TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Numerische Mechanik

Advanced Non-Linear Solution Techniques for Computational Contact Mechanics

Michael Hiermeier

Vollständiger Abdruck der von der Fakultät für Maschinenwesen der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr.-Ing. Boris Lohmann
Prüfer der Dissertation:	
1.	Prof. Dr.-Ing. Wolfgang A. Wall
2.	Prof. Dr. rer. nat. Michael Ulbrich
3.	Prof. Dr.-Ing. Alexander Popp

Die Dissertation wurde am 02.01.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 20.08.2020 angenommen.

Abstract

The objective of this thesis is to improve the understanding and robustness of large deformation contact simulations by developing new formulations and strategies suitable for real-world problems. For this purpose, a finite element discretization together with a mortar-like contact formulation will be in focus. The combination of both approaches currently represents one of the most promising ways to handle complex contact scenarios in a truly non-linear realm. The necessary constraint enforcement strategy will be realized with both Lagrange multiplier and penalty methods, with a clear emphasis on the first one. The focused development of a robust non-linear solution strategy is complemented by a profound mathematical foundation, since otherwise the necessary universality would not be achievable.

Therefore, the first two steps in this work consist of the classification of the state-of-the-art formulations and in the investigation of the non-linear solution paths taken by these methods. Upon this knowledge, two distinct mortar-like contact formulations are developed. The first one will be a truly variationally consistent formulation, which leads to a completely symmetric system matrix with respect to all active contact contributions. This variationally consistent frictionless contact formulation immediately opens the way to the entire classical inequality constrained optimization literature. However, since the evaluation of the necessary second order derivatives can become computational very expensive, a second approach is introduced, which neglects certain terms during the variation, but still fulfills important properties such as the balance of linear and angular momentum. It will be shown that their final converged solutions are very similar to currently well-established mortar-like contact formulations, however, the path to the solution might be severely different. To demonstrate this, a number of numerical investigations will be performed revealing drawbacks and advantages of each approach.

The next step will be to use the developed formulations as well as the newly established transition to the mathematical constrained optimization literature such that meaningful modifications can be made, which allow a significantly improved overall robustness of contact simulations. This will be achieved by introducing a small but very effective modification to the consistently linearized saddle point system of equations. The applied modification will not only improve the performance for large initial penetrations, but it will also allow the condensation of the additional Lagrange multipliers from the linear system of equation without the need for any special dual Lagrange multiplier shape functions. Nevertheless, such a modification of the linear system will influence the local convergence behavior such that an appropriate detailed analysis will become necessary. Furthermore, the often problematic choice of the complementarity parameter c_{N} will be resolved by the novel modification approach. Therefore, a suitable dynamic correction method for this parameter together with a precise study of its boundedness will be included. Additionally, an analysis of possible changes regarding the conditioning of the system matrix will follow including a comparison between the saddle-point and the condensed system of equations. This contribution, which leads to an improved robustness of the non-linear solver performance, will be concluded with novel switching conditions allowing a smooth transition back to the consistently linearized method close to the solution.

The final corner stone on the way to an improved robustness will be the consideration of advanced globally convergent optimization strategies. Thereby, a line search filter method will be chosen, which is inherently capable of inequality constraint problems, and in this thesis its ideas will be carried over to discretized dynamic contact problems with large deformations. The differ-

ent adaptions will be presented and comprehensively discussed. One example is the correction of the linear system of equations to overcome non-positive definite points on the way to the solution. Despite the fact that there are already powerful correction algorithms, the challenge will be to make them applicable to large, parallel distributed sparse systems. These systems often do not provide a directly accessible detailed inertia analysis and thus other strategies must be developed. Furthermore, special problems related to a finite element discretization will be taken into account such as locally invalid elements, which are not sufficiently represented in the global filter acceptability tests. Other points are the controlled decrease of the c_{N} parameter under rare, but critical, circumstances or the correct scaling of the filter entries and many more. Finally, more advanced topics such as the enhanced assumed strains (EAS) formulation or the consideration of dynamic contact formulations will be discussed and adapted for the use with a line search filter method.

In total, this thesis contributes with a number of multi-purpose non-linear solver tools, which simplify the daily use of mortar-like contact or other constrained methods for upcoming complex real-world problems.

