contact boundary, which is where we expect contact to occur. More precisely, while the actual zone of contact is an unknown of the problem, the model contains the assumption that it will be a subset of $\Gamma_{i,C}$. This is not a serious restriction in small deformation mechanics.

Under applied loads, the two bodies will deform and take on new configurations. As is customary in linear elasticity we will specify deformations by their displacement functions $\mathbf{u}_i : \Omega_i \to \mathbb{R}^d$, $i = 1,2$ (3.1). In the context of linear elasticity we assume these displacements to be small. In Sec. 3.1 it was shown that in the absence of contact the equilibrium displacement functions are solutions of the following boundary value problem

$$-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}_i) = \mathbf{f}_i \qquad \text{on } \Omega_i, \tag{3.23a}$$

$$\mathbf{u}_i = 0 \qquad \text{on } \Gamma_{i,D}, \tag{3.23b}$$

$$\boldsymbol{\sigma}(\mathbf{u}_i)\boldsymbol{\nu}_i = \mathbf{t}_i \qquad \text{on } \Gamma_{i,N}, \tag{3.23c}$$

for $i \in \{1,2\}$. Here, $\mathbf{f}_i : \Omega_i \to \mathbb{R}^d$ and $\mathbf{t}_i : \Gamma_{i,N} \to \mathbb{R}^d$ are prescribed volume and surface forces, respectively. The outward unit normal vectors of Ω_i are denoted by Ω_i.

In order to model the contact between the two bodies, further conditions have to be prescribed at $\Gamma_{i,C}$. To this end, we introduce a homeomorphism $\Phi : \Gamma_{1,C} \to \Gamma_{2,C}$, which we call the *contact mapping*. It forms an a priori identification of points on $\Gamma_{1,C}$ and $\Gamma_{2,C}$ which may come into contact with each other. There is a certain amount of choice in its construction; see [34]. To be specific we take Φ to be the normal projection of $\Gamma_{1,C}$ onto $\Gamma_{2,C}$ and assume that $\Gamma_{1,C}$ and $\Gamma_{2,C}$ are chosen such that this is possible.[2] The contact mapping allows to define the initial gap function $g : \Gamma_{1,C} \to \mathbb{R}$ with $g(x) = |\Phi(x) - x|$ and the relative displacement

$$\begin{aligned} [\,\cdot\,]_\Phi &: \prod_i \mathbf{H}^1(\Omega_i) \to \mathbf{H}^{1/2}(\Gamma_{1,C}), \\ [\mathbf{u}]_\Phi &= \mathbf{u}_1|_{\Gamma_{1,C}} - \mathbf{u}_2|_{\Gamma_{2,C}} \circ \Phi, \end{aligned} \tag{3.24}$$

where $\circ$ denotes function composition. We can now state the linearized nonpenetration condition

$$\langle [\mathbf{u}]_\Phi, \boldsymbol{\nu}_1 \rangle \le g, \qquad \text{for all } x \in \Gamma_{1,C}. \tag{3.25}$$

Let $\boldsymbol{\nu}_1$ have the components ν_i, $0 \le i < d$. The Kuhn–Tucker conditions for the constraint (3.25) are

$$\sigma_{\boldsymbol{\nu}_1}(\mathbf{u}_1)|_{\Gamma_{1,C}} = \sigma_{\boldsymbol{\nu}_1}(\mathbf{u}_2)|_{\Gamma_{2,C}} \circ \Phi \le 0, \tag{3.26}$$

$$\big(\langle [\mathbf{u}]_\Phi, \boldsymbol{\nu}_1 \rangle - g\big) \cdot \sigma_{\boldsymbol{\nu}_1}(\mathbf{u}_1)|_{\Gamma_{1,C}} = 0, \tag{3.27}$$

with the normal pressure $\sigma_{\boldsymbol{\nu}_1} = \sum_{i,j} \nu_i \sigma_{ij} \nu_j$ playing the role of the Lagrange multiplier. Condition (3.26) ensures that the normal stresses at the contact boundary have the character of a pure pressure. Equation (3.27) states that there can be non-vanishing normal pressure at $\Gamma_{i,C}$ only if there is contact. Setting $(\boldsymbol{\sigma}_T)_i = \sum_j \sigma_{ij}\nu_j - \sigma_{\boldsymbol{\nu}_1}\nu_i$, $i = 0, \ldots, d-1$ we also prescribe

$$\boldsymbol{\sigma}_T(\mathbf{u}_1)|_{\Gamma_{1,C}} = \boldsymbol{\sigma}_T(\mathbf{u}_2)|_{\Gamma_{2,C}} \circ \Phi = 0, \tag{3.28}$$

which is the absence of friction. See [17, 34] for detailed derivations of these conditions. For existence, uniqueness and regularity results, we refer to [17].

Define the set of admissible displacements

$$\mathcal{K} = \big\{ \mathbf{v} \in \prod_i \mathbf{H}_0^1(\Omega_i) \;\big|\; \langle [\mathbf{v}]_\Phi, \boldsymbol{\nu}_1 \rangle \le g, \quad \text{a.e.} \big\}, \tag{3.29}$$

and note that it is closed and convex. In analogy to (3.14) we can write the contact problem in minimization form [17].

Lemma 3.2.1. *Let* $\mathbf{u}$ *be a solution of* (3.23) *subject to* (3.25) *and* (3.28). *Then* $\mathbf{u} \in \mathcal{K}$ *such that*

$$J(\mathbf{u}) \le J(\mathbf{v}) \qquad \forall \mathbf{v} \in \mathcal{K}, \tag{3.30}$$

where $J : \mathbf{H}^1 \to \mathbb{R}$ *is given by* (3.11).

[2] Note that in contrast to Krause and Sander [60], we use the normal on $\Gamma_{1,C}$ instead of $\Gamma_{2,C}$.

From Lemma 3.1.1 we know that J is strictly convex and coercive on $\mathcal{K} \subset \mathbf{H}_0^1(\Omega)$. Using convex analysis [35, Prop. 1.2] we can directly show the following.

Lemma 3.2.2. *The minimization problem* (3.30) *has a unique solution.*

We further get the following equivalence [17].

Lemma 3.2.3. *Let* $\mathbf{u}$ *be a solution of* (3.30)*. Then* $\mathbf{u}$ *also solves the variational inequality*

$$\mathbf{u} \in \mathcal{K} \quad : \quad a(\mathbf{u}, \mathbf{v} - \mathbf{u}) \geq l(\mathbf{v} - \mathbf{u}) \qquad \forall \mathbf{v} \in \mathcal{K}. \tag{3.31}$$

Conversely, let $\mathbf{u}$ *be a solution of* (3.31)*. Then it also solves* (3.30)*.*

3.3 Discretization Using Mortar Elements

Having presented the discretization for the linear elasticity problem in Sec. 3.1, we now turn to the discretization of the contact conditions (3.25) and (3.28). Various approaches can be found in the literature. Nodal collocation has been popular for a long time [64]. In recent years, weak formulations have been shown to yield much better estimates [48]. In particular, the mortar approach has been used with great success because it provides great flexibility and yields stable discretizations for nonmatching grids.

Let W be the trace space of $H_0^1(\Omega_1)$ restricted to $\Gamma_{1,C}$. Assuming that $\overline{\Gamma}_{1,C}$ is a compact subset of $\partial\Omega_1 \setminus \overline{\Gamma}_{1,D}$ we have $W = H^{1/2}(\Gamma_{1,C})$. Let

$$W^+ = \left\{ w \in W \;\middle|\; w \geq 0 \text{ a.e.} \right\}$$

be the cone of positive functions in W.

The mortar method replaces the pointwise inequality in (3.29) by a set of weak constraints [48]

$$\mathcal{K}^w = \left\{ \mathbf{v} \in \prod_i \mathbf{H}_0^1(\Omega_i) \;\middle|\; \mu(\langle [\mathbf{v}]_\Phi, \boldsymbol{\nu}_1 \rangle) \leq \mu(g) \quad \forall \mu \in M^+ \right\},$$

where the mortar space M^+ is the cone of all positive functionals on W^+,

$$M^+ = \left\{ \mu : W^+ \to \mathbb{R} \;\middle|\; \mu(w) \geq 0 \right\}. \tag{3.32}$$

We assume that $\Gamma_{1,C}$ is resolved by the grid and denote by $\mathcal{V}_{1,C}$ the set of grid vertices on $\Gamma_{1,C}$. Wohlmuth [96] proposed to discretize M^+ using dual mortar basis functions $\{\theta\}$. For simplicial grids the functions $\theta_p : \Gamma_{1,C} \to \mathbb{R}$, $p \in \mathcal{V}_{1,C}$, are defined elementwise for $T \in \Gamma_{1,C}$ such that $\operatorname{supp} \theta_p = \operatorname{supp} \psi_p$ and

$$\theta_p|_T := (d\psi_p - \sum_{q \in T, q \neq p} \psi_q)\big|_T.$$

They are compactly supported, piecewise linear functions, and form a partition of unity. They are, however, discontinuous (see Fig. 3.3). Most importantly they fulfill a *biorthogonality relation* with the nodal basis functions $\{\psi\}$. Let T be a simplex in $\mathbb{R}$ or $\mathbb{R}^2$ and

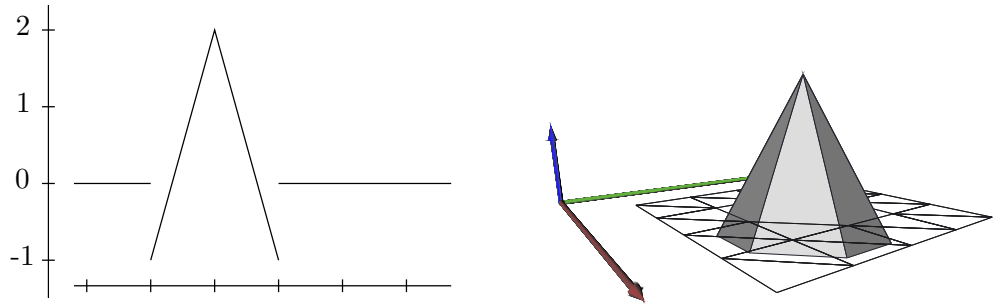


Figure 3.3: Dual mortar basis functions in 1d and 2d.

let ψ_q and θ_p be the nodal shape function and dual mortar shape function for the vertices $q, p \in T$, respectively. Then

$$\int_T \theta_p \psi_q \, ds = \delta_{pq} \int_T \psi_q \, ds.$$

Since $\Gamma_{1,C}$ is resolved by the grid we have

$$\int_{\Gamma_{1,C}} \theta_p \psi_q \, ds = \delta_{pq} \int_{\Gamma_{1,C}} \psi_q \, ds \tag{3.33}$$

for all dual mortar functions θ_p and nodal basis functions ψ_q associated to vertices $p, q \in \mathcal{V}_{1,C}$. With

$$W_h^+ = \left\{ w_h \in V_h(\Gamma_{1,C}) \;\middle|\; w_h \geq 0 \right\}$$

the discrete cone of positive traces we define the discrete positive mortar cone

$$M_h^+ = \left\{ \mu_h \in \operatorname{span}_{p \in \Gamma_{1,C}} \theta_p \;\middle|\; \int_{\Gamma_{1,C}} \mu_h w_h \, ds \geq 0, \; \forall w_h \in W_h^+ \right\}$$

and the set of discrete weakly admissible displacements

$$\mathcal{K}_h = \left\{ \mathbf{v}_h \in \prod_i \mathbf{V}_{h,0}(\Omega_i) \;\middle|\; \int_{\Gamma_{1,C}} \langle [\mathbf{v}_h]_\Phi, \boldsymbol{\nu}_1 \rangle \mu_h \, ds \leq \int_{\Gamma_{1,C}} g \mu_h \, ds \quad \forall \mu_h \in M_h^+ \right\}. \tag{3.34}$$

Note that we have discretized the functionals $\mu(\cdot) \in M^+$ by L^2-functions $\mu_h \in M_h^+$ which are functionals in the sense of $\mu_h(\cdot) = \int_{\Gamma_{1,C}} \cdot \, \mu_h \, ds$.

Let J be given by (3.11). We state the discrete minimization formulation of the mortar-discretized two-body contact problem, which is to find a $\mathbf{u}_h \in \mathcal{K}_h$ such that

$$J(\mathbf{u}_h) \leq J(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathcal{K}_h. \tag{3.35}$$

Using the same reasoning as for the continuous case (Lem. 3.2.2) we obtain existence and uniqueness of solutions.

Lemma 3.3.1. *The discrete minimization problem* (3.35) *has a unique solution.*

In complete analogy to the continuous case there is also a formulation of the contact problem as a variational inequality: find $\mathbf{u}_h \in \mathcal{K}_h$ such that

$$a(\mathbf{u}_h, \mathbf{v}_h - \mathbf{u}_h) \geq l(\mathbf{v}_h - \mathbf{u}_h) \qquad \forall \mathbf{v}_h \in \mathcal{K}_h. \tag{3.36}$$

So far only partial results concerning the a priori error of the mortar discretization for contact problems are available. For the case that $\boldsymbol{\nu}_1$ is constant on $\Gamma_{1,C}$ and $g \equiv 0$, Hüeber and Wohlmuth [48] showed that

$$\|\mathbf{u} - \mathbf{u}_h\|_1 + \|\mu - \mu_h\|_{-\frac{1}{2},\Gamma_{1,C}} \leq C h^{\frac{1}{2}+\epsilon} |\mathbf{u}|_{\frac{3}{2}+\epsilon,\Omega}$$

if $\mathbf{u} \in \mathbf{H}^{\frac{3}{2}+\epsilon}(\Omega)$, $0 < \epsilon \leq \frac{1}{2}$, and an additional technical regularity assumption holds [48, Assumption 3.1]. Here, $\mathbf{u}$ and $\mathbf{u}_h$ are the solutions of the continuous (3.31) and the discrete (3.36) problem, and $\mu \in M^+$ and $\mu_h \in M_h^+$ are the corresponding Lagrange multipliers. $\|\cdot\|_1^2 = \sum_{i=1,2} \|\cdot\|_{1,\Omega_i}^2$ is the broken $\mathbf{H}^1$-norm, and $|\cdot|_{\frac{3}{2}+\epsilon,\Omega}$ is the half-norm in $\mathbf{H}^{\frac{3}{2}+\epsilon}(\Omega)$. This result is optimal, but it only holds for straight contact boundaries. For curved interfaces and the Laplace equation $-\Delta u = f$ on two subdomains with equality constraints, Flemisch et al. [36] showed the similar result that if $u \in H^2(\Omega)$,

$$\|u - u_h\|_1 + \|\mu - \mu_h\|_{-\frac{1}{2},\Gamma_{1,C}} \leq h C(u).$$

Note, however, that the interface normal $\boldsymbol{\nu}_1$ does not appear in the problem formulation, since it is a scalar problem. Also, the constraints are *equality constraints*, that is, the multipliers μ are not restricted to the positive cone. A result for contact problems with curved, nonconforming contact boundaries seems to still be missing.

Let $v \in \mathbb{R}^{dn}$ be the vector of coefficients of a finite element function $\mathbf{v}_h \in \mathbf{V}_{h,0}(\Omega)$ with respect to the nodal basis $\{\boldsymbol{\psi}\}$. We split the vector into four subvectors v_1^I, v_1^C, v_2^I, v_2^C, where the $v_i^C, i = 1, 2$, correspond to the vertices on the contact boundaries $\Gamma_{i,C}$, and the v_i^I correspond to the remaining nodes of the respective grids. The set $\mathcal{K}_h$ of discrete weakly admissible displacements can then be written algebraically as

$$\overline{\mathcal{K}}_{\text{alg}} = \left\{ v \in \mathbb{R}^{dn} \;\middle|\; \overline{\mathsf{D}} v_1^C - \overline{\mathsf{M}} v_2^C \leq \mathsf{g} \right\}, \tag{3.37}$$

where $\mathsf{g} \in \mathbb{R}^{|\mathcal{V}_{1,C}|}$ with $\mathsf{g}_p = \int_{\Gamma_{1,C}} g \theta_p \, ds$. The matrices $\overline{\mathsf{D}}$ and $\overline{\mathsf{M}}$ have dimensions $|\mathcal{V}_{1,C}| \times d|\mathcal{V}_{1,C}|$ and $|\mathcal{V}_{1,C}| \times d|\mathcal{V}_{2,C}|$, respectively, and the entries

$$\begin{aligned}
\overline{\mathsf{D}}_{pq}^i &= \int_{\Gamma_{1,C}} \theta_p \langle \boldsymbol{\psi}_{q,i}|_{\Gamma_{1,C}}, \boldsymbol{\nu}_1 \rangle \, ds \\
&= \int_{\Gamma_{1,C}} \theta_p \psi_q \boldsymbol{\nu}_1^i \, ds, && p, q \in \mathcal{V}_{1,C}, \\
\overline{\mathsf{M}}_{pq'}^i &= \int_{\Gamma_{1,C}} \theta_p \langle \boldsymbol{\psi}_{q',i}|_{\Gamma_{2,C}} \circ \Phi, \boldsymbol{\nu}_1 \rangle \, ds \\
&= \int_{\Gamma_{1,C}} \theta_p (\psi_{q'} \circ \Phi) \boldsymbol{\nu}_1^i \, ds, && p \in \mathcal{V}_{1,C},\ q' \in \mathcal{V}_{2,C}.
\end{aligned}$$

If we replace the domain normals $\boldsymbol{\nu}_1$ in the expressions for $\overline{\mathsf{D}}$ and $\overline{\mathsf{M}}$ by the averaged vertex normals $\hat{\boldsymbol{\nu}}_p, p \in \mathcal{V}_{1,C}$, we can move them out of the integrals and obtain the modified admissible set

$$\mathcal{K}_{\text{alg}} = \left\{ v \in \mathbb{R}^{dn} \;\middle|\; N\mathsf{D}v_1^C - N\mathsf{M}v_2^C \leq \mathsf{g} \right\}, \tag{3.38}$$

which we will consider exclusively from now on. The matrix N has the dimensions $|\mathcal{V}_{1,C}| \times d|\mathcal{V}_{1,C}|$ and is block-diagonal with the entries $N_{pp} = \hat{\boldsymbol{\nu}}_p^T, p \in \mathcal{V}_{1,C}$. The matrices D and M have dimensions $d|\mathcal{V}_{1,C}| \times d|\mathcal{V}_{1,C}|$ and $d|\mathcal{V}_{1,C}| \times d|\mathcal{V}_{2,C}|$, respectively, and the entries

$$\mathsf{D}_{pq} = \mathrm{Id}_{d\times d} \int_{\Gamma_{1,C}} \theta_p \psi_q \, ds, \qquad p, q \in \mathcal{V}_{1,C}, \tag{3.39}$$

$$\mathsf{M}_{pq'} = \mathrm{Id}_{d\times d} \int_{\Gamma_{1,C}} \theta_p (\psi_{q'} \circ \Phi) \, ds, \qquad p \in \mathcal{V}_{1,C},\ q' \in \mathcal{V}_{2,C}. \tag{3.40}$$

Using (3.33) it follows that D is a diagonal matrix. The assembly of the matrix M, involving the contact mapping Φ, is difficult enough to warrant its own section. We will cover the technical details in Sec. 3.5.

From (3.38) the advantage of using dual mortar basis functions becomes evident. Since D is diagonal it is easy to invert and the constraints (3.38) can be localized. This will be used in the next section.

We can now write down the complete algebraic problem. Again there is a minimization formulation and a formulation as a variational inequality. Using the energy functional J (3.20), we say that $u \in \mathcal{K}_{\text{alg}}$ is a solution of the algebraic two-body contact problem if

$$J(u) \leq J(v) \qquad \forall v \in \mathcal{K}_{\text{alg}}. \tag{3.41}$$

Let A be the stiffness matrix given by (3.21) and b the force vector (3.22). The solution of (3.41) can also be characterized as the unique vector $u \in \mathcal{K}_{\text{alg}}$ for which

$$u^T A(v-u) \geq b(v-u), \qquad \forall v \in \mathcal{K}_{\text{alg}}.$$

This corresponds to the variational inequality for finite element functions (3.36).

3.4 The Truncated Nonsmooth Newton Multigrid Algorithm

> Wie wir sehen, kann eine gute Idee eine Menge Arbeit ersparen.
> (Deuflhard & Bornemann, Numerik II, p. 242)

Various algorithms have been proposed for constraint convex minimization problems like (3.41). Important examples are penalty methods [64], active-set methods [62], and monotone multigrid methods [95]. Penalty methods involve a parameter which has to be chosen with care. Active-set methods are expensive, unless the inner problems are solved inexactly. Then, however, they suffer from robustness problems [59]. Monotone multigrid methods (MMG) converge globally and with the asymptotic rate of linear

multigrid. This makes them very useful for biomechanical contact problems [58]. The price, however, is a fairly difficult implementation, especially if the domain boundaries are allowed to have non-constant normals. In this section we present the Truncated Nonsmooth Newton Multigrid algorithm (TNNMG), which has recently been proposed by Gräser et al. [42, 43]. It is considerably simpler to implement than the monotone multigrid method yet retains all the favorable properties. In Sec. 3.8 we compare the two methods numerically.

The underlying idea of both MMG and TNNMG is the following convergence result.

Lemma 3.4.1 ([42, Lem. 5.14])**.** *Let*

$$\mathcal{K}_{alg} = \prod_{0\le i<dn} [a_i, b_i] \qquad a_i \in \{-\infty\} \cup \mathbb{R}, \quad b_i \in \mathbb{R} \cup \{\infty\}, \tag{3.42}$$

$\mathscr{GS} : \mathcal{K}_{alg} \to \mathcal{K}_{alg}$ *be the projected Gauß–Seidel iteration operator and* $\mathscr{C} : \mathcal{K}_{alg} \to \mathcal{K}_{alg}$ *such that*

$$J(\mathscr{C}(v)) \le J(v) \tag{3.43}$$

for all $v \in \mathcal{K}_{alg}$*. Then the iteration*

$$u^{\nu+1} = \mathscr{C}(\mathscr{GS}(u^\nu))$$

is globally convergent.

In view of this lemma, the essence of designing a good multigrid method is to find a coarse grid correction operator $\mathscr{C}$ which yields 'good' corrections while retaining the monotonicity property (3.43). The monotone multigrid method, which is the spiritual father of TNNMG, restricts the corrections with certain defect obstacles in order to ensure admissibility of the iterates. Numerical experiments show that convergence can be much faster if a linear multigrid step without obstacles is used. Such a correction step, however, does not map $\mathcal{K}_{\text{alg}}$ onto $\mathcal{K}_{\text{alg}}$, and Lemma 3.4.1 does not apply. To combine both provable global convergence and fast effective convergence in a single method we propose to use the linear multigrid correction, but to project it on the defect obstacle and do a line search in this projected direction [43]. The resulting algorithm is globally convergent (Thm. 3.4.1), simple to implement, and shows very good convergence behavior (Sec. 3.8).

Applying TNNMG to multi-body contact problems involves a difficulty not present in the generic formulation of the algorithm. The set $\mathcal{K}_{\text{alg}}$, defined in (3.38), does not have the product structure (3.42). However, a special basis $\{\tilde{\boldsymbol{\psi}}\}$ of $\mathbf{V}_h$ can be chosen in which the algebraic admissible set $\widetilde{\mathcal{K}}_{\text{alg}}$ has the form (3.42). This trick was first used by Wohlmuth and Krause [95] for the monotone multigrid method.

Each iteration of TNNMG consists of four steps, which are now described in detail.

1. *Nonlinear presmoothing*
   Let O be a block-diagonal matrix with each diagonal block O_{pp} the Householder reflection which maps $\mathbf{e}_0$ onto $\hat{\boldsymbol{\nu}}_p$ if $p \in \mathcal{V}_{1,C}$ and $\mathrm{Id}_{d\times d}$ if $p \notin \mathcal{V}_{1,C}$. Note that $O_{pp}^{-1} = O_{pp}^T$. Order the nodal basis vectors $\{\boldsymbol{\psi}\}$ in a block vector

$$\{\boldsymbol{\psi}\} = \begin{pmatrix} \boldsymbol{\psi}_1^I & \boldsymbol{\psi}_1^C & \boldsymbol{\psi}_2^I & \boldsymbol{\psi}_2^C \end{pmatrix}^T$$

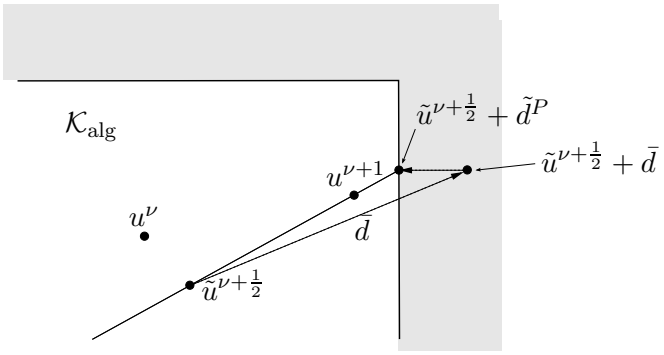


Figure 3.4: Guaranteeing monotonicity: the coarse grid correction $\bar{d}$ is projected onto $\mathcal{K}_{\text{alg}}$. Then a line search finds the minimum $u^{\nu+1}$ in the direction from $\tilde{u}^{\nu+\frac{1}{2}}$ to $\tilde{u}^{\nu+\frac{1}{2}} + \tilde{d}^P$ in $\mathcal{K}_{\text{alg}}$. Since $\tilde{J}(\tilde{u}^{\nu+\frac{1}{2}}) \leq J(u^\nu)$ by the monotonicity of the projected block Gauß–Seidel smoother and $J(u^{\nu+1}) \leq \tilde{J}(\tilde{u}^{\nu+\frac{1}{2}})$ by construction of $u^{\nu+1}$ the overall iteration step is monotone.

as for (3.37) and introduce the mortar transformation matrix

$$B = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & (\mathsf{D}^{-1}\mathsf{M})^T & 0 & I \end{pmatrix}.$$

Define the transformed basis

$$\{\tilde{\boldsymbol{\psi}}\} = OB\{\boldsymbol{\psi}\}. \tag{3.44}$$

Quantities with respect to the transformed basis $\{\tilde{\boldsymbol{\psi}}\}$ will be denoted by a tilde. In the basis (3.44), the minimization problem (3.41) takes the form

$$\tilde{J}(\tilde{u}) \leq \tilde{J}(\tilde{v}) \qquad \text{for all } \tilde{J}(\tilde{v}) \in \widetilde{\mathcal{K}}_{\text{alg}}, \tag{3.45}$$

where the new functional $\tilde{J}$ is defined by

$$\tilde{J}(\tilde{v}) = \frac{1}{2}\tilde{v}^T \widetilde{A}\tilde{v} - \tilde{b}\tilde{v}$$

with

$$\widetilde{A} = OBAB^TO^T \qquad \text{and} \qquad \tilde{b} = OBb.$$

The transformed admissible set reads

$$\widetilde{\mathcal{K}}_{\text{alg}} = \left\{ v \in \mathbb{R}^{dn} \;\middle|\; v_{p,0} \leq (\mathsf{D}^{-1}\mathsf{g})_p,\ \forall p \in \Gamma_{1,C} \right\}.$$

It has the product structure (3.42) and hence a projected Gauß–Seidel method for the transformed problem (3.45) converges [41]. Let $\tilde{u}^\nu$ be the current iterate in transformed coordinates. We do one or several projected Gauß–Seidel smoothing steps, and call the smoothed iterate $\tilde{u}^{\nu+\frac{1}{2}}$.

2. *Truncated defect problem*
We now consider the algebraic defect problem with respect to the smoothed iterate $\tilde{u}^{\nu+\frac{1}{2}}$, which is to find a $\tilde{d} \in \mathbb{R}^{dn}$ such that

$$\tilde{d}^T \widetilde{A}(\tilde{v} - \tilde{d}) \geq \tilde{b} - (\tilde{u}^{\nu+\frac{1}{2}})^T \widetilde{A}(\tilde{v} - \tilde{d}) \qquad \text{for all } \tilde{v} \in \widetilde{\mathcal{K}}_{\text{alg}}^{\nu+\frac{1}{2}},$$

with the defect obstacle

$$\widetilde{\mathcal{K}}_{\text{alg}}^{\nu+\frac{1}{2}} = \left\{ v \in \mathbb{R}^{dn} \;\middle|\; v_{p,0} \leq (\mathsf{D}^{-1}\mathsf{g})_p - (\tilde{u}^{\nu+\frac{1}{2}})_{p,0},\ \forall p \in \Gamma_{1,C} \right\}.$$

Define the current active set

$$\mathcal{V}^\bullet(\tilde{u}^{\nu+\frac{1}{2}}) = \left\{ p \in \mathcal{V}_{1,C} \;\middle|\; (\tilde{u}^{\nu+\frac{1}{2}})_{p,0} = (\mathsf{D}^{-1}\mathsf{g})_p \right\}.$$

We would like to construct a coarse grid correction $\tilde{d} \in \mathbb{R}^{dn}$ such that

$$\tilde{d}_{p,0} = 0 \qquad \text{for all } p \in \mathcal{V}^\bullet(\tilde{u}^{\nu+\frac{1}{2}}).$$

This can be achieved by truncation [56]. Define the truncation matrix $T^\nu \in \mathbb{R}^{dn \times dn}$ by

$$(T_{pq}^\nu)_{ij} = \begin{cases} 1 & p = q,\ i = j,\ p \notin \mathcal{V}^\bullet(\tilde{u}^{\nu+\frac{1}{2}}), \\ 0 & \text{else}, \end{cases}$$

and use it to set up the truncated defect problem in canonical coordinates

$$d^T \widehat{A}^\nu (v - d) \geq \hat{r}^\nu (v - d) \qquad \text{for all } v \in \mathcal{K}_{\text{alg}}^{\nu+\frac{1}{2}}, \tag{3.46}$$

with

$$\widehat{A}^\nu = (B^{-1}OT^\nu)\widetilde{A}(B^{-1}OT^\nu)^T \qquad \text{and} \qquad \hat{r}^\nu = B^{-1}OT^\nu(\tilde{b} - (\tilde{u}^{\nu+\frac{1}{2}})^T \widetilde{A}). \tag{3.47}$$

This can be computed efficiently since B^{-1} is available by the formula

$$B^{-1} = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & -(\mathsf{D}^{-1}\mathsf{M})^T & 0 & I \end{pmatrix}.$$

We do not bother to specify the appropriate obstacle set $\mathcal{K}_{\text{alg}}^{\nu+\frac{1}{2}}$ as it will not be used in the algorithm.

3. *Admissible coarse grid correction*
The next step is a linear multigrid step for the defect problem (3.46), restricted to the subspace

$$\mathcal{V}_T^\nu = \left\{ v \in \mathbb{R}^{dn} \;\middle|\; (OBv)_{p,0} = 0,\ \forall p \in \mathcal{V}^\bullet(\tilde{u}^{\nu+\frac{1}{2}}) \right\},$$

and disregarding the obstacles $\mathcal{K}_{\text{alg}}^{\nu+\frac{1}{2}}$. In other words, we consider the truncated linear defect problem

$$\widehat{A}^\nu d = \hat{r}^\nu. \tag{3.48}$$

The restriction to $\mathcal{V}_T^\nu$ ensures that (3.48) is uniquely solvable. In algorithmic terms the Gauß–Seidel update formula

$$d_i^{k+1} = \frac{1}{\widehat{A}_{ii}^\nu}\left(\hat{r}_i^\nu - \sum_{j<i}\widehat{A}_{ij}^\nu d_j^{k+1} - \sum_{j>i}\widehat{A}_{ij}^\nu d_j^k\right), \quad i = 1, 2, \ldots, dn,$$

is replaced by

$$d_i^{k+1} = \begin{cases} \frac{1}{\widehat{A}_{ii}^\nu}\left(\hat{r}_i^\nu - \sum_{j<i}\widehat{A}_{ij}^\nu d_j^{k+1} - \sum_{j>i}\widehat{A}_{ij}^\nu d_j^k\right), & \text{if } \widehat{A}_{ii}^\nu \neq 0, \\ 0 & \text{else.} \end{cases}$$

Let $\bar{d}$ be the resulting correction after one multigrid step in canonical coordinates and

$$\tilde{d} = T^\nu O^{-1} B^{-1} \bar{d}$$

the correction in transformed coordinates. Since $\tilde{d}$ may not be contained in the defect admissible set $\widetilde{\mathcal{K}}_{\text{alg}}^{\nu+\frac{1}{2}}$, we project it onto $\widetilde{\mathcal{K}}_{\text{alg}}^{\nu+\frac{1}{2}}$ in the l^2 (Euclidean) sense to obtain $\tilde{d}^P$.

4. *Line search*
The correction $\tilde{d}^P$ is admissible with respect to the defect obstacle $\widetilde{\mathcal{K}}_{\text{alg}}^{\nu+\frac{1}{2}}$, however, due to the projection step we may not have $\tilde{J}(\tilde{u}^{\nu+\frac{1}{2}} + \tilde{d}^P) \leq \tilde{J}(\tilde{u}^{\nu+\frac{1}{2}})$. To regain monotonicity of the method we perform a line search step within $\widetilde{\mathcal{K}}_{\text{alg}}^{\nu+\frac{1}{2}}$ in the direction of $\tilde{d}^P$ and define the overall new iterate as

$$\tilde{u}^{\nu+1} = \tilde{u}^{\nu+\frac{1}{2}} + \alpha \tilde{d}^P.$$

Since $\tilde{J}$ is quadratic, the optimal line search parameter α can be computed efficiently as

$$\alpha = \min\left\{\frac{(\tilde{b} - (\tilde{u}^{\nu+\frac{1}{2}})^T)\widetilde{A}\tilde{d}^P}{(\tilde{d}^P)^T\widetilde{A}\tilde{d}^P},\ \overline{\Psi}\right\},$$

with the line search obstacle

$$\overline{\Psi} = \min_{\tilde{d}_{p,0}^P > 0} \frac{(\mathsf{D}^{-1}\mathsf{g})_p - (\tilde{u}^{\nu+\frac{1}{2}})_{p,0}}{\tilde{d}_{p,0}^P}.$$

Note that, by construction, $\tilde{u}^{\nu+1} \in \widetilde{\mathcal{K}}_{\text{alg}}$ and we have an admissible next iterate. Also, $J(u^{\nu+1}) = \tilde{J}(\tilde{u}^{\nu+1}) \leq \tilde{J}(\tilde{u}^{\nu+\frac{1}{2}}) \leq \tilde{J}(\tilde{u}^\nu) = J(u^\nu)$ and hence the overall algorithm is monotone.

Using monotonicity and the property (3.42) of $\widetilde{\mathcal{K}}_{\text{alg}}$, global convergence follows from Lem. 3.4.1.

Theorem 3.4.1 ([42, Thm. 6.4])**.** *For any initial iterate $u^0 \in \mathcal{K}_{alg}$, the Truncated Nonsmooth Newton Multigrid algorithm converges to the unique minimum u of J in $\mathcal{K}_{alg}$.*

In the actual implementation the fine grid matrix uses up most of the memory. It is therefore inconvenient to store both the transformed matrix $\widetilde{A}$ and the truncated canonical matrix $\widehat{A}$ on the finest grid. It is more efficient to omit the linear fine grid smoothing and do the linear correction only on the next-coarser level. Truncation, the transformation (3.47), and the standard multigrid restriction operator R can be combined in a single special-purpose restriction operator

$$\tilde{R} = R(B^{-1})^T O^T T^\nu,$$

which is used for the transfer from the finest to the second-finest grid.

3.5 Implementing the Contact Mapping

A crucial component of the contact formulation is the homeomorphism

$$\Phi : \Gamma_{1,C} \to \Gamma_{2,C}$$

used in (3.24) to identify the nonmortar with the mortar contact boundary. For the implementation of the mortar-discretized contact problem we need to evaluate the entries

$$\mathsf{M}_{pq} = \mathrm{Id}_{d\times d} \int_{\Gamma_{1,C}} \theta_p(\psi_q \circ \Phi)\, ds, \qquad p \in \mathcal{V}_{1,C},\ q \in \mathcal{V}_{2,C},$$

of the mass matrix M (see Sec. 3.3). For this we need to be able to evaluate $\Phi(x) \in \Gamma_{2,C}$ for each $x \in \Gamma_{1,C}$. We restrict the exposition to simplicial grids in three space dimensions. The generalization to two-dimensional grids and grids involving quadrilateral faces is straightforward. Since Φ will be called many times during the assembly of M, its evaluation needs to be cheap. Also, we would not like the construction of the mapping to be of higher complexity than a multigrid cycle, which is in $O(|\mathcal{V}|)$ for first-order Lagrangian finite elements.

Our implementation of Φ has been described briefly in [60]. We will repeat it here in some more detail. For ease of notation in this section we will write Γ_1 and Γ_2 instead of $\Gamma_{1,C}$ and $\Gamma_{2,C}$. We begin with the presentation of a data structure that is suitable to hold piecewise affine homeomorphisms between triangulated surfaces in $\mathbb{R}^3$. Then, we will show how the projection of Γ_1 in normal direction onto Γ_2 can be implemented as an actual instance of a contact mapping.

As an introductory reminder, remember that a *graph* $\mathcal{G}$ is a finite set of vertices V together with a set E of unordered pairs of vertices which are called *edges*. Equip each vertex $v \in V$ with a position $p(v) \in \mathbb{R}^2$, and each edge $e = (v_0, v_1)$ with the line segment

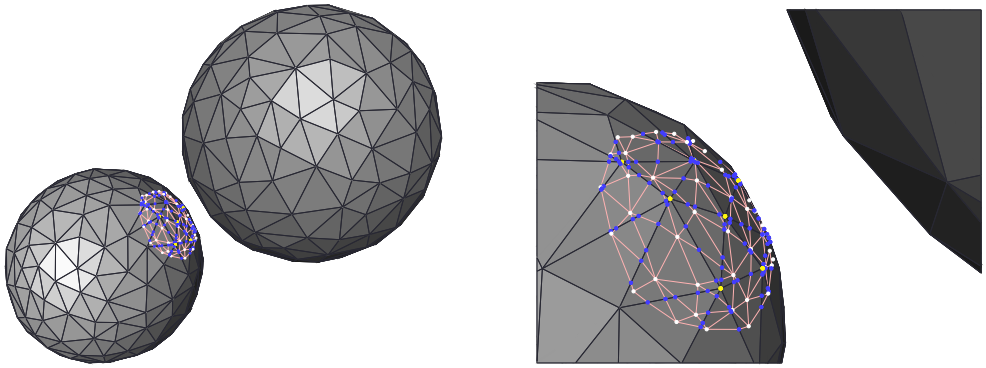


Figure 3.5: The edge graph of $\Gamma_{2,C}$ embedded in $\Gamma_{1,C}$.

from $p(v_0)$ to $p(v_1)$. If none of these segments intersect except at vertices then the graph together with the embedding into $\mathbb{R}^2$ is called a straight-line plane graph [32]. Note that each plane graph divides $\mathbb{R}^2$ into a set of *regions*. If each region except for the unbounded one is a triangle, $\mathcal{G}$ is called a *triangulation*. More generally we can define embeddings of graphs into general triangulated surfaces.

Definition 3.5.1. *Let $\mathcal{G} = (V, E)$ be a graph and $S \subset \mathbb{R}^3$ a triangulated surface. With each vertex $v \in V$ associate a position $p(v) \in S$, and with each edge $e = (v_0, v_1) \in E$ associate a set of open line segments $\eta_e = \{(e_0, \bar{e}_0), \ldots, (e_{n_e}, \bar{e}_{n_e})\}$. If*

- *$p(v_0) = e_0$, $p(v_1) = \bar{e}_{n_e}$, $\bar{e}_i = e_{i+1}$,*
- *for each $(e_i, \bar{e}_i)$ there exists a triangle T of S such that $(e_i, \bar{e}_i) \subset \overline{T}$,*
- *for any two segments $(e_i, \bar{e}_i)$ and $(e'_j, \bar{e}'_j)$ we have $(e_i, \bar{e}_i) \cap (e'_j, \bar{e}'_j) = \emptyset$,*

then (p, η) with $\eta = \{\eta_e \mid e \in E\}$ is called a piecewise straight embedding of $\mathcal{G}$ in S.

A graph embedded into a triangulated surface S subdivides S into regions. In general these regions do not coincide with the triangles of S.

Let G_1, G_2 be two simplicial grids that resolve Γ_1 and Γ_2, respectively. Hence Γ_1 and Γ_2 are triangulated surfaces and we assume that their embeddings into $\mathbb{R}^3$ are homeomorphic. This means that there exist functions $\Xi : \Gamma_1 \to \Gamma_2$ which are continuous and have a continuous inverse. We are interested in a data structure that can store piecewise linear homeomorphisms from Γ_1 to Γ_2.

Definition 3.5.2. *Let N, M be triangulated surfaces. A function $\Xi : N \to M$ is called piecewise linear if for each pair of triangles $T_N \in N$ and $T_M \in M$ the restriction of Ξ to $T_N \cap \Xi^{-1}(T_M)$ is an affine function.*

Denote by $\mathcal{V}_i$, $i \in \{1, 2\}$, the set of vertices of the triangulated surface Γ_i, by $\mathcal{E}_i$ its set of edges, and by $\mathcal{F}_i$ its set of triangles. The vertices $\mathcal{V}_2$ and the edges $\mathcal{E}_2$ form the *edge graph*

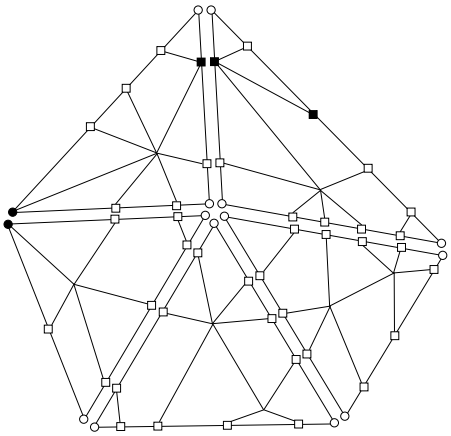


Figure 3.6: Implementation of a piecewise linear mapping Ξ as a graph on Γ_1. Graph node types are *corner* (●), *touching* (■), *intersection* (□), *ghost* (○), and *inner* (no symbol).

$\mathcal{G}_2$ of the mortar boundary. Let $\Xi : \Gamma_1 \to \Gamma_2$ be a piecewise linear homeomorphism. The preimage $\Xi^{-1}(\mathcal{G}_2)$ of $\mathcal{G}_2$ under Ξ is a graph embedded in Γ_1 with piecewise straight edges (Fig. 3.5). This graph embedding contains all relevant information about Ξ. Therefore, the mapping $\Xi : \Gamma_1 \to \Gamma_2$ can be stored using a data structure for graphs on Γ_1.

Remark 3.5.1. We actually allow the slightly more general case of functions Ξ whose domain of definition is a subset of Γ_1 (Fig. 3.5). See, however, Remark 3.5.3.

The foundation of the data structure for piecewise straight homeomorphisms $\Xi : \Gamma_1 \to \Gamma_2$ is a data structure for the domain surface Γ_1. We store a set of vertices $v_i \in \mathbb{R}^3$, $0 \leq i < |\mathcal{V}_1|$, and a set of triangles represented as triples (j_0, j_1, j_2) of vertex indices. Additionally, for each triangle $T \in \mathcal{F}_1$ we store a plane graph consisting of a set of nodes, each storing its local position on T, its position on Γ_2, and a list of all of its neighbors. This graph encapsulates the preimage of the edge graph of $\Xi(T)$.

We introduce five different types of graph nodes. There are (Fig. 3.6):

Inner nodes: These are nodes in the interior of a triangle T.

Corner nodes: These are nodes on the corners of T which are preimages of vertices of Γ_2. If $v \in \mathcal{V}_1$ is the vertex corresponding to corner node c then there is a copy of c on each triangle $T' \in \mathcal{F}_1$ with $v \in T'$.

Touching nodes: These are nodes p on an edge $k \subset \partial T$, but not on a corner of T. If there is a second triangle T' with $k \subset \partial T'$, this adjacent triangle T' also stores a copy of p.

Intersection nodes: For any edge $e = (v_0, v_1) \in \mathcal{E}_2$, the corresponding preimage $\Xi^{-1}(e)$ is generally not contained in a single triangle of Γ_1. Starting at $\Xi^{-1}(v_0) \in T_1$ and following $\Xi^{-1}(e)$ towards $\Xi^{-1}(v_1) \in T_2$, at some point one will leave T_1 and enter another triangle (not necessarily T_2). Thus the restriction of the edge graph of Γ_2 to T_1 is not a graph in its own right, because the line segment $\Xi^{-1}(e)|_{T_1}$ does not connect two graph nodes. To remedy this, an *intersection node* is inserted at the point where $\Xi^{-1}(e)$ leaves T_1, and a corresponding one on the boundary of the adjacent triangle where the path $\Xi^{-1}(e)$ continues.

Ghost nodes: If a corner c of a triangle $T \in \mathcal{F}_1$ does not get mapped onto a vertex of Γ_2, a *ghost node* without any neighbors is added at c.

Thus, only the first three types of nodes correspond to vertices in Γ_2. Each node p stores its image $\Xi(p) \in \Gamma_2$ by storing a triangle $T \in \mathcal{F}_2$ with $\Xi(p) \in T$ and local coordinates of $\Xi(p)$ with respect to T. Finally, each triangle $T \in \mathcal{F}_1$ keeps three arrays containing the nodes on the three triangle edges in cyclic order, and each node on an edge knows its index in the corresponding array. That way, corresponding nodes on adjacent triangles are identified, and it is possible to efficiently track preimages of edges of Γ_2 across multiple triangles of Γ_1.

Note that while globally $\Xi^{-1}(\mathcal{G}_2)$ is a triangulation of its domain of definition, its restrictions to individual triangles of $\mathcal{F}_1$ need not be (see Fig. 3.6). Therefore, before evaluating Ξ we apply the triangular closure to $\Xi^{-1}(\mathcal{G}_2)$, i.e., we add graph edges such that $\Xi^{-1}(\mathcal{G}_2)$ is a triangulation on each $T \in \mathcal{F}_1$.

Given a consistent data structure as described above and a point $p \in \Gamma_1$ specified by a triangle $T \in \mathcal{F}_1$ and local coordinates ξ_0, ξ_1 on T, the mapping Ξ can be evaluated at p in two steps. First, using a point-location algorithm, the region r (with corners $c(r)$) of the graph $\Xi^{-1}(\mathcal{G}_2)$ containing p is determined, and the three barycentric coordinates $\xi_{c(r)}$ of p with respect to r are computed. Barycentric coordinates are well-defined since we applied the triangular closure and hence all regions of $\Xi^{-1}(\mathcal{G}_2)|_T$ are triangles. For the point-location we use the randomized version of the algorithm presented by Brown and Faigle [22]. It is simple to implement and its expected run-time is in $O(\sqrt{|e|})$, with $|e|$ the number of edges in the triangulation.

Then, since Ξ is piecewise linear and all corners of r store their positions on Γ_2, the image of p under Ξ can be evaluated by interpolation

$$\Xi(p) = \sum_{c(r)} \xi_{c(r)} \Xi(c(r)).$$

We now turn to the question of how to construct the contact mapping for given Γ_1 and Γ_2. In Section 3.2 we have chosen Φ to be the projection of Γ_1 onto Γ_2 in the linearly interpolated normal direction $\hat{\boldsymbol{\nu}}$ (3.17) of Γ_1. The construction consists of three steps.