Zusammenfassung

Das Ziel dieser Arbeit ist es, das Verständnis und die Robustheit von Kontaktsimulationen mit großen Deformationen zu verbesseren, indem neue Formulierungen und Strategien entwickelt werden, die für reale Problemstellungen geeignet sind. Zu diesem Zweck wird eine Finite-Elemente-Diskretisierung zusammen mit einer mortar-artigen Kontaktformulierung im Fokus stehen. Die Kombination beider Ansätze stellt derzeit einen der vielversprechendsten Wege dar, um komplizierte Kontaktszenarien in einem echten nichtlinearen Umfeld zu bewältigen. Die notwendige Behandlung von Nebenbedingungen wird sowohl mit Lagrange-Multiplikator- als auch mit Penalty-Methoden realisiert, wobei der Schwerpunkt eindeutig auf den Ersteren liegt. Die im Fokus stehende Entwicklung einer robusten nichtlinearen Lösungsstrategie erfordert eine fundierte mathematische Grundlage, da sonst die notwendige Allgemeingültigkeit nicht gewährleistbar wäre.
Die ersten beiden Schritte dieser Arbeit bestehen daher in der Klassifizierung der aktuellen Formulierungen und in der Untersuchung der nichtlinearen Lösungswege dieser Methoden. Auf der Grundlage dieser Erkenntnisse werden zwei unterschiedliche mortar-artige Kontaktformulierungen entwickelt. Die erste wird eine echt variationell konsistente Formulierung sein, die zu einer vollständig symmetrischen Systemmatrix in Bezug auf alle aktiven Kontaktbeiträge führt. Diese variationell konsistente, reibungslose Kontaktformulierung öffnet gleichzeitig den Weg zur gesamten klassischen Literatur bzgl. Optimierung unter Ungleichheitsnebenbedingungen. Da die Auswertung der notwendigen Ableitungen zweiter Ordnung jedoch sehr teuer werden kann, wird ein zweiter Ansatz eingeführt, der bestimmte Terme während der Variation vernachlässigt, aber dennoch wichtige Eigenschaften wie die Impuls- und Drehimpulserhaltung erfüllt. Es wird gezeigt, dass ihre endgültigen konvergierten Lösungen den derzeit etablierten mortar-artigen Kontaktformulierungen sehr ähnlich sind, jedoch kann der Weg zur Lösung sehr unterschiedlich sein. Um dies zu demonstrieren, werden eine Reihe von numerischen Untersuchungen durchgeführt, die die Nachteile und Vorteile der einzelnen Ansätze aufzeigen.
Im nächsten Schritt werden die entwickelten Formulierungen sowie der Übergang zur mathematischen Literatur für die Optimierung unter Nebenbedingungen so genutzt, dass sinnvolle Modifikationen vorgenommen werden können, die insgesamt eine deutlich verbesserte Robustheit der Kontaktsimulationen ermöglichen. Dies wird durch eine kleine, aber sehr effektive Modifikation des konsistent linearisierten Sattelpunkt-Gleichungssystems erreicht. Die angewandte Modifikation wird nicht nur die Leistung bei großen Anfangsdurchdringungen verbessern, sondern auch die Kondensation der zusätzlichen Lagrange-Multiplikatoren aus dem linearen Gleichungssystem ermöglichen, ohne dass spezielle duale Lagrange-Multiplikatorformfunktionen erforderlich sind. Jedoch wird eine solche Modifikation des linearen Gleichungssystems das lokale Konvergenzverhalten derart beeinflussen, dass eine entsprechende Detailanalyse erforderlich wird. Darüber hinaus wird die oft problematische Wahl des Komplementaritätsparameters c_{N} durch den neuartigen Modifikationsansatz gelöst. Dafür wird eine geeignete dynamische Korrekturmethode für diesen Parameter zusammen mit einer genauen Untersuchung seiner Beschränktheit miteinbezogen. Zusätzlich folgt eine Analyse möglicher Änderungen hinsichtlich der Kondition der Systemmatrix mit einem Vergleich zwischen dem Sattel-Punkt und dem kondensierten Gleichungssystem. Dieser Beitrag, der zu einer verbesserten Robustheit der nichtlinearen Löserleistung führt, wird mit neuartigen Umschaltbedingungen abgerundet, die einen reibungslosen Übergang zurück zu der konsistent linearisierten Methode nahe der Lösung ermöglichen.

Der letzte Grundstein auf dem Weg zu einer verbesserten Robustheit wird die Berücksichtigung fortschrittlicher global konvergenter Optimierungsstrategien sein. Dafür wird eine Line Search Filter Methode gewählt, die von Natur aus für Probleme mit Ungleichheitsnebenbedingungen geeignet ist, und in dieser Arbeit werden diese Ideen auf diskretisierte dynamische Kontaktprobleme mit großen Verformungen übertragen. Die verschiedenen Anpassungen werden vorgestellt und ausführlich diskutiert. Ein Beispiel ist die Korrektur des linearen Gleichungssystems zur Überwindung von nicht positiv definiten Punkten auf dem Weg zur Lösung. Trotz der Tatsache, dass es bereits gute Korrekturalgorithmen gibt, wird die Herausforderung darin bestehen, sie auf große, parallel verteilte, dünnbesetzte Systeme anzuwenden. Diese Systeme bieten oft keine direkt zugängliche detaillierte Analyse ihrer Trägheitseigenschaften, d.h. ihrer Eigenwertverteilung, und so müssen andere Strategien entwickelt werden. Darüber hinaus werden spezielle Probleme im Zusammenhang mit einer Finite-Elemente-Diskretisierung berücksichtigt, wie z.B. lokal ungültige Elemente, die in den globalen Filterakzeptanztests nicht ausreichend repräsentiert sind. Weitere Punkte sind die kontrollierte Reduktion des c_{N}-Parameters unter seltenen, aber kritischen Umständen oder die korrekte Skalierung der Filtereinträge und vieles mehr. Schließlich werden weiterführende Themen wie die Formulierung mit Enhanced Assumed Strains (EAS) oder die Berücksichtigung dynamischer Kontaktformulierungen diskutiert und für die Verwendung mit einer Line Search Filter Methode angepasst.
Insgesamt bringt diese Arbeit eine Reihe von Mehrzweck-Werkzeugen für nichtlineare Löser mit sich, die den täglichen Gebrauch von mortar-artigen Kontaktmethoden oder anderen Verfahren unter Nebenbedingungen für anstehende komplizierte Probleme in der realen Welt vereinfachen.

Contents

Nomenclature

General Representation of Scalars, Vectors and Tensors

Note: *A more comprehensive introduction into the general notation-specific aspects of this thesis can be found at the very beginning of Chapter 2.*

c, C	Scalar quantity
$\underline{v}, \underline{V}$	Vector
$\underline{\underline{m}}, \underline{\underline{M}}$	Dyadic tensor, or matrix

Operators and Symbols

$\nabla_{(\cdot)}$	Consistent gradient with respect to $(\cdot)$
$\nabla^2_{(\cdot)(\cdot)}$	Consistent Hessian or second order derivative matrix
$\tilde{\nabla}_{(\cdot)}$	(In)consistent gradient with respect to $(\cdot)$, see Remark 4.1
$\tilde{\nabla}^2_{(\cdot)(\cdot)}$	(In)consistent Hessian or second order derivative matrix, see Remark 4.1
$\det$	Determinant
$(\cdot)^{-1}$	Inverse
$(\cdot)^{\mathrm{T}}$	Transpose, see (2.7) for the definition in case of dyadic tensors
$\delta(\cdot)$	Virtual quantity, variation of a quantity $(\cdot)$
$(\cdot)^\flat$	Pure co-variant definition of a tensor
$\otimes$	Dyadic product
$\times$	Cross product
$\langle(\cdot),(\cdot)\rangle$	Inner product
$\underline{\underline{I}}$	Identity tensor
$\delta^{ij}, \delta_{ij}, \delta^i{}_j, \delta_i{}^j$	Kronecker delta
$\mathrm{D}_{\underline{v}}(\cdot)$	First order directional derivative in $\underline{v}$ direction, see (4.25)
$\mathrm{D}_{\underline{w}}(\mathrm{D}_{\underline{v}}(\cdot))$	Second order directional derivative in $\underline{v}$ and $\underline{w}$ directions, see (4.25)
$\dot{(\cdot)}, \ddot{(\cdot)}$	First and second time derivatives at a fixed reference position
$\check{(\cdot)}$	Prescribed quantity
Div	Material divergence operator w.r.t. reference coordinates
div	Material divergence operator w.r.t. current coordinates
$(\cdot)\cdot(\cdot)$	Dot product of two tensors
$(\cdot):(\cdot)$	Double-dot product of two tensors
$\ln(\cdot)$	Natural logarithm
$\min(\cdot), \max(\cdot)$	Minimum and maximum function