1. *Computing $\Phi^{-1}(q)$ for all $q \in \mathcal{V}_2$*
   For a given $q \in \mathcal{V}_2$ we have to find $\Phi^{-1}(q) \in \Gamma_1$ such that $q - \Phi^{-1}(q)$, as a vector in $\mathbb{R}^3$, is normal to Γ_1 in the sense of (3.17). Let T be a triangle of Γ_1 with the vertices p_0, p_1, p_2, and denote the respective vertex normal vectors by $\hat{\boldsymbol{\nu}}_0, \hat{\boldsymbol{\nu}}_1, \hat{\boldsymbol{\nu}}_2$. Assume there is a point $p \in T$ with $\Phi(p) = p + \mu\hat{\boldsymbol{\nu}}_p = q$ for some $\mu \geq 0$, and call λ_0, λ_1 the barycentric coordinates of p with respect to T,

   $$p(\lambda_0, \lambda_1) = \lambda_0 p_0 + \lambda_1 p_1 + (1 - \lambda_0 - \lambda_1) p_2. \tag{3.49}$$

   The normal $\hat{\boldsymbol{\nu}}$ at p is

   $$\hat{\boldsymbol{\nu}}(\lambda_0, \lambda_1) = \lambda_0 \hat{\boldsymbol{\nu}}_0 + \lambda_1 \hat{\boldsymbol{\nu}}_1 + (1 - \lambda_0 - \lambda_1) \hat{\boldsymbol{\nu}}_2. \tag{3.50}$$

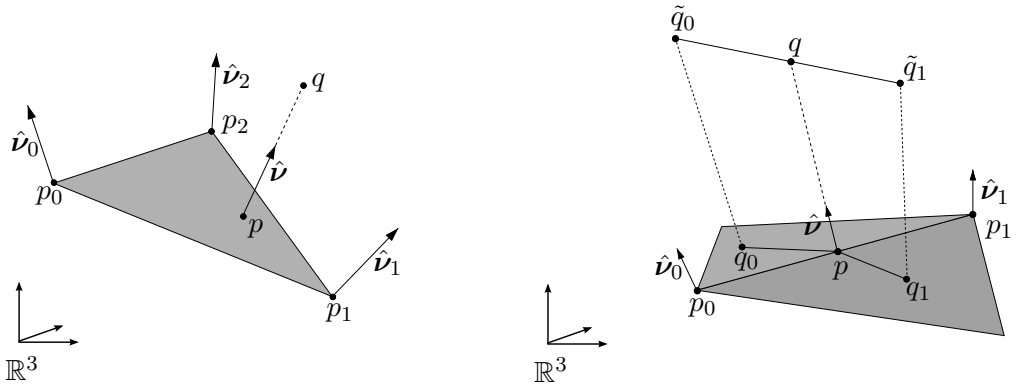


Figure 3.7: Backward normal projection from a point onto a triangle (left), and from an edge (q_0, q_1) onto an edge (p_0, p_1) (right).

Then $p = p(\lambda_0, \lambda_1)$ is a solution of

$$p(\lambda_0, \lambda_1) + \mu\hat{\boldsymbol{\nu}}(\lambda_0, \lambda_1) = q, \tag{3.51}$$

with $\lambda_0, \lambda_1, \mu \geq 0$ and $\lambda_0 + \lambda_1 \leq 1$. Inserting (3.49) and (3.50) yields the nonlinear system of equations

$$0 = p_2 - q + \lambda_0(p_0 - p_2) + \lambda_1(p_1 - p_2) + \mu\lambda_0(\hat{\boldsymbol{\nu}}_0 - \hat{\boldsymbol{\nu}}_2) + \mu\lambda_1(\hat{\boldsymbol{\nu}}_1 - \hat{\boldsymbol{\nu}}_2) + \mu\hat{\boldsymbol{\nu}}_2. \tag{3.52}$$

This system may in principle be solved analytically yielding at most two solutions. However, its apparent simplicity is deceptive, as the arithmetic expressions of the exact solutions get extremely long. It has proven easier to use a standard damped Newton algorithm [30] to solve (3.52) numerically.

If the projection $p = \Phi^{-1}(q)$ is found to be in the interior of T, an *inner node* is added to the data structure on T. If, on the other hand, it is found to be on an edge (corner) of T, then a *touching node* (*corner node*) is inserted along with corresponding copies on the other triangles sharing the edge (corner). If several distinct vertices are found which comply with the normality condition (3.51), the one with the shortest distance $|p - q|$ is chosen.

2. *Computing $\Phi(v)$ for all $v \in \mathcal{V}_1$*
   At this stage all vertices of Γ_2 appear as nodes in the graph on Γ_1. We have to add additional *ghost nodes* at the vertices of Γ_1 which are not mapped onto vertices of Γ_2. This is comparatively easy as it does not involve solving nonlinear equations. Given a vertex $v \in \mathcal{V}_1$, its exact position on Γ_2 can be found by considering the ray r in direction $\hat{\boldsymbol{\nu}}_v$ beginning in v. If r intersects more than one triangle of Γ_1 the intersection closest to v is chosen. At this point we assume that there is at least one such intersection, but see Remark 3.5.2. If v is mapped onto a vertex of Γ_2 then it has already been treated in Step 1 and the data structure already contains a node for it.

3. *Adding the edges*
   In order to enter an edge $\tilde{e} = (\tilde{q}_0, \tilde{q}_1)$ of Γ_2 into the graph on Γ_1 we try to 'walk' on Γ_1 along $\Phi^{-1}(\tilde{e})$ from $q_0 = \Phi^{-1}(\tilde{q}_0)$ to $q_1 = \Phi^{-1}(\tilde{q}_1)$. Since q_0 and q_1 will generally not be on the same triangle of Γ_1, we have to find the points where the path from q_0 to q_1 crosses edges of Γ_1. Let $T \in \mathcal{F}_1$ be the current triangle in this walking process. For an edge $e = (p_0, p_1)$ of T we have to check whether there are points $p \in e$ and $q \in \tilde{e}$ with $q - p$ normal (in the sense of (3.17)) to Γ_1 (Fig. 3.7, right). This can be formulated as a nonlinear system of equations

$$0 = p_0 - \tilde{q}_0 + \lambda(p_1 - p_0) - \mu(\tilde{q}_1 - \tilde{q}_0) + \eta\hat{\boldsymbol{\nu}}_0 + \eta\lambda(\hat{\boldsymbol{\nu}}_1 - \hat{\boldsymbol{\nu}}_0), \tag{3.53}$$

   which can be solved with a damped Newton algorithm. We have found an intersection if (3.53) has a solution with $0 \leq \lambda, \mu \leq 1$ and $0 \leq \eta$. This intersection is then inserted as an *intersection node* and the procedure is continued on the triangle which borders T on e.

Assuming that the Newton solvers for (3.52) and (3.53) terminate after a constant number of iterations and using an octree to speed up the search for ray–triangle intersections, the projection algorithm described above requires $O(|\mathcal{V}_C| \log |\mathcal{V}_C|)$ time. Here $\mathcal{V}_C = \mathcal{V}_{1,C} \cup \mathcal{V}_{2,C}$ are the vertices on the contact boundaries. Asymptotically, $|\mathcal{V}_C|$ behaves like $|\mathcal{V}|^{2/3}$. The construction of Φ therefore takes $O(|\mathcal{V}|^{2/3} \log |\mathcal{V}|^{2/3}) \subsetneq O(|\mathcal{V}|)$ time, which is less than a multigrid iteration.

Remark 3.5.2. In Sec. 3.2 it was assumed that Γ_1 and Γ_2 are such that the normal mapping Φ exists. In practice this is difficult to ensure. For example, it is easy to construct nonpathological cases with an edge $e = (p, q) \in \mathcal{E}_2$ such that both $\Phi^{-1}(p)$ and $\Phi^{-1}(q)$ are contained in Γ_1, but $\Phi^{-1}(e)$ is not. To handle this case it is necessary to check whether $\Phi^{-1}(e)$ is defined completely before inserting it. If an edge is left out the domain of definition of Φ is effectively reduced. Practically, this is not a problem as long as the reduced Γ_1 is still a superset of the true region of contact (cf. Sec. 3.2). The same remark applies, e.g., if, in Step 2, the ray casted from a vertex of Γ_1 does not hit any triangle of Γ_2.

Remark 3.5.3. Diagonality of the mass matrix D (3.39) only holds if the domain of definition of Φ is resolved by the grid. On the other hand, in (3.40) and (3.39) the same domain of integration needs to be used. Therefore, the domain of definition of Φ may need to be truncated further in order to be resolved by the grid.

Remark 3.5.4. The algorithm contains many decision procedures in two- and three-dimensional geometry. Rays need to be tested for an intersection with a given triangle in space, points need to be checked on which side of a line in the plane they are, and the like. The finite precision of normal computer arithmetic becomes a problem, because it leads to inconsistent test results. The implementor is strongly advised to avoid geometric testing as much as possible and rely on combinatorial information instead! For example, the point-location algorithm of Brown and Faigle expects the edges around a graph vertex v to be ordered cyclically. Instead of computing angles and sorting it is more stable to determine the cyclic ordering by using a graph algorithm to determine a longest path in the subgraph of all neighbors of v.

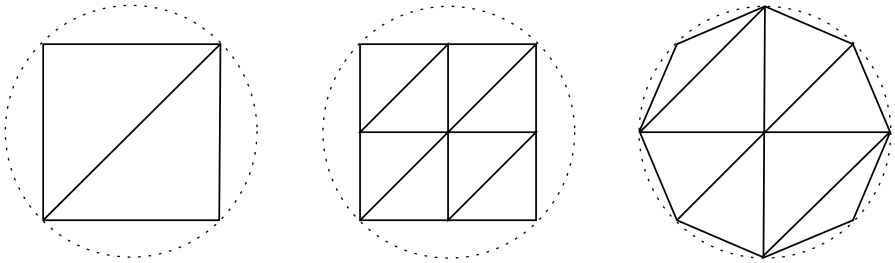


Figure 3.8: The advantage of boundary parametrizations. Left: Coarse grid with parametrization (dashed). Center: uniform refinement disregarding the boundary parametrization, the approximation of the boundary is a bad as for the coarse grid. Right: using the boundary parametrization. The geometric approximation is improved.

3.6 Creating and Using Parametrized Boundaries

Multigrid algorithms rely on a hierarchy of grids to achieve their fast convergence rates, and with the use of a posteriori error estimation techniques these hierarchies can be tailored to suite the problem at hand. Grids are usually refined by introducing new vertices at edge midpoints and suitably adding new elements [18]. For free-form geometries there is the problem, though, that grid refinement by adding edge midpoints does not improve the accuracy of the approximation of the true domain (Fig 3.8). This is unfortunate, since in biomechanical applications the problem domains have curved boundaries which cannot be approximated sufficiently by a coarse grid. Segmentation and surface extraction from CT or MRI data leads to bounding surfaces which are very highly resolved. In fact, in order to be able to construct reasonably coarse grids from them, they have to be simplified enormously. Simplification rates of 95% are not uncommon. The corresponding geometric information is lost in the process [79, Sec. 2.4].

To overcome this deficiency, various finite element software packages, for example UG [10], provide grids with parametrized boundaries. Let G_0 be a grid approximating a domain Ω in the sense that $\partial\Omega$ and ∂G_0 are homeomorphic and that $v \in \partial\Omega$ for all $v \in \mathcal{V}_0 \cap \partial G_0$, i.e., the vertices on the boundary of G_0 are contained in the boundary of Ω.

Definition 3.6.1 (Boundary parametrization)**.** *A boundary parametrization is a homeomorphism $\pi : \partial G_0 \to \partial\Omega$ such that $\pi(v) = v$ for all $v \in \mathcal{V}_0 \cap \partial G_0$.*

With a boundary parametrization at hand, a hierarchy of grids approximating Ω with increasing accuracy can be constructed as follows. When refining G_0 to obtain a new grid G_1, instead of inserting the edge midpoint $\bar{v} = (v_0 + v_1)/2$ when refining an edge $e = (v_0, v_1)$ on the boundary, the parametrization function π is used to obtain the corresponding position $\bar{v}_\pi = \pi((v_0 + v_1)/2) \in \partial\Omega$. Hence $\partial G_1 \neq \partial G_0$ in general, but we obtain a piecewise linear homeomorphism $h_{1\to 0} : \partial G_1 \to \partial G_0$. This can be extended to an arbitrary number of levels of refinement. Let G_j be the grid after j refinement steps.

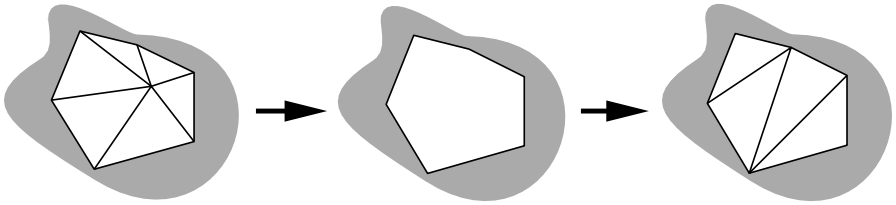


Figure 3.9: The surface is simplified by removing vertices and retriangulating the holes.

Then, when refining an edge $e = (v_0, v_1)$ of G_j the position $\bar{v}_\pi = \pi(h_{j\to 0}((v_0 + v_1)/2))$ can be chosen. We then set $h_{j+1\to 0} = h_{j\to 0} \circ h_{j+1\to j}$.

The grids in the resulting hierarchy are still logically nested, but they are not geometrically nested anymore. A multigrid convergence theory which covers this case can be found in [72]. There, the approximation property of the coarse grid spaces necessary for multigrid convergence is shown under reasonable assumptions on the parametrization function π.

For biomechanical problems, the construction of boundary parametrizations is a serious problem. Unlike in engineering applications, where CAD data provides analytical descriptions of curved boundaries, computational domains in biomechanics originate from computer tomograph scans. After segmentation and surface extraction, the description of the domain boundary is available only in form of a highly-resolved, but nevertheless piecewise linear surface [79].

In [61, 79] we have presented an algorithm for the automatic construction of boundary parametrizations from highly-resolved triangulated surfaces. It first constructs a parametrization of the high-resolution input surface over itself, and then successively coarsens the domain surface while maintaining a valid parametrization at each step. More formally, let S_T be the input surface extracted from CT data. We create a sequence of surfaces $S_0, \ldots, S_J$ and corresponding parametrization functions $\pi_i : S_i \to S_T$ such that $S_0 = S_T$ and S_J is a suitable boundary for a coarse grid.

1. Set $S_0 = S_T$, $\pi_0 = \mathrm{Id}$, and $i = 0$.
2. Choose a vertex v^* of S_i which is optimal in the sense of some suitable error criterion [79].
3. Remove all segments T_j that have v^* as a vertex, but do not discard the parametrizations $\pi_i|_{T_j}$ they carry.
4. Retriangulate the hole left by removing the T_j (Fig. 3.9). Restore the parametrization there and call the resulting surface S_{i+1}.
5. If S_{i+1} is coarse enough then terminate. If not increase i by one and go back to 2.

With a good criterion for the choice of vertices in 2. this algorithm delivers coarse surfaces that approximate the actual domain Ω well. These can be used as input for a

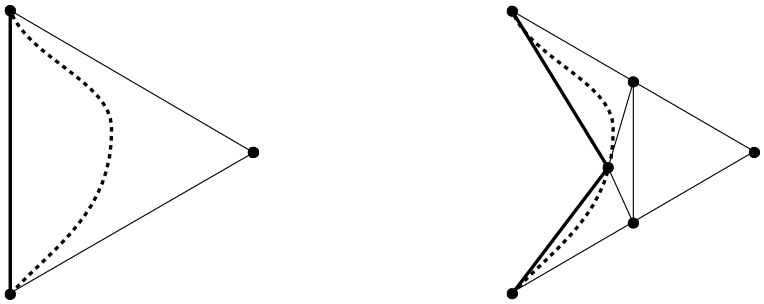


Figure 3.10: The use of boundary parametrizations can lead to severe degradation in mesh quality. Left: unrefined. Right: refined; the dashed line is the true domain boundary $\partial\Omega$.

grid generator to produce a coarse finite element grid. The final parametrization π_J needs to undergo a smoothing postprocessing before it can be used. A detailed description of the algorithm can be found in [61, 79].

Remark 3.6.1. An important ingredient of this algorithm is the data structure that stores a mapping from one triangulated surface to another. Such a data structure has already been used for contact mappings in Sec. 3.5. In fact, the same implementation is used for both applications.

It should not be concealed that the use of parametrized boundaries can lead to severe difficulties. With standard red–green grid refinement for tetrahedral grids, the aspect ratio of the elements can be bounded from above by a constant which is independent of the grid level [18]. This is not the case when a parametrization function π is used. From Fig. 3.10 it is clear that the vertex position change induced by a boundary parametrization can lead to elements of arbitrarily bad aspect ratio. Even elements with negative orientation are possible, in particular near concavities of the domain boundary. However, both the standard discretization error estimates for finite element spaces and the convergence rate of multigrid methods depend on shape regularity of the grid. Therefore grid improvement techniques may need to be used. This is a topic of great practical relevance and much research effort has been put into it. Since geometric multigrid methods rely on a hierarchy of logically nested grids, we can only consider mesh improvement algorithms which preserve the discrete topology. Grid smoothing algorithms that preserve the discrete topology usually work by solving either a sequence of local nonlinear problems [38, 53], or a single global one [78]. In both cases, the aim is to improve the quality of the grid by moving interior vertices. However, while this improves the grid quality on each individual grid level it destroys the geometric nestedness of the grids on the interior of the domain. This thwarts the improvement of the convergence rates due to the increased quality of each individual level grid. Some form of grid improvement strategy should nevertheless be applied at least on the finest grid to improve the finite element discretization error.

3.7 Hierarchical A Posteriori Error Estimation

Frequently, the solutions of contact problems in biomechanics show very localized features. This justifies the use of adaptive grid refinement. However, while a large body of theory exists for a posteriori error estimation of unconstrained problems, few people have considered obstacle problems. Wohlmuth [94] introduced an estimator for the mortar-discretized contact problem and gave references to some earlier work. Veeser [87] presented a residual-type estimator for scalar obstacle problems which additionally provides an estimate of how well the active set is approximated. Kornhuber [55] extended the hierarchical approach to elliptic problems with pointwise nonlinearities. Hierarchical error estimators have the advantage of being parameter-free. We now generalize the approach of Kornhuber to two-body contact problems.

We first present the basic idea in a scalar setting. Let $a(\cdot,\cdot)$ be a continuous, symmetric, and $H_0^1(\Omega)$-elliptic bilinear form and l a linear functional. The symbol $\|\cdot\|$ will denote the energy norm $\|v\|^2 = a(v,v)$. With $\mathcal{S} \subsetneq \mathcal{Q} \subset H_0^1(\Omega)$ two nested discrete spaces and $u_\mathcal{S}$ and $u_\mathcal{Q}$ the solutions of the elliptic problems

$$u_\mathcal{S} \in \mathcal{S} \quad : \quad a(u_\mathcal{S}, v) = l(v) \qquad \forall v \in \mathcal{S}, \tag{3.54}$$

and

$$u_\mathcal{Q} \in \mathcal{Q} \quad : \quad a(u_\mathcal{Q}, v) = l(v) \qquad \forall v \in \mathcal{Q},$$

respectively, the difference $\|u_\mathcal{S} - u_\mathcal{Q}\| = \|e_\mathcal{Q}\|$ between the two solutions will be shown to be an estimate of the true error $\|u_\mathcal{S} - u\|$. The underlying idea is the assumption that the discretization error decreases if the solution is searched for in a larger space.

Assumption 3.7.1 (Saturation assumption)**.** *Call u, $u_\mathcal{S}$, and $u_\mathcal{Q}$ the solutions of the variational problem $a(\cdot, v) = l(v)$, for all v, in the spaces $H_0^1(\Omega)$, $\mathcal{S}$, and $\mathcal{Q}$, respectively. Then there exists a $\beta < 1$ such that*

$$\|u - u_\mathcal{Q}\| \leq \beta \|u - u_\mathcal{S}\|. \tag{3.55}$$

This assumption is justified if the problem is sufficiently regular and the space $\mathcal{Q}$ is sufficiently large. On the other hand, it is shown in [18] that for any two spaces $\mathcal{S}$ and $\mathcal{Q}$ there exist functionals l such that the saturation assumption does not hold.

The error $e = u - u_\mathcal{S}$ is solves the defect equation

$$e \in H_0^1 \quad : \quad a(e, v) = l(v) - a(u_\mathcal{S}, v) \qquad \forall v \in H_0^1(\Omega).$$

Discretizing this equation in the space $\mathcal{Q}$ we get

$$e_\mathcal{Q} \in \mathcal{Q} \quad : \quad a(e_\mathcal{Q}, v) = l(v) - a(u_\mathcal{S}, v) \qquad \forall v \in \mathcal{Q}, \tag{3.56}$$

and the estimates [19, Prop. 2.1 and Thm. 2.1]

$$\sqrt{(1-\beta^2)}\|u - u_\mathcal{S}\| \leq \|e_\mathcal{Q}\| \leq \|u - u_\mathcal{S}\|. \tag{3.57}$$

Remark 3.7.1. In practice, iterative methods are commonly used to solve (3.54). They yield an approximate solution $\tilde{u}$ which contains an algebraic error $\|u_{\mathcal{S}} - \tilde{u}\|$ along with the discretization error $\|u - u_{\mathcal{S}}\|$. We will disregard this algebraic error for the sake of simplicity.

Since solving (3.56) is expensive, $a(\cdot,\cdot)$ is replaced by a different bilinear form $b(\cdot,\cdot)$ such that

$$e_b \in \mathcal{Q} \quad : \quad b(e_b, v) = l(v) - a(u_{\mathcal{S}}, v) \qquad \forall v \in \mathcal{Q} \tag{3.58}$$

is cheaper to solve than (3.56), and the new estimate e_b is equivalent to $e_{\mathcal{Q}}$.

Lemma 3.7.1. *Let $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ be symmetric, positive definite bilinear forms, and let c_0 and c_1 be positive constants such that*

$$c_0 \leq \frac{b(w,w)}{a(w,w)} \leq c_1 \tag{3.59}$$

for all $w \in \mathcal{Q}$, $w \neq 0$. Let $e_{\mathcal{Q}}$ and e_b be defined by (3.56) *and* (3.58)*, respectively. Then*

$$c_0\|e_b\| \leq \|e_{\mathcal{Q}}\| \leq c_1\|e_b\|.$$

Proof. Cf. Bank and Smith [9], Thm. 2.2. □

We now generalize this approach to two-body contact problems. We closely follow [55], where variational inequalities with pointwise nonlinearities were considered. Let Ω_i, $i \in \{1,2\}$, be polygonal domains and let G_i, $i \in \{1,2\}$, be simplicial grids for the Ω_i. For $\mathcal{S}$ we use the space $\mathbf{V}^1_{h,0}(G) = \prod_i \mathbf{V}^1_{h,0}(G_i)$ of first-order Lagrangian finite elements, where the superscript index denotes the order of the finite elements. We set $\mathcal{Q} = \mathbf{V}^2_{h,0}(G)$, the space of continuous second-order Lagrangian finite elements. This is the canonical choice; Bornemann et al. [19] discuss various alternatives.

Remember from Sec. 3.2 that the continuous linear two-body contact problem can be written as the variational inequality

$$\mathbf{u} \in \mathcal{K} \quad : \quad a(\mathbf{u}, \mathbf{v} - \mathbf{u}) \geq l(\mathbf{v} - \mathbf{u}) \qquad \forall \mathbf{v} \in \mathcal{K}, \tag{3.60}$$

with $a(\cdot,\cdot)$ as defined in (3.12) a continuous, symmetric, and $\mathbf{H}^1_0(\Omega)$-elliptic bilinear form and the linear functional l given by (3.13). The admissible set

$$\mathcal{K} = \left\{ \mathbf{v} \in \prod_i \mathbf{H}^1_0(\Omega_i) \;\middle|\; \langle [\mathbf{v}]_\Phi, \boldsymbol{\nu}_1 \rangle \leq g \right\}$$

is closed and convex. Discretizing (3.60) using $\mathbf{V}^1_h(G)$ we obtained the finite-dimensional problem (3.36)

$$\mathbf{u}_{\mathcal{S}} \in \mathcal{K}_h \quad : \quad a(\mathbf{u}_{\mathcal{S}}, \mathbf{v}_h - \mathbf{u}_{\mathcal{S}}) \geq l(\mathbf{v}_h - \mathbf{u}_{\mathcal{S}}) \qquad \forall \mathbf{v}_h \in \mathcal{K}_h, \tag{3.61}$$

with $\mathcal{K}_h$ given by (3.34). The error $\mathbf{e} = \mathbf{u} - \mathbf{u}_{\mathcal{S}}$ is the unique solution of the defect problem

$$\mathbf{e} \in \widetilde{\mathcal{K}} \quad : \quad a(\mathbf{e}, \mathbf{v} - \mathbf{e}) \geq r(\mathbf{v} - \mathbf{e}) \qquad \forall \mathbf{v} \in \widetilde{\mathcal{K}}, \tag{3.62}$$

where we have used the residual

$$\begin{aligned} r &: \mathbf{H}^1(\Omega) \to \mathbb{R} \\ r(\mathbf{v}) &= l(\mathbf{v}) - a(\mathbf{u}_{\mathcal{S}}, \mathbf{v}) \end{aligned}$$

and the continuous defect obstacle

$$\widetilde{\mathcal{K}} = \Big\{ \mathbf{v} \in \prod_i \mathbf{H}_0^1(\Omega_i) \;\Big|\; \mu(\langle [\mathbf{v}]_\Phi, \boldsymbol{\nu}_1 \rangle) \le \mu(g - \langle [\mathbf{u}_{\mathcal{S}}]_\Phi, \boldsymbol{\nu}_1 \rangle),\ \forall \mu \in M^+ \Big\}.$$

Define $\mathbf{e}_{\mathcal{Q}}$ as the unique solution of the discrete defect problem

$$\mathbf{e}_{\mathcal{Q}} \in \widetilde{\mathcal{K}}_{\mathcal{Q}} \quad : \quad a(\mathbf{e}_{\mathcal{Q}}, \mathbf{v} - \mathbf{e}_{\mathcal{Q}}) \ge r(\mathbf{v} - \mathbf{e}_{\mathcal{Q}}) \qquad \forall \mathbf{v} \in \widetilde{\mathcal{K}}_{\mathcal{Q}}, \tag{3.63}$$

with the discrete quadratic defect obstacle

$$\widetilde{\mathcal{K}}_{\mathcal{Q}} = \Big\{ \mathbf{v} \in \mathbf{V}_{h,0}^2(G) \;\Big|\; \int_{\Gamma_{1,C}} \langle [\mathbf{v}]_\Phi, \boldsymbol{\nu}_1 \rangle \mu_h \, ds \le \int_{\Gamma_{1,C}} (g - \langle [\mathbf{u}_{\mathcal{S}}]_\Phi, \boldsymbol{\nu}_1 \rangle) \mu_h \, ds \quad \forall \mu_h \in M_{h,\mathcal{Q}}^+ \Big\}. \tag{3.64}$$

The mortar space $M_{h,\mathcal{Q}}^+$ is the cone of positive second-order scalar finite element functions on $\Gamma_{1,C}$. Using the triangle inequality and the saturation assumption (3.55) one can show that $\mathbf{e}_{\mathcal{Q}}$ does provide an estimate for the error.

Theorem 3.7.1. *Assume that $\mathbf{u}_{\mathcal{Q}}$ provides a better approximation than $\mathbf{u}_{\mathcal{S}}$ in the sense of the saturation assumption 3.7.1. Then we have the estimates*

$$(1 - \beta)\|\mathbf{u} - \mathbf{u}_{\mathcal{S}}\| \le \|\mathbf{e}_{\mathcal{Q}}\| \le (1 + \beta)\|\mathbf{u} - \mathbf{u}_{\mathcal{S}}\|. \tag{3.65}$$

Note that this is a slightly weaker result than (3.57).

In analogy to the unconstrained case we now replace (3.63) by a similar problem which is easier to solve while still yielding satisfactory error estimates. Let $b(\cdot, \cdot)$ be a symmetric bilinear form. We set $\|\mathbf{v}\|_b^2 = b(\mathbf{v}, \mathbf{v})$ the energy norm of $b(\cdot, \cdot)$, and define $\mathbf{e}_b$ as the solution of the preconditioned defect problem

$$\mathbf{e}_b \in \widetilde{\mathcal{K}}_{\mathcal{Q}} \quad : \quad b(\mathbf{e}_b, \mathbf{v} - \mathbf{e}_b) \ge r(\mathbf{v} - \mathbf{e}_b) \qquad \forall \mathbf{v} \in \widetilde{\mathcal{K}}_{\mathcal{Q}}, \tag{3.66}$$

with the quadratic discrete defect obstacle $\widetilde{\mathcal{K}}_{\mathcal{Q}}$ given by (3.64). With a suitable $b(\cdot, \cdot)$ this is an equivalent error estimator. The proof in [55, Thm. 4.1] for pointwise obstacles applies in our case as well.

Theorem 3.7.2. *Assume that the norm equivalence*

$$\gamma_0 b(\mathbf{v}, \mathbf{v}) \le a(\mathbf{v}, \mathbf{v}) \le \gamma_1 b(\mathbf{v}, \mathbf{v}), \qquad \mathbf{v} \in \operatorname{span}\{\mathbf{e}_{\mathcal{Q}}, \mathbf{e}_b\}, \tag{3.67}$$

holds with positive constants γ_0, γ_1. Then we have the estimates

$$c_0 \|\mathbf{e}_b\|_b^2 \le \|\mathbf{e}_{\mathcal{Q}}\|^2 \le c_1 \|\mathbf{e}_b\|_b^2 \tag{3.68}$$

with $c_0 = (\gamma_0^{-1} + 2\gamma_1(1 + \gamma_0^{-1}))^{-1}$ and $c_1 = \gamma_1 + 2\gamma_0^{-1}(1 + \gamma_1)$.

Proof. By symmetry arguments, it is sufficient to establish only the right inequality in (3.68). Inserting $\mathbf{v} = \mathbf{e}_b \in \widetilde{\mathcal{K}}_{\mathcal{Q}}$ in the original discrete defect problem (3.63), we obtain

$$\|\mathbf{e}_{\mathcal{Q}}\|^2 \leq a(\mathbf{e}_{\mathcal{Q}}, \mathbf{e}_b) + r(\mathbf{e}_{\mathcal{Q}} - \mathbf{e}_b).$$

This can be bounded using the inequality $2a(\mathbf{e}_{\mathcal{Q}}, \mathbf{e}_b) \leq \|\mathbf{e}_{\mathcal{Q}}\|^2 + \|\mathbf{e}_b\|^2$ and (3.67). We get

$$\|\mathbf{e}_{\mathcal{Q}}\|^2 \leq \gamma_1 \|\mathbf{e}_b\|_b^2 + 2r(\mathbf{e}_{\mathcal{Q}} - \mathbf{e}_b).$$

Inserting $\mathbf{v} = \mathbf{e}_{\mathcal{Q}}$ in (3.66) and using the Cauchy-Schwarz inequality, we get

$$\|\mathbf{e}_{\mathcal{Q}}\|^2 \leq \gamma_1 \|\mathbf{e}_b\|_b^2 + 2\|\mathbf{e}_b\|_b \|\mathbf{e}_{\mathcal{Q}} - \mathbf{e}_b\|_b,$$

so that it remains to show that

$$\|\mathbf{e}_{\mathcal{Q}} - \mathbf{e}_b\|_b \leq \gamma_0^{-1}(1 + \gamma_1)\|\mathbf{e}_b\|_b. \tag{3.69}$$

Insert $\mathbf{v} = \mathbf{e}_b$ in (3.63) and $\mathbf{v} = \mathbf{e}_{\mathcal{Q}}$ in (3.66). Add the two inequalities to obtain

$$a(\mathbf{e}_{\mathcal{Q}}, \mathbf{e}_b - \mathbf{e}_{\mathcal{Q}}) + b(\mathbf{e}_b, \mathbf{e}_{\mathcal{Q}} - \mathbf{e}_b) \geq 0,$$

which can be written as

$$\|\mathbf{e}_b - \mathbf{e}_{\mathcal{Q}}\|^2 \leq a(\mathbf{e}_b, \mathbf{e}_b - \mathbf{e}_{\mathcal{Q}}) - b(\mathbf{e}_b, \mathbf{e}_b - \mathbf{e}_{\mathcal{Q}}).$$

Eq. (3.69) then follows from the Cauchy-Schwarz inequality and (3.67). □

Remark 3.7.2. For the estimates (3.68) to be truly useful the constants γ_0 and γ_1 in (3.67) need to be independent of the grid size for all test functions $\mathbf{v} \in \operatorname{span}\{\mathbf{e}_{\mathcal{Q}}, \mathbf{e}_b\}$. For the unconstrained case this independence has been shown in [55, Prop. 4.1]. For problems with obstacles it is an open question.

Generalizing [55] we define the bilinear form

$$b(\mathbf{v}, \mathbf{w}) = \sum_{\substack{p \in \mathcal{V}^{\mathcal{Q}} \\ 0 \leq i,j < d}} v_i(p) w_j(p) a(\boldsymbol{\psi}_{p,i}^{\mathcal{Q}}, \boldsymbol{\psi}_{p,j}^{\mathcal{Q}}),$$

where $\boldsymbol{\psi}_{p,i}^{\mathcal{Q}} \in \mathbf{V}_h^2(G)$ denotes the vector-valued second-order nodal basis function at p in the i-th coordinate direction and $v_i(p)$ denotes the value of the i-th component of $\mathbf{v}$ at p. Note that the matrix corresponding to $b(\cdot, \cdot)$ is block-diagonal. However, by the constraint $\mathbf{e}_b \in \widetilde{\mathcal{K}}_{\mathcal{Q}}$ there is still coupling between degrees of freedom on $\Gamma_{1,C}$ and $\Gamma_{2,C}$. Set $n_{i,C}^{\mathcal{Q}} = |\mathcal{V}_{i,C}^{\mathcal{Q}}|$, $i \in \{1, 2\}$, and for a coefficient vector $v \in R^{d|\mathcal{V}^{\mathcal{Q}}|}$ let $v_1^C \in \mathbb{R}^{n_{1,C}^{\mathcal{Q}}}$ and $v_2^C \in \mathbb{R}^{n_{2,C}^{\mathcal{Q}}}$ be the vectors of coefficients corresponding to Lagrange points on $\Gamma_{1,C}$ and $\Gamma_{2,C}$, respectively. The set $\widetilde{\mathcal{K}}_{\mathcal{Q}}$ of admissible second-order defect corrections takes the algebraic form

$$\widetilde{\mathcal{K}}_{\mathcal{Q}}^{\mathrm{alg}} = \left\{ v \in \mathbb{R}^{d|\mathcal{V}^{\mathcal{Q}}|} \;\middle|\; \mathsf{M}_{\mathcal{Q}}^{\mathrm{n}} v_1^C - \mathsf{M}_{\mathcal{Q}}^{\mathrm{m}} v_2^C \leq \mathsf{g}^{\mathcal{Q}} \right\}.$$

The mass matrices $\mathsf{M}^{\mathrm{n}}_{\mathcal{Q}} \in \mathbb{R}^{n^{\mathcal{Q}}_{1,C} \times d n^{\mathcal{Q}}_{1,C}}$ and $\mathsf{M}^{\mathrm{m}}_{\mathcal{Q}} \in \mathbb{R}^{n^{\mathcal{Q}}_{1,C} \times d n^{\mathcal{Q}}_{2,C}}$ have a $1 \times d$ block structure and the entries

$$(\mathsf{M}^{\mathrm{n}}_{\mathcal{Q}})^i_{pq} = \int_{\Gamma_{1,C}} \psi^{\mathcal{Q}}_p \langle \boldsymbol{\psi}^{\mathcal{Q}}_{q,i}, \boldsymbol{\nu}_1 \rangle \, ds \qquad p, q \in \mathcal{V}^{\mathcal{Q}}_{1,C},$$

$$(\mathsf{M}^{\mathrm{m}}_{\mathcal{Q}})^i_{pq'} = \int_{\Gamma_{1,C}} \psi^{\mathcal{Q}}_p \langle \boldsymbol{\psi}^{\mathcal{Q}}_{q',i} \circ \Phi, \boldsymbol{\nu}_1 \rangle \, ds \qquad p \in \mathcal{V}^{\mathcal{Q}}_{1,C},\ q' \in \mathcal{V}^{\mathcal{Q}}_{2,C}.$$

The coefficients of the discrete second-order weak obstacle $\mathsf{g}^{\mathcal{Q}}_p \in \mathbb{R}^{n^{\mathcal{Q}}_{1,C}}$ are

$$\mathsf{g}^{\mathcal{Q}}_p = \int_{\Gamma_{1,C}} \psi^{\mathcal{Q}}_p (g - \langle [\mathbf{u}_{\mathcal{S}}]_\Phi, \boldsymbol{\nu}_1 \rangle) \, ds \qquad p \in \mathcal{V}^{\mathcal{Q}}_{1,C}.$$

A biorthogonality condition equivalent to (3.33) does not hold since we have chosen the second-order nodal basis $\{\psi^{\mathcal{Q}}\}$ for the discretization of the mortar space $M^+_{h,\mathcal{Q}}$. Nevertheless, an interior-point solver like IPOpt [90] can solve systems like (3.66) with good convergence and wall-time behavior. Exemplary iterations numbers are given below.

So far we have described global properties of the error estimator $\mathbf{e}_b$. As shown in [55], it is also suitable as a local refinement indicator. Assume that there is a hierarchical splitting

$$\mathcal{Q} = \mathcal{S} \oplus \mathcal{W},$$

and decompose a given $\mathbf{e}_b \in \widetilde{\mathcal{K}}_{\mathcal{Q}}$ by setting

$$\mathbf{e}_b = \mathbf{e}^{\mathcal{S}}_b \oplus \mathbf{e}^{\mathcal{W}}_b.$$

The local contributions

$$\eta_p = \sum_{i,j=0}^{d-1} \big(\mathbf{e}^{\mathcal{W}}_b(p)\big)_i \big(\mathbf{e}^{\mathcal{W}}_b(p)\big)_j a(\boldsymbol{\psi}^{\mathcal{W}}_{p,i}, \boldsymbol{\psi}^{\mathcal{W}}_{p,j}), \qquad p \in \mathcal{V}^{\mathcal{W}},$$

can be used as local indicators. On a simplicial grid the set $\mathcal{V}^{\mathcal{W}}$ corresponds to the edge midpoints. If η_p exceeds a certain limit for a given $p \in \mathcal{V}^{\mathcal{W}}$ all elements containing p are marked for refinement.

We demonstrate the applicability of the proposed error estimator with a numerical example. Consider a 2d Hertzian contact problem with a coarse grid as depicted in Fig. 3.11. The boundary of the upper grid carries a parametrization function which lets the domain approach a half-sphere with increasing refinement. The block is clamped at the bottom while the half-sphere receives downward displacement conditions of 2.5 length units on the horizontal diameter. Both objects are modelled with a St. Venant–Kirchhoff material with $E = 17 \cdot 10^6$ pressure units and $\nu = 0.3$.

In order to validate the output of the error estimator we compute a reference solution $\mathbf{u}^*$ by numerically solving the problem on a grid G^* which has been created with ten steps of uniform refinement. For the actual measurements we restart from the coarse grid. On each level we solve the contact problem, estimate the error, and refine the 30% of the

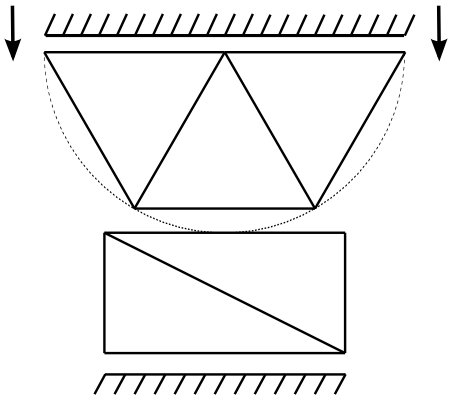


Figure 3.11: Coarse grid with symbolized boundary conditions. The upper body's boundary is parametrized to approximate a half sphere (dashed line).

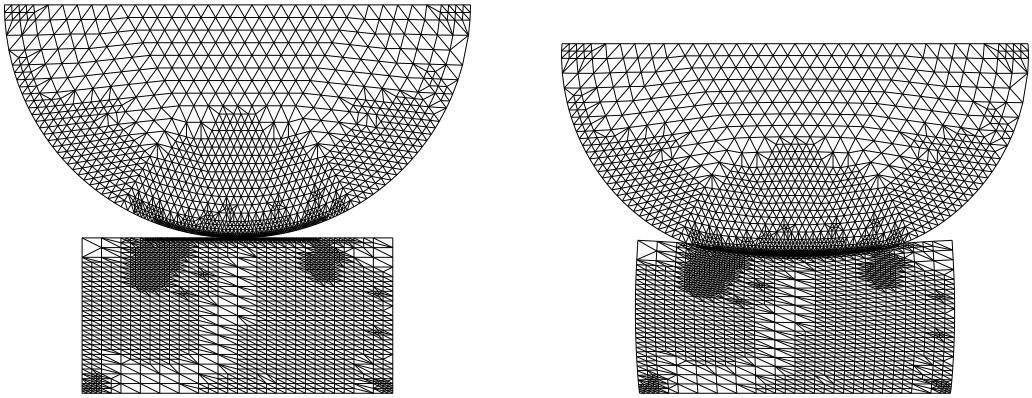


Figure 3.12: Resulting grids after ten refinement steps. Left: reference configuration. Right: deformed configuration.

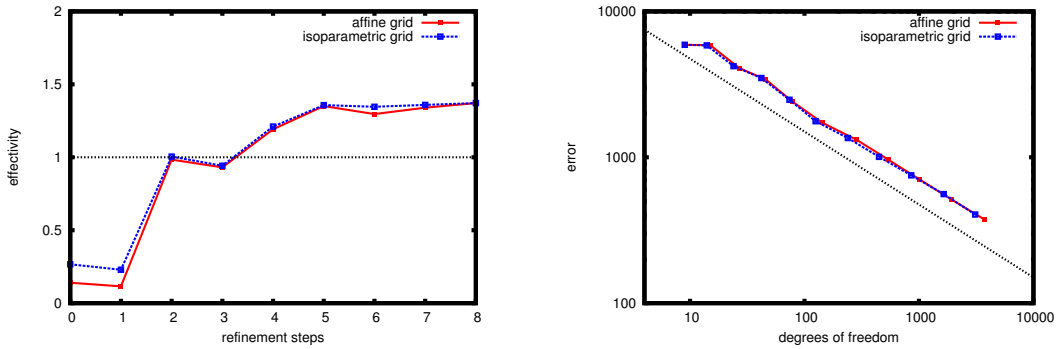


Figure 3.13: Left: the ratio of true error to estimated error. Right: true error per number of grid nodes. The dotted line illustrates the slope of $-1/2$ which is optimal for a 2d problem.

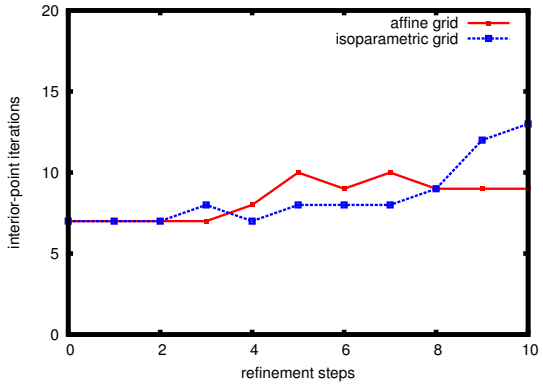


Figure 3.14: Number of iterations of the interior-point solver used to solve (3.66) on each refinement step.

elements with the highest local error. Let $\mathbf{u}_i$ be the solution after i refinement steps and let $e_i = \|\mathbf{e}_b\|_b$ be its estimated error. The true error with respect to the reference solution $\mathbf{u}^*$ is given by $\|\mathbf{u}^* - \mathbf{u}_i \circ h_i^{-1}\|$, where h_i is the canonical homeomorphism from the domain covered by G_i to the domain covered by G^*, and $\|\cdot\|$ is the energy norm of the bilinear form $a(\cdot,\cdot)$ on G^*. Fig. 3.13, left, shows the effectivity of the hierarchical error estimator, i.e., $\|\mathbf{e}_b\|_b / \|\mathbf{u}^* - \mathbf{u}_i \circ h_i^{-1}\|$ for each refinement level. The effectivity rate stays close to 1, and it does not exceed 1.5 anytime during the measurement.

Remark 3.7.3. When boundary parametrizations are used to approximate domains which are not resolved by the grid G_i it seems natural to improve the approximation quality of the quadratic space $\mathcal{Q}$ by using an isoparametric grid for its definition. We have found that in our examples the use of isoparametric grids did not lead to appreciable improvements of the estimation quality.

Fig. 3.13, right, shows the behavior of the true error $\|\mathbf{u}^* - \mathbf{u}_i \circ h_i^{-1}\|$ with increasing number of degrees of freedom for both types of approximating spaces. On the doubly logarithmic plot both graphs show a linear behavior, with a slope of $-1/2$ as is optimal for a 2d first-order element space.

In order to be practically usable the numerical solution of (3.66) must remain cheap even for increasing mesh size. We used the interior-point solver IPOpt [90], which was set up to iterate until the optimality error (a certain scaled residual, see [90, Sec. 2.1]) dropped below 10^{-8}. Fig. 3.14 shows the number of iterations needed to reach this precision. For this problem, the number appears to be independent of the mesh size. The wall-time needed for one iteration was comparable to one multigrid iteration for the contact problem. For three-dimensional problems solving (3.66) is much cheaper than solving the contact problem.

Remark 3.7.4. We have implicitly assumed that the grid hierarchies for the two domains Ω_1 and Ω_2 consist of the same number of levels. The finite elements spaces $\mathbf{V}_h(G_1)$ and $\mathbf{V}_h(G_2)$ have been treated as a single product space $\mathbf{V}_h(G_1 \cup G_2)$ for which the nonlocal basis (3.44) is then chosen, and a multigrid hierarchy of coarse spaces for the product $\mathbf{V}_h(G_1 \cup G_2)$ has been used. If there is an unequal number of grid levels due to adaptive refinement, the shorter hierarchy needs to be extended artificially using suitable zero or identity restriction operators.

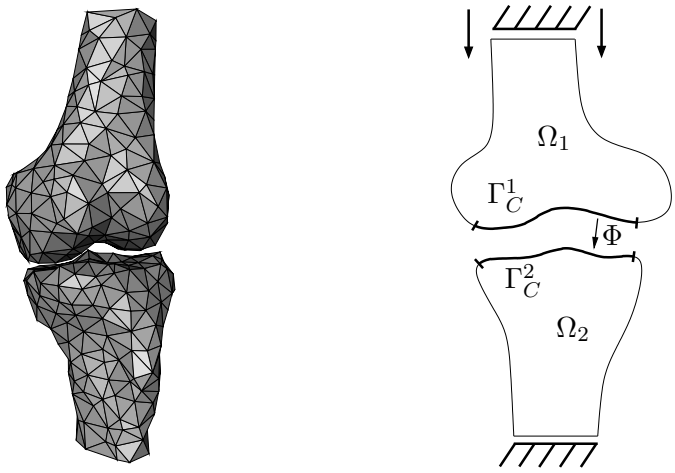


Figure 3.15: Two-body contact problem.

3.8 Contact between the Human Femur and Tibia

We close the chapter by giving a numerical example showing that the good theoretical properties of Truncated Nonsmooth Newton Multigrid (TNNMG) can be observed in practice. We will compare TNNMG with a monotone multigrid solver (MMG), which is currently the fastest globally convergent solver for two-body contact problems. We also compare it with a linear multigrid method for an equivalent linear problem, where the contact has been emulated by a Neumann boundary condition, constructed using a priori knowledge of the solution. As expected the same asymptotic convergence rate can be seen in all three cases.

As an example geometry we chose the left distal femur and proximal tibia from the Visible Human data set [3]. The data was segmented and a high-resolution boundary surface was extracted. The femur surface consisted of 7236 vertices and 14468 triangles, and the tibia surface of 7453 vertices and 14902 triangles. They were simplified as described in Sec. 3.6 to yield coarse surfaces with 268 vertices and 532 triangles for the femur and 224 vertices and 444 triangles for the tibia. The Amira [85] grid generator produced two tetrahedral grids with 378 and 306 vertices, and 1328 and 1044 elements, respectively (Fig. 3.15, left). The worst aspect ratio present in a grid was 17.1.

We modeled bone with an isotropic, homogeneous, linear elastic material with $E = 17$ GPa and $\nu = 0.3$. The bottom section of the proximal tibia was clamped and a downward displacement of 6 mm was prescribed on the upper section of the femur (see Fig. 3.15, right). The part of the femur usually covered with articular cartilage was marked as the nonmortar contact boundary, but note that the actual nonmortar boundary was smaller, because the normal projection Φ could only be constructed on a part of that boundary (Sec. 3.5). The mortar contact boundary was detected automatically as the image of Φ.

As a benchmark we constructed the following linear problem. Let $\mathbf{u}_h \in \mathcal{K}_h$ be the

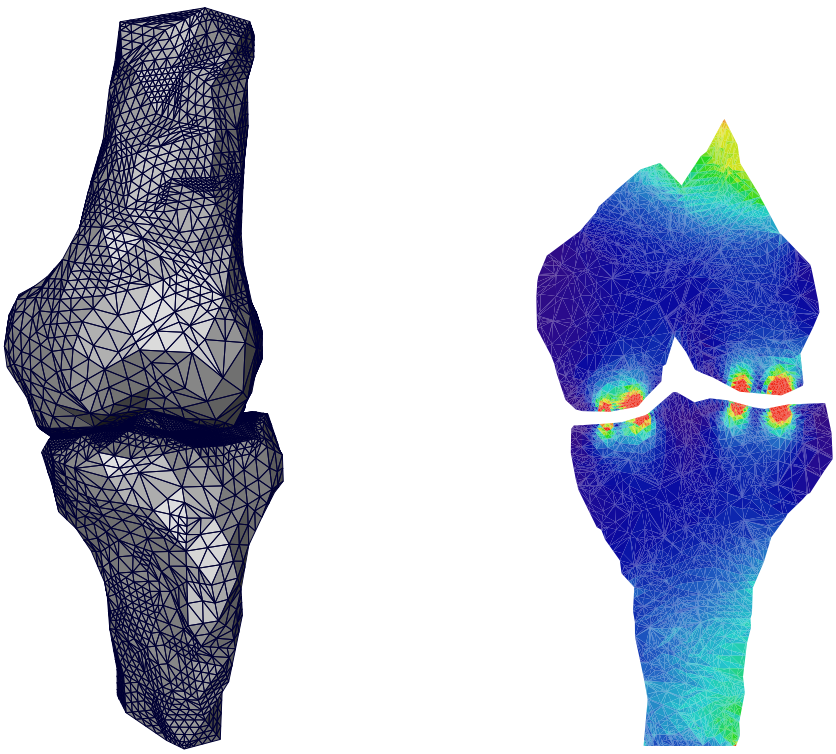


Figure 3.16: Deformed grid and cut through the von-Mises stress field without boundary parametrization.

solution of the contact problem

$$a(\mathbf{u}_h, \mathbf{v}_h - \mathbf{u}_h) \geq l(\mathbf{v}_h - \mathbf{u}_h), \qquad \text{for all } \mathbf{v}_h \in \mathcal{K}_h.$$

Defining the residual $r(\cdot) = l(\cdot) - a(\mathbf{u}_h, \cdot)$, the function $\mathbf{u}_h$ is also the solution of the *linear* problem

$$a(\mathbf{u}_h, \mathbf{v}_h) = l(\mathbf{v}_h) - r(\mathbf{v}_h) \qquad \text{for all } \mathbf{v}_h \in \mathbf{V}_{h,0}. \tag{3.70}$$

In this formulation the contact is emulated by the Neumann force term $r(\cdot)$, which can be computed solving the original contact problem.

In a first series of measurements we tested the solvers on a grid hierarchy obtained by adaptive refinement without a boundary parametrization. In an application of the classic refinement loop we solved on the coarsest grid, estimated the error as described in Sec. 3.7, refined the grid, prolonged the solution and proceeded on the next level. Fig. 3.17, shows the iteration history on the fourth grid level. We started once at zero and once at the solution of the previous level. As expected, all three solvers showed the same asymptotic convergence rate. In fact, the Truncated Nonsmooth Newton Multigrid method needed only very few iterations more than the linear multigrid on the linear problem to solve the contact problem. This is impressive considering that one iteration

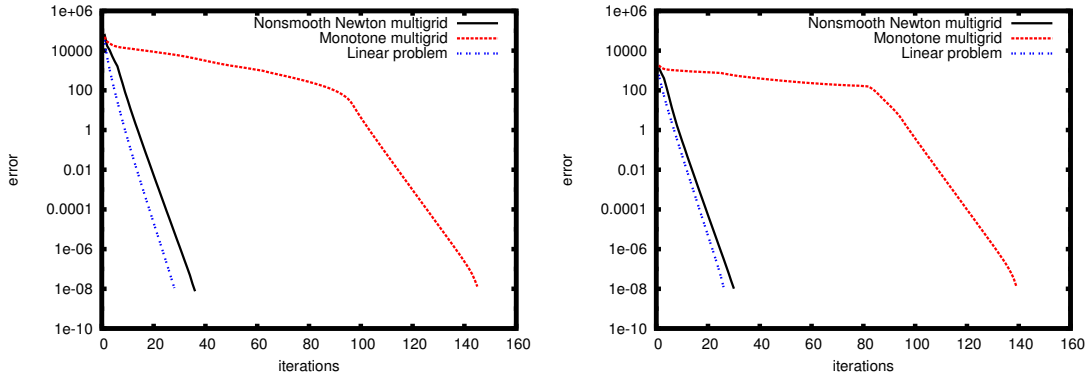


Figure 3.17: Error per iteration. Computation without boundary parametrization, obstacle directions are normal to the nonmortar boundary. Left: starting from zero. Right: starting from the solution on the next-coarser grid.

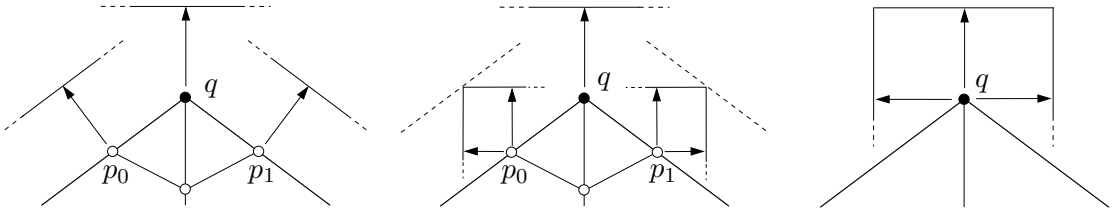


Figure 3.18: Construction of the coarse grid obstacles for a monotone multigrid method in two space dimensions. Left: three fine grid nodes p_0, p_1, q; the center one q also exists on the next-coarser grid. At each node, the normal direction is different and there is an obstacle in this normal direction. Center: When expressed in the coordinate system of q, the admissible sets of p_0 and p_1, which were half-spaces before, now shrink to quarter-spaces. Right: coarse grid obstacle for q. Due to the coordinate changes the coarse obstacle for q now has a restrictive tangential component.

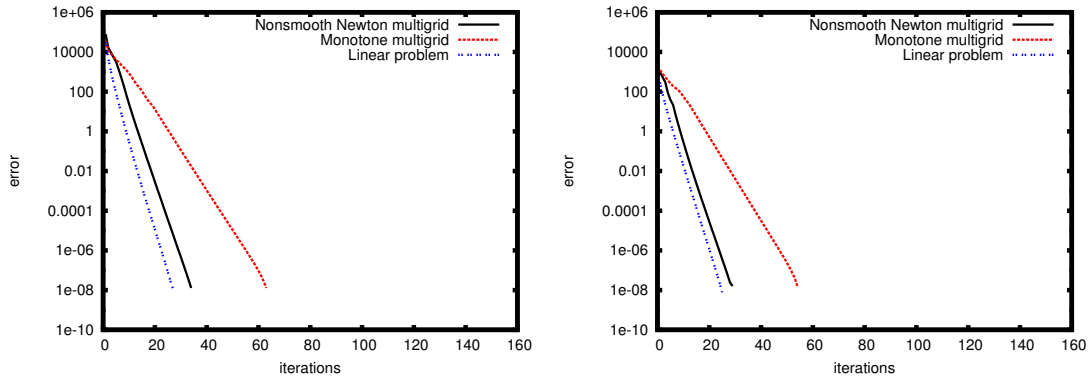


Figure 3.19: Error per iteration. Computation without boundary parametrization, all obstacles point in the direction of the negative z-axis. Left: starting from zero. Right: starting from the solution on the next-coarser grid.

of TNNMG is hardly more costly than a linear multigrid iteration. The monotone multigrid method, on the other hand needed remarkably many iterations to enter the asymptotic phase. This can be explained as follows. In order to make sure that coarse grid corrections do not lead to configurations which violate the obstacle, the coarse grid corrections of MMG are subjected to certain coarse grid defect obstacles [58, 95]. For a given coarse grid vertex p and corresponding nodal basis functions $\boldsymbol{\psi}_{p,i}$, $0 \leq i < d$, the obstacle at p is constructed from all obstacles at the next-finer level which belong to vertices in the support of the $\boldsymbol{\psi}_{p,i}$. In the basis $\{\tilde{\boldsymbol{\psi}}\}$ (3.44), also used for the monotone multigrid method, the obstacles are box constraints, but with respect to the local coordinate systems $O_{pp}, p \in \mathcal{V}$. These vary from node to node if the domain normals $\boldsymbol{\nu}_p$ vary. The process of constructing a single set of box constraints for the coarse grid vertex p involves certain basis transformations [58, 95], which can severely restrict the coarse grid corrections. This is illustrated in Fig. 3.18.

To provide further evidence to our hypothesis that the varying obstacle directions are responsible for the excessively long preasymptotic phase of the monotone multigrid method, we repeated the same set of computations. However, we replaced the domain normals $\boldsymbol{\nu}_p$ which make up the matrix N appearing in the definition (3.38) of the admissible set by the constant vector $\mathbf{e}_2$. This is not a severe change, because in the test problem the normals on $\Gamma_{1,C}$ do all point roughly in the direction of $\mathbf{e}_2$. Since no coarse grid correction is lost due to coordinate system transformations we expected the monotone multigrid method to enter the asymptotic phase quicker. It can be seen in Fig. 3.19 that this is indeed the case. The preasymptotic phase is considerably shortened, and the correct active set is now found after 13 iterations, instead of after 82 iterations.

Qualitatively, all these considerations also hold true when parametrized boundaries are used (Figs. 3.21 and 3.22). From the high-resolution surface we constructed a boundary parametrization as described in Sec. 3.6. After each refinement step we moved all new vertices onto the high-resolution surface. This decreased the mesh quality markedly. In fact, the tibia grid even contained one tetrahedron with negative volume. We tried to

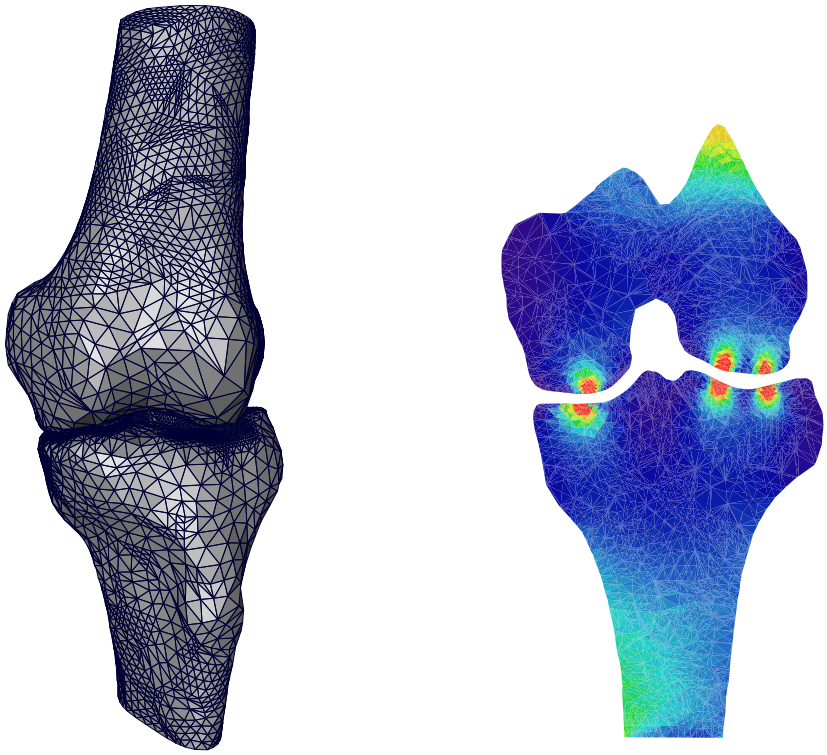


Figure 3.20: Deformed grid and cut through the von-Mises stress field with boundary parametrization.

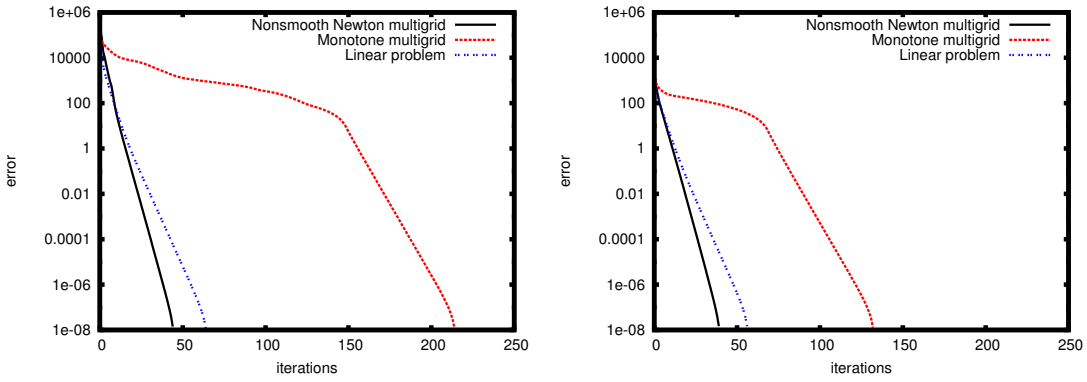


Figure 3.21: Error per iteration. Computation with boundary parametrization, obstacle directions are normal to the nonmortar boundary. Left: starting from zero. Right: starting from the solution on the next-coarser grid.

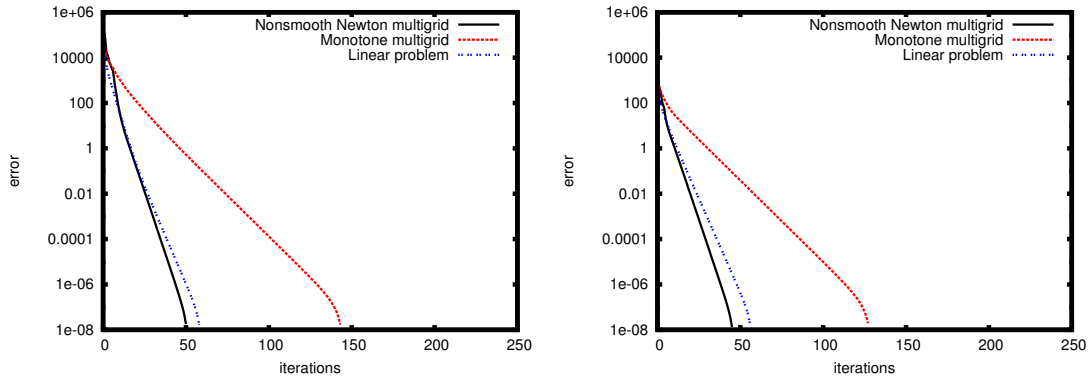


Figure 3.22: Error per iteration. Computation with boundary parametrization, all obstacles point in the direction of the negative z-axis. Left: starting from zero. Right: starting from the solution on the next-coarser grid.

improve the situation by applying the local mesh untangling and smoothing algorithms of Freitag [38], with no improvement of the convergence rate worth the extra effort. The examples in this section have been computed on unsmoothed meshes. Fortunately, the solvers were able to cope quite well with this problem. While the asymptotic convergence rates deteriorated noticeably (from ≈ 0.45 to ≈ 0.55), they stayed within a range that kept the solvers usable. Again, with the obstacles pointing in the directions of the contact boundary normals the preasymptotic phase of the monotone multigrid solver was very long. However, with the boundary parametrization turned on this effect was not as marked as before. This may be due to the fact that the surface normals vary 'more smoothly' when the refined grid approaches a smooth surface. Again, the long preasymptotic phase disappeared when the obstacles were forced to be parallel. As in the case without parametrized boundary, the TNNMG algorithm performed extremely well. There was no noticeable preasymptotic phase even with the obstacles pointing in surface normal direction. In several cases it was even slightly faster than the linear multigrid algorithm for the linear problem. This is probably caused by the line search step, which leads to a slightly larger decrease of energy at each step when compared to a standard linear multigrid method.

One has to mention that TNNMG is much simpler to implement than a monotone multigrid method for contact problems. First of all, in a monotone multigrid method there are obstacles on all grid levels, and a suitable obstacle restriction algorithm needs to be implemented. Moreover, the problem needs to be treated in the transformed basis $\{\tilde{\psi}\}$ (3.44) on all grid levels in order to ensure Gauß–Seidel convergence. This basis has to be constructed for the coarser levels, which in particular requires the assembly of the mass matrices D and M for all levels and a set of specially constructed multigrid prolongation operators. None of this is necessary for TNNMG.

Not having to use the coupling matrices D and M on the coarser grid levels leads to another subtle simplification. Remember that D is diagonal because the dual mortar basis functions are biorthogonal (3.33) with respect to the standard first-order Lagrangian

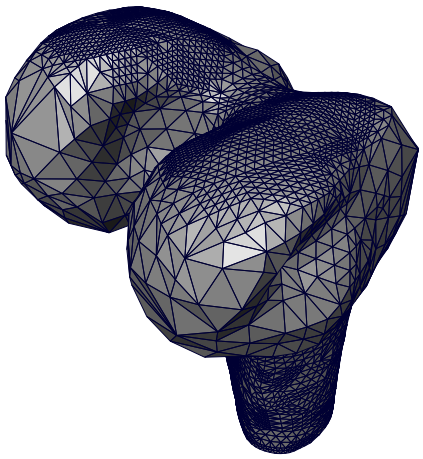


Figure 3.23: View of the refinement of the femoral condyles.

basis functions $\{\psi\}$. Since (3.33) only holds for integration over entire boundary segments, when using the monotone multigrid method the nonmortar boundary $\Gamma_{1,C}$ has to be resolved by the grid on all levels. Depending on the resolution of the lower grid levels this can be a severe restriction on the choice of $\Gamma_{1,C}$, and it further complicates the code that constructs the contact mapping Φ. However no such restrictions arise with TNNMG. Last, but not least, for TNNMG a linear solver can be used to solve the problems on the coarsest grid. For the monotone multigrid method an interior-point method has to be used [58].

In conclusion it can be said that the Truncated Nonsmooth Newton Multigrid method performs extremely well on the biomechanical example problem. It allows the solution of a two-body contact problem with roughly the same number of iterations as a linear multigrid solver on a corresponding linear problem. At the same time one iteration of TNNMG is not more costly than one linear multigrid iteration. The Truncated Nonsmooth Newton Multigrid clearly outperforms the more complicated monotone multigrid method, which seems unsuitable for contact problems on domains with varying boundary normals.

4 Cosserat Rods as Models for Ligaments

In this chapter we present our ligament model. After a brief review of some results from differential geometry we introduce Cosserat rods in Sec. 4.2. Cosserat rods are a well-known concept used to model long slender structures. We introduce a new finite element discretization for the nonlinear rod configuration space. Also we present a globally convergent method to find local minima of the hyperelastic energy functional. We close the section with some numerical results.

4.1 Riemannian Manifolds, Lie Groups, and SO(3)

A *manifold* M is a space that looks locally like a Euclidean space. More formally, it is a topological space together with a collection $\{\mu\}$ of one-to-one mappings of open subsets of $\mathbb{R}^n$ onto subsets of M. These mappings are called *charts*, and n is the dimension of the manifold. It is required that each $x \in M$ is represented in at least one chart. M is called *differentiable* if for any two charts μ_1 and μ_2 with overlapping image in M the coordinate change $\mu_2^{-1} \circ \mu_1$ is a C^∞-function on its domain of definition.

As an example consider S^2, the unit sphere in $\mathbb{R}^3$, which is a two-dimensional differentiable manifold. Let T_{sp} and T_{np} be the tangent planes at the south pole x_{sp} and north pole x_{np}, respectively. Then the stereographic projections $\mu_{\text{sp}} : T_{\text{sp}} \to S^2 \setminus \{x_{\text{np}}\}$ and $\mu_{\text{np}} : T_{\text{np}} \to S^2 \setminus \{x_{\text{sp}}\}$ with respect to the opposite poles form a collection of charts which covers all of S^2. For a general introduction to Riemannian geometry see, e.g., the book by do Carmo [33].

By the Whitney embedding theorem all differentiable n-dimensional manifolds can be considered subsets of some $\mathbb{R}^m$ with $m \geq 2n + 1$. For a point $x \in M$, we call the set of all vectors which are tangent to M at x the *tangent space* at x and denote it by T_xM. The tangent space is a vector space of the same dimension as M. The disjoint union of the tangent spaces of all $x \in M$ is called the *tangent bundle* $TM = \cup_{x \in M} T_xM$. It can be given the structure of a $2n$-dimensional manifold [33]. In the case of S^2, for a point $x \in S^2$ the tangent space is given by $T_xS^2 = \{v \in \mathbb{R}^3 \mid \langle x, v\rangle = 0\}$, which is a two-dimensional subspace of $\mathbb{R}^3$. The tangent bundle TS^2 is four-dimensional. Indeed, any $p \in TS^2$ can be specified by giving the base point $x \in S^2$ such that $p \in T_xS^2$ and coordinates of p in T_xS^2.

A *Riemannian metric* g on a manifold M is a family $\{g_x(\cdot, \cdot) \,|\, x \in M\}$ of scalar products which varies smoothly in M. The pair (M, g) of the manifold together with the metric is called a *Riemannian manifold*. Using the metric it is possible to define a norm $\|\cdot\|_g : TM \to \mathbb{R}_0^+$ by setting $\|v\|_g = \sqrt{g_x(v, v)}$ for $v \in T_xM$. Manifolds such as S^2 which are subsets of some $\mathbb{R}^m$ can inherit a Riemannian metric from the surrounding Euclidean space.

Let $c : [a, b] \to M$ be a curve on M and for each $t \in [a, b]$ denote by $c'(t) \in T_{c(t)}M$ its velocity vector. The length of the curve is given by $\mathcal{L}(c) = \int_{[a,b]} \|c'(t)\|_g \, ds$. A curve c is called a *geodesic* if c minimizes $\mathcal{L}$ locally. This means that for all $\bar{a} < \bar{b}$, $\bar{a}, \bar{b} \in [a, b]$ and $\bar{b} - \bar{a}$ small enough the curve c restricted to $[\bar{a}, \bar{b}]$ is the shortest curve from $c(\bar{a})$ to $c(\bar{b})$ on M. A curve $c : [a, b] \to M$ is called a constant-speed geodesic motion on M if it is a geodesic and there exists a $\kappa \in \mathbb{R}$ such that $c_\kappa : [\kappa^{-1}a, \kappa^{-1}b] \to M, c_\kappa(t) = c(\kappa t)$ is parametrized by arc length.

Intuitively, geodesics generalize the notion of straight lines. Let $q \in M$ and $v \in T_qM$. Then there is a unique geodesic γ parametrized by arc length such that $\gamma(0) = q$ and $\gamma'(0) = v$. The mapping

$$\exp_q : U \subset T_qM \to M, \qquad \exp_q v = \gamma(1) \tag{4.1}$$

is called the *exponential map*. It maps a neighborhood of $0 \in T_qM$ onto a neighborhood of q in M. From the implicit function theorem it follows that the exponential map is C^∞ on a neighborhood of $0 \in T_qM$ [33, Prop. 3.2.9]. In particular we have $\lim_{v\to 0} \exp_q v \to q$. Locally, the curves of the form $c : [-\epsilon, \epsilon] \to M$, $c(t) = exp_q tv$ for $q \in M$, $v \in T_qM$ are geodesics, and all geodesics can be written in this form. A manifold M is called (geodesically) complete if for each pair $p, q \in M$ there exists at least one geodesic that joins p with q. Geodesically complete manifolds are metric spaces where the distance $\mathrm{dist}(x, y)$ between two points $x, y \in M$ is given by the length of the shortest geodesic between x and y. We have the following important result [33, Lem. 3.3.7].

Lemma 4.1.1. *For any $p \in M$ there exists an open neighborhood W of p such that $\mathrm{dist}(p, \cdot) : W \to \mathbb{R}$ depends differentiably on its second argument.*

On the two-dimensional unit sphere S^2 together with the Riemannian metric induced by the restriction of the Euclidean scalar product in $\mathbb{R}^3$ the geodesics are arcs of great circles. For two points $x, y \in S^2$ not opposite each other there are precisely two geodesics, but only one of them minimizes the distance between x and y.

Let M, N be two manifolds of dimension m and n, respectively. A mapping $f : M \to N$ is called differentiable in $x \in M$ if the mapping $\mu_N^{-1} \circ f \circ \mu_M : \mathbb{R}^m \to \mathbb{R}^n$ is differentiable at $\mu_M^{-1}(x)$ for suitable charts μ_N and μ_M. The mapping is called differentiable if this holds for all $x \in M$. Then at each point $x \in M$ there is a linear mapping $D : T_xM \to T_{f(x)}N$ which maps, for each curve c on M through x, the tangent vector of c at x onto the tangent vector of $f(c)$ at $f(x)$. This mapping D is called the derivative of f.

If M and N are two manifolds, their Cartesian product $M \times N$ consisting of tuples (p, q) with $p \in M, q \in N$ is again a manifold. The tangent spaces have the structure

$$T_{(p,q)}(M \times N) = T_pM \oplus T_qN, \tag{4.2}$$

where $\oplus$ denotes the direct sum of vector spaces. If M and N are equipped with Riemannian metrics g and h, respectively, the product manifold can be turned into a Riemannian manifold by using the metric $\hat{g}_{(p,q)} = g_p + h_q$. If both M and N are complete then so is $M \times N$. If $(p, q) \in M \times N$ and $v \in T_pM$, $w \in T_qN$, then

$$\exp_{(p,q)}(v, w) = (\exp_p v, \exp_q w). \tag{4.3}$$

Construction of manifolds through k-fold Cartesian products is done by induction.

A *Lie group* is a manifold G that has a group structure consistent with its manifold structure in the sense that group multiplication

$$\chi : G \times G \to G; \quad (g, h) \to gh$$

is a C^∞ map. The tangent space T_eG at the identity e of G is called its *Lie algebra* $\mathfrak{g}$.

An important Lie group is the group of rotations in $\mathbb{R}^3$. It is also called the *special orthogonal group* and denoted by SO(3). It can be represented by the set of all 3×3 matrices Q with $Q^TQ = \mathrm{Id}$ and $\det Q = 1$. As a manifold, SO(3) is three-dimensional and embedded in the Euclidean space $\mathbb{R}^{3\times3}$ of all 3×3 matrices. Furthermore, SO(3) is compact in $\mathbb{R}^{3\times3}$.

Using this compactness of SO(3) and the theorem of Hopf and Rinow [33, Thm. 7.2.8] the following lemma can be shown.

Lemma 4.1.2. *SO*(3) *is geodesically complete.*

We have the following characterization of the tangent spaces of SO(3).

Lemma 4.1.3. *Let* $\mathbb{A}^3$ *be the space of antisymmetric* 3×3 *matrices and let* $p \in SO(3)$. *Then*

$$T_pSO(3) = p\mathbb{A}^3 := \{m \in \mathbb{R}^{3\times3} \,|\, m = p\hat{v},\ \hat{v} \in \mathbb{A}^3\}. \tag{4.4}$$

Proof. Let $Q : [a, b] \to \mathrm{SO}(3)$ be a curve on SO(3). Differentiating $\mathrm{Id} = Q^T(t)Q(t)$ we get by the product rule

$$(Q^T)'(t)Q(t) = -Q^T(t)Q'(t).$$

If $Q(t) = p$ then this equality holds if $Q'(t) = pA$ for some $A \in \mathbb{A}^3$. Hence we have shown that $T_p\mathrm{SO}(3) \subset p\mathbb{A}^3$. Since both $p\mathbb{A}^3$ and $T_p\mathrm{SO}(3)$ are three-dimensional vector spaces we even have equality. □

In view of this lemma we will frequently write $p\hat{v}$ to denote an element of $p\mathbb{A}^3$. From Lemma 4.1.3 follows in particular that the Lie algebra $\mathfrak{so}(3)$ of SO(3) is the vector space of antisymmetric 3×3 matrices $\mathbb{A}^3$. We may identify $\mathfrak{so}(3)$ with $\mathbb{R}^3$ via the hat map $\hat{}: \mathbb{R}^3 \to \mathfrak{so}(3)$ setting

$$v = (v_1, v_2, v_3) \to \hat{v} = \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix}. \tag{4.5}$$

The canonical Euclidean metric $g(A, B) = \mathrm{tr}(A^TB)$ of $\mathbb{R}^{3\times3}$ induces on SO(3) the metric

$$g_q(q\hat{v}_1, q\hat{v}_2) = \mathrm{tr}(\hat{v}_1^T\hat{v}_2) = \langle v_1, v_2\rangle, \qquad q \in \mathrm{SO}(3),$$

see [4]. Using a general result about Lie group homomorphisms [84, Chap. 10, Prop. 9], the corresponding exponential map can be written as

$$\exp_q q\hat{v} = q \exp \hat{v}, \tag{4.6}$$

where exp (without the subscript) is used to denote the exponential map of SO(3) at the identity.

For computations on SO(3) the *unit quaternions* $\mathbb{H}_{|1|}$ form a set of suitable coordinates [31]. Quaternions are quadruples of real numbers $q = (q_1, q_2, q_3, q_4)$. Together with the multiplication $p = q\tilde{q}$,

$$\begin{aligned} p_1 &= q_4\tilde{q}_1 - q_3\tilde{q}_2 + q_2\tilde{q}_3 + q_1\tilde{q}_4, \\ p_2 &= q_3\tilde{q}_1 + q_4\tilde{q}_2 - q_1\tilde{q}_3 + q_2\tilde{q}_4, \\ p_3 &= -q_2\tilde{q}_1 + q_1\tilde{q}_2 + q_4\tilde{q}_3 + q_3\tilde{q}_4, \\ p_4 &= -q_1\tilde{q}_1 - q_2\tilde{q}_2 - q_3\tilde{q}_3 + q_4\tilde{q}_4, \end{aligned}$$

they form a noncommutative algebra $\mathbb{H}$. In this algebra, the inverse element can be expressed as

$$q^{-1} = \frac{\bar{q}}{|q|^2},$$

where $\bar{q} = (-q_1, -q_2, -q_3, q_4)$ is the element *conjugate* to q and $|q| = \sqrt{\sum_i q_i^2}$ is the absolute value. The unit quaternions $\mathbb{H}_{|1|}$ are the subset of $\mathbb{H}$ for which $|q| = 1$. They form a double covering of SO(3). In particular, the mapping

$$\mathbf{A}(q) = \begin{pmatrix} q_1^2 - q_2^2 - q_3^2 + q_4^2 & 2(q_1q_2 - q_3q_4) & 2(q_1q_3 + q_2q_4) \\ 2(q_1q_2 + q_3q_4) & -q_1^2 + q_2^2 - q_3^2 + q_4^2 & 2(-q_1q_4 + q_2q_3) \\ 2(q_1q_3 - q_2q_4) & 2(q_1q_4 + q_2q_3) & -q_1^2 - q_2^2 + q_3^2 + q_4^2 \end{pmatrix}$$

is a two-to-one mapping from $\mathbb{H}_{|1|}$ onto SO(3) which maps the multiplication in $\mathbb{H}$ onto the group multiplication in SO(3). In other words, for all $p, q \in \mathbb{H}_{|1|}$ we have $pq = \mathbf{A}(p)\mathbf{A}(q)$, where the first product is a quaternion product and the second product the standard matrix one.

In quaternion coordinates there is a closed-form expression for the exponential map of SO(3). Let $\hat{v} \in \mathbb{A}^3 = \mathfrak{so}(3)$ and let $v = (v_1, v_2, v_3)$ be the vector corresponding to $\hat{v}$ by the hat map (4.5). Then $q = \exp \hat{v} \in \mathbb{H}_{|1|}$ can be computed as

$$q_j = \frac{v_j}{|v|} \sin \frac{|v|}{2} \quad \text{for } j = 1, 2, 3, \qquad \text{and} \qquad q_4 = \cos \frac{|v|}{2}. \tag{4.7}$$

Setting $\exp 0 = (0, 0, 0, 1)$ the function exp is C^∞ at 0 as expected from the general theory [67, page 249].

Remark 4.1.1. Due to the factor $|v|^{-1}$, the numerical evaluation of exp is unstable close to zero. Grassia [44] recommends the following strategy. For v with $|v| \leq \sqrt[4]{\epsilon}$, ($\epsilon$ the machine precision), use the first two terms of the series expansion

$$\frac{\sin \frac{|v|}{2}}{|v|} = \frac{1}{2} + \frac{|v|^2}{48} - \frac{|v|^4}{160} + \dots$$

In the context of finite machine precision this representation is exact, since the remainder term is less than ϵ when $|v| \leq \sqrt[4]{\epsilon}$.

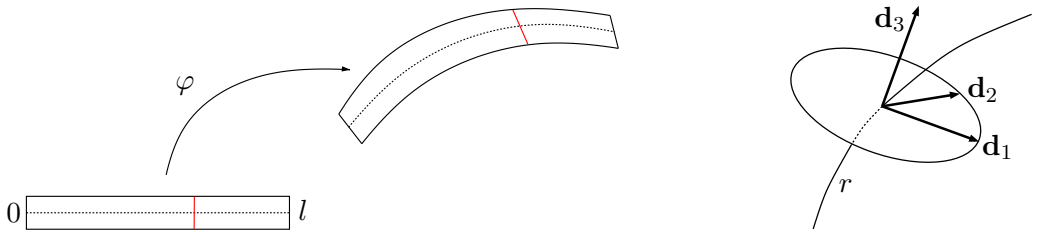


Figure 4.1: Kinematics of Cosserat rods. Left: under deformation, rod cross-sections remain planar, but not necessarily orthogonal to the centerline. Right: the cross-section orientation is represented by three director vectors.

4.2 Cosserat Rods

In this section we briefly present Cosserat rods, which model the large deformation behavior of long, slender objects [28]. The theory of Cosserat rods can be derived by suitably constraining or approximating three-dimensional continuum mechanics models. For the sake of brevity we omit these derivations and merely state the resulting model. For an in-depth presentation see the book by Antman [7] and the references it contains.

Define

$$\mathrm{SE}(3) = \mathbb{R}^3 \times \mathrm{SO}(3),$$

the *special Euclidean group* in $\mathbb{R}^3$. A Cosserat rod is described by a map

$$\begin{array}{rcl} \varphi & : & [0,l] \to \mathrm{SE}(3) \\ s & \to & (\mathbf{r}, q), \end{array} \tag{4.8}$$

which we will assume to be as smooth as necessary. Here $l \in \mathbb{R}$ is the reference length of the rod. The first component $\mathbf{r}$ of φ determines the position of the *centerline* of the rod. The second component q determines the orientation of an idealized cross-section $\mathcal{A}(s)$ (Fig. 4.1, left) which may have an arbitrary shape. This orientation is represented by three pairwise orthonormal vectors $\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3 \in \mathbb{R}^3$, which are called *directors* (Fig. 4.1, right). When quaternions are used as coordinates on SO(3), the expressions for the three directors are

$$\mathbf{d}_1(q) = \begin{pmatrix} q_1^2 - q_2^2 - q_3^2 + q_4^2 \\ 2(q_1 q_2 + q_3 q_4) \\ 2(q_1 q_3 - q_2 q_4) \end{pmatrix}, \tag{4.9a}$$

$$\mathbf{d}_2(q) = \begin{pmatrix} 2(q_1 q_2 - q_3 q_4) \\ -q_1^2 + q_2^2 - q_3^2 + q_4^2 \\ 2(q_2 q_3 + q_1 q_4) \end{pmatrix}, \tag{4.9b}$$

$$\mathbf{d}_3(q) = \begin{pmatrix} 2(q_1 q_3 + q_2 q_4) \\ 2(q_2 q_3 - q_1 q_4) \\ -q_1^2 - q_2^2 + q_3^2 + q_4^2 \end{pmatrix}. \tag{4.9c}$$

When a rod deforms it undergoes *strains*. These are described by two strain functions $\mathbf{v}, \mathbf{u} : [0, l] \to \mathbb{R}^3$ which are defined by the relations

$$\mathbf{v}(s) = \mathbf{r}'(s), \qquad s \in [0, l],$$

and

$$\mathbf{d}_k'(s) = \mathbf{u}(s) \times \mathbf{d}_k(s), \qquad k = 1, 2, 3, \qquad s \in [0, l],$$

where the prime denotes derivation with respect to s. In order to make these strain measures invariant under rigid-body motions they are expressed in the local coordinate systems spanned by the directors $\mathbf{d}_k$. We introduce the new vectors

$$\mathsf{v} = (\mathsf{v}_1, \mathsf{v}_2, \mathsf{v}_3) = \big(\langle \mathbf{v}, \mathbf{d}_1 \rangle, \langle \mathbf{v}, \mathbf{d}_2 \rangle, \langle \mathbf{v}, \mathbf{d}_3 \rangle\big) \tag{4.10}$$

and

$$\mathsf{u} = (\mathsf{u}_1, \mathsf{u}_2, \mathsf{u}_3) = \big(\langle \mathbf{u}, \mathbf{d}_1 \rangle, \langle \mathbf{u}, \mathbf{d}_2 \rangle, \langle \mathbf{u}, \mathbf{d}_3 \rangle\big). \tag{4.11}$$

In the context of rod mechanics we will always use sans serif characters to denote quantities in coordinates of the director frame. The components v_1 and v_2 are interpreted as the shear strains, while v_3 is the stretching strain. The components u_1 and u_2 are the bending strains, and u_3 the strain related to torsion. Using quaternion coordinates the components u_k can be written as

$$\mathsf{u}_k = 2\mathbf{B}_k(q)q', \tag{4.12}$$

where the linear mappings $\mathbf{B}_k : \mathbb{H} \to \mathbb{H}$ are defined as

$$\begin{aligned} \mathbf{B}_1 q &= (q_4, q_3, -q_2, -q_1) \\ \mathbf{B}_2 q &= (-q_3, q_4, q_1, -q_2) \\ \mathbf{B}_3 q &= (q_2, -q_1, q_4, -q_3). \end{aligned}$$

These mappings can be interpreted such that for a small $\epsilon \in \mathbb{R}$, a change in q by $\epsilon \mathbf{B}_k(q)$ produces a rotation about the $\mathbf{d}_k$ axis by an angle of 2ϵ [31].

For a meaningful theory deformations have to preserve the orientation of the material. In particular it should be impossible to compress any part of a rod with positive rest length to zero length. The simplest condition is

$$\mathsf{v}_3 = \langle \mathbf{v}, \mathbf{d}_3 \rangle > 0. \tag{4.13}$$

A more involved treatment which takes the finite cross-sectional area of the rod into account is given by Antman [7]. As (4.13) only rules out very extreme configurations, we will disregard it to simplify our treatment.

The forces and moments acting across a material cross-section $\mathcal{A}(s)$, $s \in [0, l]$ are implicitly averaged to yield a resultant force $\mathbf{n}(s) \in \mathbb{R}^3$ and a resultant moment $\mathbf{m}(s) \in$

$\mathbb{R}^3$ about $\mathbf{r}(s) \in \mathbb{R}^3$. Then balance of moments and forces implies the equilibrium equations

$$\begin{aligned} \mathbf{n}' + \mathbf{f} &= 0, \qquad \text{on } [0,l], \\ \mathbf{m}' + \mathbf{r}' \times \mathbf{n} + \mathbf{l} &= 0, \qquad \text{on } [0,l], \end{aligned}$$

where $\mathbf{f} : [0,l] \to \mathbb{R}^3$ is an external force and $\mathbf{l} : [0,l] \to \mathbb{R}^3$ an external moment [7]. The components of the net forces $\mathbf{n}$ and moments $\mathbf{m}$ with respect to the local coordinate systems spanned by the directors are denoted n_i and m_i, $i \in \{1,2,3\}$, respectively. We refer to m_1, m_2 as the bending moments and to m_3 as the twisting moment. The components n_1 and n_2 are the shear forces and n_3 the tension.

Forces and moments are linked to the strain by constitutive relations which describe the properties of specific materials. Analogously to the continuum mechanics case described in Sec. 3.1, a material is called *hyperelastic* if there exists an energy function $W(\mathsf{w},\mathsf{z},s)$ with $\mathsf{w} = (\mathsf{w}_1,\mathsf{w}_2,\mathsf{w}_3)$, $\mathsf{z} = (\mathsf{z}_1,\mathsf{z}_2,\mathsf{z}_3)$ such that

$$\mathsf{m} = \frac{\partial W}{\partial \mathsf{w}}(\mathsf{u} - \hat{\mathsf{u}}, \mathsf{v} - \hat{\mathsf{v}}, s), \qquad \mathsf{n} = \frac{\partial W}{\partial \mathsf{z}}(\mathsf{u} - \hat{\mathsf{u}}, \mathsf{v} - \hat{\mathsf{v}}, s).$$

Here, $\hat{\mathsf{u}}$ and $\hat{\mathsf{v}}$ are the components of strain in a reference configuration $\hat{\varphi} : [0,l] \to \mathrm{SE}(3)$. We assume this reference configuration to be stress-free by requiring that

$$\frac{\partial W}{\partial \mathsf{w}}(\mathbf{0},\mathbf{0},s) = \frac{\partial W}{\partial \mathsf{z}}(\mathbf{0},\mathbf{0},s) = \mathbf{0}.$$

Further we take the strain-energy function W to be convex, coercive, and as smooth as needed by the analysis. The function W is called *coercive* if for all $s \in [0,l]$

$$\frac{W(\mathsf{w},\mathsf{z},s)}{\sqrt{|\mathsf{w}|^2 + |\mathsf{z}|^2}} \to \infty \quad \text{as} \quad |\mathsf{w}|^2 + |\mathsf{z}|^2 \to \infty$$

(see [7, 51]). The rod is called *uniform* if neither the energy function W nor the reference strains $\hat{\mathsf{u}}$ and $\hat{\mathsf{v}}$ explicitly depend on s.

The simplest choice for a rod material is the linear elastic material. In this case, the energy is a quadratic function of the strains

$$W(\mathsf{w},\mathsf{z},s) = \frac{1}{2} \begin{pmatrix} \mathsf{w} \\ \mathsf{z} \end{pmatrix}^T \mathbf{W}(s) \begin{pmatrix} \mathsf{w} \\ \mathsf{z} \end{pmatrix}, \tag{4.14}$$

where $\mathbf{W}(s) \in \mathbb{R}^{6\times 6}$ is symmetric and positive definite for all $s \in [0,l]$. General linear material models contain 21 free parameters. A further simplification takes the matrix $\mathbf{W}$ to be diagonal. Then the energy density W takes the form

$$W(\mathsf{u} - \hat{\mathsf{u}}, \mathsf{v} - \hat{\mathsf{v}}) = \frac{1}{2} \sum_{i=1}^{3} K_i (\mathsf{u}_i - \hat{\mathsf{u}}_i)^2 + \frac{1}{2} \sum_{i=1}^{3} A_i (\mathsf{v}_i - \hat{\mathsf{v}}_i)^2, \tag{4.15}$$

with scalar parameters K_i, A_i, $i \in \{1,2,3\}$. For the stresses and moments we get

$$\mathsf{m}_i = K_i(\mathsf{u}_i - \hat{\mathsf{u}}_i) \qquad \text{and} \qquad \mathsf{n}_i = K_i(\mathsf{v}_i - \hat{\mathsf{v}}_i), \tag{4.16}$$

again with $i \in \{1,2,3\}$.

If the rod models a solid body whose diameter is small compared to its length, and whose material is homogeneous and isotropic, these parameters can be interpreted as follows [65]. For a cross-section $\mathcal{A}(s)$ let $|\mathcal{A}(s)|$ be its surface area, and for any point $p \in \mathcal{A}(s)$ let $(\xi,\zeta) \in \mathbb{R}^2$ be its coordinates with respect to the coordinate system spanned by $\mathbf{d}_1$ and $\mathbf{d}_2$. Then

$$A_1 = A_2 = G|\mathcal{A}|, \quad A_3 = E|\mathcal{A}| \tag{4.17}$$

with Young's modulus E and shear modulus $G = E/(2+2\nu)$, which contains the Poisson ratio $\nu \in (0, \frac{1}{2})$. Further,

$$K_1 = EJ_1, \quad K_2 = EJ_2, \quad K_3 = GJ_3, \tag{4.18}$$

where

$$J_1 = \int_{\mathcal{A}} \zeta^2 \, dx \qquad \text{and} \qquad J_2 = \int_{\mathcal{A}} \xi^2 \, dx$$

are the *second moments of area* of the cross-section and $J_3 = J_1 + J_2$ is the *polar moment of inertia*. These moments describe how the shape of the cross-section influences the deformation behavior of the rod. For a circular cross-section of radius r we have $J_1 = J_2 = \frac{\pi}{4} r^4$.

Remark 4.2.1. The linear approximation (4.15) may appear over-simplified for the modeling of ligaments. Note, though, that ligaments do indeed show a linear elastic behavior beyond a small toe region (see Sec. 2.3). Also bear in mind that additional parameters as they appear, e.g., in (4.14) may be very difficult to measure (Sec. 2.5). From the mathematical point of view the kinematic rod model can be combined with any sort of material law. See, e.g., [7, Chap. 8] for a viscoelastic law which may be modified to yield the quasilinear viscoelasticity of Fung ([40] and Sec. 2.3).