$\mathrm{minimize}(\cdot)$	Find the minimum of a function
argmin	Arguments of the minimum
$\mathrm{d}(\cdot)$	Infinitesimal increment of a quantity
$\Delta(\cdot)$	Increment or direction
$(\cdot)'$	First order derivative w.r.t. a scalar
$(\cdot)''$	Second order derivative w.r.t. a scalar
$\|(\cdot)\|$	ℓ_2-norm, if not explicitly defined differently
$\mathrm{diag}[(\cdot)]$	Diagonal matrix with the vector $(\cdot)$ on the diagonal
A	Assembly operator
$\|(\cdot)\|$	If $(\cdot)$ is a set, the cardinality of this set; otherwise, the absolute value
$(\cdot)\|^{\mathcal{A}}_{\mathcal{SM}}$	Selecting vector entries located at degrees of freedom for which the vector $\tilde{\nabla}_{\underline{d}}\tilde{g}^{\mathcal{A}}_{\mathrm{N}}\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}}$ contains values unequal to zero
$\sin(\cdot), \cos(\cdot)$	Sine and cosine functions
$\mathrm{null}(\cdot)$	Compute the null space of a matrix
$\Delta[\cdot]_{ij}$	Denoting $[\cdot]_i - [\cdot]_j$
vol_{Δ}	Volume of a tetrahedron

Superscripts and Subscripts

Note: *Many of these can occur either as super- or as subscripts depending on the surrounding context and the actual variable* $(\cdot)$.

$(\cdot)^{[b]}$	Body index, mostly $b \in \{1, 2\}$
$(\cdot)^{\star}$	Value at the solution
$(\cdot)^{(e)}$	Quantity associated to an element
$(\cdot)_{\mathrm{N}}$	Quantity in normal direction
$(\cdot)_{\tau}$	Quantity in tangential direction
$(\cdot)^{\mathrm{h}}$	Explicit marking of discrete quantities (for convenience often dropped)
$(\cdot)^{\{n\}}$	Quantity associated to the discrete point in time t_n; equivalently for t_{n+1} etc.
$(\cdot)^{\{k\}}$	Quantity associated to the Newton iteration k; equivalently for $k+1$ etc.
$(\cdot)^{\{k,l\}}$	Quantity associated to the line search iteration l of Newton iteration k
$(\cdot)_{\mathrm{g}\alpha}$	Quantity associated to the Generalized-α method
$(\cdot)^{\mathcal{A}}$	Contributions associated to the active set
$(\cdot)^{\mathcal{I}}$	Contributions associated to the inactive set
$(\cdot)^{\mathrm{SOC}}$	Second order correction quantities
$(\cdot)_{\mathrm{elb}}$	Element load balancing
$(\cdot)^{\mathrm{nf}}$	Near field
$(\cdot)^{\mathrm{ff}}$	Far field

Kinematics

$\varphi_t(\underline{X}, t)$	Mapping operator from reference to current configuration
$\varphi_t^{-1}(\underline{X}, t)$	Pull-back operator from current to reference configuration
$\underline{x}$	Spatial position in current configuration
$\underline{X}$	Spatial position in reference configuration
$\underline{u}$	Displacement
$\dot{\underline{u}}$	Velocity
$\ddot{\underline{u}}$	Acceleration
$\underline{\underline{F}}$	Deformation gradient
$\underline{\underline{R}}$	Rotation tensor
$\underline{\underline{U}}$	Right stretch tensor
$\underline{\underline{V}}$	Left stretch tensor
$\underline{\underline{C}}$	Right Cauchy–Green tensor
$\underline{\underline{E}}$	Green–Lagrange strain
$\underline{\underline{L}}$	Material velocity gradient
Λ	Principal elongation
v	Volume in current configuration
V_0	Volume in reference configuration
a	Area in current configuration
A_0	Area in reference configuration
$\underline{n}$	Normal in current configuration
$\underline{N}$	Normal in reference configuration
$\underline{e}^1, \underline{e}^2, \underline{e}^3$	Base vectors of Cartesian coordinate system

Stresses and Material Properties

E	Young's modulus
ν	Poisson's ratio
$\lambda_{\mathrm{nH}}, \mu_{\mathrm{nH}}$	Lamé constants
ϱ	Material density in current configuration
ϱ_0	Material density in reference configuration
$\dot{\varrho}, \dot{\varrho}_0$	Material density rate in current/reference configuration
$\underline{t}$	Surface force in current configuration
$\underline{t}_0$	Surface force in reference configuration
$\underline{b}$	Body force in current configuration
$\underline{b}_0$	Body force in reference configuration
$\underline{\underline{\sigma}}$	Cauchy stress tensor
$\underline{\underline{P}}$	First Piola–Kirchhoff stress tensor
$\underline{\underline{\mathfrak{P}}}$	First Piola–Kirchhoff stress function for pure elastic materials, see Section 2.1.3
$\underline{\underline{S}}$	Second Piola–Kirchhoff stress tensor

Ψ	Strain energy function
Ψ_{nH}	Strain energy function for the coupled neo-Hookean material law
Ψ_{nHlog}	Strain energy function for the logarithmic neo-Hookean material law
Ψ_{tv}	Additive part of a simple orthotropic, transversely isotropic material law
I_1, I_2, I_3	Scalar valued invariants of the right Cauchy–Green tensor
I_4, I_5	Pseudo invariants of the right Cauchy–Green tensor
$\underline{a}$	Fiber direction
$\alpha_{\mathrm{tv}}, \beta_{\mathrm{tv}}, \gamma_{\mathrm{tv}}$	Auxiliary variables for Ψ_{tv}
$E_{\parallel}$	Young's modulus in fiber direction
$E_{\perp}$	Young's modulus in the plane normal to the fiber direction
$G_{\perp\parallel}$	Shear modulus in a plane parallel to the fiber direction
$\nu_{\perp\parallel}$	Poisson's ratio for tension in fiber direction
$\nu_{\perp\perp}$	Poisson's ratio in a plane orthogonal to the fiber direction

Domains and Boundaries

Ω	Current domain
Ω_0	Reference domain
Γ_σ	Neumann boundary zone in reference configuration
Γ_u	Dirichlet boundary zone in reference configuration
Γ_{c}	Contact boundary zone in reference configuration
γ_{c}	Contact boundary zone in current configuration

Function Spaces

$\mathcal{W}$	Weighting function space
$\mathcal{H}^1$	Sobolev function space
$\mathcal{H}^{1/2}$	Trace space of $\mathcal{W}$
$\mathcal{H}^{-1/2}$	Dual space of $\mathcal{H}^{1/2}$
$\mathcal{U}$	Solution function space
$\mathcal{M}_+$	Lagrange multiplier function space

General Spatial and Temporal Discretization

$\mathcal{E}$	Set of all elements
$\underline{\underline{N}}$	Shape function matrix
N	Single shape function
$\underline{d}$	Discrete (nodal) displacement vector

$\underline{v}$	Discrete (nodal) velocity
$\underline{a}$	Discrete (nodal) acceleration
ξ_i	Parametric directions/coordinates
$\underline{f}_{\text{ext}}$	Nodal external force vector
$\underline{f}_{\text{int}}$	Nodal internal force vector
$\underline{\underline{K}}$	Stiffness matrix
$\underline{\underline{C}}$	Damping matrix
$\underline{\underline{M}}$	Mass matrix
$\underline{r}$	Residual vector
$\beta_{\text{g}\alpha}, \gamma_{\text{g}\alpha}$	Parameters of the Newmark-β method
$\alpha_{\text{m}}, \alpha_{\text{f}}$	Parameters of the Generalized-α method
ρ_∞	Spectral radius
$j^{(e)}$	Element Jacobian determinant

Structural Dynamics

t	Time
$t_{n-1}, t_n, \ldots$	Discrete points in time
Δt	Discrete time step
$\mathscr{K}$	Kinetic energy
$\mathscr{P}_{\text{ext}}$	Rate of external mechanical work
$\mathscr{P}_{\text{int}}$	Rate of internal mechanical work, stress power
$\mathscr{U}_{\text{tot}}$	Total potential energy
$\mathscr{U}_{\text{ext}}$	Potential energy of external loading
$\mathscr{U}_{\text{int}}$	Total strain energy
$\mathscr{V}_t, \mathscr{V}_b$	Auxiliary potentials for surface and volume loading
$\mathcal{L}$	Lagrangian density function