In analogy to the continuum mechanics case, the stable equilibrium configurations of a Cosserat rod with a hyperelastic material law can be characterized as the minima of an energy functional j defined by

$$j(\varphi) = \int_{[0,l]} W\big(\mathsf{u}(\varphi) - \hat{\mathsf{u}}, \mathsf{v}(\varphi) - \hat{\mathsf{v}}, s\big) \, ds, \tag{4.19}$$

in a space of functions $[0,l] \to \mathrm{SE}(3)$ of appropriate regularity [82]. From the coerciveness of W follows the coerciveness of j as a function of u and v. Hence level sets of j are bounded. The problem of finding equilibrium configurations can therefore be written as an optimization problem. Let C_D^* be the space of all functions $[0,l] \to \mathrm{SE}(3)$ that are sufficiently regular and that fulfill given Dirichlet boundary conditions. We want to find a $\varphi^* \in C_D^*$ such that

$$j(\varphi^*) \leq j(\varphi) \qquad \text{for all } \varphi \in C_D^*. \tag{4.20}$$

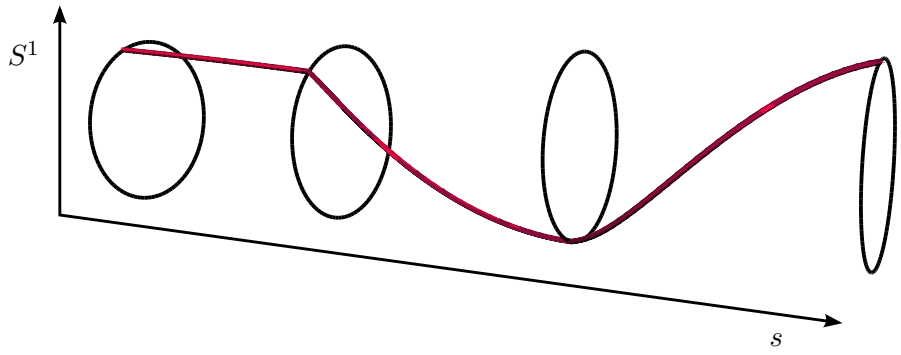


Figure 4.2: First-order geodesic finite element function on the unit circle S^1.

Existence and regularity of solutions to this problem have been shown by Seidman and Wolfe [82]. The main difficulty is the condition (4.13) on the preservation of orientation, since it is a strict inequality and hence the admissible set is open. Seidman and Wolfe also showed that solutions of nonlinear rod problems are generally not unique.

4.3 Geodesic Finite Element Spaces

When rod problems are to be solved numerically, special care has to be taken to choose a suitable discretization. The concept of linear finite elements has to be generalized since it does not make sense in a nonlinear space. In this section we introduce geodesic finite element spaces as a generalization of first-order finite element spaces to problems involving functions whose range space is a nonlinear manifold. This concept provides a geometric view on the interpolation formula of Sansour and Wagner [80], and generalizes their work to smooth Riemannian manifolds. To our knowledge geodesic finite elements have not appeared in the literature before. With the application to rod problems in mind we stick to one-dimensional domain spaces. However, an extension to higher dimensions is conceivable and may be useful for nonlinear shell models or micropolar materials. We first discuss geodesic finite element spaces for general Riemannian manifolds and then concentrate on SE(3), the configuration manifold of a Cosserat rod cross-section.

Let M be a Riemannian manifold which is geodesically complete, and consider continuous functions from an interval $[0, l]$ to M. Introduce a one-dimensional grid G on $[0, l]$ by subdividing the interval in finitely many subintervals $[l_i, l_{i+1}]$ of not necessarily the same size. Call n the number of grid vertices and $h := \max_i |l_{i+1} - l_i|$ the maximum element size. Note that for any first-order finite element function ϕ_h defined on G with the range space $\mathbb{R}^m, m \geq 1$, the graph of ϕ_h is a connected series of line segments in $\mathbb{R}^{m+1}$. This, together with the observation that geodesics are the natural generalization of straight lines on manifolds, motivates the definition of geodesic finite element spaces.

Definition 4.3.1 (Geodesic finite elements)**.** *Let G be a one-dimensional grid on $[0, l]$ and M a Riemannian manifold that is geodesically complete. We call $\phi_h : [0, l] \to M$*

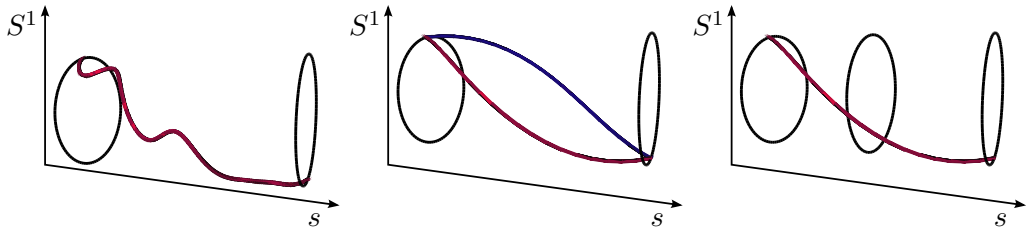


Figure 4.3: Left: a smooth function $\phi : [0, l] \to S^1$ with $\phi(0) = 0$ and $\phi(l) = \pi$. Center: nodal interpolation on the grid $G = [0, l]$ yields two minimizing geodesics. Right: if the grid is fine enough the nodal interpolation is unique.

a first-order geodesic finite element function for M if it is continuous and, for each element $[l_i, l_{i+1}]$ of G, $\phi_h(s)$ is a constant-speed geodesic motion on M on a minimizing geodesic. The space of all such functions will be denoted by V_h^M.

Example 1. Let $M = S^1$ be the unit sphere in $\mathbb{R}^2$ with a coordinate system given by the angle α. Then $V_h^{S^1}$ is the set of all continuous functions $\phi_h : [0, l] \to S^1$ such that the restriction of ϕ_h to an element $[l_i, l_{i+1}]$ is of the form

$$\phi_h|_{[l_i, l_{i+1}]}(s) = \alpha_i + m\left(\frac{s - l_i}{l_{i+1} - l_i}\right),$$

with $|m| \leq \pi$ (see Fig. 4.2).
Example 2. If $M = \mathbb{R}^m$ for some $m \geq 1$ then V_h^M is precisely the standard m-valued first-order finite element space.

We will now explore a few properties of geodesic finite element spaces. Note first that V_h^M is a linear space if and only if M is. Therefore, unless M is a linear space, there is no such thing as a basis of V_h^M. In particular, there is no nodal basis.

The finite element method makes much use of the natural isomorphism between finite element functions and coefficient vectors. Let V_h be a space of first-order finite element functions mapping into $\mathbb{R}^m$ and let v_h be a function in V_h. The coefficient vector corresponding to v_h is obtained by pointwise evaluation at the grid vertices. Let $C([0, l]; \mathbb{R}^m)$ be the space of continuous functions mapping $[0, l]$ to $\mathbb{R}^m$. For a given grid G we denote the pointwise evaluation operator by

$$\mathcal{E} : C([0, l]; \mathbb{R}^m) \to \mathbb{R}^{m \times n}, \qquad (\mathcal{E}(v_h))_i = v_h(l_i) \in \mathbb{R}^m,\ i = 0, \ldots, n - 1.$$

Its inverse, the prolongation $\mathcal{E}^{-1}$ is set-valued and maps a coefficient vector v in $\mathbb{R}^{m \times n}$ to the set of all continuous functions that have v as their pointwise evaluation at the grid vertices. However, the prolongation $\mathcal{E}^{-1}v \subset C([0, l]; \mathbb{R}^m)$ contains only a single element in V_h for each $v \in \mathbb{R}^{m \times n}$.

This may not be true for general manifolds M, where one would hope for an isomorphism between V_h^M and M^n, the n-fold product of M. While there is always a unique straight line segment between two points in a Euclidean space, there may be more than

one minimizing geodesic between two given points on M. As a simple example consider again $M = S^1$ with a coordinate system given by the angle α, a grid G consisting of a single element $[l_0, l_1]$, and the coefficient vector $\bar{\phi} \in (S^1)^2$ such that $\bar{\phi}_0 = 0$ and $\bar{\phi}_1 = \pi$. Then there are two minimizing geodesics from $\bar{\phi}_0$ to $\bar{\phi}_1$, namely the one in clockwise and the one in counterclockwise direction, both with length π (Fig. 4.3, center). Hence there are two functions ϕ_h^+ and ϕ_h^- in $V_h^{S^1}$ such that $\mathcal{E}(\phi_h^+) = \bar{\phi}$ and $\mathcal{E}(\phi_h^-) = \bar{\phi}$.

On certain manifolds it can be shown that minimizing geodesics are always unique. An example are the hyperbolic spaces H^d [33, Prop. 8.3.1], and of course the linear spaces. If minimizing geodesics are not unique, we can show a local property which is sufficient for practical applications. We need the following classical result [33, Thm. 3.3.7 and Rem. 3.3.8].

Lemma 4.3.1. *For each $p \in M$ there is a nonempty neighborhood U of p in M such that for all $q, q' \in U$ there is a unique minimizing geodesic from q to q'.*

Example: For two points $\alpha, \beta \in S^1$ the minimizing geodesic from α to β is unique if $\operatorname{dist}(\alpha, \beta) < \pi$.

The radius of the largest geodesic ball $B(p)$ at p with $B(p) \subset U$ is called the *injectivity radius* at p and denoted by $\operatorname{inj}(p)$. The infimum of the injectivity radii over all of M is called the injectivity radius of M and denoted by $\operatorname{inj}(M)$.

Let $\phi : [0, l] \to M$. We say that ϕ is Lipschitz continuous with Lipschitz constant L if

$$\operatorname{dist}(\phi(a), \phi(b)) \leq L|a - b|$$

holds for all $a, b \in [0, l]$.

Lemma 4.3.2. *Let $\operatorname{inj}(M) > 0$ and $\phi : [0, l] \to M$ be Lipschitz continuous with Lipschitz constant L. Let G be a grid with maximum element size $h < \operatorname{inj}(M)/L$. Then, setting $\bar{\phi} = \mathcal{E}(\phi) \in M^n$, the inverse of $\mathcal{E}$ at $\bar{\phi}$ has only a single element in V_h^M. If $h < \epsilon \operatorname{inj}(M)/L$ for some $\epsilon \in (0, 1)$, the inverse of $\mathcal{E}$ has only a single element in V_h^M for all $\tilde{\phi}$ in a neighborhood of $\bar{\phi}$.*

Proof. We note first that by Lemma 4.3.1 a minimizing geodesic joining a and b is unique if $\operatorname{dist}(a, b) < \operatorname{inj}(M)$. If h is less than $\operatorname{inj}(M)/L$ and

$$\bar{\phi} = \mathcal{E}(\phi) \in M^n, \qquad \bar{\phi}_i = \phi(l_i), \qquad 0 \leq i < n$$

we get

$$\operatorname{dist}(\bar{\phi}_i, \bar{\phi}_{i+1}) \leq L|l_i - l_{i+1}| \leq Lh \leq \operatorname{inj}(M)$$

for all $0 \leq i < n - 1$. Hence there is a unique minimizing geodesic from $\bar{\phi}_i$ to $\bar{\phi}_{i+1}$ and a unique prolongation of $\bar{\phi}$ into V_h^M.

Let now G be such that $h = \epsilon \operatorname{inj}(M)/L$ with $\epsilon \in (0, 1)$. Then $\operatorname{dist}(\bar{\phi}_i, \bar{\phi}_{i+1}) \leq \epsilon \operatorname{inj}(M)$ for all $0 \leq i < n-1$. Define $\epsilon^* = \frac{(1-\epsilon)}{2} \operatorname{inj}(M)$ and set $B_{\epsilon^*}(\bar{\phi}_i)$ and $B_{\epsilon^*}(\bar{\phi}_{i+1})$ the geodesic balls of radius ϵ^* around $\bar{\phi}_i$ and $\bar{\phi}_{i+1}$ in M, respectively. Then for any $\tilde{\phi}_i \in B_{\epsilon^*}(\bar{\phi}_i)$ and $\tilde{\phi}_{i+1} \in B_{\epsilon^*}(\bar{\phi}_{i+1})$ we have, by the triangle inequality,

$$\operatorname{dist}(\tilde{\phi}_i, \tilde{\phi}_{i+1}) \leq \operatorname{dist}(\tilde{\phi}_i, \bar{\phi}_i) + \epsilon \operatorname{inj}(M) + \operatorname{dist}(\bar{\phi}_{i+1}, \tilde{\phi}_{i+1}) \leq \operatorname{inj}(M).$$

Hence if $B_{\epsilon^*}(\bar{\phi})$ is the geodesic ball in M^n of radius ϵ^* around $\bar{\phi} \in M^n$ there is a unique prolongation for all $\tilde{\phi} \in B_{\epsilon^*}(\bar{\phi})$ into V_h^M. □

This lemma implies that for a given continuous problem we can always find a grid fine enough such that we can disregard the distinction between V_h^M and M^n in the vicinity of the solution. In this vicinity V_h^M inherits the manifold structure of M^n, because short geodesics depend differentiably on their endpoints (Lemma 4.1.1). If M is a Lie group V_h^M also locally inherits the Lie group properties of M^n.

We will now focus on geodesic finite element functions mapping to SO(3). As a first result we can compute the injectivity radius of SO(3) exactly.

Lemma 4.3.3. *Let $p, q \in SO(3)$. Then $\operatorname{dist}(p,q) \leq \pi$. If $\operatorname{dist}(p,q) < \pi$ then there is a unique minimizing geodesic from p to q. If $\operatorname{dist}(p,q) = \pi$ then there are precisely two minimizing geodesics from p to q.*

Proof. Let $q \neq p$. Geodesics from p to q are images of straight lines under the exponential map $\exp_p$. By Lemmas 4.1.2 and 4.1.3 there is at least one unit vector $\hat{v} \in \mathfrak{so}(3)$ such that $\exp_p tp\hat{v} = p \exp t\hat{v} = q$ for some $t \in \mathbb{R}^+$. We first show that $\hat{v}$ must be unique in the sense that any other unit tangent vector $\hat{\bar{v}}$ with $p \exp \bar{t}\hat{\bar{v}} = q$ must be collinear to $\hat{v}$. Indeed, let $\hat{\bar{v}} \in \mathfrak{so}(3)$ be a second linearly independent unit vector such that $p \exp \bar{t}\hat{\bar{v}} = q$, with $\bar{t} \in \mathbb{R}^+$. Consider Rodrigues' formula, which is an explicit formula for the exponential map on SO(3) using orthogonal matrices as coordinates [67]

$$\begin{aligned} \exp t\hat{v} &= \exp \begin{pmatrix} 0 & -v_3 t & v_2 t \\ v_3 t & 0 & -v_1 t \\ -v_2 t & v_1 t & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix} \sin t + (\mathrm{Id} - vv^T) \cos t + vv^T. \end{aligned} \tag{4.21}$$

It is easy to check that $v = (v_1, v_2, v_3)$ is an eigenvector of $\exp t\hat{v}$ to the eigenvalue 1. Indeed, unless $t = 0$ (i.e., unless $\exp t\hat{v}$ is the identity) it is the only one. Therefore, from the assumptions above it follows that v and $\bar{v}$ are both eigenvectors of $p^{-1}q$ interpreted as a matrix for the eigenvalue 1. This however contradicts their assumed linear independence. Hence $\hat{\bar{v}} = \pm\hat{v}$ and all geodesics from p to q are generated by $\exp t\hat{v}$ with $t \in \mathbb{R}$.

Consider now $\hat{v}$ fixed and $\exp t\hat{v}$ as a function from $\mathbb{R}$ to SO(3). By (4.21) it is periodic with the period 2π. Hence there exists a $t^* \in (0, 2\pi)$ such that $p \exp t\hat{v} = q$ if and only if $t = t^* + 2i\pi$ for $i \in \mathbb{Z}$. If $t^* < \pi$ then $c_+ = \{p \exp t\hat{v} \mid 0 \leq t \leq t^*\}$ is the unique minimizing geodesic from p to q of length t^*. If $t^* > \pi$ then $c_- = \{p \exp t\hat{v} \mid t^* - 2\pi \leq t \leq 0\}$ is the unique minimizing geodesic from p to q of length $2\pi - t^* < \pi$. If $t^* = \pi$ then both c_+ and c_- have the same length π and are minimizing geodesics. □

For actual computations the geodesic finite element functions obtained by prolongation of coefficient vectors need to be evaluated at the quadrature points. Using quaternion

coordinates and the formulas (4.7) for the exponential map on SO(3) explicit expressions for the geodesic interpolation on SO(3) can be derived. More formally, we give a closed-form expression for

$$\mathsf{q} : [0, \delta] \times \mathrm{SO}(3) \times \mathrm{SO}(3) \to \mathrm{SO}(3),$$

where $[0, \delta]$ is an interval and

$$\mathsf{q}(\cdot, p, q) : [0, \delta] \to \mathrm{SO}(3)$$

is a constant-speed parametrization of a minimizing geodesic with $\mathsf{q}(0, p, q) = p$ and $\mathsf{q}(\delta, p, q) = q$.

For any geodesic that connects p to q there is a tangent vector $p\hat{v} \in T_p\mathrm{SO}(3)$ such that $\exp_p p\hat{v} = q$. Using that $\exp_p$ is a local diffeomorphism we get the analytical expression $\hat{v} = \exp^{-1}(p^{-1}q)$. We can use (4.7) to obtain that the geodesic distance between p and q is $|\hat{v}| = |v| = 2 \arccos(p^{-1}q)_4$ and that

$$v_j = \frac{(p^{-1}q)_j |v|}{\sin \frac{|v|}{2}}, \qquad j \in \{1, 2, 3\}.$$

The interpolation between p and q along the connecting geodesic induced by $\hat{v}$ is then

$$\mathsf{q}(s, p, q) = \exp_p \frac{s}{\delta} p\hat{v} = p \exp \frac{s}{\delta} \hat{v} = p \exp \frac{s}{\delta} [\exp^{-1} p^{-1}q]. \tag{4.22}$$

Note that $\mathsf{q}(s, p, q) = \mathsf{q}(\delta - s, q, p)$ can be shown by direct calculation.

Finite element computations also require the tangent vector of a geodesic from p to q at quadrature points $s \in [0, \delta]$. By (4.22) and the chain rule we get

$$\left.\frac{\partial \mathsf{q}(\cdot, p, q)}{\partial s}\right|_s = p \left.\frac{\partial \exp}{\partial v}\right|_{v = \frac{s}{\delta}\hat{v}} \cdot \frac{\hat{v}}{\delta}. \tag{4.23}$$

The explicit formula for the partial derivatives $\partial \exp / \partial v$ of the exponential map can be found in Appendix A.

Functions describing configurations of Cosserat rods map intervals onto SE(3). Using the results about products of spaces we know that this is a Riemannian manifold with tangent spaces

$$T_{(r,q)}\mathrm{SE}(3) = T_r\mathbb{R}^3 \oplus T_q\mathrm{SO}(3). \tag{4.24}$$

The metric is given by

$$\begin{aligned} g_{(r,q)}(\cdot, \cdot) &: T_{(r,q)}\mathrm{SE}(3) \times T_{(r,q)}\mathrm{SE}(3) \to \mathbb{R} \\ g_{(r,q)}\big((w_1, q\hat{v}_1), (w_2, q\hat{v}_2)\big) &= \langle w_1, w_2 \rangle + \langle v_1, v_2 \rangle \end{aligned} \tag{4.25}$$

and the exponential map is

$$\begin{aligned} \exp_{(r,q)} &: T_{(r,q)}\mathrm{SE}(3) \to \mathrm{SE}(3) \\ \exp_{(r,q)}(w, q\hat{v}) &= (r + w, q \exp \hat{v}). \end{aligned}$$

The following result is straightforward.

Lemma 4.3.4. *The Riemannian manifold SE(3) is geodesically complete.*

Using that minimizing geodesics in Euclidean spaces are always unique together with Lem. 4.3.3 we can determine the injectivity radius of SE(3).

Lemma 4.3.5. *The injectivity radius of the special Euclidean group is* $\operatorname{inj}(SE(3)) = \pi$.

We now write down the discrete version of the rod minimization problem (4.20). Let G be a grid on $[0, l]$ with n vertices and maximum element size h. Let $\varphi_0, \varphi_l \in \mathrm{SE}(3)$ be given Dirichlet boundary data. We want to find a $\varphi_h^* \in V_{h,D}^{\mathrm{SE}(3)}$ such that

$$j(\varphi_h^*) \leq j(\varphi_h) \qquad \text{for all } \varphi_h \in V_{h,D}^{\mathrm{SE}(3)}, \tag{4.26}$$

where

$$V_{h,D}^{\mathrm{SE}(3)} = \left\{ \varphi_h \in V_h^{\mathrm{SE}(3)} \;\middle|\; \varphi_h(0) = \varphi_0,\ \varphi_h(l) = \varphi_l \right\}$$

is the geodesic finite element space for SE(3) fulfilling given Dirichlet conditions, and

$$j = \int_{[0,l]} W(\mathsf{u} - \hat{\mathsf{u}}, \mathsf{v} - \hat{\mathsf{v}}, s)\, ds$$

is the hyperelastic energy functional (4.19).

We can also write down an algebraic formulation of the same problem. Let h be sufficiently small. We want to find a $\bar{\varphi}^* \in \mathrm{SE}(3)^n$ with $\bar{\varphi}_0^* = \varphi_0$ and $\bar{\varphi}_{n-1}^* = \varphi_l$ such that

$$j(\bar{\varphi}^*) \leq j(\bar{\varphi}) \qquad \text{for all } \bar{\varphi} \in \mathrm{SE}(3)^n \text{ with } \bar{\varphi}_0 = \varphi_0 \text{ and } \bar{\varphi}_{n-1} = \varphi_l \tag{4.27}$$

with

$$j(\bar{\varphi}) = \int_{[0,l]} W\big(\mathsf{u}(\mathcal{E}^{-1}(\bar{\varphi})) - \hat{\mathsf{u}}, \mathsf{v}(\mathcal{E}^{-1}(\bar{\varphi})) - \hat{\mathsf{v}}, s\big)\, ds. \tag{4.28}$$

Note that, by Lemma 4.3.2, the minimization problem (4.27) is restricted to those $\bar{\varphi}$ for which $\mathcal{E}^{-1}$ is single-valued.

Using Lemma 4.1.1 and the fact that $\mathbf{u}$, $\mathbf{v}$, and W are all differentiable we can show the following.

Lemma 4.3.6. *The functional* $j : SE(3)^n \to \mathbb{R}$ *depends differentiably on its arguments.*

Remark 4.3.1. Rod and beam models are well-known to exhibit *locking*, i.e., a bad approximation property of the discrete problem for coarse grids [21]. This problem can be solved by using a mixed discretization. For problems which do not couple translational and rotational strains (i.e., the matrix $\mathbf{W}$ in (4.14) is block-diagonal), it is also possible to use the much simpler selective integration method [73].

4.4 Riemannian Trust-Region Solvers

In this section we present a trust-region solver for the algebraic minimization problem (4.27). Trust-region solvers are a standard tool for nonconvex optimization in Euclidean spaces [26]. Absil et al. generalized them to optimization problems on Riemannian manifolds and used them successfully for problems in numerical linear algebra [5]. To our knowledge no trust-region algorithm for Cosserat rods has been published. However, Simo and Vu-Quoc [83] used the related idea of a Newton method on the nonlinear rod configuration space. Adler et al. [6] used a similar method for the simulation of the human spine.

There are several reasons why to choose a trust-region solver rather than other, seemingly simpler algorithms such as the one presented in [46]. First of all, they are very fast. From the underlying Newton idea trust-region solvers inherit local superlinear convergence. This is of particular importance when solving static rod problems as part of an outer iterative scheme, since then a good starting iterate is frequently available. Secondly, they are globally convergent, see Theorem 4.4.1. Last but not least, they are very flexible. It is, for example, quite easy to extend them to handle certain inequality constraints. That way contact problems involving rods such as ligaments wrapping around bone can be treated.

We first present the basic trust-region algorithm on Euclidean spaces. Let $J : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable, bounded from below and such that $J(x) \to \infty$ when $\|x\| \to \infty$. We look for a $x^* \in \mathbb{R}^n$ such that

$$J(x^*) \leq J(x) \qquad \text{for all } x \in \mathbb{R}^n. \tag{4.29}$$

Starting from an initial iterate $x_0 \in \mathbb{R}^n$, a trust-region solver will, at the ν-th iteration, fit a quadratic model $m_\nu : \mathbb{R}^n \to \mathbb{R}$ to J at x_ν by setting

$$m_\nu(v) = J(x_\nu) + \langle \nabla J(x_\nu), v \rangle + \frac{1}{2} v^T \nabla^2 J(x_\nu) v. \tag{4.30}$$

If v_ν is the minimum of m_ν then $x_{\nu+1} = x_\nu + v_\nu$ is expected to be a better iterate than x_ν. As m_ν can only be supposed to be a good approximation of J in a neighborhood of x_ν, the search for a minimum of m_ν is restricted to

$$K_\nu^{\mathrm{tr}} = \left\{ v \in \mathbb{R}^n \;\middle|\; \|v\| \leq \rho_\nu \right\} \qquad \rho_\nu > 0, \tag{4.31}$$

the name-giving *trust-region*. Its radius ρ_ν depends on how far we trust the quadratic functional m_ν to be a good approximation of J around x_ν. The norm in (4.31) is arbitrary and may even differ from iteration to iteration [26]. The restriction of the minimization problem for (4.30) to K_ν^{tr} also ensures the existence of a minimum of the subproblem when (4.30) is not convex.

Let v_ν be a minimum of m_ν on K_ν^{tr}. The quality of a correction step is estimated by comparing the functional decrease and the model decrease. If the quotient

$$\eta_\nu = \frac{J(x_\nu) - J(x_\nu + v_\nu)}{m_\nu(0) - m_\nu(v_\nu)} \tag{4.32}$$

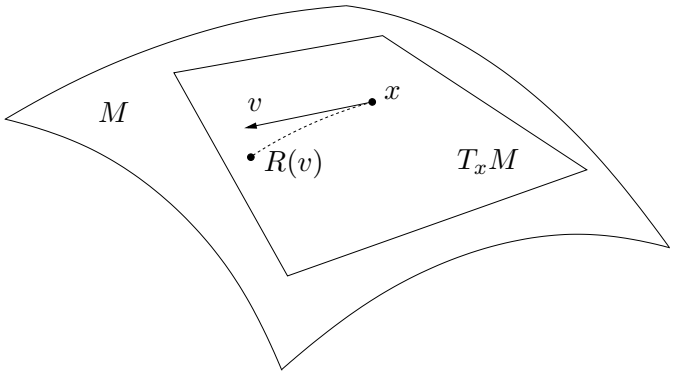


Figure 4.4: The retraction mapping $R_x : T_xM \to M$.

is smaller than a fixed value η_1, then the step is rejected, and v_ν is recomputed for a smaller ρ. If not, the step is accepted. If η_ν is larger than a second value η_2, the trust region radius is enlarged for the next step. Common values are $\eta_1 = 0.01$ and $\eta_2 = 0.9$ [26].

The trust-region algorithm converges globally to a first-order critical point of J. Saddle points of J are numerically unstable and therefore limit points are practically always minima. When close enough to a solution, the trust-region constraint (4.31) becomes inactive and the local quadratic convergence of a Newton solver is recovered. The book by Conn et al. [26] contains the detailed convergence theory.

If J is defined on a nonlinear manifold instead of a Euclidean space, the concept of a local quadratic model (4.30) of the objective function has to be revised. Also, as there is no canonical addition defined on a manifold, the update procedure $x_{\nu+1} = x_\nu + v_\nu$ needs to be replaced by something more general. We will now briefly present the generalization of the trust-region algorithm to Riemannian manifolds as introduced by Absil et al. [5]. In the next section we will show how the general algorithm specializes when the underlying space is the set $\mathrm{SE}(3)^n$ of configurations of an algebraic Cosserat rod problem.

Let (M, g) be a smooth Riemannian manifold with metric g. The basic idea of the Riemannian trust-region algorithm is that in a neighborhood of a point $x \in M$ the objective function can be transported onto the tangent space T_xM at x. There, a vector space trust-region subproblem can be solved and the result transported back onto M. The notion of transporting onto T_xM is made formal by the introduction of *retractions*.

Definition 4.4.1 (Retraction)**.** *A retraction on a manifold M is a continuously differentiable mapping*

$$R : TM \to M$$

with the following properties.

1. *$R_x(0_x) = x$, where 0_x is the zero element of T_xM, and R_x is the restriction of R to T_xM.*

2. *$DR_x(0) = Id_{T_xM}$, the identity mapping on T_xM, with the canonical identification $T_{0_x}T_xM \simeq T_xM$.*

The exponential map (4.1) is always a retraction. Alternatives have been considered [6], since $\exp_x$ is not always easy to evaluate. When dealing with SO(3) this is not a problem, since the explicit formula (4.7) is available. We will hence always set $R_x = \exp_x$.

A Riemannian trust-region step now proceeds as follows. Let $x_\nu \in M$ be the current iterate. We use R_{x_ν} to locally lift the objective function J onto the tangent space $T_{x_\nu}M$ to obtain

$$\begin{aligned} \hat{J}_\nu &: T_{x_\nu}M \to \mathbb{R} \\ \hat{J}_\nu(v) &= J(R_{x_\nu}(v)). \end{aligned}$$

The Riemannian structure g turns $T_{x_\nu}M$ into a Banach space with the norm $\|\cdot\|_{g_{x_\nu}} = \sqrt{g_{x_\nu}(\cdot,\cdot)}$. There, the trust-region subproblem reads

$$v_\nu = \arg \min_{\substack{v \in T_{x_\nu}M \\ \|v\|_{g_{x_\nu}} \leq \rho_\nu}} m_{x_\nu}(v) \tag{4.33}$$

with

$$m_\nu(v) = \hat{J}_\nu(0_{x_\nu}) + d\hat{J}_\nu(0_{x_\nu})[v] + \frac{1}{2} d^2\hat{J}_\nu(0_{x_\nu})[v,v]. \tag{4.34}$$

Here, $d\hat{J}_\nu(0_{x_\nu}) : T_{x_\nu}M \to \mathbb{R}$ is the differential of $\hat{J}_\nu$ at $0_{x_\nu} \in T_{x_\nu}M$ and $d^2\hat{J}_\nu(0_{x_\nu}) : T_{x_\nu}M \times T_{x_\nu}M \to \mathbb{R}$ the second-order differential operator of $\hat{J}_\nu$ at $0_{x_\nu} \in T_{x_\nu}M$. Note that the problem is independent of a specific coordinate system on $T_{x_\nu}M$. Alternatively, the quadratic functional (4.34) can be written without differential forms to read

$$m_\nu(v) = \hat{J}_\nu(0_{x_\nu}) + g_{x_\nu}(\nabla\hat{J}_\nu(0_{x_\nu}), v) + \frac{1}{2} g_{x_\nu}(\nabla^2\hat{J}_\nu(0_{x_\nu})v, v).$$

The gradient $\nabla\hat{J}_\nu(0_{x_\nu}) \in T_{x_\nu}M$ is the unique vector with $d\hat{J}_\nu(0_{x_\nu})[v] = g_{x_\nu}(\nabla\hat{J}_\nu(0_{x_\nu}), v)$, $\forall v \in T_{x_\nu}M$, and the Hessian $\nabla^2\hat{J}_\nu(0_{x_\nu}) : T_{x_\nu}M \to T_{x_\nu}M$ is such that $d^2\hat{J}_\nu(0_{x_\nu})[v,\cdot] = g_{x_\nu}(\nabla^2\hat{J}_\nu(0_{x_\nu})v, \cdot)$, again for all $v \in T_{x_\nu}M$. The solution $v_\nu \in T_{x_\nu}M$ of (4.33) generates the new iterate through the retraction

$$x_{\nu+1} = R_{x_\nu}(v_\nu).$$

Finally, acceptance of the new iterate and regulation of the trust-region radius ρ is handled as in the Euclidean case by looking at the quotient (4.32).

The main problem for the convergence analysis is the fact that at each iterate x_ν a new lifted objective function $\hat{J}_{x_\nu}$ is considered. Absil et al. [4] proved the following theorem.

Theorem 4.4.1 (Global convergence)**.** *Let $\{x_\nu\}$ be the sequence of iterates generated by the Riemannian trust-region algorithm with $\eta_1 \in [0, \frac{1}{4})$, and starting from an arbitrary initial iterate $x_0 \in M$. Suppose that the objective function J and the retraction*

R are smooth and the Riemannian manifold M is compact. Further assume that the approximate solutions v_ν to the local problems (4.33) *fulfill the* Cauchy descent criterion

$$m_{x_\nu}(0) - m_{x_\nu}(v_\nu) \geq c\|\nabla J(x_\nu)\| \min\left(\rho_\nu, \frac{\|\nabla J(x_\nu)\|}{\|\nabla^2 \hat{J}_\nu(0_{x_\nu})\|}\right), \tag{4.35}$$

where $\|\nabla^2 \hat{J}_\nu(0_{x_\nu})\|$ is the operator norm of the Hessian as a mapping $T_{x_\nu}M \to T_{x_\nu}M$ and c a positive constant. Then

$$\lim_{\nu\to\infty} \nabla J(x_\nu) = 0$$

holds.

Also, depending on the stopping criterion used for the inner solver we get superlinear or even quadratic local convergence [4].

Theorem 4.4.2 (Local convergence)**.** *Suppose that the objective function J and the retraction R are smooth and the Riemannian manifold M is compact. Let $x^* \in M$ be a nondegenerate local minimizer of J, i.e., $\nabla J(x^*) = 0$ and $g_{x^*}(\nabla^2 J(x^*)\cdot, \cdot)$ is positive definite. Then there exists a neighborhood V of x^* such that, for all $x_0 \in V$, the sequence $\{x_\nu\}$ generated by the Riemannian trust-region algorithm with the Steihaug–Toint algorithm [4] as the inner solver converges to x^*. Furthermore, there exists a $c > 0$ such that, for all sequences $\{x_\nu\}$ generated by the algorithm converging to x^*, there exists a $K > 0$ such that for all $\nu > K$*

$$\operatorname{dist}(x_{\nu+1}, x^*) \leq c(\operatorname{dist}(x_\nu, x^*))^{\min(\theta+1,2)}$$

where $\theta > 0$ depends on the stopping criterion of the solver for the quadratic subproblems.

4.5 A Trust-Region Solver for the Cosserat Rod with Hyperelastic Material

We will now apply the general Riemannian trust-region algorithm of the previous section to the discrete Cosserat rod problem (4.26). Call φ^* the continuous solution of a given rod problem and assume the grid to be fine enough such that there is a neighborhood V of $\mathcal{E}(\varphi^*) \in \mathrm{SE}(3)^n$ with a bijection between $V_h^{\mathrm{SE}(3)}$ and $\mathrm{SE}(3)^n$ in V. The existence of such a neighborhood is guaranteed by Lemma 4.3.2. We further suppose that if the initial iterate φ_0 is in V then so are all subsequent iterates. In particular, by Theorem 4.4.2, this is fulfilled if φ_0 is close enough to the solution. We can then replace the discrete problem (4.26) by the algebraic problem (4.27) and work on the manifold $M = \mathrm{SE}(3)^n$.

We denote elements of $\mathrm{SE}(3)^n$ by

$$(r, q) = \prod_{i=0}^{n-1} (r_i, q_i).$$

At any $(r, q) \in \mathrm{SE}(3)^n$, by (4.24) the tangent space is

$$T_{(r,q)}\mathrm{SE}(3)^n = \bigoplus_{i=0}^{n-1} \big(T_{r_i}\mathbb{R}^3 \oplus T_{q_i}\mathrm{SO}(3)\big) = \bigoplus_{i=0}^{n-1} \big(\mathbb{R}^3 \oplus q_i\mathbb{A}^3\big), \tag{4.36}$$

where the spaces $q_i\mathbb{A}^3$ have been defined in (4.4). The exponential map $\exp_{(r,q)} : T_{(r,q)}\mathrm{SE}(3)^n \to \mathrm{SE}(3)^n$ can be written as

$$\exp_{(r,q)}(w, q\hat{v}) = \prod_{i=0}^{n-1} \big(\exp_{r_i} w_i,\, \exp_{q_i} q_i\hat{v}_i\big) = \prod_{i=0}^{n-1} \big(r_i + w_i,\, q_i \exp \hat{v}_i\big), \tag{4.37}$$

where we have used (4.3) together with (4.6). Since the closed-form expressions (4.7) for the exponential map are available and easy to evaluate, we use (4.37) as the retraction.

We consider the algebraic rod energy functional (4.28)

$$\begin{aligned} j &: \mathrm{SE}(3)^n \to \mathbb{R} \\ j(\bar{\varphi}) &= \int_{[0,l]} W\big(\mathsf{u}(\mathcal{E}^{-1}(\bar{\varphi})) - \hat{\mathsf{u}}, \mathsf{v}(\mathcal{E}^{-1}(\bar{\varphi})) - \hat{\mathsf{v}}, s\big)\, ds. \end{aligned} \tag{4.38}$$

For $s \in [l_i, l_{i+1}]$, the mapping $\mathcal{E}^{-1} : \mathrm{SE}(3)^n \to V_h^{\mathrm{SE}(3)}$ is given element-wise by

$$\mathcal{E}^{-1}(\bar{\varphi})(s) = \left(\sum_{\substack{j\in\{0,1,2\} \\ k\in\{i,i+1\}}} (\bar{\varphi}_r)_{k,j}\boldsymbol{\psi}_{k,j}(s),\ \mathsf{q}\left(\frac{s - l_i}{l_{i+1} - l_i}, (\bar{\varphi}_q)_i, (\bar{\varphi}_q)_{i+1}\right)\right),$$

where q is the geodesic interpolation (4.22) with $\delta = 1$.

We now lift j onto the tangent bundle of $\mathrm{SE}(3)^n$. Let $(r_\nu, q_\nu) \in \mathrm{SE}(3)^n$ be the current iterate. Then, using the retraction (4.37), the lift of j onto $T_{(r_\nu,q_\nu)}\mathrm{SE}(3)^n$ is

$$\begin{aligned} \hat{\jmath}_\nu &: T_{(r_\nu,q_\nu)}\mathrm{SE}(3)^n \to \mathbb{R} \\ \hat{\jmath}_\nu(w, q_\nu\hat{v}) &= j(\exp_{(r_\nu,q_\nu)}(w, q_\nu\hat{v})) = j(r_\nu + w, q_\nu \exp \hat{v}). \end{aligned} \tag{4.39}$$

In order to obtain the quadratic model m_ν (4.34), we need to compute the differential $d\hat{\jmath}_\nu$ and the second derivative $d^2\hat{\jmath}_\nu$ of the lifted functional $\hat{\jmath}_\nu$. Using (4.25) we get

$$d\hat{\jmath}_\nu(0)[(w, q_\nu\hat{v})] = g_\nu(\nabla\hat{\jmath}_\nu, (w, q_\nu\hat{v})) = \langle \nabla_w\hat{\jmath}_\nu, w\rangle + \langle \nabla_v\hat{\jmath}_\nu, v\rangle$$

for all $(w, q_\nu\hat{v}) \in T_{(r_\nu,q_\nu)}\mathrm{SE}(3)^n$, where $\nabla_w \in \mathbb{R}^{3n}$ and $\nabla_v \in \bigoplus_i q_i\mathbb{A}^3$ denote the gradients with respect to w and v, respectively. Using (4.38) and (4.39) we get

$$\nabla_w\hat{\jmath} = \int_{[0,l]} \nabla_w W\big(\mathsf{u}(\mathcal{E}^{-1}(\bar{\varphi})) - \hat{\mathsf{u}}, \mathsf{v}(\mathcal{E}^{-1}(\bar{\varphi})) - \hat{\mathsf{v}}, s\big)\, ds. \tag{4.40}$$

and likewise for $\nabla_v\hat{\jmath}$. Let w_i^j be the coefficient of w pertaining to the i-th grid vertex and the canonical coordinate direction j. Then the coefficients of $\nabla_w W$ and $\nabla_v W$ are given

by

$$\frac{\partial W}{\partial w_i^j} = \sum_{m=1}^{3} A_m\big(\mathsf{v}_m(q) - \hat{\mathsf{v}}_m(s)\big)\frac{\partial}{\partial w_i^j}\mathsf{v}_m(r+\hat{r}, q\exp\hat{v}) \tag{4.41}$$

$$\begin{aligned}\frac{\partial W}{\partial v_i^j} &= \sum_{m=1}^{3} K_m\big(\mathsf{u}_m(q) - \hat{\mathsf{u}}_m(s)\big)\frac{\partial}{\partial v_i^j}\mathsf{u}_m(q\exp\hat{u}) \\ &\quad + \sum_{m=1}^{3} A_m\big(\mathsf{v}_m(q) - \hat{\mathsf{v}}_m(s)\big)\frac{\partial}{\partial v_i^j}\mathsf{v}_m(r+\hat{r}, q\exp\hat{v}).\end{aligned} \tag{4.42}$$

The derivatives of the strain measures u and v using the geodesic interpolation formulas of Sec. 4.3 are somewhat technical, and can be found in Appendix A. Deriving (4.41) and (4.42) once more we get the coefficients of the Hessian matrix

$$\frac{\partial^2 W}{\partial r_i^j\,\partial r_k^l} = \sum_{m=1}^{3} A_m\left[\frac{\partial \mathsf{v}_m}{\partial r_i^j}\cdot\frac{\partial \mathsf{v}_m}{\partial r_k^l} + (\mathsf{v}_m - \hat{\mathsf{v}}_m)\frac{\partial^2 \mathsf{v}_m}{\partial r_i^j\,\partial r_k^l}\right]$$

$$\frac{\partial^2 W}{\partial r_i^j\,\partial v_k^l} = \sum_{m=1}^{3} A_m\left[\frac{\partial \mathsf{v}_m}{\partial r_i^j}\cdot\frac{\partial \mathsf{v}_m}{\partial v_k^l} + (\mathsf{v}_m - \hat{\mathsf{v}}_m)\frac{\partial^2 \mathsf{v}_m}{\partial r_i^j\,\partial v_k^l}\right]$$

$$\begin{aligned}\frac{\partial^2 W}{\partial v_i^j\,\partial v_k^l} &= \sum_{m=1}^{3} K_m\left[\frac{\partial \mathsf{u}_m}{\partial v_i^j}\cdot\frac{\partial \mathsf{u}_m}{\partial v_k^l} + (\mathsf{u}_m - \hat{\mathsf{u}}_m)\frac{\partial^2 \mathsf{u}_m}{\partial v_i^j\,\partial v_k^l}\right] \\ &\quad + \sum_{m=1}^{3} A_m\left[\frac{\partial \mathsf{v}_m}{\partial v_i^j}\cdot\frac{\partial \mathsf{v}_m}{\partial v_k^l} + (\mathsf{v}_m - \hat{\mathsf{v}}_m)\frac{\partial^2 \mathsf{v}_m}{\partial v_i^j\,\partial v_k^l}\right],\end{aligned}$$

and the remaining terms by symmetry of the Hessian. In principle these expressions can also be computed analytically. However, in order to avoid the very unwieldy formulas which result we opted for an approximation of the Hessian matrix by a finite difference scheme. This is covered by the convergence theory which allows approximations to the Hessian matrix.

In summary, we obtain the inner trust-region problem

$$(w_\nu, q_\nu\hat{v}_\nu) = \arg\min m_\nu((w, q_\nu\hat{v})), \qquad \|(w, q_\nu\hat{v})\|_{g_{(r_\nu,q_\nu)}} \le \rho_\nu, \tag{4.43}$$

where the minimization is over all $(w, q_\nu\hat{v}) \in T_{(r_\nu,q_\nu)}\mathrm{SE}(3)^n$ and

$$m_\nu((w, q_\nu\hat{v})) = \hat{\jmath}_\nu(0) + d\hat{\jmath}_\nu(0)[(w, q_\nu\hat{v})] + \frac{1}{2}d^2\hat{\jmath}_\nu(0)[(w, q_\nu\hat{v}), (w, q_\nu\hat{v})]. \tag{4.44}$$

The functional m_ν is quadratic, but not necessarily convex. Being a minimization problem of a continuous function on a compact set, Problem (4.43) has at least one solution $(w_\nu, q_\nu\hat{v}_\nu) \in T_{(r_\nu,q_\nu)}\mathrm{SE}(3)^n$. The next iterate is then given by

$$(r_{\nu+1}, q_{\nu+1}) = \exp_{(r_\nu,q_\nu)}(w_\nu, q_\nu\hat{v}_\nu) = (r_\nu + w_\nu, q_\nu\exp\hat{v}_\nu).$$

Various solvers have been proposed for problems like (4.43). In their article on Riemannian trust-region algorithms, Absil et al. used the Steihaug-Toint algorithm [4], and their local convergence result relies on the use of this solver. With future generalizations to higher-dimensional problems in mind we instead apply a multigrid solver to the inner quadratic problems. For this note that the trust-region convergence theory allows norms other that $\|\cdot\|_{g_{x_\nu}}$ for the definition of the trust-region. We therefore introduce a generalization of the maximum norm on the tangent bundle of $\mathrm{SE}(3)^n$ by defining

$$\begin{aligned} \|\cdot\|_{\infty,T\mathrm{SE}(3)^n} &: T\mathrm{SE}(3)^n \to \mathbb{R}_0^+ \\ T_{(r,q)}\mathrm{SE}(3)^n \ni (w, q\hat{v}) &\to \max\{\|w\|_\infty, \|v\|_\infty\}. \end{aligned} \tag{4.45}$$

Using this norm, the trust-region subproblem (4.43) reads

$$(w_\nu, q_\nu\hat{v}_\nu) = \arg\min(m_\nu((w, q_\nu\hat{v})) \qquad \|(w, q_\nu\hat{v})\|_{\infty,T\mathrm{SE}(3)^n} \le \rho_\nu, \tag{4.46}$$

with m_ν given as before. The new trust-region

$$K^{\mathrm{tr}}_{\infty,\nu} = \left\{(w, q_\nu\hat{v}) \in T_{(r_\nu,q_\nu)}\mathrm{SE}(3)^n \;\middle|\; \|(w, q_\nu\hat{v})\|_{\infty,T\mathrm{SE}(3)^n} \le \rho_\nu\right\} = [-\rho_\nu, \rho_\nu]^{6n}$$

has the tensor product structure (3.42) needed to apply Lemma 3.4.1. Hence it seems natural to apply a monotone multigrid method (MMG) to the inner problems (4.46). The monotone multigrid method is known as an efficient and globally convergent algorithm for quadratic convex minimization problems on sets having the product structure (3.42). Even though the functions (4.44) may be nonconvex we expect MMG to perform well on the inner problems (4.46) for the following reasons. At a given outer iteration ν if m_ν is convex then, by Thm. 3.4.1, the monotone multigrid method will converge to the solution of (4.46). If m_ν is not convex we cannot prove that MMG produces sufficient energy decrease. To make sure that the monotone multigrid method fulfills the Cauchy descent criterion (4.35) we precede each multigrid iteration with a gradient descent step. This *gradient-prefixed* monotone multigrid method trivially fulfills the Cauchy descent criterion.

Lemma 4.5.1. *Let $J : \mathbb{R}^n \to \mathbb{R}$ be a quadratic, possibly nonconvex functional, and let $\mathcal{K} = \{x \in \mathbb{R}^n \mid x_i \in [a_i, b_i]\}$ be compact. Then the gradient-prefixed monotone multigrid method fulfills the Cauchy descent criterion.*

Since MMG is a descent method which produces very good iterates on convex problems we also expect it to produces large energy decrease on nonconvex problems.

The convergence analysis given in [4] for the case of a Riemannian-norm trust-region and the Steihaug-Toint algorithm for the solution of the inner problems can be generalized to the maximum norm trust-region, but the details are beyond the scope of this work.

Remark 4.5.1. The Truncated Nonsmooth Newton Multigrid method of Sec. 3.4 cannot be used to solve the problems (4.43) with the norm (4.45). Unlike the monotone multigrid method it solves unconstrained coarse grid minimization problems which do not have a solution if the functional m_ν is not convex.

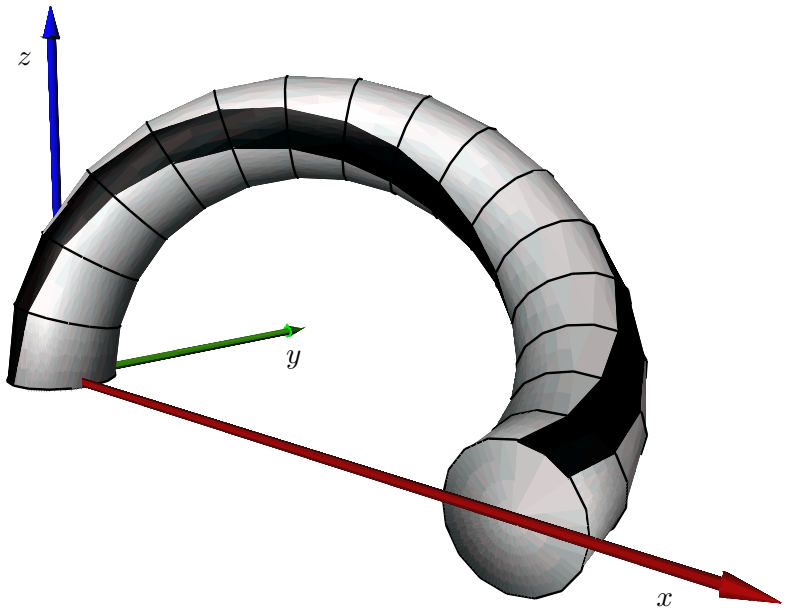


Figure 4.5: Solution of the example rod problem using a grid with 16 elements.

Remark 4.5.2. The contributions from $T\mathbb{R}^{3n}$ and $T\mathrm{SO}(3)^n$ of the corrections $(w, q\hat{v}) \in T\mathrm{SE}(3)^n$ can differ vastly in their scaling. While the $q\hat{v} \in T_q\mathrm{SO}(3)^n$ are infinitesimal rotations which are invariant under scaling the $w \in T_r\mathbb{R}^{3n}$ are corrections to the position of the centerline. As such, they depend on the overall size of the rod. It may therefore be advantageous to replace the norm (4.45) by a scaled norm which compensates for this disparity.

4.6 Numerical Results

We close this chapter by giving a short example demonstrating the efficiency of the trust-region solver for rod problems. We are particularly interested in pure Dirichlet problems since that is what will need to be solved as part of a Dirichlet–Neumann algorithm in Chapters 5 and 6.

Consider a rod of unit length which is completely straight and aligned with the z-axis in its unstressed state. The cross-section is circular with a radius of 0.05 length units. We choose a linear material law with parameters $E = 2.5 \cdot 10^5$ pressure units and $\nu = 0.3$. Using the formulas (4.17) and (4.18), this leads to approximately $A = (755, 755, 1963)$ and $K = (1.23, 1.23, 0.94)$.

We clamp the first end of the rod at the origin. The second endpoint of the centerline is placed at $(1/2, 0, 0)$, with the cross-section such that $\mathbf{d}_1 = (1, 0, 0)$, $\mathbf{d}_2 = (0, 0, 1)$, and $\mathbf{d}_3 = (0, -1, 0)$. The solution configuration for these boundary conditions contains stretching, shear, bending, and torsion (see Fig. 4.5).

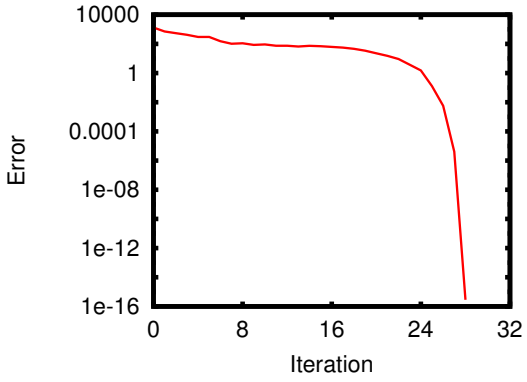


Figure 4.6: Error $|\varphi_\nu|_*$ per iteration ν for a grid consisting of 4096 elements.

levels	2	3	4	5	6	7	8	9	10	11	12
elements	4	8	16	32	64	128	256	512	1024	2048	4096
overall it.	16	14	25	19	24	26	19	29	34	30	30
unsuccessful it.	2	1	2	1	3	4	2	3	5	1	1

Table 4.1: Number of overall iterations and unsuccessful iterations of the trust-region solver per grid size.

To measure the convergence speed we first computed a reference solution φ^* by letting the solver iterate until the maximum norm of the correction $\|\exp_{\varphi_\nu}^{-1} \varphi_{\nu+1}\|_{\infty,T\mathrm{SE}(3)^n}$ dropped below 10^{-12}. We then computed the Hessian matrix H^* of the lifted energy functional $\hat{\jmath}_*$ at the last iterate φ^*. Assuming φ^* to be close to a minimum of j we get convexity of $\hat{\jmath}_*$ and hence H^* is positive definite. It therefore creates an energy norm on $T_{\varphi^*}\mathrm{SE}(3)^n$ which we denote by $\|\cdot\|_{H^*}$. Revisiting now the iteration history $\varphi_0, \ldots, \varphi^*$ we transport each φ_ν onto $T_{\varphi^*}\mathrm{SE}(3)^n$ using the inverse exponential map at φ^* and define the error of φ_ν as

$$|\varphi_\nu|_* = \|\exp_{\varphi^*}^{-1} \varphi_\nu\|_{H^*}.$$

All experiments in this section used reduced integration to avoid locking problems [73]. The inner problems where solved to a precision of 10^{-13} by tracking the H^1-norm of the relative corrections. The initial trust-region had a radius of 1 and the parameters η_1 and η_2 where set to 0.01 and 0.9, respectively.

We first have a look at the convergence behavior of the trust-region solver itself. In view of Theorems 4.4.1 and 4.4.2 we expect global convergence and local convergence at least close to quadratic. Fig. 4.6 shows the error per iteration on a grid of 4096 elements. The initial iterate is the unstressed reference configuration where the rod is straight along the z-axis. One can see good global convergence in spite of this fairly remote starting iterate. The sharp drop starting around the 24th iteration confirms the predictions of Theorem 4.4.2 concerning fast local convergence.

Next we investigate the behavior of the convergence with respect to grid size. In analogy to [91] one would hope for asymptotically grid-independent convergence. Starting from a one-level grid with two elements of equal size we used uniform refinement to create

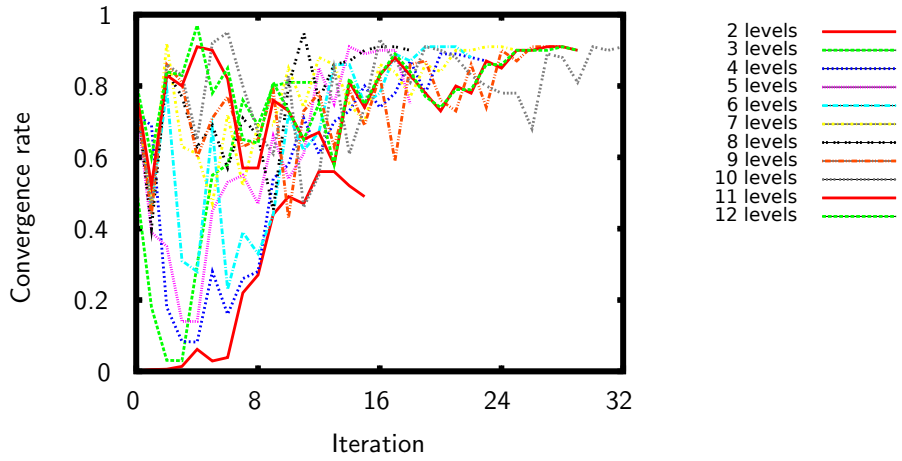


Figure 4.7: Convergence rates for the inner multigrid solver per outer iteration on grid hierarchies of up to 12 levels.

a set of test grids. These grids have between two and twelve levels and correspondingly the range of element numbers goes from four to 4096. We ran the same problem as in the previous paragraph on each of these grids. Table 4.1 shows the number of trust-region iterations. They appear to be bounded from above, marking another desirable property of the algorithm. We must mention though that we noted a dependence of the trust-region convergence on the rod thickness, with the iteration numbers deteriorating for decreasing rod thickness.

Finally, the convergence rates of the inner multigrid solver are of interest. These are plotted in Fig. 4.7. They were computed as follows. At a given trust-region iteration ν let $w^0, \dots, w^k$ be the sequence of multigrid iterates and let

$$\eta_i = \frac{\|w^{i+1} - w^i\|}{\|w^i - w^{i-1}\|} \tag{4.47}$$

be the approximate convergence rate at multigrid iteration i. The norm used here is the energy norm of a Laplace problem on $T_{\varphi_\nu}\mathrm{SE}(3)^n$, which is equivalent to the H^1-norm. The average approximate convergence rate is then

$$\eta^\nu = \left(\prod_{i=0}^{k-1} \eta_i \right)^{1/k}.$$

From Fig. 4.7 it is unclear whether the multigrid convergence rates are bounded away from 1. In particular they are not as good as one would expect. This will be investigated further in the future. Again we noted a dependence on the rod thickness.

Remark 4.6.1. In view of Lemma 3.4.1 one would expect the trust-region algorithm to also work well with an inexact inner solver. Initial numerical experiments confirm this.

5 Coupling Rods and Three-Dimensional Objects

In the two previous chapters we have covered the modeling of bones and ligaments. Bones were described using three-dimensional linear elasticity, while for the ligaments we opted for one-dimensional Cosserat rods. This chapter treats the coupling of these two models. The anatomical bone–ligament connections, or *insertions*, have been described in Sec. 2.4 as rigid junctions. We propose a way to model such rigid junctions between a three-dimensional and a one-dimensional object, and we prove existence of solutions for the coupled problem under certain symmetry conditions. As it turns out, the main difficulty for the analysis is not the difference in dimensions but the nonlinearity of the rod problem. Further, we introduce a Dirichlet–Neumann algorithm for the solution of the coupled problem and numerically investigate its properties with a series of test problems. Simulation results for an actual human knee involving bones and ligaments will be given in Sec. 6.2.

Lagnese et al. [63] have studied the coupling of beams to plates extensively. In their work, however, the main focus is on the linearized equations. Modeling of 3d-2d junctions between linear elastic objects of different dimensions using a method of asymptotic expansion has been carried out by Ciarlet et al. [24]. Monaghan et al. [69] describe a 3d-1d coupling between linear elastic elements in the discrete setting, while Formaggia et al. [37] couple 3d and 1d variants of the Navier–Stokes equations in a simulation of blood circulation. We are not aware of previous work on the coupling of three-dimensional linear elastic objects to Cosserat rods.

5.1 Homogeneous Coupling in Nonlinear Elasticity

In order to motivate our choice of coupling conditions we start by deriving the corresponding conditions for the coupling of two d-dimensional nonlinear elastic objects. Let Ω be a bounded, open, connected domain in $\mathbb{R}^d$ with $d \in \{2, 3\}$. Its boundary $\partial\Omega$ is supposed to be Lipschitz and to consist of two disjoint parts Γ_N and Γ_D such that $\partial\Omega = \overline{\Gamma}_N \cup \overline{\Gamma}_D$ and Γ_D has positive $(d-1)$-dimensional measure. We use $\boldsymbol{\nu}$ to denote the outward unit normal of Ω. For any displacement function $\mathbf{u} \in \mathbf{H}^1(\Omega)$ we set

$$\mathbf{E} = \frac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^T + \nabla\mathbf{u}^T\nabla\mathbf{u})$$

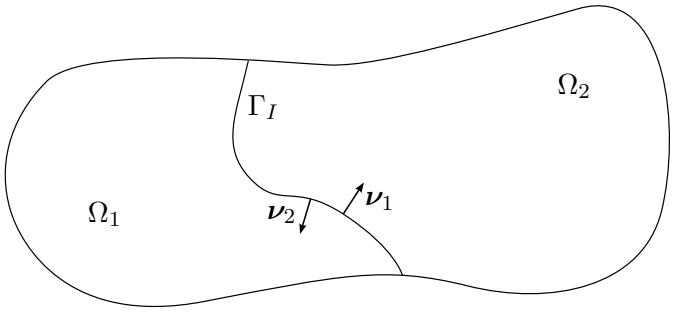


Figure 5.1: Coupling between two domains of equal dimension.

the *nonlinear* strain tensor (cf. Sec 3.1) and $\boldsymbol{\sigma} : \Omega \to \mathbb{S}^d$ the second Piola–Kirchhoff stress tensor of a differentiable, but possibly nonlinear material law

$$\begin{aligned} \tilde{\boldsymbol{\sigma}} &: \mathbb{S}^d \times \Omega \to \mathbb{S}^d \\ \tilde{\boldsymbol{\sigma}}(\mathbf{E}(x), x) &= \boldsymbol{\sigma}(x). \end{aligned}$$

In an abuse of notation we will frequently write $\boldsymbol{\sigma}(\mathbf{v})$ instead of $\tilde{\boldsymbol{\sigma}}(\mathbf{E}(\mathbf{v})(x), x)$ for $\mathbf{v} \in \mathbf{H}^1(\Omega)$. The boundary value problem of elasticity in its strong form is (Sec. 3.1)

$$-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}) = \mathbf{f} \qquad \text{in } \Omega, \tag{5.1a}$$

$$\mathbf{u} = 0 \qquad \text{on } \Gamma_D, \tag{5.1b}$$

$$\boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu} = \mathbf{t} \qquad \text{on } \Gamma_N. \tag{5.1c}$$

The vector field $\mathbf{f} : \Omega \to \mathbb{R}^d$ describes the volume forces, and $\mathbf{t} : \Gamma_N \to \mathbb{R}^d$ the surface forces. Condition (5.1b) clamps the body at the Dirichlet boundary. We write $\mathbf{H}^1_0(\Omega)$ to denote the space of d-valued first-order Sobolev functions on Ω which are zero in the sense of traces on Γ_D. Multiplying (5.1a) by test functions $\mathbf{v} \in \mathbf{H}^1_0(\Omega)$ and applying Green's formula we obtain the weak formulation

$$\text{find } \mathbf{u} \in \mathbf{H}^1_0(\Omega) \quad : \quad a(\mathbf{u}, \mathbf{v}) = l(\mathbf{v}), \qquad \text{for all } \mathbf{v} \in \mathbf{H}^1_0(\Omega), \tag{5.2}$$

with

$$a(\mathbf{v}, \mathbf{w}) = \int_\Omega \boldsymbol{\sigma}(\mathbf{v}) : \boldsymbol{\nabla}\mathbf{w} \, dx,$$

and $l(\cdot)$ given by (3.13). Note that $a(\cdot, \cdot)$ is nonlinear in its first argument. However, when $\boldsymbol{\sigma}$ is given by the linear material law (3.8) and the linear strain tensor $\boldsymbol{\varepsilon}$ is used instead of $\mathbf{E}$ then $a(\cdot, \cdot)$ reduces to the bilinear form $a(\cdot, \cdot)$ defined in (3.12).

Now suppose that Ω consists of two nonoverlapping, connected subdomains Ω_1, Ω_2 (Fig. 5.1). We denote by $\Gamma_I = \overline{\Omega}_1 \cap \overline{\Omega}_2$ the *interface*, assuming that it is a sufficiently smooth $(d-1)$-dimensional manifold, and that $\overline{\Gamma}_I \cap \overline{\Gamma}_D = \emptyset$. With these conditions the space of traces of functions in $\mathbf{H}^1_0(\Omega)$ on Γ_I is $\mathbf{H}^{1/2}(\Gamma_I)$. As in Sec. 3.2 we use $\mathbf{u}_i$, $\mathbf{f}_i$,

$\mathbf{t}_i$ to denote the restrictions of $\mathbf{u}$, $\mathbf{f}$, $\mathbf{t}$ to Ω_i for $i \in \{1,2\}$, and $\boldsymbol{\nu}_i$ for the outward unit normal of Ω_i. Further, for $i \in \{1,2\}$ we define

$$a_i(\mathbf{v},\mathbf{w}) = \int_{\Omega_i} \boldsymbol{\sigma}(\mathbf{v}) : \boldsymbol{\nabla}\mathbf{w}\, dx \qquad \text{and} \qquad l_i(\mathbf{v}) = \int_{\Omega_i} \mathbf{f}\mathbf{v}\, dx + \int_{\Gamma_N \cap \partial\Omega_i} \mathbf{t}\mathbf{v}\, ds.$$

The weak problem (5.2) is equivalent to solving separate problems on Ω_i, $i \in \{1,2\}$, together with certain coupling conditions. This is made formal by the following lemma. Denote by $\mathbf{H}^1_0(\Omega_i)$ the space of d-valued first-order Sobolev functions on Ω_i which are zero in the sense of traces on $\Gamma_D \cap \partial\Omega_i$ and by $\mathbf{H}^1_*(\Omega_i)$ the subspace of $\mathbf{H}^1_0(\Omega_i)$ of functions which are zero on Γ_I.

Lemma 5.1.1. *The weak nonlinear elasticity problem* (5.2) *is equivalent to finding* $\mathbf{u}_i \in \mathbf{H}^1_0(\Omega_i)$, $i = 1,2$, *such that*

$$a_i(\mathbf{u}_i,\mathbf{v}_i) = l_i(\mathbf{v}_i) \qquad \text{for all } \mathbf{v}_i \in \mathbf{H}^1_*(\Omega_i),\ i = 1,2 \tag{5.3a}$$

$$\mathbf{u}_1 = \mathbf{u}_2 \qquad \text{on } \Gamma_I \tag{5.3b}$$

$$\sum_{i=1}^{2} a_i(\mathbf{u}_i, \mathcal{R}_i\boldsymbol{\mu}) = \sum_{i=1}^{2} l_i(\mathcal{R}_i\boldsymbol{\mu}) \qquad \text{for all } \boldsymbol{\mu} \in \mathbf{H}^{1/2}(\Gamma_I), \tag{5.3c}$$

where $\mathcal{R}_i$ *denotes any possible continuous extension operator from* $\mathbf{H}^{1/2}(\Gamma_I)$ *to* $\mathbf{H}^1_0(\Omega_i)$.

Proof. We generalize the proof of Lemma 1.2.1 in [77]. Let first $\mathbf{u}$ be a solution of (5.2), and set $\mathbf{u}_i = \mathbf{u}|_{\Omega_i}$, $i = 1,2$. Then $\mathbf{u}_i \in \mathbf{H}^1_0(\Omega_i)$ and (5.3b) is satisfied by the trace theorem. We also get (5.3a) since any test function $\mathbf{v}_i \in \mathbf{H}^1_*(\Omega_i)$ can be extended by zero to a test function in $\mathbf{H}^1_0(\Omega)$ [14, Lem. 3.2.3]. Next, for $\boldsymbol{\mu} \in \mathbf{H}^{1/2}(\Gamma_I)$ define $\mathcal{R}\boldsymbol{\mu}$ as

$$\mathcal{R}\boldsymbol{\mu} = \begin{cases} \mathcal{R}_1\boldsymbol{\mu} & \text{in } \Omega_1 \\ \mathcal{R}_2\boldsymbol{\mu} & \text{in } \Omega_2. \end{cases}$$

Then the function $\mathcal{R}\boldsymbol{\mu}$ belongs to $\mathbf{H}^1_0(\Omega)$ [14, Lem. 3.2.3], and hence by (5.2) we have

$$a(\mathbf{u}, \mathcal{R}\boldsymbol{\mu}) = l(\mathcal{R}\boldsymbol{\mu}),$$

which is equivalent to (5.3c). To see the other direction, let the pair $\mathbf{u}_i \in \mathbf{H}^1_0(\Omega_i)$, $i = 1,2$, be a solution of (5.3), and set

$$\mathbf{u} = \begin{cases} \mathbf{u}_1 & \text{in } \Omega_1 \\ \mathbf{u}_2 & \text{in } \Omega_2. \end{cases}$$

From (5.3b) it follows that $\mathbf{u} \in \mathbf{H}^1_0(\Omega)$ [14, Lem. 3.2.3]. Take a test function $\mathbf{v} \in \mathbf{H}^1_0(\Omega)$ and set $\boldsymbol{\mu} = \mathbf{v}|_{\Gamma_I} \in \mathbf{H}^{1/2}(\Gamma_I)$. Then $(\mathbf{v}|_{\Omega_i} - \mathcal{R}_i\boldsymbol{\mu}) \in \mathbf{H}^1_*(\Omega_i)$ and from (5.3a) and (5.3c) follows that

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) &= \sum_{i=1}^{2} \Big[a_i(\mathbf{u}_i, \mathbf{v}|_{\Omega_i} - \mathcal{R}_i\boldsymbol{\mu}) + a_i(\mathbf{u}_i, \mathcal{R}_i\boldsymbol{\mu}) \Big] \\ &= \sum_{i=1}^{2} \Big[l_i(\mathbf{v}|_{\Omega_i} - \mathcal{R}_i\boldsymbol{\mu}) + l_i(\mathcal{R}_i\boldsymbol{\mu}) \Big] \\ &= l(\mathbf{v}), \end{aligned}$$

which is (5.2). □

A similar result holds when there are additional inequality constraints arising, for example, from the modeling of contact. The proof, however, is more complicated (cf. [14, Prop. 3.2.4]).

When solutions of (5.3) are smooth enough they also solve a corresponding strong boundary value problem.

Lemma 5.1.2. *Let* $\mathbf{u}_i$*,* $i = 1, 2$*, be a solution of* (5.3) *and sufficiently regular. Then* $\mathbf{u}_i$ *also solves the boundary value problem*

$$\begin{aligned} -\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}_i) &= \mathbf{f}_i && \text{in } \Omega_i, \\ \mathbf{u}_i &= 0 && \text{on } \Gamma_D \cap \partial\Omega_i, \\ \boldsymbol{\sigma}(\mathbf{u}_i)\boldsymbol{\nu}_i &= \mathbf{t}_i && \text{on } \Gamma_N \cap \partial\Omega_i, \end{aligned} \tag{5.4a}$$

for $i = 1, 2$ *together with*

$$\mathbf{u}_1 = \mathbf{u}_2 \qquad \text{on } \Gamma_I, \tag{5.4b}$$

$$\boldsymbol{\sigma}(\mathbf{u}_1)\boldsymbol{\nu}_1 = -\boldsymbol{\sigma}(\mathbf{u}_2)\boldsymbol{\nu}_2 \qquad \text{on } \Gamma_I. \tag{5.4c}$$

Proof. We show only that the weak continuity of normal stresses (5.3c) follows from the strong form (5.4c). Multiply (5.4c) by a test function $\boldsymbol{\mu} \in \mathbf{H}^{1/2}(\Gamma_I)$ and integrate over Γ_I to obtain

$$\int_{\Gamma_I} \boldsymbol{\sigma}(\mathbf{u}_1)\boldsymbol{\nu}_1 \boldsymbol{\mu}\, ds = -\int_{\Gamma_I} \boldsymbol{\sigma}(\mathbf{u}_2)\boldsymbol{\nu}_2 \boldsymbol{\mu}\, ds.$$

Let $\mathcal{R}_i$ be some continuous extension operator from $\mathbf{H}^{1/2}(\Gamma_I)$ to $\mathbf{H}_0^1(\Omega_i)$. Multiply (5.4a) by $\mathcal{R}_i\boldsymbol{\mu}$, integrate over Ω_i, and use Green's formula to get

$$0 = a_i(\mathbf{u}, \mathcal{R}_i\boldsymbol{\mu}) - l_i(\mathcal{R}_i\boldsymbol{\mu}) - \int_{\Gamma_I} \boldsymbol{\sigma}(\mathbf{u}_i)\boldsymbol{\nu}_i \boldsymbol{\mu}\, ds,$$

for $i = 1, 2$. Combining these two equalities gives (5.3c). □

In conclusion, given sufficient regularity, the boundary value problem (5.1) is equivalent to the strong coupled formulation (5.4).

5.2 Heterogeneous Coupling Conditions

According to the general theory of Quarteroni and Valli [77], coupling conditions for elliptic problems can be written formally as

$$\Phi(\mathbf{u}_1) = \Phi(\mathbf{u}_2), \qquad \Psi(\mathbf{u}_1) = \Psi(\mathbf{u}_2), \tag{5.5}$$

for two functionals Φ, Ψ. The first equality will generally involve the primal variables, which in our case are the displacements. The second equality relates the dual variables,

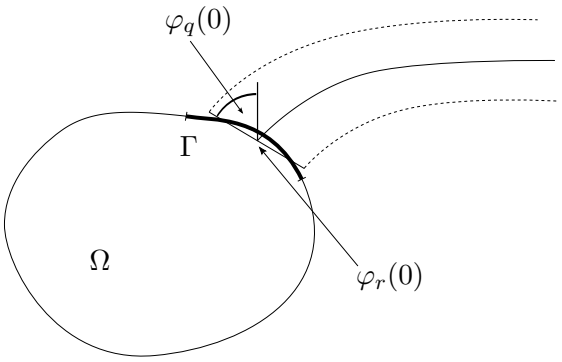


Figure 5.2: Coupling between a two-dimensional domain and a rod.

i.e., the stresses, to each other. The case of nonlinear elasticity described in the previous section is covered by this theory with

$$\Phi(\mathbf{v}) = \mathbf{v}|_{\Gamma_I} \qquad \text{and} \qquad \Psi(\mathbf{v}) = \boldsymbol{\sigma}(\mathbf{v})\boldsymbol{\nu}_1.$$

With this choice (5.5) corresponds to the strong conditions (5.4b) and (5.4c).

We will now derive conditions for the coupling of a linear elastic three-dimensional object and a Cosserat rod. These heterogeneous conditions will be special instances of the general equations (5.5). We use the results of the previous section as a motivation. From now on we consider the case $d = 3$ exclusively. However, all results also hold for $d = 2$. Let $\Omega \subset \mathbb{R}^3$ be an open, bounded, connected domain with a sufficiently smooth boundary. We denote the outward unit normal by $\boldsymbol{\nu}$. The boundary $\partial\Omega$ is supposed to consist of three disjoint parts Γ_D, Γ_N, and Γ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}$. We assume that Γ_D and Γ have positive two-dimensional measure. The three-dimensional object represented by Ω will couple with the rod across Γ, which we call the coupling boundary. Consider also a Cosserat rod defined on the interval $[0, l]$. The boundary of the one-dimensional parameter domain $[0, l]$ consists of the two points 0 and l, and the respective domain normals are $\boldsymbol{\nu}_0 = -1$ and $\boldsymbol{\nu}_l = 1$. To be specific, we pick 0 as the coupling boundary. Let $\varphi : [0, l] \to \mathrm{SE}(3)$ be a rod configuration. For any $s \in [0, l]$ we denote by $\varphi_r(s) \in \mathbb{R}^3$ the position of the centerline at s and by $\varphi_q(s) \in \mathrm{SO}(3)$ the orientation of the cross-section at s. We further call $\mathbf{n}(s) \in \mathbb{R}^3$ the total force transmitted across the rod cross-section at s and $\mathbf{m}(s) \in \mathbb{R}^3$ the total moment about $\varphi_r(s)$ transmitted across the cross-section at s. We assume a stress-free configuration $\hat{\varphi} : [0, l] \to \mathrm{SE}(3)$ such that $\hat{\varphi}_r(0) = |\Gamma|^{-1} \int_\Gamma x \, ds$, i.e., the coupling interface of the rod in its stress-free state is placed at the center of gravity of the coupling interface of Ω (Fig. 5.2).

We begin by looking for coupling conditions for the dual variables. For the linear elastic 3d object these are the normal stresses $\boldsymbol{\sigma}\boldsymbol{\nu}$ on Γ_I that appear in (5.4c). The corresponding quantities in the rod are the total force $\mathbf{n}(0)\boldsymbol{\nu}_0$ and the total moment $\mathbf{m}(0)\boldsymbol{\nu}_0$ about $\varphi_r(0)$ transmitted in normal direction across the cross-section at $s = 0$. To relate these quantities note that in the full-dimensional case it follows from the pointwise condition (5.4c) that the total force and moment transmitted across the interface are

preserved. In formulas,

$$\int_{\Gamma_I} \boldsymbol{\sigma}(\mathbf{u}_1)\boldsymbol{\nu}_1 \, ds = -\int_{\Gamma_I} \boldsymbol{\sigma}(\mathbf{u}_2)\boldsymbol{\nu}_2 \, ds \tag{5.6}$$

and

$$\int_{\Gamma_I} (x - x_0) \times \boldsymbol{\sigma}(\mathbf{u}_1)\boldsymbol{\nu}_1 \, ds = -\int_{\Gamma_I} (x - x_0) \times \boldsymbol{\sigma}(\mathbf{u}_2)\boldsymbol{\nu}_2 \, ds \tag{5.7}$$

for any fixed $x_0 \in \mathbb{R}^3$. These are six equations in six unknowns, namely the three components of the total force and the three components of the total moment transmitted across Γ_I. Choosing $x_0 = \varphi_r(0)$, these quantities are exactly the dual variables of the rod problem. We therefore replace one side of the equations (5.6) and (5.7) with the corresponding quantities of the rod and get

$$\int_{\Gamma} \boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu} \, ds = -\mathbf{n}(0)\boldsymbol{\nu}_0$$
$$\int_{\Gamma} (x - \varphi_r(0)) \times (\boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu}) \, ds = -\mathbf{m}(0)\boldsymbol{\nu}_0.$$

These are our heterogeneous coupling conditions for the dual variables. Note that we do not assume that Γ has the same shape or area as the rod cross-section at $s = 0$.

We now turn our attention to coupling conditions for the primal variables. From (5.4b) we easily get that

$$\frac{1}{|\Gamma_I|} \int_{\Gamma_I} (\mathbf{u}_1(x) + x) \, ds = \frac{1}{|\Gamma_I|} \int_{\Gamma_I} (\mathbf{u}_2(x) + x) \, ds, \tag{5.8}$$

which is the equality of the center of gravity of Γ_I transformed under $\mathbf{u}_1$ and $\mathbf{u}_2$. The direct equivalent to this in a rod problem is the position of the centerline, and we obtain a first primal condition for the heterogeneous problem

$$\frac{1}{|\Gamma|} \int_{\Gamma} (\mathbf{u}(x) + x) \, ds = \varphi_r(0). \tag{5.9}$$

To obtain a complete set of primal conditions we also need to relate the orientations at the interface. This requires some technical preparations. Using the deformation gradient $\boldsymbol{\nabla}(\mathbf{u} + \mathrm{Id})$ (3.2) we first define the average deformation of the interface boundary Γ

$$\mathcal{F}(\mathbf{u}) = \frac{1}{|\Gamma|} \int_{\Gamma} \boldsymbol{\nabla}(\mathbf{u}(x) + x) \, ds. \tag{5.10}$$

As long as $\mathbf{u}$ is sufficiently well-behaved the matrix $\mathcal{F}(\mathbf{u})$ has a positive determinant. Using the following lemma it can then be split up into a rotation and a stretching.

Lemma 5.2.1 (Polar decomposition, [25])**.** *Let A be a quadratic, nonsingular matrix. Then there is a unique decomposition*

$$A = OH, \tag{5.11}$$

with O orthogonal and H symmetric and positive definite.

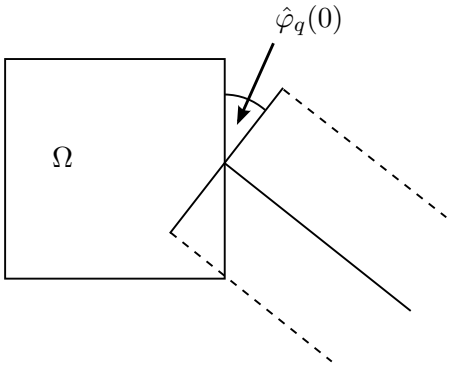


Figure 5.3: In the stress-free configuration the rod may meet the body at an arbitrary spatial angle $\hat{\varphi}_q(0)$.

The rotational part of the deformation gradient is known as the continuum rotation [71]. We define the average orientation of Γ induced by a deformation $\mathbf{u}$ as the rotational part of $\mathcal{F}(\mathbf{u})$.

Definition 5.2.1 (Average orientation)**.** *Let $\mathbf{u} \in \mathbf{H}^1(\Omega)$ be such that the polar decomposition*

$$\mathcal{F}(\mathbf{u}) = O_\Gamma(\mathbf{u})H(\mathbf{u})$$

exists. Then we call $O_\Gamma(\mathbf{u})$ the average orientation of Γ under $\mathbf{u}$.

Note that the average orientation $O_\Gamma(\mathbf{u})$ reproduces rigid body motions of the interface in the sense that $O_\Gamma(\mathbf{u}) = Q$ if $\mathbf{u}(x) + x = Qx + \mathbf{a}$ for a $Q \in \mathrm{SO}(3)$ and a vector $\mathbf{a} \in \mathbb{R}^3$ in a neighborhood of Γ. In particular, if $\mathbf{u} \equiv 0$ then $O_\Gamma = \mathrm{Id}$.

The average orientation $O_\Gamma(\mathbf{u})$ can now be set in relation to $\varphi_q(0)$, the orientation of the cross-section at $s = 0$. We require the coupling condition to be fulfilled by the stress-free configuration $\mathbf{u} = 0$, $\varphi = \hat{\varphi}$. This leads to the condition

$$O_\Gamma(\mathbf{u})\hat{\varphi}_q(0) = \varphi_q(0). \tag{5.12}$$

This is an equation in the three-dimensional space $\mathrm{SO}(3)$. Together with (5.9) we get six independent conditions for the six primal variables.

The factor $\hat{\varphi}_q(0) \in \mathrm{SO}(3)$ in (5.12) can be seen as a free parameter in the coupling conditions which specifies the spatial angle at which the rod meets the three-dimensional object (Fig. 5.3). It is part of the problem description.

For later reference we will write the primal coupling conditions in an alternative form. Introduce the averaging operator $\mathrm{Av} : \mathbf{H}^1(\Omega) \to \mathrm{SE}(3)$ by setting

$$\mathrm{Av}(\mathbf{u}) = \Big(\frac{1}{|\Gamma|}\int_\Gamma (\mathbf{u}(x) + x)\, ds,\, O_\Gamma(\mathbf{u})\hat{\varphi}_q(0)\Big), \tag{5.13}$$

where we have used $(\cdot,\cdot)$ to denote elements of the product space $\mathrm{SE}(3) = \mathbb{R}^3 \times \mathrm{SO}(3)$. Then (5.9) and (5.12) can be written concisely as

$$\mathrm{Av}(\mathbf{u}) = \varphi(0).$$

We now state the entire heterogeneous coupling problem in its strong form. For the 3d object let $\mathbf{f}_{3\mathrm{d}} : \Omega \to \mathbb{R}^3$ and $\mathbf{t} : \Gamma_N \to \mathbb{R}^3$ be volume and surface force fields,

respectively. For the rod, let there be body force and torque fields $\mathbf{f}_{\text{rod}} : [0,l] \to \mathbb{R}^3$ and $\mathbf{l} : [0,l] \to \mathbb{R}^3$, respectively, a Dirichlet value $\varphi_D \in \mathrm{SE}(3)$, and a stress-free rod configuration $\hat{\varphi} : [0,l] \to \mathrm{SE}(3)$. We look for functions $\mathbf{u} : \Omega \to \mathbb{R}^3$ and $\varphi : [0,l] \to \mathrm{SE}(3)$ such that

$$-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}) = \mathbf{f}_{\text{3d}} \qquad \text{in } \Omega \tag{5.14a}$$

$$\mathbf{m}' + \mathbf{r}' \times \mathbf{n} + \mathbf{l} = 0 \qquad \text{on } [0,l] \tag{5.14b}$$

$$-\mathbf{n}' = \mathbf{f}_{\text{rod}} \qquad \text{on } [0,l] \tag{5.14c}$$

Eq. (5.14a) is the equilibrium equation of linear elasticity on Ω (3.5) and Eqs. (5.14b) and (5.14c) are the rod equations presented on page 69. The solutions are subject to the boundary conditions

$$\mathbf{u} = 0 \qquad \text{on } \Gamma_D \tag{5.14d}$$

$$\boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu} = \mathbf{t} \qquad \text{on } \Gamma_N \tag{5.14e}$$

$$\varphi(l) = \varphi_D, \tag{5.14f}$$

and the coupling conditions

$$\int_\Gamma \boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu}\, ds = -\mathbf{n}(0)\boldsymbol{\nu}_0 \tag{5.14g}$$

$$\int_\Gamma (x - \varphi_r(0)) \times (\boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu})\, ds = -\mathbf{m}(0)\boldsymbol{\nu}_0 \tag{5.14h}$$

$$\operatorname{Av}(u) = \varphi(0). \tag{5.14i}$$

This is the problem that will be investigated in the rest of this chapter. Note that since the coupling conditions relate only finite-dimensional quantities, a discrete formulation of this problem can be obtained by replacing the continuous problems (5.14a)–(5.14c) by their corresponding discrete ones.

5.3 A Dirichlet–Neumann Algorithm

In this section we present a Dirichlet–Neumann algorithm for the coupled problem (5.14). This algorithm is of double interest. First of all, it is used to solve actual numerical problems; see Sections 5.5 and 6.2. Additionally, its formulation as an abstract operator $\mathscr{DN} : \mathrm{SE}(3) \to \mathrm{SE}(3)$ is used in Sec. 5.4 to show the existence of solutions of (5.14) for the special case of certain symmetric problems.

Dirichlet–Neumann algorithms belong to the family of nonoverlapping Schwarz methods. Given a partition of the domain into two nonoverlapping subdomains, they alternately solve a Dirichlet problem on one domain and a Neumann problem on the other. For linear problems it is well known that convergence can be expected only if the algorithm is suitably damped. The book by Quarteroni and Valli [77] contains an in-depth treatment.

A Dirichlet–Neumann method for the full-dimensional problem (5.4) can be interpreted as an iterative method for the Steklov–Poincaré equation in the trace space $\mathbf{H}^{1/2}(\Gamma_I)$. This is the space of configurations of the coupling boundary Γ_I of the homogeneous problem. For the heterogeneous problem we adopt a similar view. In this case, the interface is the single point $s = 0$, and the configuration space there is SE(3), the set of configurations of a single rod cross-section.

Consider the setting of the previous section. Each iteration of the Dirichlet–Neumann algorithm consists of three steps: solving a Dirichlet problem for the rod, solving a Neumann problem for the body, and a damped update. Since the interface space SE(3) is a nonlinear Riemannian manifold the damping will be along geodesics. Let $\lambda^0 \in \mathrm{SE}(3)$ be the initial interface value and $k \geq 0$ the iteration number. In more detail, the steps are as follows.

1. *Dirichlet problem for the Cosserat rod*
   Let $\lambda^k \in \mathrm{SE}(3)$ be the current interface value. Find a solution φ^{k+1} of the Dirichlet rod problem

$$\begin{aligned} (\mathbf{m}^{k+1})' + (\mathbf{r}^{k+1})' \times \mathbf{n}^{k+1} + \mathbf{l} &= 0 && \text{on } [0,l] \\ (\mathbf{n}^{k+1})' + \mathbf{f}_{\mathrm{rod}} &= 0 && \text{on } [0,l] \\ \varphi^{k+1}(0) &= \lambda^k \\ \varphi^{k+1}(l) &= \varphi_D. \end{aligned}$$

   If there is more than one solution the subdomain solver is free to pick any one of them.

2. *Neumann problem for the 3d object*
   The new rod iterate φ^{k+1} exerts a resultant force $\mathbf{n}^{k+1}(0)\boldsymbol{\nu}_0$ and a resultant moment $\mathbf{m}^{k+1}(0)\boldsymbol{\nu}_0$ across its cross-section at $s = 0$. We construct a Neumann data field $\boldsymbol{\tau}^{k+1} : \Gamma \to \mathbb{R}^3$ such that

$$\int_\Gamma \boldsymbol{\tau}^{k+1}(x)\, ds = -\mathbf{n}^{k+1}(0)\boldsymbol{\nu}_0 \tag{5.15}$$

   and

$$\int_\Gamma (x - \varphi_r^{k+1}(0)) \times \boldsymbol{\tau}^{k+1}(x)\, ds = -\mathbf{m}^{k+1}(0)\boldsymbol{\nu}_0. \tag{5.16}$$

   An algorithm for this is given at the end of the section. We then solve the three-dimensional linear elasticity problem with Neumann data $\boldsymbol{\tau}^{k+1}$ on Γ

$$\begin{aligned} -\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}^{k+1}) &= \mathbf{f}_{3\mathrm{d}} && \text{in } \Omega && \text{(5.17a)} \\ \boldsymbol{\sigma}(\mathbf{u}^{k+1})\boldsymbol{\nu} &= \boldsymbol{\tau}^{k+1} && \text{on } \Gamma && \text{(5.17b)} \\ \mathbf{u}^{k+1} &= 0 && \text{on } \Gamma_D && \text{(5.17c)} \\ \boldsymbol{\sigma}(\mathbf{u}^{k+1})\boldsymbol{\nu} &= \mathbf{t} && \text{on } \Gamma_N. && \text{(5.17d)} \end{aligned}$$

3. *Damped geodesic update*
   From the solution $\mathbf{u}^{k+1}$ we compute the average interface displacement and orientation $\mathrm{Av}(\mathbf{u}^{k+1})$ as defined in (5.13). The new interface value λ^{k+1} is then computed as a geodesic combination in SE(3) of the old value λ^k and $\mathrm{Av}(\mathbf{u}^{k+1})$. Let $\theta \in (0, \infty)$ be a damping parameter[1] and $\lambda^k = (\lambda_r^k, \lambda_q^k)$. We set λ_r^{k+1} as the affine combination

$$\lambda_r^{k+1} = \theta \,\mathrm{Av}_r(\mathbf{u}^{k+1}) + (1-\theta)\lambda_r^k, \tag{5.18a}$$

   where we have used $\mathrm{Av}_r(\mathbf{u}^{k+1})$ to denote the projection of $\mathrm{Av}(\mathbf{u}^{k+1})$ onto $\mathbb{R}^3$. The term λ_q^{k+1} is computed by interpolating along a minimizing geodesic on SO(3) through λ_q^k and $\mathrm{Av}_q(\mathbf{u}^{k+1}) = O_\Gamma(\mathbf{u}^{k+1})\hat{\varphi}_q(0)$. By Lem. 4.3.3 the minimizing geodesic is unique if

$$\mathrm{dist}\left(\lambda_q^k,\, O_\Gamma(\mathbf{u}^{k+1})\hat{\varphi}_q(0)\right) < \pi.$$

   We can use the interpolation formula (4.22) to get

$$\lambda_q^{k+1} = \exp_{\lambda_q^k} \theta\big[\exp_{\lambda_q^k}^{-1} O_\Gamma(\mathbf{u}^{k+1})\hat{\varphi}_q(0)\big]. \tag{5.18b}$$

This concludes the description of the Dirichlet–Neumann method for the multidimensional coupling problem (5.14). However, several algorithmic questions remain. The first is how the average orientation $O_\Gamma : \mathbf{H}^1(\Omega) \to \mathrm{SO}(3)$ can be computed. We use the following result, which follows directly from the properties of the singular value decomposition [86].

Lemma 5.3.1. *Let A be a quadratic, nonsingular matrix with the polar decomposition*

$$A = OH$$

and the singular value decomposition

$$A = P\Sigma Q^T,$$

with P and Q^T orthogonal matrices and Σ a diagonal matrix. Then $O = PQ^T$ and $H = Q\Sigma Q^T$.

Let $\mathcal{F}(\mathbf{u}^k)$ be the average interface deformation (5.10) and $\mathcal{F}(\mathbf{u}^k) = P_k\Sigma_k Q_k^T$ its singular value decomposition. Then

$$O_\Gamma(\mathbf{u}^k) = P_k Q_k^T.$$

Efficient algorithms for the computation of the singular value decomposition are available in the literature; see, e.g., [74].

The second question is how to construct suitable fields of Neumann data $\boldsymbol{\tau}^k$ that satisfy the conditions (5.15) and (5.16). Let us drop the index k for simplicity. In principle, any function $\boldsymbol{\tau} : \Gamma \to \mathbb{R}^3$ of sufficient regularity fulfilling (5.15) and (5.16) can

[1] Note that we also allow overrelaxation.

be used as Neumann data in (5.17b). We first show that functions fulfilling (5.15) and (5.16) exist in $\mathbf{L}^2(\Gamma)$. For $\mathbf{n}, \mathbf{m} \in \mathbb{R}^3$ we define the sets

$$C_{\mathbf{n}} = \left\{ \mathbf{f} \in \mathbf{L}^2(\Gamma) \;\middle|\; \int_\Gamma \mathbf{f}\, ds = \mathbf{n} \right\}$$

and

$$C_{\mathbf{m}} = \left\{ \mathbf{f} \in \mathbf{L}^2(\Gamma) \;\middle|\; \int_\Gamma (x - \varphi_r(0)) \times \mathbf{f}\, ds = \mathbf{m} \right\}.$$

We restrict ourselves to the case that Γ is flat and that $\varphi_r(0)$ is contained in Γ.

Lemma 5.3.2. *Let Γ be contained in an affine subspace of $\mathbb{R}^3$ and suppose that $\varphi_r(0) \in \Gamma$. Then $C_{\mathbf{n}} \cap C_{\mathbf{m}}$ is nonempty for all $\mathbf{n}, \mathbf{m} \in \mathbb{R}^3$.*

Proof. Without loss of generality we assume that Γ is contained in $\mathbb{R}^2 \times \{0\}$ and that $\varphi_r(0) = 0$. Define the four sets

$$\begin{aligned} \Gamma_\xi^+ &= \{x \in \Gamma \mid x_0 > 0\}, \qquad \Gamma_\xi^- = \{x \in \Gamma \mid x_0 < 0\}, \\ \Gamma_\zeta^+ &= \{x \in \Gamma \mid x_1 > 0\}, \qquad \Gamma_\zeta^- = \{x \in \Gamma \mid x_1 < 0\}. \end{aligned}$$

Since $0 \in \Gamma$ none of them is empty. Define the scalar functions $\psi_\xi, \psi_\zeta \in L^2(\Gamma)$

$$\psi_\xi(x) = \begin{cases} |\Gamma_\xi^+|^{-1} & \text{if } x \in \Gamma_\xi^+ \\ -|\Gamma_\xi^-|^{-1} & \text{else,} \end{cases} \qquad \psi_\zeta(x) = \begin{cases} |\Gamma_\zeta^+|^{-1} & \text{if } x \in \Gamma_\zeta^+ \\ -|\Gamma_\zeta^-|^{-1} & \text{else,} \end{cases}$$

and use them to define the vector-valued functions $\Psi_0, \Psi_1, \Psi_2 \in \mathbf{L}^2(\Gamma)$

$$\Psi_0 = (0, 0, \psi_\xi)^T, \qquad \Psi_1 = (0, 0, \psi_\zeta)^T, \qquad \Psi_2 = (0, \psi_\xi, 0)^T.$$

All three functions transmit zero total force

$$\int_\Gamma \Psi_i\, ds = 0, \qquad i \in \{0, 1, 2\}. \tag{5.19}$$

For the angular moments around 0 we get

$$\begin{aligned} \int_\Gamma x \times \Psi_0(x)\, ds &= \int_\Gamma \begin{pmatrix} x_0 \\ x_1 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ \psi_\xi \end{pmatrix} ds = \int_\Gamma \begin{pmatrix} \psi_\xi x_1 \\ -\psi_\xi x_0 \\ 0 \end{pmatrix} ds \\ &= \int_{\Gamma_\xi^+} \begin{pmatrix} x_1 |\Gamma_\xi^+|^{-1} \\ -x_0 |\Gamma_\xi^+|^{-1} \\ 0 \end{pmatrix} ds + \int_{\Gamma_\xi^-} \begin{pmatrix} -x_1 |\Gamma_\xi^-|^{-1} \\ x_0 |\Gamma_\xi^-|^{-1} \\ 0 \end{pmatrix} ds \\ &= \int_{\Gamma_\xi^+} \begin{pmatrix} x_1 |\Gamma_\xi^+|^{-1} \\ -|x_0| |\Gamma_\xi^+|^{-1} \\ 0 \end{pmatrix} ds + \int_{\Gamma_\xi^-} \begin{pmatrix} -x_1 |\Gamma_\xi^-|^{-1} \\ -|x_0| |\Gamma_\xi^-|^{-1} \\ 0 \end{pmatrix} ds. \end{aligned}$$

Similarly we get

$$\int_\Gamma x \times \Psi_1(x)\, ds \;=\; \int_{\Gamma_\zeta^+} \begin{pmatrix} |x_1||\Gamma_\zeta^+|^{-1} \\ x_0|\Gamma_\zeta^+|^{-1} \\ 0 \end{pmatrix} ds + \int_{\Gamma_\zeta^-} \begin{pmatrix} |x_1||\Gamma_\zeta^-|^{-1} \\ x_0|\Gamma_\zeta^-|^{-1} \\ 0 \end{pmatrix} ds$$

and

$$\int_\Gamma x \times \Psi_2(x)\, ds \;=\; \int_{\Gamma_\xi^+} \begin{pmatrix} 0 \\ 0 \\ |x_0||\Gamma_\xi^+|^{-1} \end{pmatrix} ds + \int_{\Gamma_\xi^-} \begin{pmatrix} 0 \\ 0 \\ |x_0||\Gamma_\xi^-|^{-1} \end{pmatrix} ds.$$

The three vectors

$$\begin{pmatrix} \alpha \\ -\beta \\ 0 \end{pmatrix} := \int_\Gamma x \times \Psi_0\, ds, \quad \begin{pmatrix} \gamma \\ \delta \\ 0 \end{pmatrix} := \int_\Gamma x \times \Psi_1\, ds, \quad \begin{pmatrix} 0 \\ 0 \\ \beta \end{pmatrix} := \int_\Gamma x \times \Psi_2\, ds \tag{5.20}$$

are nonzero because $\beta \neq 0$ and $\gamma \neq 0$. They form a basis of $\mathbb{R}^3$ if

$$\alpha\delta \neq \beta\gamma.$$

This is the case, for example, if Γ is symmetric with respect to reflections about the xz or yz planes, since then α or δ is zero.

Let $\mathbf{f}$ be a function in $C_\mathbf{n}$. Then

$$\int_\Gamma x \times \mathbf{f}\, ds = \mathbf{m_f}$$

for some $\mathbf{m_f} \in \mathbb{R}^3$ which will in general not be equal to $\mathbf{m}$. However, since the vectors (5.20) form a basis of $\mathbb{R}^3$, there are coefficients $\kappa_i \in \mathbb{R}^3$, $i \in \{0, 1, 2\}$ such that

$$\int_\Gamma x \times \Big(\mathbf{f} + \sum_{i=0}^{2} \kappa_i \Psi_i\Big)\, ds = \mathbf{m}. \tag{5.21}$$

Define the term in parentheses as the new function

$$\bar{\mathbf{f}} = \mathbf{f} + \sum_{i=0}^{2} \kappa_i \Psi_i.$$

By (5.21) we have $\bar{\mathbf{f}} \in C_\mathbf{m}$. On the other hand, using (5.19) we have

$$\int_\Gamma \bar{\mathbf{f}}\, ds = \int_\Gamma \Big(\mathbf{f} + \sum_{i=0}^{2} \kappa_i \Psi_i\Big)\, ds = \mathbf{n},$$

and hence $\bar{\mathbf{f}} \in C_\mathbf{n} \cap C_\mathbf{m}$. □

The theory of Cosserat rods assumes that forces and moments are transmitted evenly across cross-sections. We therefore construct $\boldsymbol{\tau}$ to be 'as constant as possible'. More formally, we introduce the constancy functional

$$\begin{aligned} T &: \mathbf{L}^2(\Gamma) \times \mathbb{R}^3 \to \mathbb{R}, \\ T(\mathbf{f}, c) &= \int_\Gamma \|\mathbf{f}(x) - c\|^2 \, ds, \end{aligned}$$

and construct $\boldsymbol{\tau}$ as the solution of the minimization problem

$$(\boldsymbol{\tau}, c_{\boldsymbol{\tau}}) = \arg \min_{\substack{\mathbf{f} \in \mathbf{L}^2(\Gamma) \\ c \in \mathbb{R}^3}} T(\mathbf{f}, c) \tag{5.22}$$

under the constraints that

$$\int_\Gamma \mathbf{f}(x) \, ds = -\mathbf{n}(0)\boldsymbol{\nu}_0 \qquad \text{and} \qquad \int_\Gamma (x - \varphi_r(0)) \times \mathbf{f}(x) \, ds = -\mathbf{m}(0)\boldsymbol{\nu}_0. \tag{5.23}$$

To show that this problem is well posed we first prove a few properties of T.

Lemma 5.3.3. *The functional T is continuous, coercive, and strictly convex on $C_{\mathbf{n}} \times \mathbb{R}^3$ for all $\mathbf{n} \in \mathbb{R}^3$.*

Proof. We only show coercivity and convexity. Assume a sequence $(\mathbf{f}_k, c_k) \in C_{\mathbf{n}} \times \mathbb{R}^3$ such that $\lim_{k\to\infty}(\|\mathbf{f_k}\| + \|c_k\|) = \infty$. Then

$$\begin{aligned} T(\mathbf{f}_k, c_k) &= \|\mathbf{f_k}\|^2_{\mathbf{L}^2(\Gamma)} + \|c_k\|^2 - 2\int_\Gamma \langle \mathbf{f}_k, c_k \rangle \, ds = \|\mathbf{f_k}\|^2_{\mathbf{L}^2(\Gamma)} + \|c_k\|^2 - 2\langle \mathbf{n}, c_k \rangle \\ &\geq \|\mathbf{f_k}\|^2_{\mathbf{L}^2(\Gamma)} + \|c_k\|^2 - 2\|\mathbf{n}\|\|c_k\| = \|\mathbf{f_k}\|^2_{\mathbf{L}^2(\Gamma)} + \|c_k\|(\|c_k\| - 2\|\mathbf{n}\|) \end{aligned}$$

Hence $\lim_{k\to\infty} T(\mathbf{f}_k, c_k) = \infty$ and T is coercive in $C_{\mathbf{n}} \times \mathbb{R}^3$.

We now show that T is strictly convex on $C_{\mathbf{n}} \times \mathbb{R}^3$. Pick $\boldsymbol{\tau}^0, \boldsymbol{\tau}^1 \in C_{\mathbf{n}}$ and $c^0, c^1 \in \mathbb{R}^3$ such that either $\boldsymbol{\tau}^0 \neq \boldsymbol{\tau}^1$ on a set of positive measure or $c^0 \neq c^1$ or both. Then, for $t \in (0, 1)$,

$$\begin{aligned} &T(t(\boldsymbol{\tau}^0, c^0) + (1-t)(\boldsymbol{\tau}^1, c^1)) \\ &\qquad = tT(\boldsymbol{\tau}^0, c^0) + (1-t)T(\boldsymbol{\tau}^1, c^1) - \frac{1}{2}t(1-t)\int_\Gamma \|\boldsymbol{\tau}^0 - c^0 - (\boldsymbol{\tau}^1 - c^1)\|^2 \, ds. \end{aligned}$$

The functional T is convex because the integral is nonnegative. To show that we have even strict convexity note that

$$\int_\Gamma \|\boldsymbol{\tau}^0 - c^0 - (\boldsymbol{\tau}^1 - c^1)\|^2 \, ds = \int_\Gamma \|\boldsymbol{\tau}^0 - \boldsymbol{\tau}^1\|^2 \, ds + \int_\Gamma \|c^0 - c^1\|^2 \, ds > 0$$

where we have used that $\boldsymbol{\tau}^0, \boldsymbol{\tau}^1 \in C_{\mathbf{n}}$. □

Lemma 5.3.4. *For all $\mathbf{n}, \mathbf{m} \in \mathbb{R}^3$ the minimization problem (5.22) has a unique solution on the set $(C_{\mathbf{n}} \cap C_{\mathbf{m}}) \times \mathbb{R}^3$.*

Proof. We show that $(C_\mathbf{n} \cap C_\mathbf{m}) \times \mathbb{R}^3$ is closed in $\mathbf{L}^2(\Gamma) \times \mathbb{R}^3$. For this let $\mathbf{f}_k \in C_\mathbf{n}$, $k = 0, 1, \ldots$, be a sequence with $\lim \mathbf{f}_k = \mathbf{f}$. This means that $\lim \int_\Gamma |\mathbf{f}_k - \mathbf{f}|^2 = 0$ and also $\lim \int_\Gamma |\mathbf{f}_k - \mathbf{f}| = 0$. However, since

$$\int_\Gamma |\mathbf{f}_k - \mathbf{f}|\, ds \geq \Big| \int_\Gamma \mathbf{f}_k\, ds - \int_\Gamma \mathbf{f}\, ds \Big| = \Big| \mathbf{n} - \int_\Gamma \mathbf{f}\, ds \Big| \geq 0$$

we get $\int_\Gamma \mathbf{f}\, ds = \mathbf{n}$ and hence $C_\mathbf{n}$ is closed. Showing that $C_\mathbf{m}$ is also closed in $\mathbf{L}^2(\Gamma)$ proceeds similarly, noting that

$$\begin{aligned} 0 \leq \Big| \mathbf{m} - \int_\Gamma (x - \varphi_r(0)) \times \mathbf{f}\, ds \Big| \\ &= \Big| \int_\Gamma (x - \varphi_r(0)) \times \mathbf{f}_k\, ds - \int_\Gamma (x - \varphi_r(0)) \times \mathbf{f}\, ds \Big| \\ &\leq \int_\Gamma |(x - \varphi_r(0)) \times (\mathbf{f}_k - \mathbf{f})|\, ds \leq \max |x - \varphi_r(0)| \int_\Gamma |\mathbf{f}_k - \mathbf{f}|\, ds. \end{aligned}$$

Hence $(C_\mathbf{n} \cap C_\mathbf{m}) \times \mathbb{R}^3$ is closed in $\mathbf{L}^2(\Gamma) \times \mathbb{R}^3$. By Lem. 5.3.3, T is continuous, coercive, and strictly convex on $C_\mathbf{n} \times \mathbb{R}^3$, and thus in particular on $(C_\mathbf{n} \cap C_\mathbf{m}) \times \mathbb{R}^3$. Therefore T has a unique minimum there [35]. □

For later reference we introduce the operator

$$\Upsilon : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbf{L}^2(\Gamma) \tag{5.24}$$

with

$$\Upsilon(\mathbf{n}(0)\boldsymbol{\nu}_0, \mathbf{m}(0)\boldsymbol{\nu}_0) = \boldsymbol{\tau}$$

where $\boldsymbol{\tau}$ is the solution of (5.22) subject to (5.23).

We now look how $\boldsymbol{\tau}$ can be computed in a discrete setting. Let Ω be discretized by a grid which resolves Γ. Let further $\mathbf{V}_h(\Gamma)$ be the space of 3-valued first-order finite element functions on the grid restricted to Γ and n_Γ be the number of grid vertices in Γ. We are looking for a function $\boldsymbol{\tau}_h \in \mathbf{V}_h(\Gamma)$ such that

$$(\boldsymbol{\tau}_h, c_h) = \arg \min_{\substack{\mathbf{f}_h \in \mathbf{V}_h(\Gamma) \\ c \in \mathbb{R}^3}} T(\mathbf{f}_h, c) \tag{5.25}$$

subject to

$$\int_\Gamma \mathbf{f}_h(x)\, ds = -\mathbf{n}(0)\boldsymbol{\nu}_0 \qquad \text{and} \qquad \int_\Gamma (x - \varphi_r(0)) \times \mathbf{f}_h(x)\, ds = -\mathbf{m}(0)\boldsymbol{\nu}. \tag{5.26}$$

We denote by ψ_i the i-th scalar hat function in the scalar finite element space $V_h(\Gamma)$ and by $\boldsymbol{\psi}_{i,j} = \psi_i \mathbf{e}_j \in \mathbf{V}_h(\Gamma)$ the i-th vector-valued hat function in the direction of the j-th canonical basis vector $\mathbf{e}_j$. Express $\boldsymbol{\tau}_h$ in the nodal basis as

$$\boldsymbol{\tau}_h = \sum_{i=0}^{n_\Gamma - 1} \sum_{j=0}^{2} \tau_j^i \boldsymbol{\psi}_{i,j}. \tag{5.27}$$

Then the functional T restricted to $\mathbf{V}_h(\Gamma)$ has the algebraic form

$$\begin{aligned} T &: \mathbb{R}^{3n_\Gamma} \times \mathbb{R}^3 \to \mathbb{R} \\ T(\tau, c) &= \int_\Gamma \Big\| \sum_{i,j} \tau_i^j \boldsymbol{\psi}_{i,j} - c \Big\|^2 \, ds. \end{aligned}$$

Inserting the ansatz (5.27) into (5.26) we obtain two linear systems of constraints

$$\mathbf{N}\tau = -\mathbf{n}(0)\boldsymbol{\nu}_0 \qquad \text{and} \qquad \mathbf{M}\tau = -\mathbf{m}(0)\boldsymbol{\nu}_0.$$

Both $\mathbf{N}$ and $\mathbf{M}$ are $1 \times n_\Gamma$ block matrices. Each entry of $\mathbf{N}$ is a 3×3 block

$$(\mathbf{N})_{0i} = \mathrm{Id}_{3\times 3} \int_\Gamma \psi_i \, ds, \qquad 0 \le i < n_\Gamma,$$

whereas each entry of $\mathbf{M}$ is a block

$$(\mathbf{M})_{0i} = \begin{pmatrix} \mu_{0,0}^i & \mu_{1,0}^i & \mu_{2,0}^i \\ \mu_{0,1}^i & \mu_{1,1}^i & \mu_{2,1}^i \\ \mu_{0,2}^i & \mu_{1,2}^i & \mu_{2,2}^i \end{pmatrix}, \qquad 0 \le i < n_\Gamma,$$

with

$$\mu_{j,l}^i = \Big(\int_\Gamma (x - \varphi_r^k(0)) \times \boldsymbol{\psi}_{i,j} \, ds \Big)_l, \qquad 0 \le i < n_\Gamma, \quad j,l \in \{0,1,2\}.$$

Lemma 5.3.5. *The minimization problem* (5.25) *subject to* (5.26) *has a unique solution.*

Proof. By Lem. 5.3.3 the functional T is continuous, coercive, and strictly convex on $C_\mathbf{n} \times \mathbb{R}^3$. Hence it is so, in particular, on $(C_\mathbf{n} \cap C_\mathbf{m} \cap \mathbf{V}_h(\Gamma)) \times \mathbb{R}^3$, which we assume to be nonempty. Also, $(C_\mathbf{n} \cap C_\mathbf{m} \cap \mathbf{V}_h(\Gamma)) \times \mathbb{R}^3$ is closed because $(C_\mathbf{n} \cap C_\mathbf{m}) \times \mathbb{R}^3$ is closed (Lem. 5.3.4) and $\mathbf{V}_h(\Gamma) \times \mathbb{R}^3$ is finite-dimensional. Therefore, the minimization problem (5.25) subject to (5.26) has a unique solution [35]. □

A minimization problem of this type can be solved, e.g., with an interior-point method like IPOpt [90].

5.4 Existence of Solutions of the Heterogeneous Problem

In this section we prove that the heterogeneous coupling problem (5.14) does have at least one solution if the reference configuration and the boundary conditions exhibit certain symmetries. Since the rod problem by itself may admit more than a single solution [82], we cannot hope to show uniqueness of solutions for the coupled problem.

The proof is based on a fixed point argument. The damped Dirichlet–Neumann algorithm of the preceding section is written as an operator

$$\begin{aligned} \mathscr{DN}_\theta &: \mathrm{SE}(3) \to \mathrm{SE}(3) \\ \mathscr{DN}_\theta \lambda^k &= \lambda^{k+1} \end{aligned}$$

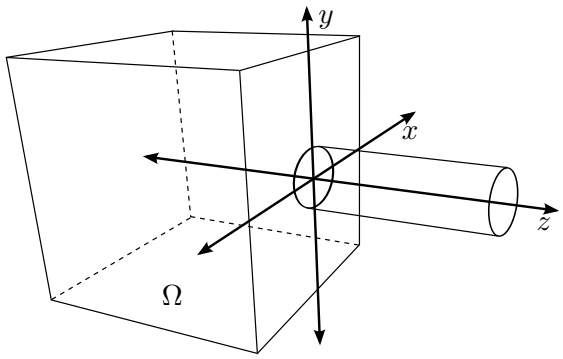


Figure 5.4: Geometric setting of the existence result.

on the configuration space SE(3) of the rod cross-section at $s = 0$. We interpret $\mathscr{DN}_\theta$ as an operator chain

$$\mathscr{DN}_\theta \;:\; \mathrm{SE}(3) \xrightarrow{\mathrm{DtN}} \mathbb{R}^3 \times \mathbb{R}^3 \xrightarrow{\Upsilon} \mathbf{L}^2(\Gamma) \xrightarrow{\mathfrak{E}} \mathbf{H}^1(\Omega) \xrightarrow{\mathrm{Av}} \mathrm{SE}(3) \xrightarrow{\theta} \mathrm{SE}(3),$$

where $\mathrm{DtN} : \mathrm{SE}(3) \to \mathbb{R}^3 \times \mathbb{R}^3$ is the set-valued Dirichlet-to-Neumann operator of the rod problem, which maps the Dirichlet value at $s = 0$ of a rod problem to the set of corresponding Neumann values of the solutions. The Neumann field operator $\Upsilon : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbf{L}^2(\Gamma)$ has been defined in (5.24), $\mathfrak{E} : \mathbf{L}^2(\Gamma) \to \mathbf{H}^1(\Omega)$ is the solution operator for the linear elasticity problem (5.17) with Neumann data on Γ, Av is the averaging operator defined in (5.13), and θ symbolizes the geodesic damping (5.18). We show that under certain symmetry assumptions $\mathscr{DN}_\theta$ is a single-valued, continuous contraction on the restricted cross-section space

$$\widetilde{\mathrm{SE}}(3) = \big\{(\lambda_r, \lambda_q) \in \mathbb{R}^3 \times \mathrm{SO}(3) \;\big|\; (\lambda_r)_0 = 0, (\lambda_r)_1 = 0, \lambda_q = \mathrm{Id}\big\}.$$

This is the space of parallel translations of the rod cross-section at $s = 0$ in the z-direction (Fig. 5.4) and can be identified with $\mathbb{R}$. By Banach's fixed point theorem $\mathscr{DN}_\theta$ then has a fixed point, which is unique in $\widetilde{\mathrm{SE}}(3)$, and which is shown to generate a solution of (5.14).

We begin by formally stating the symmetry conditions.

Definition 5.4.1 (Symmetry)**.** *Define*

$$S^{yz} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \textit{and} \qquad S^{xz} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

the reflection matrices about the yz and xz planes. A set $\Sigma \subseteq \mathbb{R}^3$ is called symmetric with respect to reflections about the xz and yz planes*, or* symmetric *for short, if*

$$x \in \Sigma \Leftrightarrow S^{yz}x \in \Sigma \qquad \textit{and} \qquad x \in \Sigma \Leftrightarrow S^{xz}x \in \Sigma.$$

A function $\mathbf{v} : \Sigma \to \mathbb{R}^3$ *is called symmetric if*

$$\mathbf{v}(x) = S^{yz}\mathbf{v}(S^{yz}x) \text{ a.e. on } \Sigma \qquad \text{and} \qquad \mathbf{v}(x) = S^{xz}\mathbf{v}(S^{xz}x) \text{ a.e. on } \Sigma. \tag{5.28}$$

To motivate this definition of symmetry note that if $\Omega \subset \mathbb{R}^3$ is a symmetric domain and $\mathbf{u} : \Omega \to \mathbb{R}^3$ is a symmetric deformation of Ω, then the deformed domain $\mathbf{u}(\Omega) = \{\mathbf{u}(x) + x \mid x \in \Omega\}$ is also symmetric.

The proof proceeds in three steps. Consider the setting of Sec. 5.2. We first show that, under symmetry assumptions, $\widetilde{\mathrm{SE}}(3)$ is an invariant subspace of $\mathscr{QN}_\theta$. Then we show that $\mathscr{QN}_\theta$ is Lipschitz continuous on $\widetilde{\mathrm{SE}}(3)$. This allows to conclude that there must be a fixed point λ^* of $\mathscr{QN}_\theta$ in $\widetilde{\mathrm{SE}}(3)$ which is then shown to induce solutions of the coupled problem (5.14).

Lemma 5.4.1. *Consider a rod of rest length l which is straight in its unstressed configuration, i.e.,* $\hat{\mathrm{u}}(s) = (0,0,0)$ *and* $\hat{\mathrm{v}}(s) = (0,0,1)$ *for all* $0 \le s \le l$*. Suppose that neither volume forces nor volume moments act on this rod and let* $l_0, l_1 \in \mathbb{R}$*,* $l_0 < l_1$*. Then the problem*

$$\mathbf{m}' + \mathbf{r}' \times \mathbf{n} = 0 \qquad \text{on } [0,l], \tag{5.29a}$$

$$\mathbf{n}' = 0 \qquad \text{on } [0,l], \tag{5.29b}$$

subject to the Dirichlet boundary conditions

$$\varphi(0) = \varphi_{D,0} = ((0,0,l_0), Id), \qquad \varphi(l) = \varphi_{D,1} = ((0,0,l_1), Id), \tag{5.29c}$$

has at least one solution φ^* that is axially symmetric. We call φ^* the trivial solution. Furthermore, there is a function $n : \mathbb{R} \to \mathbb{R}$ such that $(0,0,n(l_0)) \times (0,0,0) \in \mathrm{DtN}\,\varphi_{D,0}$ for all $l_0 < l_1$.

Proof. Direct computations show that the axially symmetric function

$$\varphi_r^*(s) = \frac{s}{l} l_1 \mathbf{e}_2 + \Big(1 - \frac{s}{l}\Big) l_0 \mathbf{e}_2, \qquad \varphi_q^*(s) = \mathrm{Id},$$

solves problem (5.29). Consider this as a function in l_0 only. Using the definition of the strain (4.10) and (4.11) and the diagonal rod material law (4.15) we get

$$\mathrm{DtN}\,\varphi_{D,0} \ni (\mathbf{n}^*(0)\boldsymbol{\nu}_0, \mathbf{m}^*(0)\boldsymbol{\nu}_0) = ((0,0,n),(0,0,0)),$$

for some $n \in \mathbb{R}$ which depends on l_0. □

Note that the trivial solution φ^* may be unstable, i.e., it may be a stationary point but not a minimum of the rod energy j.

We now introduce a selection of DtN restricted to $\widetilde{\mathrm{SE}}(3)$ by setting

$$\widetilde{\mathrm{DtN}}\lambda = \begin{cases} (0,0,n) \times (0,0,0) & \text{if } (0,0,n) \times (0,0,0) \in \mathrm{DtN}\,\lambda \text{ for some } n \in \mathbb{R}, \\ \mathrm{DtN}\,\lambda & \text{else.} \end{cases}$$

In other words, for Dirichlet values that admit a trivial solution we consider only this trivial solution. For the rest of this section, we will write DtN instead of $\widetilde{\mathrm{DtN}}$ in a conscious abuse of notation.

Lemma 5.4.2. *Let $\Omega \subset \mathbb{R}^3$ be a symmetric domain and $\Gamma_N \subset \partial\Omega$ a symmetric part of its boundary such that $\Gamma_D = \partial\Omega \setminus \Gamma_N$ has a positive two-dimensional measure. Set $\boldsymbol{\sigma}(\mathbf{u}) = \mathbf{C} : \boldsymbol{\varepsilon}(\mathbf{u})$, and let $\mathbf{f} : \Omega \to \mathbb{R}^3$, $\mathbf{t} : \Gamma_N \to \mathbb{R}^3$ be symmetric. Then the solution of the boundary value problem*

$$
\begin{aligned}
-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}) &= \mathbf{f} && \text{in } \Omega, && (5.30a)\\
\boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu} &= \mathbf{t} && \text{on } \Gamma_N, && (5.30b)\\
\mathbf{u} &= 0 && \text{on } \partial\Omega \setminus \Gamma_N, && (5.30c)
\end{aligned}
$$

is also symmetric.

Proof. It was shown in Sec. 3.1 that Problem (5.30) has a unique solution $\mathbf{u}$. We show that the reflection $\mathbf{u}^{yz}(x) := S^{yz}\mathbf{u}(S^{yz}x)$ also solves (5.30). By the uniqueness of $\mathbf{u}$ then follows $\mathbf{u} = \mathbf{u}^{yz}$ and hence $\mathbf{u}$ is symmetric with respect to reflection about the yz plane.

We first show that $\mathbf{u}^{yz}$ solves the equilibrium equation (5.30a). Indeed, for any $x \in \Omega$ we have

$$
\begin{aligned}
-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}^{yz}(x)) &= -\operatorname{div} \boldsymbol{\sigma}(S^{yz}\mathbf{u}(S^{yz}x))\\
&= -\operatorname{div}\Big[\mathbf{C} : \frac{1}{2}\big(\boldsymbol{\nabla}(S^{yz}\mathbf{u}(S^{yz}x)) + (\boldsymbol{\nabla}(S^{yz}\mathbf{u}(S^{yz}x)))^T\big)\Big]\\
&= -S^{yz}\operatorname{div}\Big[\mathbf{C} : \frac{1}{2}\big(\boldsymbol{\nabla}\mathbf{u}(S^{yz}x) + (\boldsymbol{\nabla}\mathbf{u}(S^{yz}x))^T\big)\Big],
\end{aligned}
$$

where we have used that S^{yz} is diagonal. Now we use that $\mathbf{u}$ is a solution of (5.30a) to obtain

$$
-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}^{yz}(x)) = S^{yz}\mathbf{f}(S^{yz}x) = \mathbf{f}(x),
$$

because $\mathbf{f}$ is symmetric. Similar reasoning shows that $\mathbf{u}^{yz}$ fulfills the boundary conditions (5.30b) and (5.30c). Therefore the unique solution $\mathbf{u}$ of (5.30) is symmetric with respect to the yz plane. Showing the same for $\mathbf{u}^{xz}(x) := S^{xz}\mathbf{u}(S^{xz}x)$ proves the assertion. □

We show next that both the averaging operator Av and the Neumann value operator Υ preserve symmetry.

Lemma 5.4.3. *In addition to the assumptions of Lemma 5.4.2 let Γ be a symmetric subset of $\partial\Omega$. We have*

$$
\operatorname{Av}(\mathbf{u}) = ((0, 0, d_{\mathbf{u}}), Id) \tag{5.31}
$$

for some $d_{\mathbf{u}} \in \mathbb{R}$.

Proof. By Lemma 5.4.2 u_0 is antisymmetric with respect to x_0 and u_1 is antisymmetric with respect to x_1. The integrals of antisymmetric functions over symmetric domains vanish and we get

$$
\operatorname{Av}_r(\mathbf{u}) = \frac{1}{|\Gamma|}\int_\Gamma \mathbf{u}(x) + x\, ds = (0, 0, d_{\mathbf{u}}).
$$

To prove that $\mathrm{Av}_q(\mathbf{u}) = \mathrm{Id}$ we first show that $\mathcal{F}(\mathbf{u}) = \int_\Gamma \boldsymbol{\nabla}(\mathbf{u} + x)\, ds$ is a diagonal matrix. For this we prove that all off-diagonal elements of $\boldsymbol{\nabla}\mathbf{u}$ are antisymmetric either with respect to x_0 or x_1. We begin by considering $(\boldsymbol{\nabla}\mathbf{u})_{01} = \partial u_0/\partial x_1$ and get

$$\begin{aligned}\frac{\partial}{\partial x_1}u_0(-x_0,x_1,x_2) &= \lim_{\delta\to 0}\frac{u_0(-x_0,x_1+\delta,x_2)-u_0(-x_0,x_1,x_2)}{\delta}\\ &= \lim_{\delta\to 0}\frac{-u_0(x_0,x_1+\delta,x_2)+u_0(x_0,x_1,x_2)}{\delta}\\ &= -\frac{\partial}{\partial x_1}u_0(x_0,x_1,x_2).\end{aligned}$$

Similarly one can show that $\partial u_i/\partial x_j$ is antisymmetric for $i \in \{0,1\}$, $j \in \{0,1,2\}$, $i \neq j$. The partial derivatives $\partial u_2/\partial x_0$ and $\partial u_2/\partial x_1$ are antisymmetric by the following reasoning

$$\begin{aligned}\frac{\partial}{\partial x_0}u_2(-x_0,x_1,x_2) &= \lim_{\delta\to 0}\frac{u_2(-x_0+\delta,x_1,x_2)-u_2(-x_0,x_1,x_2)}{\delta}\\ &= \lim_{\delta\to 0}\frac{u_2(x_0-\delta,x_1,x_2)-u_2(x_0,x_1,x_2)}{\delta}\\ &= \lim_{\delta\to 0}\frac{u_2(x_0+\delta,x_1,x_2)-u_2(x_0,x_1,x_2)}{-\delta}\\ &= -\frac{\partial}{\partial x_0}u_2(x_0,x_1,x_2).\end{aligned}$$

Hence $\mathcal{F}(\mathbf{u})$ is diagonal and its polar decomposition is

$$\mathcal{F}(\mathbf{u}) = O_\Gamma(\mathbf{u})H(\mathbf{u}) = \begin{pmatrix}\pm 1 & & \\ & \pm 1 & \\ & & \pm 1\end{pmatrix}\begin{pmatrix}\alpha_0 & & \\ & \alpha_1 & \\ & & \alpha_2\end{pmatrix},$$

with $\alpha_i > 0$, $0 \le i < 3$. Since $\det \mathcal{F}(\mathbf{u}) > 0$ by assumption (see (5.10)), either $O_\Gamma(\mathbf{u})$ is the identity or two of its three diagonal entries are -1. To see that it can only be the identity note that the reasoning above holds for all symmetric subsets $\widetilde{\Gamma}$ of Ω. Let $\widetilde{\Gamma}_t$ be a continuous family of subsets of Ω such that $\widetilde{\Gamma}_0 = \Gamma_D$ and $\widetilde{\Gamma}_1 = \Gamma$. We have $O_{\widetilde{\Gamma}_0}(\mathbf{u}) = \mathrm{Id}$, because $\mathbf{u}$ is clamped at Γ_D. For fixed $\mathbf{u}$ the entries of $O_{\widetilde{\Gamma}_t}(\mathbf{u})$ depend continuously on t and take only integer values. Hence $O_{\widetilde{\Gamma}_1}(\mathbf{u}) = O_\Gamma(\mathbf{u}) = \mathrm{Id}$. □

Lemma 5.4.4. *Let $n \in \mathbb{R}$ and let $\varphi_r(0) = (0,0,l_0)$ with $l_0 \in \mathbb{R}$. Then*

$$\Upsilon\big((0,0,n),\,(0,0,0)\big) = (0,0,-n/|\Gamma|)^T.$$

In particular,Υ is Lipschitz continuous on $\{0\}^2 \times \mathbb{R} \times \{0\}^3$.

Proof. The constant function $\mathbf{f} := (0,0,-n/|\Gamma|)^T$ fulfills the conditions (5.23) because

$$\int_\Gamma \mathbf{f}\, ds = \begin{pmatrix}0\\0\\-n\end{pmatrix} \qquad \text{and} \qquad \int_\Gamma \left(x - \begin{pmatrix}0\\0\\l_0\end{pmatrix}\right) \times \mathbf{f}\, ds = \begin{pmatrix}0\\0\\0\end{pmatrix}$$

by the symmetry of Γ. To see that $\mathbf{f}$ is a minimum of T note that T takes only nonnegative values. However $T\big(\mathbf{f},\, (0,0,-n/|\Gamma|)^T\big) = 0$, and hence $\mathbf{f}$ must be a minimum. □

The preceding results allow to conclude that $\widetilde{\mathrm{SE}}(3)$ is an invariant subspace of $\mathscr{DN}_\theta$.

Lemma 5.4.5. *Let the sets Ω, Γ_N, Γ as well as the volume force field $\mathbf{f} : \Omega \to \mathbb{R}^3$ and the surface traction $\mathbf{t} : \Gamma_N \to \mathbb{R}^3$ by symmetric in the sense of Def. 5.4.1. Let the rod reference configuration $\hat{\varphi}$ be straight and the rod Dirichlet value φ_D be in $\widetilde{SE}(3)$. Then $\widetilde{SE}(3)$ is an invariant subspace of the damped Dirichlet–Neumann operator $\mathscr{DN}_\theta$.*

In the remainder of this section we take the assumptions of Lemma 5.4.5 to hold. Proceeding in several steps we now show that $\mathscr{DN} := \mathscr{DN}_1$ is Lipschitz continuous on $\widetilde{\mathrm{SE}}(3)$.

Lemma 5.4.6. *The Dirichlet-to-Neumann map $\widetilde{\mathrm{DtN}}$ is Lipschitz continuous on $\widetilde{SE}(3)$.*

Proof. By Lemma 5.4.1 the rod problem (5.29) has the solution

$$\varphi_r(s) = \frac{s}{l} l_1 \mathbf{e}_2 + \Big(1 - \frac{s}{l}\Big) l_0 \mathbf{e}_2, \qquad \varphi_q(s) = \mathrm{Id}, \qquad s \in [0, l].$$

For the stresses and moments we get (cf. (4.16))

$$\mathbf{n}(s) = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} \varphi_r'(s) - \mathbf{e}_2 = A_3 \Big(\frac{l_1 - l_0}{l} - 1\Big)\mathbf{e}_2, \qquad \mathbf{m}(s) = 0.$$

Hence in particular $\mathbf{n}(0)\boldsymbol{\nu}_0 = -A_3[(l_1 - l_0)/l - 1]\mathbf{e}_2$, $\mathbf{m}(0)\boldsymbol{\nu}_0 = 0$, and these are affine functions of l_0. This proves the assertion. □

Lemma 5.4.7. *The operator $\mathfrak{E} : \mathbf{L}^2(\Gamma) \to \mathbf{H}^1(\Omega)$ mapping Neumann data on Γ to solutions of the linear elasticity problem on Ω is Lipschitz continuous.*

Proof. Let $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in \mathbf{L}^2(\Gamma)$ be two fields of Neumann boundary data and let $\mathbf{u}_1, \mathbf{u}_2 \in \mathbf{H}^1_0(\Omega)$ be the corresponding solutions of the linear elasticity problem in weak form, i.e.,

$$a(\mathbf{u}_1, \mathbf{v}) = \int_\Gamma \boldsymbol{\tau}_1 \mathbf{v}\, ds + \int_\Omega \mathbf{f}\mathbf{v}\, dx + \int_{\Gamma_N} \mathbf{t}\mathbf{v}\, ds \qquad \text{for all } \mathbf{v} \in \mathbf{H}^1_0(\Omega), \quad (5.32a)$$

$$a(\mathbf{u}_2, \mathbf{v}) = \int_\Gamma \boldsymbol{\tau}_2 \mathbf{v}\, ds + \int_\Omega \mathbf{f}\mathbf{v}\, dx + \int_{\Gamma_N} \mathbf{t}\mathbf{v}\, ds \qquad \text{for all } \mathbf{v} \in \mathbf{H}^1_0(\Omega), \quad (5.32b)$$

where $a(\cdot,\cdot)$ is the bilinear form defined in (3.12). Subtracting (5.32a) from (5.32b) yields

$$a(\mathbf{u}_2 - \mathbf{u}_1, \mathbf{v}) = \int_\Gamma (\boldsymbol{\tau}_2 - \boldsymbol{\tau}_1)\mathbf{v}\, ds.$$

From that follows, using the ellipticity (3.15) of $a(\cdot,\cdot)$, the Cauchy-Schwarz inequality, and the trace theorem,

$$\begin{aligned}\gamma\|\mathbf{u}_2-\mathbf{u}_1\|^2_{\mathbf{H}^1(\Omega)} &\leq a(\mathbf{u}_2-\mathbf{u}_1,\mathbf{u}_2-\mathbf{u}_1)\\ &= \int_\Gamma(\boldsymbol{\tau}_2-\boldsymbol{\tau}_1)(\mathbf{u}_2-\mathbf{u}_1)\,ds\\ &\leq \|\boldsymbol{\tau}_2-\boldsymbol{\tau}_1\|_{\mathbf{L}^2(\Gamma)}\cdot\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{L}^2(\Gamma)}\\ &\leq \|\boldsymbol{\tau}_2-\boldsymbol{\tau}_1\|_{\mathbf{L}^2(\Gamma)}\cdot\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{H}^{1/2}(\Gamma)}\\ &\leq C\|\boldsymbol{\tau}_2-\boldsymbol{\tau}_1\|_{\mathbf{L}^2(\Gamma)}\cdot\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{H}^1(\Omega)},\end{aligned}$$

and hence

$$\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{H}^1(\Omega)}\leq\frac{C}{\gamma}\|\boldsymbol{\tau}_2-\boldsymbol{\tau}_1\|_{\mathbf{L}^2(\Gamma)}.$$

□

Lemma 5.4.8. *The averaging operator* $\mathrm{Av}:\mathbf{H}^1(\Omega)\to SE(3)$ *is Lipschitz continuous on the set of symmetric functions.*

Proof. Consider the two component functions $\mathrm{Av}_r:\mathbf{H}^1(\Omega)\to\mathbb{R}^3$ and $\mathrm{Av}_q:\mathbf{H}^1(\Omega)\to\mathrm{SO}(3)$ separately. By Lemma 5.4.3, $\mathrm{Av}_q(\mathbf{u})=\mathrm{Id}$ if $\mathbf{u}$ is symmetric and hence it is trivially Lipschitz continuous. It remains to show that Av_r is Lipschitz continuous to obtain the same property for Av. Let $\mathbf{u}_1,\mathbf{u}_2\in\mathbf{H}^{1/2}(\Gamma)$. Then

$$\begin{aligned}\|\mathrm{Av}_r(\mathbf{u}_2)-\mathrm{Av}_r(\mathbf{u}_1)\| &\leq |\Gamma|^{-1}\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{L}^1(\Gamma)}\\ &\leq |\Gamma|^{-1}\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{L}^2(\Gamma)}\cdot\|1\|_{\mathbf{L}^2(\Gamma)}\\ &= |\Gamma|^{-1/2}\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{L}^2(\Gamma)}\\ &\leq |\Gamma|^{-1/2}\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{H}^{1/2}(\Gamma)}\\ &\leq C|\Gamma|^{-1/2}\|\mathbf{u}_2-\mathbf{u}_1\|_{\mathbf{H}^1(\Omega)},\end{aligned}$$

where we have used Hölder's inequality and the trace theorem. □

Combining Lemmas 5.4.6–5.4.8 and Lemma 5.4.4 we get the following result.

Lemma 5.4.9. $\mathscr{DN}$ *is Lipschitz continuous on* $\widetilde{SE}(3)$.

Before we show that the damped Dirichlet–Neumann operator $\mathscr{DN}_\theta$ has a fixed point on $\widetilde{\mathrm{SE}}(3)$ we need a technical lemma. It shows that the Dirichlet–Neumann method exhibits a certain alternating behavior. Note that the identification of $\widetilde{\mathrm{SE}}(3)$ with $\mathbb{R}$ allows us to treat elements of $\widetilde{\mathrm{SE}}(3)$ as elements of $\mathbb{R}$.

Lemma 5.4.10. *For all* $\lambda,\mu\in\mathbb{R}(=\widetilde{SE}(3))$, *we have* $\mathscr{DN}\lambda<\mathscr{DN}\mu$ *if and only if* $\lambda>\mu$.

Proof. Let $\lambda > \mu$. By Lemma 5.4.6

$$\mathrm{DtN}\,\lambda = (\mathbf{n}_\lambda(0)\boldsymbol{\nu}_0,\ \mathbf{m}_\lambda(0)\boldsymbol{\nu}_0) = \Big(- A_3\Big[\frac{l_1-\lambda}{l} - 1\Big]\mathbf{e}_2,\ (0,0,0)\Big),$$

and similarly for $\mathrm{DtN}\,\mu$. Hence $(\mathbf{n}_\lambda(0)\boldsymbol{\nu}_0)_2 > (\mathbf{n}_\mu(0)\boldsymbol{\nu}_0)_2$. Denote by

$$\boldsymbol{\tau}_\lambda = \Upsilon(\mathrm{DtN}\,\lambda) = \frac{1}{|\Gamma|}\begin{pmatrix} 0 \\ 0 \\ -(\mathbf{n}_\lambda(0)\boldsymbol{\nu}_0)_2 \end{pmatrix}$$

(and similarly for $\boldsymbol{\tau}_\mu$) the corresponding constant Neumann data functions (Lemma 5.4.4), and let $\mathbf{u}_\lambda$, $\mathbf{u}_\mu$ be the corresponding solutions of the linear elasticity problem, i.e.,

$$\begin{aligned} a(\mathbf{u}_\lambda, \mathbf{v}) &= \int_\Gamma \boldsymbol{\tau}_\lambda \mathbf{v}\, ds + \int_\Omega \mathbf{f}\mathbf{v}\, dx + \int_{\Gamma_N} \mathbf{t}\mathbf{v}\, ds \qquad \text{for all } \mathbf{v} \in \mathbf{H}^1_0(\Omega), \\ a(\mathbf{u}_\mu, \mathbf{v}) &= \int_\Gamma \boldsymbol{\tau}_\mu \mathbf{v}\, ds + \int_\Omega \mathbf{f}\mathbf{v}\, dx + \int_{\Gamma_N} \mathbf{t}\mathbf{v}\, ds \qquad \text{for all } \mathbf{v} \in \mathbf{H}^1_0(\Omega). \end{aligned}$$

Subtracting the second from the first equality we get

$$a(\mathbf{u}_\lambda - \mathbf{u}_\mu, \mathbf{v}) = \int_\Gamma (\boldsymbol{\tau}_\lambda - \boldsymbol{\tau}_\mu)\mathbf{v}\, ds \qquad \text{for all } \mathbf{v} \in \mathbf{H}^1_0(\Omega),$$

and, in particular, setting $\mathbf{v} = \mathbf{u}_\lambda - \mathbf{u}_\mu$ and using the ellipticity of $a(\cdot,\cdot)$ we obtain

$$\begin{aligned} 0 &\le \int_\Gamma (\boldsymbol{\tau}_\lambda - \boldsymbol{\tau}_\mu)(\mathbf{u}_\lambda - \mathbf{u}_\mu)\, ds \\ &= \frac{(\mathbf{n}_\mu(0)\boldsymbol{\nu}_0)_2 - (\mathbf{n}_\lambda(0)\boldsymbol{\nu}_0)_2}{|\Gamma|} \int_\Gamma (\mathbf{u}_\lambda - \mathbf{u}_\mu)_2\, ds. \end{aligned}$$

Hence from $\mu < \lambda$ follows $\int_\Gamma (\mathbf{u}_\lambda)_2\, ds < \int_\Gamma (\mathbf{u}_\mu)_2\, ds$. However from symmetry we know that $(\mathrm{Av}_r(\mathbf{u}))_0 = (\mathrm{Av}_r(\mathbf{u}))_1 = 0$, $\mathrm{Av}_q(\mathbf{u}) = \mathrm{Id}$, and hence

$$\begin{aligned} \mathscr{M}\lambda &= \mathrm{Av}(\mathbf{u}_\lambda) = \Big(|\Gamma|^{-1}\int_\Gamma (\mathbf{u}_\lambda + x)_2\, ds,\ \mathrm{Id}\Big) \\ &< \Big(|\Gamma|^{-1}\int_\Gamma (\mathbf{u}_\mu + x)_2\, ds,\ \mathrm{Id}\Big) = \mathrm{Av}(\mathbf{u}_\mu) = \mathscr{M}\mu. \end{aligned}$$

□

We are now ready to show our main result.

Theorem 5.4.1. *For $\theta \in \mathbb{R}$ sufficiently small the operator $\mathscr{M}_\theta$ has a unique fixed point λ^* on $\widetilde{SE}(3)$. The pair of functions $\mathbf{u} : \Omega \to \mathbb{R}^3$ with*

$$\begin{aligned} -\operatorname{div}\boldsymbol{\sigma}(\mathbf{u}) &= \mathbf{f}_{3d} && \text{in } \Omega \\ \mathbf{u} &= 0 && \text{on } \Gamma_D \\ \boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu} &= \mathbf{t} && \text{on } \Gamma_N \\ \mathrm{Av}(\mathbf{u}) &= \lambda^* \end{aligned}$$

and $\varphi : [0, l] \to SE(3)$ *with*

$$\begin{aligned} \mathbf{m}' + \mathbf{r}' \times \mathbf{n} &= 0 \qquad \text{on } [0, l] \\ \mathbf{n}' &= 0 \qquad \text{on } [0, l] \\ \varphi(0) &= \lambda^* \\ \varphi(l) &= \varphi_D \end{aligned}$$

is a solution of (5.14). *In fact, it is the unique symmetric solution.*

Proof. We show that $\mathscr{DN}_\theta$ is a contraction on $\widetilde{\mathrm{SE}}(3)$ if θ is small enough. Existence of a fixed point λ^* then follows from Banach's fixed point theorem. Use again the identification of $\widetilde{\mathrm{SE}}(3)$ with $\mathbb{R}$. Then, for any $\lambda, \mu \in \mathbb{R}$ we have

$$\begin{aligned} |\mathscr{DN}_\theta\lambda - \mathscr{DN}_\theta\mu|^2 &= |\theta(\mathscr{DN}\lambda - \mathscr{DN}\mu) + (1-\theta)(\lambda - \mu)|^2 \\ &= \theta^2|\mathscr{DN}\lambda - \mathscr{DN}\mu|^2 + (1-\theta)^2|\lambda - \mu|^2 \\ &\quad + 2\theta(1-\theta)(\mathscr{DN}\lambda - \mathscr{DN}\mu)(\lambda - \mu) \\ &= \theta^2|\mathscr{DN}\lambda - \mathscr{DN}\mu|^2 + (1-\theta)^2|\lambda - \mu|^2 \\ &\quad - 2\theta(1-\theta)|\mathscr{DN}\lambda - \mathscr{DN}\mu||\lambda - \mu|, \end{aligned}$$

where the last equality follows from Lem. 5.4.10. By Lem. 5.4.9, $\mathscr{DN}$ is Lipschitz continuous with a positive constant L. Hence we get the bound

$$|\mathscr{DN}_\theta\lambda - \mathscr{DN}_\theta\mu|^2 \leq [\theta^2 L^2 + (1-\theta)^2 - 2\theta(1-\theta)L]|\lambda - \mu|^2.$$

Simple algebra shows that the term in brackets is less than one if

$$0 < \theta < \frac{2}{L+1}.$$

We now show that fixed points of $\mathscr{DN}_\theta$ induce solutions of (5.14). This holds even for fixed points $\lambda^* \notin \widetilde{\mathrm{SE}}(3)$. Let λ^* be a fixed point of $\mathscr{DN}_\theta$. Then there is a pair $(\mathbf{n}^*, \mathbf{m}^*) \in \mathbb{R}^3 \times \mathbb{R}^3$ such that

$$(\mathbf{n}^*, \mathbf{m}^*) \in \mathrm{DtN}\,\lambda^* \tag{5.33}$$

and

$$\lambda^* = \mathrm{Av}(\mathfrak{E}(\Upsilon(\mathbf{n}^*, \mathbf{m}^*))). \tag{5.34}$$

Eq. (5.33) implies that there is a function $\varphi : [0, l] \to \mathrm{SE}(3)$ such that

$$\begin{aligned} \mathbf{m}' + \mathbf{r}' \times \mathbf{n} &= 0 \qquad \text{on } [0, l] \\ \mathbf{n}' &= 0 \qquad \text{on } [0, l] \\ \varphi(0) &= \lambda^* \qquad\qquad (5.35a) \\ \varphi(l) &= \varphi_D \\ (\mathbf{n}(0)\boldsymbol{\nu}_0, \mathbf{m}(0)\boldsymbol{\nu}_0) &= (\mathbf{n}^*, \mathbf{m}^*). \qquad\qquad (5.35b) \end{aligned}$$

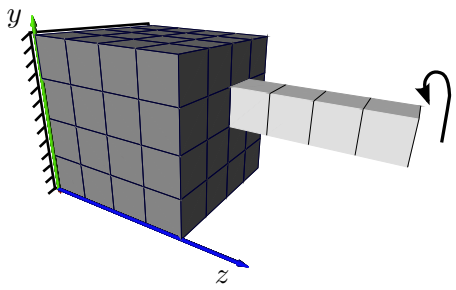


Figure 5.5: Geometric setting of the first numerical example.

Eq. (5.34) in turn means that there is a function $\mathbf{u} : \Omega \to \mathbb{R}^3$ with

$$\begin{aligned} -\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}) &= \mathbf{f} && \text{in } \Omega \\ \mathbf{u} &= 0 && \text{on } \Gamma_D \\ \boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu} &= \mathbf{t} && \text{on } \Gamma_N \\ \operatorname{Av}(\mathbf{u}) &= \lambda^* && && \text{(5.36a)} \\ \boldsymbol{\sigma}(\mathbf{u})\boldsymbol{\nu} &= \Upsilon(\mathbf{n}^*, \mathbf{m}^*) && \text{on } \Gamma. && \text{(5.36b)} \end{aligned}$$

Eqs. (5.35a) and (5.36a) together show that the functions φ and $\mathbf{u}$ fulfill condition (5.14i). From Eqs. (5.35b) and (5.36b) follows that φ and $\mathbf{u}$ fulfill conditions (5.14g) and (5.14h). Conversely, let $\mathbf{u}, \varphi$ be a solution of the coupled problem (5.14). Then $\lambda^* := \varphi(0)$ is a fixed point of $\mathscr{DN}_\theta$. Uniqueness of the solution pair $(\mathbf{u}, \varphi)$ among the symmetric functions follows from the uniqueness of the fixed point λ^* in $\widetilde{\mathrm{SE}}(3)$. □

We immediately get the following corollary.

Lemma 5.4.11. *Let Ω, Γ_D, Γ_N, Γ be symmetric sets; and the volume force field $\mathbf{f}$, and the surface traction $\mathbf{t}$ be symmetric functions. Let the rod reference configuration $\hat{\varphi}$ be straight and the rod Dirichlet value φ_D be of the form $(0, 0, l_1) \times Id$. Let the starting iterate λ^0 be in $\widetilde{SE}(3)$. If the rod subdomain solver implements $\widetilde{\mathrm{DtN}}$, i.e., if it always yields the trivial solution if there is one, and if θ is sufficiently small, the damped Dirichlet–Neumann iteration $\mathscr{DN}_\theta$ converges to the unique symmetric solution of* (5.14).

5.5 Numerical Results

We close this chapter with a few simple numerical examples showing the properties of the Dirichlet–Neumann solver. Consider the unit cube $\Omega = [0, 1]^3$ with a rod attached to the face $[0, 1]^2 \times \{1\}$ (Fig. 5.5). The rod has a quadratic cross-section of edge length $1/4$, a rest length of 1, and is completely straight in its unstressed state. The cross-section at $s = 0$ attaches to the cube at $(5/8, 5/8, 1)$ and the directors at $s = 0$ are set to $\mathbf{d}_1 = (1, 0, 0)$, $\mathbf{d}_2 = (0, 1, 0)$, $\mathbf{d}_3 = (0, 0, 1)$. This corresponds to $\hat{\varphi}_q(0) = \mathrm{Id}$ in the coupling condition (5.12). The interface Γ on the cube Ω is chosen to be $\Gamma = [1/2, 3/4]^2 \times \{1\}$, which is the part of $\partial\Omega$ covered by the rod cross-section in the unstressed state. Note

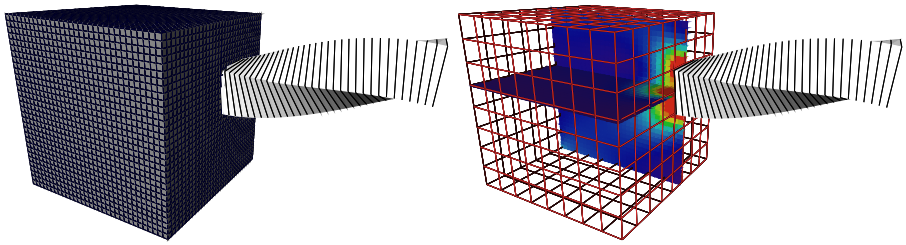


Figure 5.6: Left: resulting deformation after three uniform refinement steps. Right: with cut through the von Mises stress field.

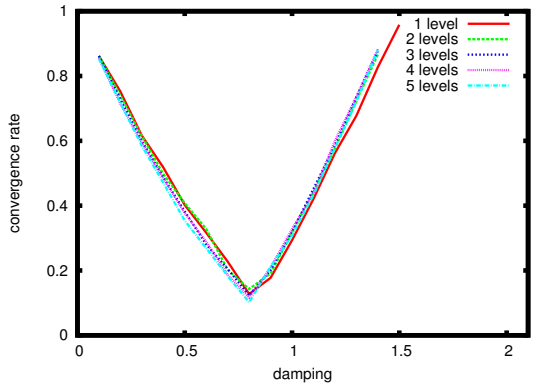


Figure 5.7: Convergence rates as a function of the damping parameter θ for the problem in Fig. 5.5.

that we have deliberately chosen a slightly asymmetric configuration to demonstrate that the symmetry assumptions in Sec. 5.4 are not relevant for actual computations.

We model the cube as a linear elastic material with parameters $E = 10^9$ PU (pressure units) and $\nu = 0.3$. It is discretized with a uniform coarse grid of $4 \times 4 \times 4$ hexahedral elements which is to be uniformly refined up to four times for a final grid with 274625 vertices. Note that Γ aligns with a coarse grid element boundary. We prescribe homogeneous Dirichlet boundary conditions on $[0,1]^2 \times \{0\}$ and homogeneous Neumann conditions everywhere else except for Γ. The rod is modeled with the linear law given in (4.15) and the parameters $E = 10^9$ PU and $\nu = 0.3$. With the square cross-section of edge-length $1/4$ we obtain $J_1 = J_2 = \frac{1}{12}(\frac{1}{4})^4 = 3072^{-1}$ and $J_3 = \frac{1}{6}(\frac{1}{4})^4 = 1536^{-1}$ (see p. 70). The rod is discretized using a uniform coarse grid of 4 elements. At the far end ($s = 1$) we prescribe a movement of $1/4$ LU (length units) in the positive y-direction and a torsion of 90° counterclockwise when viewed in the negative z-direction. This induces all forms of bending, torsion, and shear that can be represented by the rod model.

We solved the combined problem using the Dirichlet–Neumann algorithm described in Sec. 5.3. At each iteration, a Dirichlet problem had to be solved for the rod and a mixed Dirichlet–Neumann problem had to be solved for the cube. For the rod we used the Riemannian trust-region solver of Chap. 4, whereas the linear elasticity problem on the cube was solved using a linear multigrid method. For the rod, the trust-region algorithm iterated until the absolute size of the correction $\|\exp_{\varphi_i}^{-1} \varphi_{i+1}\|_{\infty,T\mathrm{SE}(3)^n}$ dropped below

10^{-12}. In this region rounding errors prevented further improvement. At each trust-region step, the inner monotone multigrid algorithm stopped when a relative correction of 10^{-13} in the H^1-norm was reached. The linear multigrid solver for the cube was set to iterate until the energy norm of the relative correction $\|\mathbf{u}_{i+1} - \mathbf{u}_i\|/\|\mathbf{u}_i\|$ dropped below 10^{-13}. We used IPOpt [90] to solve the minimization problems (5.25) needed to evaluate Υ. Fig. 5.6 shows the resulting deformation and a cut through the von Mises stress field of the cube.

We started the Dirichlet–Neumann iteration at the stress-free configuration $\mathbf{u} = 0$, $\varphi = \hat{\varphi}$. The convergence rate of the Dirichlet–Neumann solver was measured by first iterating until the maximum norm of the correction $\|\mathbf{u}_{\nu+1} - \mathbf{u}_\nu\|_\infty + \|\exp_{\varphi_\nu}^{-1} \varphi_{\nu+1}\|_{\infty, T\mathrm{SE}(3)^n}$ dropped below 10^{-9}. Even though this is several orders of magnitudes away from the nominal machine precision of about 10^{-16}, it is about as far as the solver would go before failing to converge further due to rounding errors. The result $(\mathbf{u}_{\mathrm{ref}}, \varphi_{\mathrm{ref}})$ was then used as a reference solution. Using the list of all iterates we defined the overall error at step ν as

$$e_\nu^2 = \|\mathbf{u}_\nu - \mathbf{u}_{\mathrm{ref}}\|_A^2 + \|\exp_{\varphi_{\mathrm{ref}}}^{-1} \varphi_\nu\|_{H^*}^2,$$

where $\|\cdot\|_A$ is the energy norm corresponding to the linear elasticity stiffness matrix, and $\|\cdot\|_{H^*}$ is the energy norm with respect to the Hessian matrix H^* of the rod energy functional at the reference solution φ_{ref}.

We have measured the convergence rate as a function of the damping parameter θ for up to five levels of uniform refinement. Fig. 5.7 gives the results. For all practical purposes the convergence rate is independent of the grid size. It is well known that mesh independence of the convergence rates holds asymptotically for linear Dirichlet–Neumann algorithms with full-dimensional subdomains [77]. Absolute mesh independence as observed here appears plausible when considering that the interface space SE(3) has a fixed dimension which does not grow with mesh refinement. For practical applications this mesh independence is important, since the optimal damping parameter θ^* can be determined cheaply using a coarse grid. Then all computations involving large and highly-refined grids can be performed using this optimal θ^*. Unfortunately, Sec. 6.2 will show that this mesh-independent behavior is not always obtained. The reason for this is unclear.

To show that our modeling approach does also cover situations where the rod does not meet the 3d object at a right angle we give a second example. Consider the setting as above, with the only difference that now the rod meets the cube at a 45° angle (Fig. 5.8). In terms of the coupling conditions set forth in Sec. 5.2 this means that the factor $\hat{\varphi}_q(0)$ appearing in Cond. (5.12) has a value different from the identity. More precisely, $\hat{\varphi}_q(0)$ is such that $\mathbf{d}_1(\hat{\varphi}_q(0)) = (1, 0, 0)$, $\mathbf{d}_2(\hat{\varphi}_q(0)) = (0, 1/\sqrt{2}, 1/\sqrt{2})$, and $\mathbf{d}_3(\hat{\varphi}_q(0)) = (0, -1/\sqrt{2}, 1/\sqrt{2})$. For the boundary conditions we set $\varphi_r(1) = (0.625, 0.875, 2)$ and the orientation $\varphi_q(1)$ such that $\mathbf{d}_1 = (0, 1, 0)$, $\mathbf{d}_2 = (-1, 0, 0)$, $\mathbf{d}_3 = (0, 0, 1)$. This is the same value as in the previous example. We get the result given in Fig. 5.9. The plot in Fig. 5.10 shows that an identical convergence behavior as before can be observed with this modified coupling.

In the first two examples we have seen a very good numerical behavior. We now give

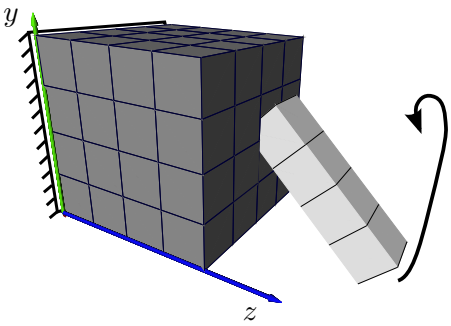


Figure 5.8: Geometric setting of the second numerical example.

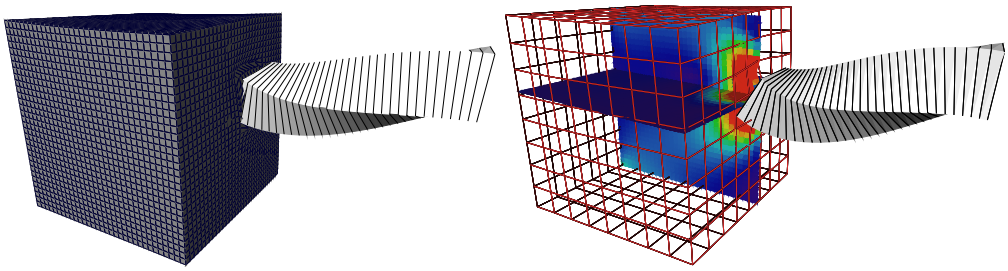


Figure 5.9: Left: resulting deformation after three steps of refinement. Right: cut through the von Mises stress field.

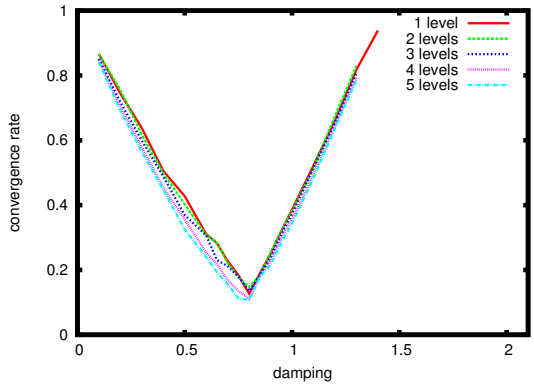


Figure 5.10: Convergence rates as a function of the damping parameter θ for the problem in Fig. 5.8.

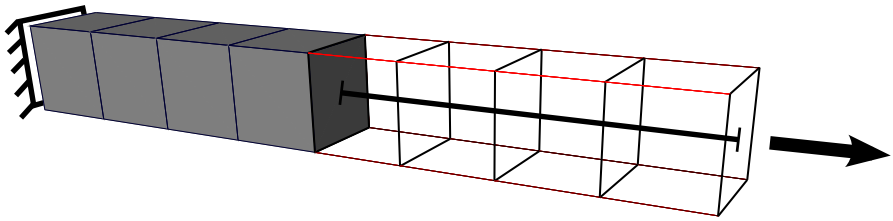


Figure 5.11: A coupling problem showing poor numerical behavior.

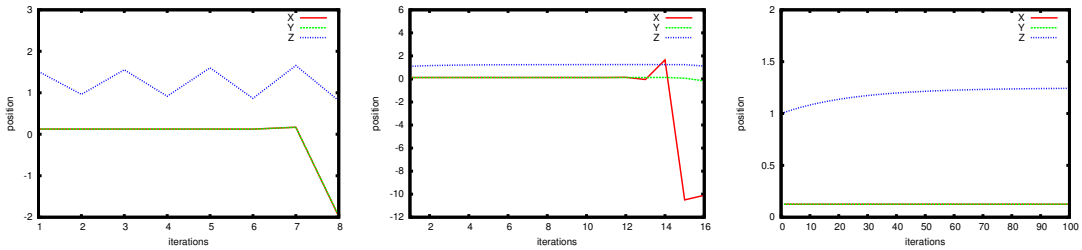


Figure 5.12: The position $\lambda_r^\nu = (\varphi_\nu)_q(0)$ of the interface cross-section for the example of Fig. 5.11 for three damping parameters θ. Left: $\theta = 1.05$, center: $\theta = 0.2$, right: $\theta = 0.02$. In the left and right plots, the graphs for x and y are identical.

a third example which shows that the convergence rates display a certain dependence on the geometry. Consider the setting in Fig. 5.11, where the cube has been replaced by a 'beam' with the same length and cross-section as the Cosserat rod and discretized by four hexahedral elements. Both objects get the same material parameters $E = 10^6$ PU and $\nu = 0.3$. The beam is clamped at its far end and a pure displacement of $(0, 0, 0.5)$ LU is applied to the far end of the rod. From an application point of view this is certainly not a pathological setting, yet for this case the Dirichlet–Neumann algorithm converges poorly. The reason seems to be related to the lateral instability of the structure. Consider Fig. 5.12, where we have plotted the position $\lambda_r^\nu \in \mathbb{R}^3$ of the cross-section at $s = 0$ as a function of the iteration number ν for three different damping parameters θ. Not surprisingly, the method diverges for large θ. For smaller θ, the z-component of λ_r^ν converges nicely, yet there is still divergence for the other two components. Once θ is small enough for all three components to converge it is far to small for the z-component to converge with a reasonable rate. In this case the method converges, but with a rate of only 0.96.

Dependence of Dirichlet–Neumann convergence rates on the problem geometry is a well-known fact. An elucidating example for this in 1d is given in [77, p. 12]. Attempts to increase the robustness of the algorithm are beyond the scope of this work and are left to further research. Sec. 6.2 shows good convergence rates can nevertheless be obtained for real-world problems.

6 Software Issues and Numerical Results

6.1 The Distributed and Unified Numerics Environment (DUNE)

The numerical applications in this thesis are demanding on the software framework. Necessary features include free-form simplicial grids in three space dimensions with local grid refinement, arbitrary boundary parametrizations, second-order isoparametric elements, one-dimensional grids, and the possibility to use several of them at once together with several three-dimensional grids in a single application.

Previous work [60, 95] on two-body contact problems used the UG system [10] for its implementation. UG is well known to be very flexible and offers most features listed above, with the notable exception of one-dimensional grids and isoparametric elements. On the other hand, UG is also known to have a very slow linear algebra and to be very difficult to use, mainly due to an arcane scripting language, little documentation, and bad debugger support.

The last years have seen the development of the DUNE system [1, 11, 12]. The aim of the DUNE project has been to offer more flexibility, efficiency, and productivity to the developers of grid-based PDE applications by addressing a fundamental dilemma. Virtually all PDE software available comes with a grid implementation hardwired into the code. It is, however, impossible for any single implementation to fully satisfy all application writers' demands. For example, one application may need a lean, fast implementation of a structured grid for, say, a flow problem. This implementation is then evidently unable to support local mesh refinement. More general grid managers, on the other hand, may support local refinement, but can never yield the space and time performance of a dedicated structured grid implementation.

DUNE solves this problem by fully separating the grid data structure from the applications that use it. Based on a mathematical definition of a grid, an abstract interface

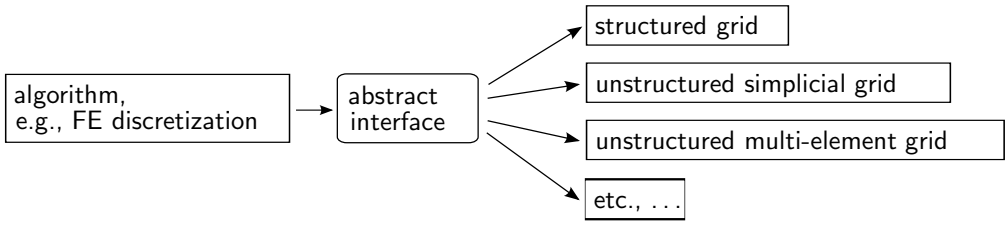


Figure 6.1: Separating the application from the grid implementation by an abstract interface.

realized as a set of C++ classes separates the two (see Fig. 6.1). This allows to change the grid implementation used by an application at any time during code development. It is hence possible to choose the optimal grid implementation for a given application at very little additional cost. It is also possible to freely mix different grid implementations in a single application, a possibility that we have made use of in this thesis. Existing grid managers can be made available through the grid interface with a reasonable amount of work. Also, separating the grid from the data frees the users to choose any linear algebra implementation he or she desires. Compared to, for example, the built-in UG linear algebra this can lead to considerable speedups.

The grid interface is realized using the generic programming techniques of C++. As a result, compilers can optimize away most of the interface. Hence the added flexibility comes at very little run-time cost. A comparison between a DUNE implementation and its equivalent implementation using a hardwired grid manager can be found in [11, Sec. 4.2].

We will now briefly sketch the DUNE definition of a grid. Readers interested in the details can find them in [12]. We will then comment on a few technical aspects of the integration of the UG grid manager `UGGrid`, a step that makes the full power of UG grids available while at the same time easing its use considerably. The section closes describing a few miscellaneous other features of DUNE that have proven valuable for this work.

The DUNE Grid Interface

The DUNE grid interface was designed to support geometric multigrid and locally adaptive algorithms, and hence its notion of a 'grid' directly contains a hierarchical structure. A DUNE grid consists of a finite set of *level grids*, which are connected by a *father relation*. Each level grid in turn consists of an *entity complex* E together with a *geometric realization* M.

The entity complex embodies the topological properties of the level grid. Calling V a finite set of *vertices*, a d-dimensional entity complex consists of the set $E^d = V$, a set E^{d-1} of subsets of E^d called *edges*, a set E^{d-2} of subsets of E^{d-1} called *faces*, and so on. This recursion stops at E^0 whose elements (in the set-theoretic sense) are called *elements* (in the finite element sense).

A mapping R associates to each $e \in E^c$ its reference element, which is a $(d-c)$-dimensional convex polytope in $\mathbb{R}^{d-c}$. The geometric realization M is a family of maps $m(e)$ which for each $e \in E$ contains a map from its reference element $R(e)$ to a w-dimensional Euclidean space which is called the *world space*. Fig. 6.2 illustrates the relationship between the entity complex, the reference elements, and the geometric realization.

Two level grids $G_i = (E_i, M_i)$ and $G_{i+1} = (E_{i+1}, M_{i+1})$ are connected by a father relation. This relation associates a father element in E_i^0 to each element in E_{i+1}^0. Furthermore it associates father entities to some entities of lower dimension. For example, some edges have father edges. The definition includes the case of nonconforming level grids.

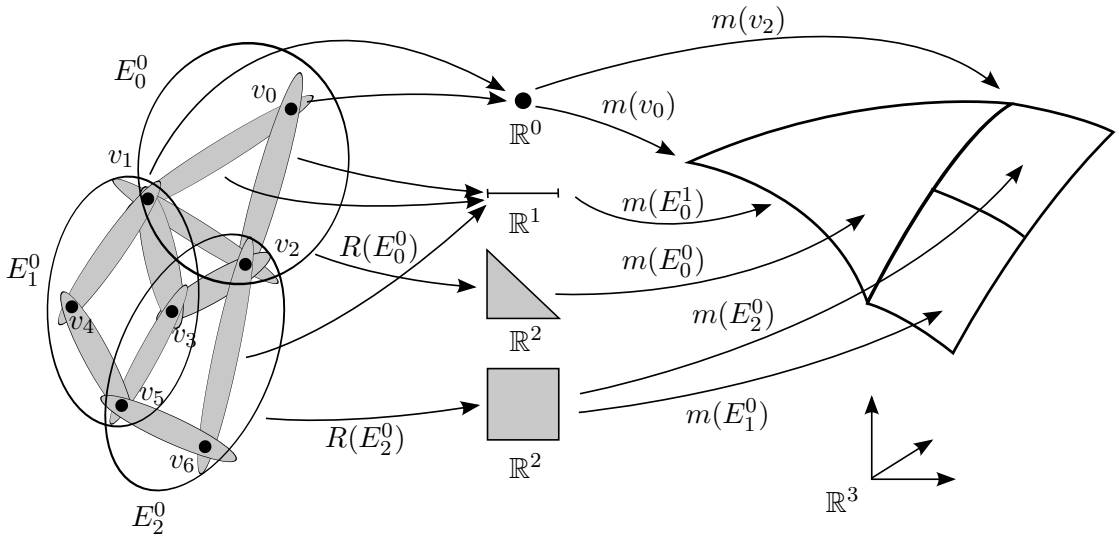


Figure 6.2: A two-dimensional entity complex (left), the reference elements (center), and a geometric realization in $\mathbb{R}^3$ (right).

Together with the father relation, the grid entities form a forest structure. Under certain conditions on the geometric realizations, the leafs of this forest form the *leaf grid*. This is the natural grid for nonhierarchical methods on a locally adaptive grid.

A consequence of the strict separation of grids and algorithms is that numerical data cannot be stored directly in the grid data structure. To connect grids and data the grid interface provides *index maps*. These associate indices with each entity in the entity complexes. The level and leaf index sets provide consecutive indices which can be used to address arrays, but which change if the grid is modified. The indices provided by the *persistent index maps* do not change during the lifetime of an entity. A fortiori they are not consecutive and can only be used to address associative arrays with logarithmic time complexity.

There are several additional features which we only mention in passing. *Intersections* provide more information about the intersections between neighboring elements. This is particularly useful for finite volume methods. Also, the necessary methods for dynamic distributed computing are provided. The interested reader is again referred to [11, 12] for details.

The abstract definition carries over fairly directly into C++ classes. The implementation uses wrapper classes which delegate method calls to engine classes provided as template parameters which do the actual work [88]. Access to the entities is provided by STL-style iterators. See [1] for an up-to-date class documentation.

The `UGGrid` Grid Manager

The UG kernel consists of about 300.000 lines of C code, of which about 150.000 constitute the actual grid manager. The rest consists of linear algebra and solvers, the scripting language interpreter, and graphics output. The code is highly portable; the list of supported platforms contains more than 25 entries. UG uses a hand-written build system. Despite the lack of language support, the internal structure very much reflects an object-oriented design. At various places the code uses preprocessor techniques extensively to force the compiler to create efficient code. In particular, the grid dimension (2 or 3) is a preprocessor constant and hence the entire kernel has to be recompiled in order to change it.

In order to make UG available through the DUNE grid interface, the first step was to replace the old hand-written build system by a new one based on the GNU AutoTools.[1] Some minor adjustments allowed to compile the entire UG kernel as C++. In order to avoid name clashes the kernel was included in the C++ namespace `UG`, and the dimension-dependent parts were additionally included into a nested namespace called `D2` or `D3`, depending on the current dimension setting. Together with a suitable library naming scheme, the 2d and 3d UG libraries could thus be linked together without name clashes. To make the necessary declarations available in DUNE for both dimension settings at the same time, the UG headers are included twice into the DUNE wrapping header `uggrid.hh`, once with the preprocessor dimension flag set to 2 and once to 3. As a result it is possible to use two-dimensional and three-dimensional UG grids together in a single DUNE application, which is impossible using UG alone.

UG has the ability to use boundary parametrizations for grid refinement. The data structures for the automatically created parametrizations (Sec. 3.6) had previously been added to the UG code in an ad-hoc way [61]. Within the DUNE framework this functionality is now cleanly separated between the grid manager, which does the actual refinement, and an external library which is queried for world positions when new boundary nodes are created.

Despite these changes backward compatibility has been preserved, and at the time of writing it is still possible to run normal UG applications using the modified kernel. The modifications are available as a set of patches on the DUNE project homepage [1].

Further Relevant Features

There are a few additional features of the DUNE system which have proven useful for this thesis. The `OneDGrid` grid manager provides unstructured grids with local refinement and coarsening in one space dimension. It has been written from scratch and is used for the ligament models. As it fully complies with the DUNE interface no specific application-side code had to be written for its use.

As a separate part of DUNE, the *Iterative Solver Template Library* (ISTL, [16]) offers efficient linear algebra data structures for finite element applications. Those frequently have a certain amount of structure, which the ISTL data structures can exploit. For

[1] This work was done by Thimo Neubauer.

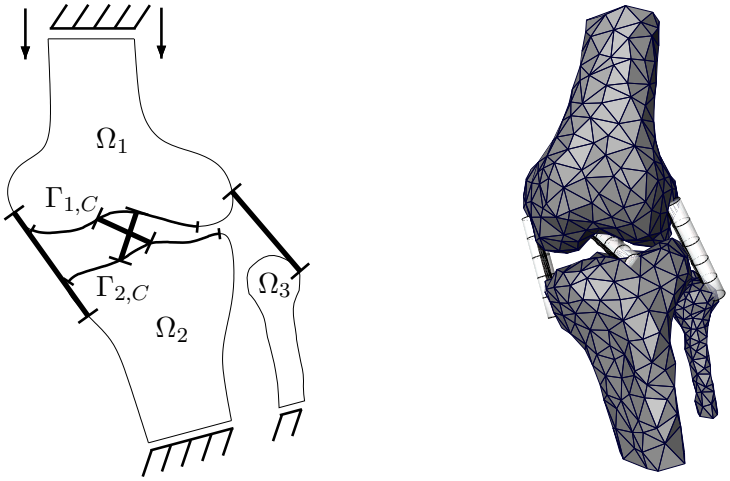


Figure 6.3: Left: problem setting. Tibia and fibula are rotated 15° in valgus direction to put additional stress on the MCL. This can happen, e.g., in a sports accident. Right: the corresponding coarse grids.

example, stiffness matrices for elasticity problems are sparse matrices with each entry a dense 3×3 block. The discretization of Cosserat rod problems with geodesic finite elements (Sec. 4.3) leads to block-band matrices with each block a dense 6×6 matrix. ISTL handles these types comfortably and efficiently by allowing arbitrary nestings of dense, sparse, band, and other matrices.

In Sec. 3.7 grids with isoparametric elements have been used. However, currently no DUNE grid manager provides this feature directly. Instead, we have implemented isoparametric elements as a DUNE meta grid. Meta grids provide views on other grids. In this case, the DUNE meta grid `PolynomialGrid` passes on all queries not related to geometry to the underlying host grid. Calls to the geometry are intercepted and isoparametric elements are simulated using information about the domain boundary available from the grid managers. Since meta grids implement the entire DUNE grid interface they can be used in precisely the same way as regular grids at all times.

6.2 Two-Body Contact and Ligaments

In this closing section we provide simulation results for a knee model which combines three bones of the knee with the four main ligaments. The model includes the distal femur and proximal tibia and fibula bones modeled as three-dimensional linear elastic objects, and the anterior and posterior cruciate ligaments (ACL and PCL, respectively), and the medial and lateral collateral ligaments (MCL and LCL, respectively), which are modeled as Cosserat rods. The model combines the contact problems of Chap. 3 and the heterogeneous coupling of Chap. 5. To obtain a test case where the contact stresses do not entirely predominate the stresses created in the bone by pulling ligaments, we

	fine surfaces		coarse surfaces	
bone	vertices	triangles	vertices	triangles
femur	7236	14468	268	532
tibia	7453	14902	224	444
fibula	1822	3640	126	248

Table 6.1: Sizes of boundary surfaces.

	coarse grids			fine grids			
bone	vertices	elements	aspect ratio	vertices	elements	aspect ratio	inv.
femur	268	1328	14.2	31980	170493	104480	2
tibia	224	1044	17.1	14956	77117	$3.15 \cdot 10^7$	5
fibula	126	368	12.1	2185	9996	113304	14

Table 6.2: Sizes and quality of the coarse and refined grids. Aspect ratio is the largest ratio of the radii of circumsphere and insphere occurring in a grid. Inv. denotes the number of elements with incorrect orientation.

applied a valgus rotation of 15° to tibia and fibula (Fig. 6.3). This leads to a high strain in the MCL and can be interpreted as an imminent MCL rupture.

From the Visible Human data set [3], high-resolution boundary surfaces of the regions of interest were extracted. Using the algorithm described in Sec. 3.6, these were simplified to yield coarse approximations of the geometries, and corresponding boundary parametrizations. From these coarse surfaces tetrahedral grids were built using the Amira grid generator [85]. Finally we applied the aforementioned valgus rotation. Fig. 6.3, right, shows the coarse bone grids, while Table 6.1 and Table 6.2 report on the sizes of the surfaces and grids.

We modeled bone with an isotropic, homogeneous, linear elastic material with $E = 17$ GPa and $\nu = 0.3$. The distal horizontal sections of tibia and fibula were clamped, and a prescribed downward displacement of 2 mm was applied to the upper section of the femur. The part of the femur usually covered with articular cartilage was marked as the nonmortar contact boundary, but note that by construction of the contact mapping Φ, the actual nonmortar boundary was smaller (Sec. 3.5).

The four ligaments ACL, PCL, MCL, and LCL were each modeled by a single Cosserat rod with a circular cross-section of radius 5 mm. The first three ligaments connect the femur to the tibia, while the LCL connects the femur to the fibula. We chose the linear material law (4.15) with parameters $E = 330$ MPa (as suggested by [93, Table 1]) and $\nu = 0.3$. For each rod $\varphi_{\text{lig}} : [0, l_{\text{lig}}] \to \text{SE}(3)$ we chose $s = 0$ to be the proximal and $s = l_{\text{lig}}$ the distal end, where $\text{lig} \in \{\text{ACL}, \text{PCL}, \text{MCL}, \text{LCL}\}$. On the bones, the insertion sites $\Gamma_{\text{bone,lig}} \subset \partial\Omega_{\text{bone}}$, $\text{bone} \in \{\text{femur}, \text{tibia}, \text{fibula}\}$, were set manually based on [75]. For simplicity we chose the sets $\Gamma_{\text{bone,lig}}$ to be resolved by the coarsest grids. Fig. 6.4 shows the insertion sites in the knee model.

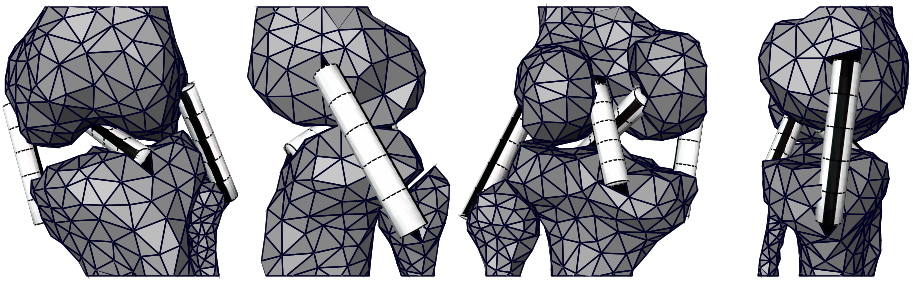


Figure 6.4: Insertion sites of the four ligaments included in the model (marked in black).

ligament	ACL	PCL	MCL	LCL
length	54.5 mm	35.0 mm	40.1 mm	62.7 mm

Table 6.3: Approximate lengths of the ligaments in their stress-free configurations.

The end positions $(\varphi_{\text{lig}})_r(0), (\varphi_{\text{lig}})_r(l_{\text{lig}})$ of the rod centerlines were set to the centers of gravity of the insertion sites on the bones, e.g.,

$$(\varphi_{\text{MCL}})_r(0) = \frac{1}{|\Gamma_{\text{femur,MCL}}|} \int_{\Gamma_{\text{femur,MCL}}} x \, ds$$

and

$$(\varphi_{\text{MCL}})_r(l_{\text{MCL}}) = \frac{1}{|\Gamma_{\text{tibia,MCL}}|} \int_{\Gamma_{\text{tibia,MCL}}} x \, ds.$$

We modeled all ligaments to be straight in their stress-free configurations and to have a length 8% shorter than the distances between their insertions sites when the knee is straight as in Sec. 3.8. Such a value is cited in the literature as a reasonable value for the in situ strain [92]. Table 6.3 gives the stress-free lengths for all four ligaments. In particular, this also determines the factors $(\hat{\varphi}_{\text{lig}})_q(0)$ and $(\hat{\varphi}_{\text{lig}})_q(l_{\text{lig}})$ in (5.12).

We solved the combined problem using the Dirichlet–Neumann algorithm described in Chap. 5. At each iteration, a pure Dirichlet problem had to be solved for each of the four rods and a mixed Dirichlet–Neumann problem with contact between the femur and the tibia had to be solved for the bones. The 3d contact problem for the three bones was solved using the Truncated Nonsmooth Newton Multigrid algorithm of Sec. 3.4. For the ligaments we used the Riemannian trust-region solver of Chap. 4. The TNNMG solver was set to iterate until the energy norm of the relative correction $\|\mathbf{u}_{i+1} - \mathbf{u}_i\|/\|\mathbf{u}_i\|$ dropped below 10^{-12}. For each rod, the trust-region algorithm iterated until the absolute size of the correction $\|\exp_{\varphi_i}^{-1} \varphi_{i+1}\|_{\infty,T\text{SE}(3)^n}$ dropped below 10^{-8}, where $\|\cdot\|_{\infty,T\text{SE}(3)^n}$ is the normed defined in (4.45). There, rounding errors prevented further improvement. At each trust-region step, the inner multigrid algorithm stopped when a relative correction

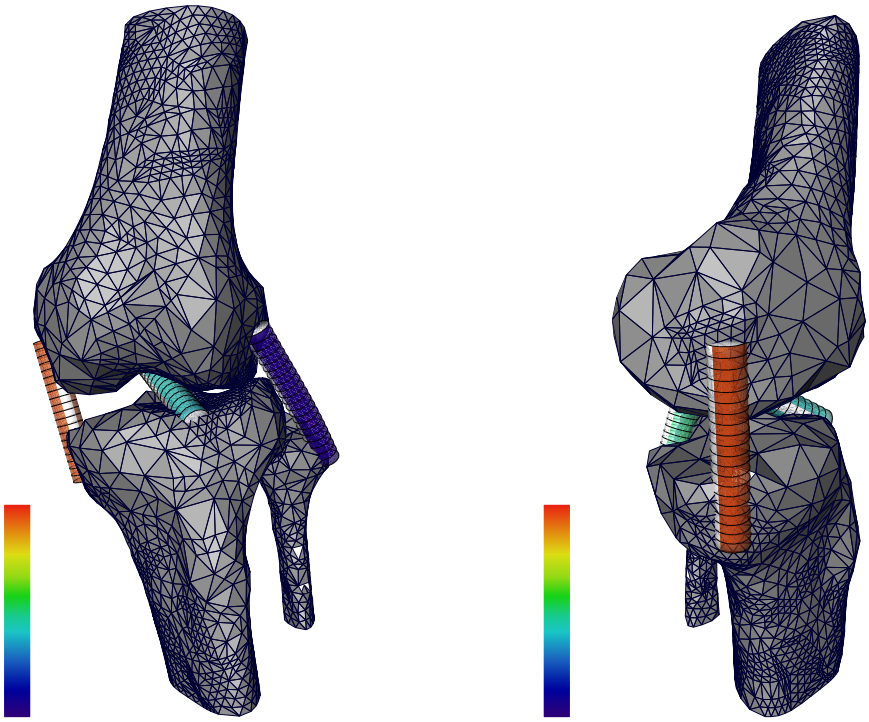


Figure 6.5: Deformed grid after two adaptive refinement steps. Note how the error estimator reacts to the pull of the MCL at its femoral insertion.

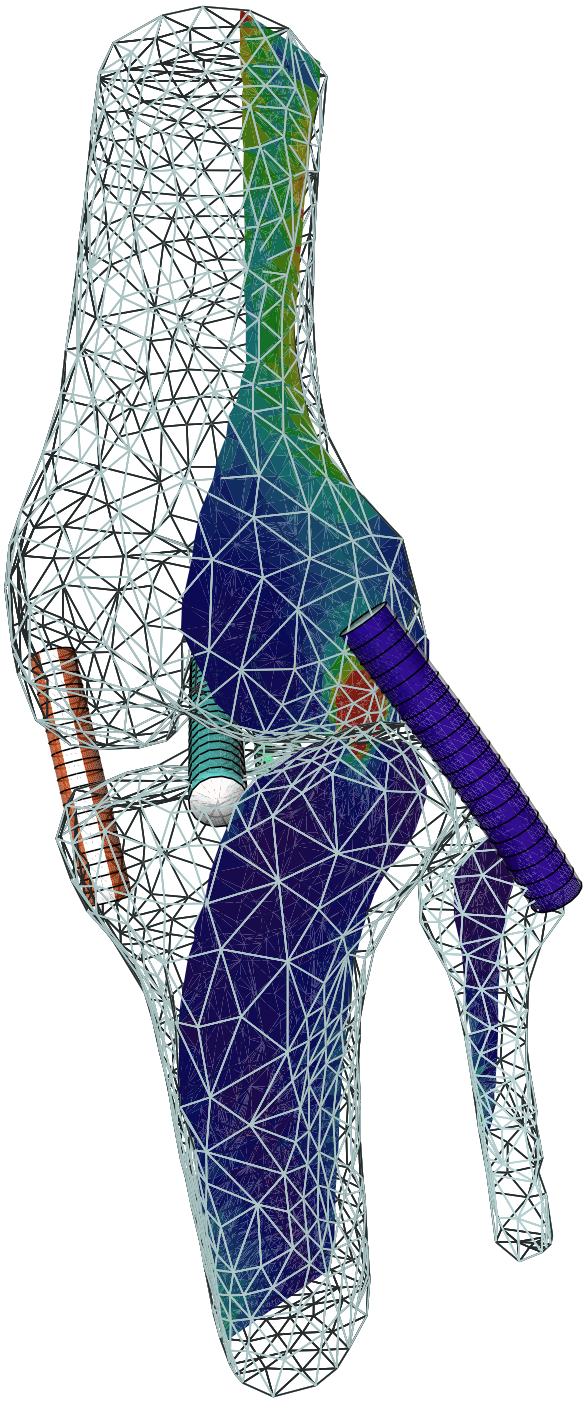


Figure 6.6: Two sagittal cuts through the von Mises stress field. The left cut goes through the area of contact between femur and tibia.

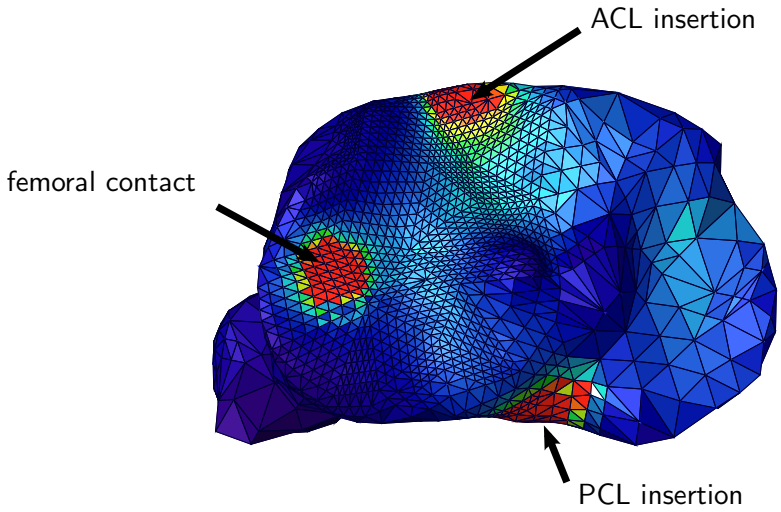


Figure 6.7: View onto the tibial plateau. One can see the pressure created by the cruciate ligaments and the contact with the femur.

of 10^{-12} in the H^1-norm was reached. We used IPOpt [90] to solve the minimization problems (5.25). Fig. 6.5 shows the deformed configuration on a grid obtained by two steps of adaptive refinement. Fig. 6.6 additionally shows cuts through the von Mises stress field. Finally, in Fig. 6.7 a caudal view onto the tibial plateau can be seen, which is colored according to the von Mises stress. The peaks due to contact and the pull of the cruciate ligaments can be clearly observed.

The convergence rate of the Dirichlet–Neumann solver was measured by first iterating until the maximum norm of the correction

$$e_i^\infty = \max_{\text{bones,ligaments}} \left\{ \|\mathbf{u}_{\text{bone}}^{i+1} - \mathbf{u}_{\text{bone}}^{i}\|_\infty, \|\exp_{\varphi_{\text{lig}}^i}^{-1} \varphi_{\text{lig}}^{i+1}\|_{\infty, T\mathrm{SE}(3)^n} \right\}$$

of one Dirichlet–Neumann step dropped below 10^{-9}. There, rounding errors prevented further improvement. The result $\mathbf{u}^{\text{ref}}, \varphi^{\text{ref}}$ was then used as a reference solution. In a subsequent step the computation was started again, and the iterates were compared to the reference solution. At each step i we defined the overall error as

$$e_i^2 = \sum_{\text{bones}} \|\mathbf{u}_{\text{bone}}^i - \mathbf{u}_{\text{bone}}^{\text{ref}}\|_A^2 + \sum_{\text{ligaments}} \|\varphi_{\text{lig}}^i - \varphi_{\text{lig}}^{\text{ref}}\|_{H^*}^2,$$

where $\|\cdot\|_A$ is the energy norm of the linear elasticity problem. The norm for the rod corrections is the energy norm of the Hessian matrix H^* of the rod energy functional j at the reference solution.

We measured the Dirichlet–Neumann convergence rates with grids containing up to four levels. Bone grids were refined adaptively using the error estimator presented in Sec. 3.7. Rod grids in turn were refined uniformly, as the low number of degrees of freedom in the rod problems did not justify the extra effort for local adaptivity.

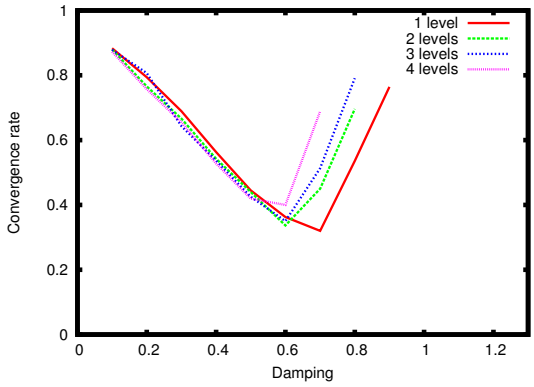


Figure 6.8: Convergence rates of the Dirichlet–Neumann method as a function of the damping parameter for up to four grid levels.

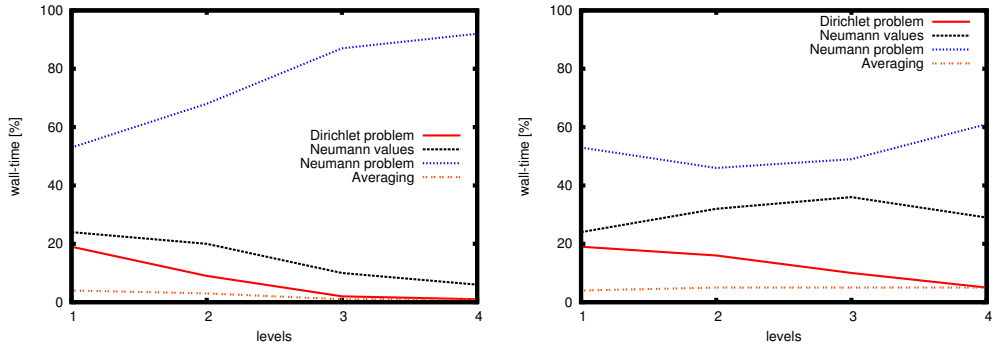


Figure 6.9: Wall-time needed for four substeps of a Dirichlet–Neumann iteration. Numbers are in percent of the total wall-time.

On each new level we started the computation from the reference configuration, i.e., $\mathbf{u}_{\text{bone}} = 0$, $\varphi_{\text{lig}} = \hat{\varphi}_{\text{lig}}$ rather then from the solution on the previous level. That way identical initial iterates for all grid refinement levels were obtained. Fig. 6.8 shows the Dirichlet–Neumann convergence rates plotted as a function of the damping parameter θ for up to four levels of refinement. When compared to Figs. 5.7 and 5.10, the perfect grid independence of the rates is lost. For each further level of refinement, the optimal convergence rate is slightly worse than the previous and obtained for a slightly lower damping parameter. This behavior seems typical for Dirichlet–Neumann methods (see, e.g., [14]). Nevertheless the optimal convergence rates stay around 0.4. This makes the algorithm well usable in practice.

We have also measured the wall-time behavior of the Dirichlet–Neumann solver. Within one iteration four substeps can be distinguished that need a relevant amount of time. These are the solution of the Dirichlet and Neumann problems, the construction of the Neumann value fields $\boldsymbol{\tau}^k$, and the evaluation of the averaging operator Av. Fig. 6.9, left, gives the percentage of wall-time for each of the four steps for grids of different sizes. Note that these timings are approximate as the implementation is not optimized for speed. The cost for the solution of the Neumann contact problem domi-

nates by far. The main reason for this is the bad grid quality due to adaptive refinement and the parametrized boundary, which leads to bad multigrid convergence rates. For comparison we also give the same timings for a grid without boundary parametrization (Fig. 6.9, right). The run-time cost is much more equilibrated, and of course the overall time for one iteration is lower. Ensuring good multigrid convergence rates on grids with parametrized boundaries remains one of the most important open problems in this context.

A The Derivatives of the Strains of a Cosserat Rod

In this appendix we derive closed-form expressions for the gradients of the strains u and v of a discrete Cosserat rod model with respect to its geodesic finite element coefficients. These gradients appear in the expression of the quadratic models of the lifted energy functional $\hat{\jmath}$ (4.39).

Remember that for a given rod configuration $\varphi : [0, l] \to \mathrm{SE}(3)$, $\varphi(s) = (\mathbf{r}(s), q(s))$, we have

$$\mathsf{v} = (\mathsf{v}_1, \mathsf{v}_2, \mathsf{v}_3), \qquad \mathsf{v}_k = \langle \mathbf{r}', \mathbf{d}_k \rangle, \qquad k \in \{1, 2, 3\},$$

the stretching and shear strain and

$$\mathsf{u} = (\mathsf{u}_1, \mathsf{u}_2, \mathsf{u}_3), \qquad \mathsf{u}_k = 2\mathbf{B}_k(q)q', \qquad k \in \{1, 2, 3\},$$

the strain due to bending and torsion. Let G be a one-dimensional grid on the interval $[0, l]$ with n vertices. If φ_h is a geodesic finite element approximation of φ on G, and $s \in \mathbb{R}$ is contained in the element $[l_i, l_{i+1}]$ for some $i \in \{0, \dots, n-2\}$, then the strains $\mathsf{v}(s)$ and $\mathsf{u}(s)$ of φ_h only depend on s and the coefficients (r_i, q_i) and (r_{i+1}, q_{i+1}) of φ_h. We need to compute the derivatives of $\mathsf{v}(s)$ and $\mathsf{u}(s)$ with respect to these coefficients to be able to compute the gradients $\nabla_w \hat{\jmath}$ and $\nabla_v \hat{\jmath}$ defined in (4.40).

We begin by computing the first and second derivatives of the exponential map of $\mathrm{SO}(3)$ at the identity

$$\exp : \mathfrak{so}(3) \to \mathrm{SO}(3).$$

We are using unit quaternions as coordinates on $\mathrm{SO}(3)$ and we identify $\mathfrak{so}(3)$ with $\mathbb{R}^3$ using the hat map (4.5). We therefore interpret the exponential map as

$$\exp : \mathbb{R}^3 \to \mathbb{H}_{|1|}.$$

In (4.7) we gave the closed-form expressions

$$(\exp v)_i = \Big(\sin \frac{|v|}{2}\Big) \frac{v_i}{|v|}, \quad i \in \{1, 2, 3\} \qquad \text{and} \qquad (\exp v)_4 = \cos \frac{|v|}{2}.$$

First note that

$$\frac{\partial |v|}{\partial v_i} = \frac{v_i}{|v|}, \qquad i \in \{1, 2, 3\},$$

and let δ_{im} be the Kronecker symbol. For $i, m \in \{1,2,3\}$ we have

$$\frac{\partial(\exp v)_m}{\partial v_i} = \frac{1}{2}\cos\frac{|v|}{2}\cdot\frac{v_i v_m}{|v|^2} + \sin\frac{|v|}{2}\cdot\left(\frac{\delta_{im}}{|v|} - \frac{v_i v_m}{|v|^3}\right)$$

and

$$\frac{\partial(\exp v)_4}{\partial v_i} = -\frac{1}{2}\sin\frac{|v|}{2}\cdot\frac{v_i}{|v|}.$$

In particular we are interested in the derivatives at $v = 0$. We know that exp is continuous there [33, Prop. 3.2.9]. This is confirmed by the following lemma.

Lemma A.0.1. *For all $i \in \{1,2,3\}$, $m \in \{1,2,3,4\}$*

$$\lim_{|v|\to 0} \left.\frac{\partial(\exp v)_m}{\partial v_i}\right|_{v=0} = \frac{\delta_{im}}{2}.$$

Proof. Let first $m \neq 4$. Then

$$\begin{aligned}\left.\frac{\partial(\exp v)_m}{\partial v_i}\right|_{v=0} &= \lim_{|v|\to 0}\left(\frac{1}{2}\cos\frac{|v|}{2} - \sin\frac{|v|}{2}\cdot|v|^{-1}\right)\frac{v_i v_m}{|v|^2} + \lim_{|v|\to 0}\sin\frac{|v|}{2}\cdot\frac{\delta_{im}}{|v|} \\ &= \frac{\delta_{im}}{2},\end{aligned}$$

since the term in parentheses converges to zero and $v_i v_m |v|^{-2}$ is bounded from above by 1. For $m = 4$ we have

$$\left.\frac{\partial(\exp v)_4}{\partial v_i}\right|_{v=0} = -\frac{1}{2}\lim_{|v|\to 0}\sin\frac{|v|}{2}\cdot\frac{v_j}{|v|} = 0.$$

□

The second derivatives of exp are

$$\begin{aligned}\frac{\partial^2(\exp v)_m}{\partial v_i\,\partial v_j} &= -\frac{1}{4}\frac{v_i v_j v_m}{|v|^3}\sin\frac{|v|}{2} \\ &\quad + \left(\delta_{ij}\frac{v_m}{|v|^2} + \delta_{jm}\frac{v_i}{|v|^2} + \delta_{im}\frac{v_j}{|v|^2} - 3\frac{v_i v_j v_m}{|v|^4}\right)\left(\frac{1}{2}\cos\frac{|v|}{2} - \frac{\sin\frac{|v|}{2}}{|v|}\right)\end{aligned} \tag{A.1}$$

for $i, j, m \in \{1,2,3\}$, and

$$\frac{\partial^2(\exp v)_4}{\partial v_i\,\partial v_j} = -\frac{1}{4}\cos\frac{|v|}{2}\cdot\frac{v_i v_j}{|v|^2} - \frac{1}{2}\sin\frac{|v|}{2}\cdot\left(\frac{\delta_{ij}}{|v|} - \frac{v_i v_j}{|v|^3}\right).$$

Expression (A.1) can be continuously extended to $v = 0$. First let $a_i \in \mathbb{N}_0$, $1 \leq i \leq k$ and $n = \sum_{i=1}^k a_i$ and pick a $w \in \mathbb{R}^k$. Then for each component w_j of w

$$|w_j|^2 = w_j^2 \leq \sum_{i=1}^k w_i^2 = |w|^2$$

and hence

$$\left|\frac{\prod_i w_i^{a_i}}{|w|^n}\right| = \frac{\prod_i |w_i|^{a_i}}{|w|^n} \leq \frac{\prod_i |w|^{a_i}}{|w|^n} = \frac{|w|^n}{|w|^n} = 1. \tag{A.2}$$

Lemma A.0.2. *For all $i, j \in \{1,2,3\}$, $m \in \{1,2,3,4\}$*

$$\lim_{|v|\to 0} \frac{\partial^2 (\exp v)_m}{\partial v_i \, \partial v_j} = -\frac{\delta_{ij}\delta_{m4}}{4}$$

holds.

Proof. Let first $m \neq 4$ and consider the expression for $\frac{\partial^2 (\exp v)_m}{\partial v_i \, \partial v_j}$ given in (A.1). By (A.2), we have

$$\lim_{|v|\to 0} \Big[-\frac{1}{4}\frac{v_i v_j v_m}{|v|^3} \sin\frac{|v|}{2}\Big] = 0.$$

For the second line of (A.1) note that while the term in the second parenthesis converges to zero for $|v| \to 0$, it is easy to choose sequences $\{v^k \in \mathbb{R}^3 \mid k = 0, 1, \dots\}$ such that the first term diverges. We expand the term in the second parenthesis as a series

$$\begin{aligned} \frac{1}{2}\cos\frac{|v|}{2} - \frac{\sin\frac{|v|}{2}}{|v|} &= \frac{1}{2} - \frac{|v|^2}{16} + O(|v|^4) - \frac{1}{2} + \frac{|v|^2}{48} - O(|v|^4) \\ &= -\frac{|v|^2}{24} + O(|v|^4), \end{aligned}$$

and note that the term denoted by $O(|v|^4)$ has the form

$$\sum_{i=4}^{\infty} b_i |v|^i = |v|^4 \sum_{i=0}^{\infty} b_{i+4} |v|^{i+4},$$

for scalar coefficients b_i. As a scalar polynomial in the variable $|v|$, the absolute value of $\sum_{i=0}^{\infty} b_{i+4}|v|^{i+4}$ is bounded on any closed interval $[-\rho, \rho]$, $0 < \rho \in \mathbb{R}$ by a constant $C = C(\rho)$. Hence,

$$\left|\frac{1}{2}\cos\frac{|v|}{2} - \frac{\sin\frac{|v|}{2}}{|v|}\right| \leq \frac{|v|^2}{24} + C|v|^4$$

holds for all v with $|v| < \rho$, and

$$\begin{aligned} \lim_{|v|\to 0} \left| \left(\delta_{ij}\frac{v_m}{|v|^2} + \delta_{jm}\frac{v_i}{|v|^2} + \delta_{im}\frac{v_j}{|v|^2} - 3\frac{v_i v_j v_m}{|v|^4}\right)\left(\frac{1}{2}\cos\frac{|v|}{2} - \frac{\sin\frac{|v|}{2}}{|v|}\right)\right| \\ \leq \left| \left(\delta_{ij}\frac{v_m}{|v|} + \delta_{jm}\frac{v_i}{|v|} + \delta_{im}\frac{v_j}{|v|} - 3\frac{v_j v_l v_m}{|v|^3}\right)\right| \left| \left(\frac{|v|}{24} + C|v|^3\right)\right|. \end{aligned}$$

Using again (A.2) the term in the first parenthesis is bounded and hence the limit is zero.

The coefficients for $m = 4$ can be evaluated at $v = 0$ directly without using limits

$$\left.\frac{\partial^2(\exp v)_4}{\partial v_i\,\partial v_j}\right|_{v=0} = -\frac{1}{2}\frac{\partial}{\partial v_j}\left(\sin\frac{|v|}{2}\cdot\frac{v_i}{|v|}\right)\Bigg|_{v=0} = -\frac{1}{2}\left.\frac{\partial(\exp v)_i}{\partial v_j}\right|_{v=0} = -\frac{\delta_{ij}}{4}.$$

□

For later reference we also compute here the derivatives of the inverse exponential map

$$\exp^{-1} : \mathbb{H}_{|1|} \to \mathbb{R}^3.$$

Note first that the derivative $D\exp$ evaluated at $v \in \mathfrak{so}(3)$ is a linear mapping from $T_v\mathfrak{so}(3)$ to $T_{\exp v}\mathrm{SO}(3)$. In coordinates $D\exp$ maps $\mathbb{R}^3$ to $T_{\exp v}\mathbb{H}_{|1|}$, which is a three-dimensional linear subspace of $\mathbb{R}^4$. Hence $D\exp$ is represented as the 4×3 matrix $\partial\exp/\partial v$. Differentiating now the identity

$$\exp^{-1}(\exp v) = v$$

and using the chain rule we get

$$\mathrm{Id}_{3\times 3} = \left.\frac{\partial\exp^{-1}}{\partial q}\right|_{q=\exp v}\cdot\frac{\partial\exp v}{\partial v}.$$

Consequently, we obtain $\partial\exp^{-1}/\partial q$ as the Moore–Penrose pseudo inverse of $\partial\exp/\partial v$

$$\frac{\partial\exp^{-1}}{\partial q} = \Big(\frac{\partial\exp}{\partial v}\Big)^{+} := \Big[\Big(\frac{\partial\exp}{\partial v}\Big)^{T}\Big(\frac{\partial\exp}{\partial v}\Big)\Big]^{-1}\Big(\frac{\partial\exp}{\partial v}\Big)^{T}.$$

It is well-defined because by the local bijectivity of exp the columns of $\partial\exp/\partial v$ span the three-dimensional tangent space $T_{\exp v}\mathbb{H}_{|1|}$ and hence the matrix

$$\Big(\frac{\partial\exp}{\partial v}\Big)^{T}\Big(\frac{\partial\exp}{\partial v}\Big) \in \mathbb{R}^{3\times 3}$$

is invertible.

As the next step we consider the derivatives of the geodesic finite element interpolation formula (4.22)

$$\mathsf{q}(s, q^0, q^1) = q^0\exp\big[\tfrac{s}{\delta}\exp^{-1}((q^0)^{-1}q^1)\big]. \tag{A.3}$$

Let $q^0, q^1 \in \mathrm{SO}(3)$ be such that $\mathrm{dist}(q^0, q^1) < \pi$. Then, by Lemma 4.3.3 there exist neighborhoods $V_0, V_1 \in \mathrm{SO}(3)$ of q^0 and q^1, respectively, such that q is well-defined on $[0,\delta]\times V_0\times V_1$. Keep $s \in [0,\delta]$ fixed and regard (A.3) as a function $\mathsf{q}_s : \mathrm{SO}(3)\times\mathrm{SO}(3) \to \mathrm{SO}(3)$. In order to compute the derivatives with respect to q^0 and q^1 we introduce normal coordinates around q^0 and q^1 using the exponential map and the canonical isomorphisms between individual tangent spaces of $\mathrm{SO}(3)$ and $\mathbb{R}^3$. With these coordinates q_s has the form

$$\begin{aligned}\mathsf{q}_s(\cdot,\cdot) &: \mathbb{R}^3\times\mathbb{R}^3 \to \mathbb{H}_{|1|}\\ \mathsf{q}_s(w^0, w^1) &= q^0\exp w^0\exp\big[\tfrac{s}{\delta}\exp^{-1}((q^0\exp w^0)^{-1}q^1\exp w^1)\big],\end{aligned}$$

and the derivatives at $0 = \exp_{p^0}^{-1} p^0 \in \mathbb{R}^3$ and $0 = \exp_{p^1}^{-1} p^1 \in \mathbb{R}^3$ correspond to the derivatives at p^0 and p^1, respectively.

We first compute the derivatives of q_s with respect to w^1. To shorten the notation we set $w^0 = 0$, because we are only interested in the derivatives at $w^0 = 0$, $w^1 = 0$. For abbreviation introduce $\xi(w^0, w^1) = \exp^{-1}((q^0 \exp w^0)^{-1} q^1 \exp w^1)$. The derivatives of q_s with respect to w_j^1, $j \in \{1, 2, 3\}$ are

$$\begin{aligned}
\left.\frac{\partial \mathsf{q}_s}{\partial w_j^1}\right|_{w^0=0} &= q^0 \frac{\partial}{\partial w_j^1} \exp\left[\frac{s}{\delta}\xi(0, w^1)\right] = q^0 \left.\frac{\partial \exp}{\partial v}\right|_{v=\frac{s}{\delta}\xi(0,w^1)} \frac{\partial}{\partial w_j^1}\left[\frac{s}{\delta}\xi(0, w^1)\right] \\
&= q^0 \left.\frac{\partial \exp}{\partial v}\right|_{v=\frac{s}{\delta}\xi(0,w^1)} \frac{s}{\delta} \left.\frac{\partial \exp^{-1}}{\partial q}\right|_{q=(q^0)^{-1}q^1 \exp w^1} (q^0)^{-1} q^1 \frac{\partial \exp w^1}{\partial w_j^1}.
\end{aligned}$$

The multiplications occurring in this expression are quaternion multiplications whenever two elements of $\mathbb{H}_{|1|}$ are involved, and regular matrix–vector or matrix–matrix multiplications otherwise.

Rather than computing the derivatives of q_s with respect to w^0 directly we note that $\mathsf{q}(s, q^0, q^1) = \mathsf{q}(\delta - s, q^1, q^0)$ and hence, for $j \in \{1, 2, 3\}$,

$$\begin{aligned}
\left.\frac{\partial \mathsf{q}_s}{\partial w_j^0}\right|_{w^1=0} &= \frac{\partial}{\partial w_j^0} \mathsf{q}(\delta - s, q^1, q^0 \exp w^0) \\
&= q^1 \left.\frac{\partial \exp}{\partial v}\right|_{v=\frac{(\delta-s)}{\delta}\exp^{-1}((q^1)^{-1}q^0 \exp w^0)} \\
&\quad \frac{(\delta - s)}{\delta} \left.\frac{\partial \exp^{-1}}{\partial q}\right|_{q=(q^1)^{-1}q^0 \exp w^0} (q^1)^{-1} q^0 \frac{\partial \exp w^0}{\partial w_j^0}.
\end{aligned}$$

Evaluating the two previous expressions at 0 we obtain

$$\begin{aligned}
\left.\frac{\partial \mathsf{q}_s}{\partial w_j^0}\right|_{w^0,w^1=0} &= q^1 \left.\frac{\partial \exp}{\partial v}\right|_{v=\frac{(\delta-s)}{\delta}\exp^{-1}((q^1)^{-1}q^0)} \\
&\quad \frac{(\delta - s)}{\delta} \left.\frac{\partial \exp^{-1}}{\partial q}\right|_{q=(q^1)^{-1}q^0} (q^1)^{-1} q^0 \left.\frac{\partial \exp w^0}{\partial w_j^0}\right|_{w^0=0}, \\
\left.\frac{\partial \mathsf{q}_s}{\partial w_j^1}\right|_{w^0,w^1=0} &= q^0 \left.\frac{\partial \exp}{\partial v}\right|_{v=\frac{s}{\delta}\exp^{-1}((q^0)^{-1}q^1)} \\
&\quad \frac{s}{\delta} \left.\frac{\partial \exp^{-1}}{\partial q}\right|_{q=(q^0)^{-1}q^1} (q^0)^{-1} q^1 \left.\frac{\partial \exp w^1}{\partial w_j^1}\right|_{w^1=0}.
\end{aligned}$$

For the evaluation of the rotational strain measure u (4.12) we also need to compute the derivatives of the interpolated velocity vector $\partial \mathsf{q}(s, q^0, q^1)/\partial s$ with respect to q^0 and

q^1. Remember the interpolation formula (4.23)

$$\frac{\partial \mathsf{q}(s,q^0,q^1)}{\partial s} = q^0 \frac{\partial \exp}{\partial v}\bigg|_{v=\frac{s}{\delta}\exp^{-1}((q^0)^{-1}q^1)} \frac{\exp^{-1}((q^0)^{-1}q^1)}{\delta}.$$

We keep s fixed and interpret this as a mapping

$$\frac{\partial \mathsf{q}_s}{\partial s} : \mathrm{SO}(3)\times\mathrm{SO}(3) \to T_{\mathsf{q}(s,q^0,q^1)}\mathrm{SO}(3).$$

Again we use normal coordinates around q^0 and q^1 and obtain the representation

$$\begin{aligned} \frac{\partial \mathsf{q}_s}{\partial s} &: \mathbb{R}^3\times\mathbb{R}^3 \to T_{\mathsf{q}(s,q^0,q^1)}\mathbb{H}_{|1|}\subset\mathbb{R}^4 \\ \frac{\partial \mathsf{q}_s(w^0,w^1)}{\partial s} &= q^0\exp w^0 \frac{\partial \exp}{\partial v}\bigg|_{v=\frac{s}{\delta}\xi(w^0,w^1)} \frac{\xi(w^0,w^1)}{\delta}. \end{aligned}$$

First we compute the variation of $\partial \mathsf{q}_s/\partial s$ with respect to w^1, again setting $w^0=0$,

$$\begin{aligned} \frac{\partial^2 \mathsf{q}_s}{\partial s\,\partial w_j^1} &= q^0 \frac{\partial^2 \exp}{\partial v^2}\bigg|_{v=\frac{s}{\delta}\xi(0,w^1)} \frac{\partial}{\partial w_j^1}\Big[\frac{s}{\delta}\xi(0,w^1)\Big]\frac{\xi(0,w^1)}{\delta} \\ &\quad + q^0 \frac{\partial \exp}{\partial v}\bigg|_{v=\frac{s}{\delta}\xi(0,w^1)} \frac{1}{\delta}\frac{\partial \exp^{-1}}{\partial q}\bigg|_{q=(q^0)^{-1}q^1\exp w^1} (q^0)^{-1}q^1 \frac{\partial \exp w^1}{\partial w_j^1}. \end{aligned} \tag{A.4}$$

Note that $\partial^2\exp/\partial v^2$ is a 3×3 matrix for each of the four components of exp as a map onto $\mathbb{H}_{|1|}$. The two subsequent factors $\frac{\partial}{\partial w_j^1}\Big[\frac{s}{\delta}\xi(0,w^1)\Big]$ and $\frac{\xi(0,w^1)}{\delta}$ are both three-vectors and the overall product is to be understood as a vector–matrix–vector multiplication for each of the four component matrices of $\partial^2\exp/\partial v^2$.

For the derivative with respect to q^0 we use the symmetry relation

$$\frac{\partial \mathsf{q}}{\partial s}(s,q^0,q^1) = -\frac{\partial \mathsf{q}}{\partial s}(\delta - s, q^1, q^0)$$

and get, noting that $-\xi(w^0,w^1) = \exp^{-1}((q^1\exp w^1)^{-1}q^0\exp w^0)$,

$$\begin{aligned} \frac{\partial^2 \mathsf{q}_s}{\partial s\,\partial w_j^0} &= -q^1 \frac{\partial^2 \exp}{\partial v^2}\bigg|_{v=-\frac{(\delta-s)}{\delta}\xi(w^0,0)} \frac{\partial}{\partial w_j^0}\Big[-\frac{(\delta-s)}{\delta}\xi(w^0,0)\Big]\frac{-\xi(w^0,0)}{\delta} \\ &\quad - q^1 \frac{\partial \exp}{\partial v}\bigg|_{v=-\frac{(\delta-s)}{\delta}\xi(w^0,0)} \frac{1}{\delta}\frac{\partial \exp^{-1}}{\partial q}\bigg|_{q=(q^1)^{-1}q^0\exp w^0} (q^1)^{-1}q^0 \frac{\partial \exp w^0}{\partial w_j^0}. \end{aligned} \tag{A.5}$$

Evaluating the expressions (A.4) and (A.5) at $w^1 = 0$ and $w^0 = 0$, respectively, we get

$$\begin{aligned}
\left.\frac{\partial^2 \mathsf{q}_s}{\partial s\, \partial w_j^0}\right|_{w^0,w^1=0} &= q^1 \left.\frac{\partial^2 \exp}{\partial v^2}\right|_{v=\frac{(s-\delta)}{\delta}\xi(0,0)} \frac{\partial}{\partial w_j^0}\Big[\frac{(s-\delta)}{\delta}\xi(w^0,0)\Big]\frac{\xi(0,0)}{\delta} \\
&\quad - q^1 \left.\frac{\partial \exp}{\partial v}\right|_{v=\frac{(s-\delta)}{\delta}\xi(0,0)} \frac{1}{\delta}\left.\frac{\partial \exp^{-1}}{\partial q}\right|_{q=(q^1)^{-1}q^0} (q^1)^{-1}q^0 \left.\frac{\partial \exp w^0}{\partial w_j^0}\right|_{w^0=0} \\
\left.\frac{\partial^2 \mathsf{q}_s}{\partial s\, \partial w_j^1}\right|_{w^0,w^1=0} &= q^0 \left.\frac{\partial^2 \exp}{\partial v^2}\right|_{v=\frac{s}{\delta}\xi(0,0)} \frac{\partial}{\partial w_j^1}\Big[\frac{s}{\delta}\xi(0,w^1)\Big]\frac{\xi(0,0)}{\delta} \\
&\quad + q^0 \left.\frac{\partial \exp}{\partial v}\right|_{v=\frac{s}{\delta}\xi(0,0)} \frac{1}{\delta}\left.\frac{\partial \exp^{-1}}{\partial q}\right|_{q=(q^0)^{-1}q^1} (q^0)^{-1}q^1 \left.\frac{\partial \exp w^1}{\partial w_j^1}\right|_{w^1=0}.
\end{aligned}$$

Using these results we can now evaluate the strains $\mathsf{v}(s)$ and $\mathsf{u}(s)$ of a Cosserat rod discretized by first-order geodesic finite element functions. Let G be a one-dimensional grid on the interval $[0, l]$ with n vertices $0 = l_0 < l_1 < \ldots < l_{n-1} = l$, and let $\varphi_h : [0, l] \to \mathrm{SE}(3)$ be a geodesic finite element function on G, with the coefficient vector $\bar{\varphi} \in \mathrm{SE}(3)^n$. Let $[l_i, l_{i+1}]$, $0 \le i < n-2$, be an element in G. For $s \in [l_i, l_{i+1}]$ we have

$$\varphi_h(s) = \left(\sum_{\substack{j\in\{1,2,3\}\\ k\in\{i,i+1\}}} (\bar{\varphi}_r)_{k,j} \boldsymbol{\psi}_{k,j}(s),\ \mathsf{q}\Big(\frac{s-l_i}{l_{i+1}-l_i}, (\bar{\varphi}_q)_i, (\bar{\varphi}_q)_{i+1}\Big)\right),$$

and for $s \in (l_i, l_{i+1})$ we have

$$\varphi_h'(s) = \left(\sum_{\substack{j\in\{1,2,3\}\\ k\in\{i,i+1\}}} (\bar{\varphi}_r)_{k,j} \frac{\partial \boldsymbol{\psi}_{k,j}}{\partial s}(s),\ \frac{\partial \mathsf{q}}{\partial s}\Big(\frac{s-l_i}{l_{i+1}-l_i}, (\bar{\varphi}_q)_i, (\bar{\varphi}_q)_{i+1}\Big)\right).$$

Both quantities depend only on $(\bar{\varphi}_h)_i \in \mathrm{SE}(3)$ and $(\bar{\varphi}_h)_{i+1} \in \mathrm{SE}(3)$. Consequently, $\mathsf{v}(s)$ and $\mathsf{u}(s)$ depend only on $(\bar{\varphi}_h)_i$ and $(\bar{\varphi}_h)_{i+1}$ as well.

Consider $s \in (l_i, l_{i+1})$ fixed. For $j, m \in \{1, 2, 3\}$ and $k \in \{i, i+1\}$, the derivative of v with respect to the finite element coefficients of $\bar{\varphi}$ is

$$\frac{\partial}{\partial r_j^k}\mathsf{v}_m\big(\mathbf{r}(s),\, \mathsf{q}_s(q^i, q^{i+1})\big) = \Big\langle \frac{\partial \boldsymbol{\psi}_{k,j}}{\partial s},\, \mathbf{d}_m\big(\mathsf{q}_s(q^i, q^{i+1})\big)\Big\rangle$$

and

$$\frac{\partial \mathsf{v}_m}{\partial w_j^k} = \Big\langle \frac{\partial \mathbf{r}}{\partial s},\, \frac{\partial}{\partial w_j^k}\mathbf{d}_m\big(\mathsf{q}_s(w^i, w^{i+1})\big)\Big\rangle = \Big\langle \frac{\partial \mathbf{r}}{\partial s},\, \frac{\partial \mathbf{d}_m}{\partial q}\cdot\frac{\partial}{\partial w_j^k}\mathsf{q}_s(w^i, w^{i+1})\Big\rangle.$$

The brackets denote the scalar product in $\mathbb{R}^3$. The derivatives $\partial \mathbf{d}_m / \partial q$ of the director

vectors follow directly from their definition (4.9),

$$\begin{aligned}
\frac{\partial \mathbf{d}_1}{\partial q} &= 2\begin{pmatrix} q_1 & -q_2 & -q_3 & q_4 \\ q_2 & q_1 & q_4 & q_3 \\ q_3 & -q_4 & q_1 & -q_2 \end{pmatrix} \\
\frac{\partial \mathbf{d}_2}{\partial q} &= 2\begin{pmatrix} q_2 & q_1 & -q_4 & -q_3 \\ -q_1 & q_2 & -q_3 & q_4 \\ q_4 & q_3 & q_2 & q_1 \end{pmatrix} \\
\frac{\partial \mathbf{d}_3}{\partial q} &= 2\begin{pmatrix} q_3 & q_4 & q_1 & q_2 \\ -q_4 & q_3 & q_2 & -q_1 \\ -q_1 & -q_2 & q_3 & q_4 \end{pmatrix}.
\end{aligned}$$

As u is independent of the centerline the derivatives $\partial \mathsf{u}/\partial r_j^k$ are zero. The derivatives with respect to the rotation are, again for $j, m \in \{1,2,3\}$ and $k \in \{i, i+1\}$,

$$\begin{aligned}
\frac{\partial}{\partial w_j^k}\mathsf{u}_m(s) &= 2\frac{\partial}{\partial w_j^k}\Big\langle \mathbf{B}_m\big(\mathsf{q}_s(w^i, w^{i+1})\big),\, \frac{\partial}{\partial s}\mathsf{q}_s(w^i, w^{i+1})\Big\rangle \\
&= 2\Big\langle \mathbf{B}_m\Big(\frac{\partial \mathsf{q}_s(w^i, w^{i+1})}{\partial w_j^k}\Big), \frac{\partial \mathsf{q}_s(w^i, w^{i+1})}{\partial s}\Big\rangle \\
&\quad + 2\Big\langle \mathbf{B}_m\big(\mathsf{q}_s(w^i, w^{i+1})\big), \frac{\partial^2 \mathsf{q}_s(w^i, w^{i+1})}{\partial s\, \partial w_j^k}\Big\rangle.
\end{aligned}$$

Unlike above, the brackets in this expression denote the scalar product in $\mathbb{R}^4$.

Bibliography

[1] DUNE – Distributed and Unified Numerics Environment. http://dune-project.org/.

[2] Wikipedia. http://wikipedia.org/wiki/In_silico.

[3] The Visible Human Project. http://www.nlm.nih.gov/research/visible/visible_human.html.

[4] P.-A. Absil, C. G. Baker, and K. A. Gallivan. Convergence analysis of Riemannian trust-region methods. Technical report, http://www.optimization-online.org/DB_HTML/2006/06/1416.html, 2006.

[5] P.-A. Absil, C. G. Baker, and K. A. Gallivan. Trust-region methods on Riemannian manifolds. *Found. Comput. Math.*, 7(3):303–330, July 2007.

[6] R. Adler, J.-P. Dedieu, J. Margulies, M. Martens, and M. Shub. Newton's method on Riemannian manifolds and a geometric model for the human spine. *IMA J. Num. Anal.*, 22:359–390, 2002.

[7] S. S. Antman. *Nonlinear problems of elasticity*, volume 107 of *Applied mathematical sciences*. Springer Verlag, Berlin, 1991.

[8] R. Ashman, S. Cowin, W. Van Buskirk, and J. Rice. A continuous wave technique for the measurement of the elastic properties of cortical bone. *J. Biomech.*, 17: 349–361, 1984. cited in [29].

[9] R. E. Bank and R. K. Smith. A posteriori error estimates based on hierarchical bases. *SIAM J. Numer. Anal.*, 30(4):921–935, 1993.

[10] P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuß, H. Rentz–Reichert, and C. Wieners. UG – a flexible software toolbox for solving partial differential equations. *Comp. Vis. Sci*, 1:27–40, 1997.

[11] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander. A generic interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE. *Computing*, accepted.

[12] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, and O. Sander. A generic interface for parallel and adaptive scientific computing. Part I: Abstract framework. *Computing*, accepted.

[13] G. Bergmann, G. Deuretzbacher, M. Heller, F. Graichen, A. Rohlmann, J. Strauss, and G. Duda. Hip contact forces and gait patterns from routine activities. *J. Biomech.*, 34(7):859–871, 2001.

[14] H. Berninger. *Domain Decomposition Methods for Elliptic Problems with Jumping Nonlinearities and Application to the Richards Equation.* PhD thesis, Freie Universität Berlin, 2008.

[15] L. Blankevoort and H. Huiskes. Ligament–bone interaction in a three-dimensional model of the knee. *J. Biomech. Eng.*, 113(3):263–269, 1991.

[16] M. Blatt and P. Bastian. The iterative solver template library. In *Applied Parallel Computing. State of the Art in Scientific Computing*, volume 4699 of *Lecture Notes in Scientific Computing*, pages 666–675. Springer Verlag, 2007.

[17] P. Boieri, F. Gastaldi, and D. Kinderlehrer. Existence, uniqueness, and regularity results for the two-body contact problem. *Appl. Math. Opt.*, 15:251–277, 1987.

[18] F. Bornemann, B. Erdmann, and R. Kornhuber. Adaptive multilevel methods in three space dimensions. *Int. J. Numer. Meth. Engrg.*, 36:3187–3203, 1993.

[19] F. Bornemann, B. Erdmann, and R. Kornhuber. A posteriori error estimates for elliptic problems in two and three space dimensions. *SIAM Journal on Numerical Analysis*, 33(3):1188–1204, 1996.

[20] A. Boyde. Electron microscopy of the mineralizing front. *Metabolic Bone Disease and Related Research*, 2, Suppl.:69–78, 1980. cited in [29].

[21] D. Braess. *Finite Elemente.* Springer Verlag, 1991.

[22] P. J. Brown and C. T. Faigle. A robust efficient algorithm for point location in triangulations. Technical report, Cambridge University, 1997.

[23] D. Carter and W. Caler. Cycle-dependent and time-dependent bone fracture with repeated loading. *J. Biomech. Eng.*, 105:166–170, 1983. cited in [29].

[24] P. Ciarlet, H. LeDret, and R. Nzengwa. Junctions between three-dimensional and two-dimensional linearly elastic structures. *J. Math. Pures Appl.*, 68:261–295, 1989.

[25] P. G. Ciarlet. *Mathematical Elasticity.* North-Holland, 1988.

[26] A. Conn, N. Gould, and P. Toint. *Trust-Region Methods.* SIAM, 2000.

[27] R. Cooper, J. Milgram, and R. Robinson. Morphology of the osteon: An electron microscopic study. *Journal of Bone and Joint Surgery*, 48-A:1239–1271, 1966. cited in [29].

[28] E. Cosserat and F. Cosserat. Sur la statique de la ligne déformable. *Comptes Rendus de l'Académie des Sciences, Paris*, 1907.

[29] J. D. Currey. *Bones*. Princeton University Press, 2. edition, 2002.

[30] P. Deuflhard and A. Hohmann. *Numerische Mathematik I – Eine algorithmisch orientierte Einführung*. de Gruyter, 2. edition, 1993.

[31] D. J. Dichmann. *Hamiltonian Dynamics of a Spatial Elastica and the Stability of Solitary Waves*. PhD thesis, University of Maryland, 1994.

[32] R. Diestel. *Graphentheorie*. Springer Verlag, 3rd edition, 2006.

[33] M. do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

[34] C. Eck. *Existenz und Regularität der Lösungen für Kontaktprobleme mit Reibung*. PhD thesis, Universität Stuttgart, 1996.

[35] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. North-Holland, 1976.

[36] B. Flemisch, J. Melenk, and B. Wohlmuth. Mortar methods with curved interfaces. *Appl. Numer. Math.*, 54(3-4):339–361, 2005.

[37] L. Formaggia, J. Gerbeau, F. Nobile, and A. Quarteroni. On the coupling of 3D and 1D Navier–Stokes equations for flow problems in compliant vessels. *Computer Methods in Applied Mechanics and Engineering*, 191:561–582, 2001.

[38] L. Freitag. *Users Manual for Opt-MS: Local Methods for Simplicial Mesh Smoothing and Untangling*. Argonne National Laboratory, Illinois, March 1999.

[39] M. Frisen, M. Magi, L. Sonnerup, and A. Viidik. Rheological analysis of soft collagenous tissue. Part I: Theoretical considerations. *J. Biomech.*, 2:13–20, 1969. cited in [93].

[40] Y. C. Fung. *Biomechanics: Mechanical Properties of Living Tissues*. Springer Verlag, 1993.

[41] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Series in Computational Physics. Springer Verlag, 1984.

[42] C. Gräser and R. Kornhuber. Multigrid methods for obstacle problems. *J. Comp. Math.*, 2008. submitted.

[43] C. Gräser, U. Sack, and O. Sander. Truncated nonsmooth Newton multigrid methods for convex minimization problems. In *Proc. of DD18*, submitted.

[44] F. S. Grassia. Practical parametrization of rotations using the exponential map. *The Journal of Graphics Tools*, 1998.

[45] M. E. Gurtin. *An Introduction to Continuum Mechanics*. Academic Press, 1981.

[46] T. J. Healey and P. G. Mehta. Straighforward computations of spatial equilibria of geometrically exact Cosserat rods. *International Journal of Bifurcation & Chaos*, 15(3):949–965, 2005.

[47] C. Hellmich, F.-J. Ulm, and L. Dormieux. Can the diverse properties of trabecular and cortical bone be attributed to only a few tissue-independent phase properties and their interactions? *Biomechan. Model. Mechanobiol.*, 2:219–238, 2004.

[48] S. Hüeber and B. Wohlmuth. An optimal a priori error estimate for non-linear multibody contact problems. *SIAM J. Numer. Anal.*, 43(1):157–173, 2005.

[49] R. Huiskes, J. Janssen, and T. Sloof. A detailed comparison of experimental and theoretical stress-analysis of a human femur. In S. Cowin, editor, *Mechanical Properties of Bone*, volume 45 of *AMD*, pages 211–234. 1981. cited in [29].

[50] J. Kastelic, I. Palley, and E. Baer. The multicomposite ultrastructure of tendon. *Connective Tissue Research*, 6:11–23, 1978. cited in [93].

[51] S. Kehrbaum. *Hamiltonian Formulations of the Equilibrium Conditions Governing Elastic Rods: Qualitative Analysis and Effective Properties*. PhD thesis, University of Maryland, 1997.

[52] N. Kikuchi and J. Oden. *Contact Problems in Elasticity*. SIAM, 1988.

[53] P. M. Knupp. Achieving finite element mesh quality via optimization of the Jacobian matrix norm and associated quantities. II: A framework for volume mesh optimization and the condition number of the Jacobian matrix. *Int. J. Numer. Methods Eng.*, 48(8):1165–1185, 2000.

[54] C. Kober, B. Erdmann, C. Hellmich, R. Sader, and H.-F. Zeilhofer. Consideration of anisotropic elasticity minimizes volumetric rather than shear deformation in human mandible. *Comp. Meth. Biomech. Biomed. Eng.*, 9(2), 2006.

[55] R. Kornhuber. A posteriori error estimates for elliptic variational inequalities. *Computers Math. Applic.*, 31(8):49–60, 1996.

[56] R. Kornhuber. *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*. B.G. Teubner, 1997.

[57] R. Kornhuber and R. Krause. Adaptive multigrid methods for Signorini's problem in linear elasticity. *Comp. Vis. Sci*, 4:9–20, 2001.

[58] R. Kornhuber, R. Krause, O. Sander, P. Deuflhard, and S. Ertel. A monotone multigrid solver for two body contact problems in biomechanics. *Comp. Vis. Sci*, 11:3–15, 2008.

[59] R. Krause. From inexact active set strategies to nonlinear multigrid methods. volume 27 of *Lecture Notes in Applied and Computational Mechanics*, pages 13–21, 2006.

[60] R. Krause and O. Sander. Fast solving of contact problems on complicated geometries. In R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Widlund, and J. Xu, editors, *Domain Decomposition Methods in Science and Engineering*, pages 495–502. Springer Verlag, 2005.

[61] R. Krause and O. Sander. Automatic construction of boundary parametrizations for geometric multigrid solvers. *Comp. Vis. Sci*, 9:11–22, 2006.

[62] K. Kunisch and G. Stadler. Generalized Newton methods for the 2d-Signorini contact problem with friction in function space. *M2AN*, 4:827–854, 2005.

[63] J. Lagnese, G. Leugering, and E. Schmidt. *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*. Birkhäuser, 1994.

[64] T. Laursen. *Computational Contact and Impact Mechanics*. Springer Verlag, 2002.

[65] A. Love. *Treatise on the Mathematical Theory of Elasticity*. Dover Publishing, 1944.

[66] H. Lowenstam and S. Weiner. *On Biomineralization*. Oxford University Press, 1989.

[67] J. E. Marsden and T. S. Ratiu. *Introduction to Mechanics and Symmetry*. Springer Verlag, 1994.

[68] J. Matyas, M. Anton, N. Shrive, and C. Frank. Stress governs tissue phenotype at the femoral insertion of the rabbit MCL. *J. Biomech.*, 28(2):147–157, 1995.

[69] D. J. Monaghan, I. W. Doherty, D. M. Court, and C. G. Armstrong. Coupling 1D beams to 3D bodies. In *Proc. 7th Int. Meshing Roundtable*. Sandia National Laboratories, 1998.

[70] V. C. Mow, W. Y. Gu, and F. H. Chen. *Basic Orthopaedic Biomechanics and Mechano-Biology*, chapter 5: 'Structure and Function of Articular Cartilage and Meniscus'. Lippincott Williams & Wilkins, third edition, 2005.

[71] P. Neff. Finite multiplicative elastic-plastic Cosserat micropolar theory for polycrystals with grain rotations including fracture. Modelling and mathematical analysis. Technical Report 2297, Technische Universität Darmstadt, 2003.

[72] N. Neuss and C. Wieners. Criteria for the approximation property for multigrid methods in nonnested spaces. *Math. Comp.*, 73:1583–1600, 2004.

[73] A. Noor and J. Peters. Mixed models and reduced/selective integration displacement models for nonlinear analysis of curved beams. *Int. J. Numer. Methods Eng.*, 17: 615–631, 1981.

[74] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes*. Cambridge University Press, 2007.

[75] R. Putz and R. Pabst, editors. *Sobotta – Atlas der Anatomie des Menschen*. Urban & Fischer, 2000.

[76] K. Quapp and J. Weiss. Material characterization of human medial collateral ligament. *J. Biomech. Eng.*, 120:757–763, 1998. cited in [93].

[77] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Oxford Science Publications, 1999.

[78] M. Rumpf. A variational approach to optimal meshes. *Num. Math.*, 72(4):523–540, 1996.

[79] O. Sander. Constructing boundary and interface parametrizations for finite element solvers. Diplomarbeit, Institute of Computer Science, Freie Universität Berlin, 2001.

[80] C. Sansour and W. Wagner. Multiplicative updating of the rotation tensor in the finite element analysis of rods and shells — a path independent approach. Technical Report 4, Universität Karlsruhe, Institut für Baustatik, 2002.

[81] A. Schmidt and K. Siebert. *Design of Adaptive Finite Element Software: The Finite Element Toolbox ALBERTA*, volume 42 of *LNCSE*. Springer Verlag, 2005.

[82] T. Seidman and P. Wolfe. Equilibrium states of an elastic conducting rod in a magnetic field. *Archive for Rational Mechanics and Analysis*, 102(4):307–329, 1988.

[83] J. Simo and L. Vu-Quoc. A three-dimensional finite-strain rod model. Part II: Computational aspects. *Comp. Meth. in Appl. Mech. and Eng.*, 58(1):79–116, 1986.

[84] M. Spivak. *Comprehensive Introduction to Differential Geometry*. Publish or Perish Inc., 1970.

[85] D. Stalling, M. Westerhoff, and H.-C. Hege. Amira: A highly interactive system for visual data analysis. In C. Hansen and C. Johnson, editors, *The Visualization Handbook*, chapter 38, pages 749–767. Elsevier, 2005.

[86] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[87] A. Veeser. Efficient and reliable a posteriori error estimators for elliptic obstacle problems. *SIAM J. Numer. Anal.*, 39(1):146–167, 2001.

[88] T. Veldhuizen. Techniques for scientific C++. Technical Report 542, Indiana University Computer Science, 2000.

[89] A. Viidik. A rheological model for uncalcified parallel-fibred collagenous tissue. *J. Biomech.*, 1:3–11, 1968.

[90] A. Wächter and L. T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Math. Progr.*, 106 (1):25–57, 2006.

[91] M. Weiser, A. Schiela, and P. Deuflhard. Asymptotic mesh independence of Newton's method revisited. *SIAM J. Numer. Anal.*, 42(5):1830–1845, 2005.

[92] J. Weiss, J. Gardiner, B. Ellis, T. Lujan, and N. Phatak. Three-dimensional finite element modeling of ligaments: Technical aspects. *Medical Engineering & Physics*, 27:845–861, 2005.

[93] J. A. Weiss and J. C. Gardiner. Computational modeling of ligament mechanics. *Critical Reviews in Biomedical Engineering*, 29(4):1–70, 2001.

[94] B. Wohlmuth. An a posteriori error estimator for two-body contact problems on non-matching meshes. 33:25–45, 2007.

[95] B. Wohlmuth and R. Krause. Monotone methods on nonmatching grids for nonlinear contact problems. *SIAM Journal on Scientific Computing*, 25(1):324–347, 2003.

[96] B. I. Wohlmuth. *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, volume 17 of *LNCSE*. Springer Verlag, 2001.

[97] Z. Yosibash, N. Trabelsi, and C. Milgrom. Reliable simulations of the human proximal femur by high-order finite element analysis validated by experimental observations. *J. Biomech.*, 40:3688–3699, 2007.