Enhanced assumed strains

$\underline{\underline{E}}^u$	Deformation dependent Green–Lagrange strain part
$\underline{\underline{\tilde{E}}}$	Additive enhancement of the Green–Lagrange strains
$\underline{\underline{\alpha}}_{\text{eas}}$	Coefficients of the discretized Green–Lagrange strain enhancement
$\mathcal{Q}_1$	Associated mapping operator based on the shape functions for $\underline{\underline{\alpha}}_{\text{eas}}$
$\underline{\underline{\beta}}_{\text{eas}}$	Coefficients of the independently discretized second Piola–Kirchhoff stresses
$\mathcal{Q}_2$	Associated mapping operator based on the shape functions for $\underline{\underline{\beta}}_{\text{eas}}$
$\tilde{\underline{r}}_{\text{eas}}$	Element-wise residual vector w.r.t. $\underline{\underline{\alpha}}_{\text{eas}}$ coefficients
$\tilde{\underline{\underline{L}}}_{\text{eas}}$	Mixed linearization matrix w.r.t. $\underline{\underline{\alpha}}_{\text{eas}}$ coefficients and the displacements
$\underline{\underline{\tilde{D}}}_{\text{eas}}$	Linearization matrix w.r.t. $\underline{\underline{\alpha}}_{\text{eas}}$ coefficients

Frictionless Contact Mechanics

$\underline{\tau}^{[b]}_{\ i}$	Covariant base vectors of the convective coordinates
$\zeta^{[b]i}$	Contravariant convective coordinates
$\underline{\chi}$	Ray-tracing projection operator
α_χ	Auxiliary distance factor of the projection operator
g_N	Normal gap
p_N	Normal pressure
λ_N	Normal Lagrange multiplier
$\mathscr{C}$	Standard frictionless contact potential
$\tilde{\mathscr{C}}$	Standard frictionless contact energy density
$\mathscr{C}_{c_\mathrm{N}}$	Augmented frictionless contact potential
$\tilde{\mathscr{C}}_{c_\mathrm{N}}$	Augmented frictionless contact energy density
c_N	Regularization/Complementarity parameter for frictionless contact

Frictional Contact Mechanics

$\underline{v}_\mathrm{rel}$	Relative velocity
$v_\mathrm{N,rel}$	Normal component of the relative velocity
$\underline{v}_{\tau,\mathrm{rel}}$	Tangential component of the relative velocity
$\overset{\circ}{\underline{g}}_\tau$	Tangential slip rate
p_τ	Frictional shear
$\mathfrak{F}$	Friction coefficient
$\mathscr{L}_{c_\tau}$	Frictional contact Lagrangian
$\tilde{\mathscr{C}}_{c_\tau}$	Frictional contact energy density function
$\tilde{\mathscr{C}}_{c_\tau,\mathrm{stick}}$	Stick component of the frictional energy density function
$\tilde{\mathscr{C}}_{c_\tau,\mathrm{slip}}$	Slip component of the frictional energy density function
c_τ	Regularization/Complementarity parameter for frictional contact

Mortar-based Contact Methods

$\hat{\underline{n}}^{[b]}, \hat{\underline{n}}^{[b](e)}$	Outward-pointing (non-unit) normal of an element e
$\underline{n}^{[b](e)}, \underline{n}^{[b]}$	Outward-pointing unit normal of an element e
$\underline{n}^{[b](e)k}$	Outward-pointing unit normal of an element e evaluated at a slave node k
$\tilde{\underline{n}}^{[b]k}$	Averaged outward-pointing nodal (non-unit) normal at a slave node k
$\breve{\underline{n}}^{[b]k}$	Averaged outward-pointing nodal unit normal at a slave node k
$\breve{\underline{n}}^{[b]}$	Smooth C^0-continuous normal field

$\underline{\hat{n}}^{[b]}$	Smooth C^0-continuous normal field with unit length
$\mathcal{S}$	Set of all slave nodes
$\mathcal{M}$	Set of all master nodes
$\mathcal{E}^{\mathcal{S}}$	Set of all slave elements
$\tilde{g}_{\mathrm{N}}, \underline{\tilde{g}}_{\mathrm{N}}$	Weighted gap
$\hat{g}_{\mathrm{N}}, \underline{\hat{g}}_{\mathrm{N}}$	Averaged weighted gap
$A^i, A^{ii}, \underline{\underline{A}}$	Tributary area (matrix)
$\underline{\underline{\tilde{K}}}_{\mathscr{C}_{c_{\mathrm{N}}}}$	Augmentation stiffness matrix
$\underline{\underline{T}}$	Orthographic projection matrix onto a certain plain
$\underline{\underline{W}}_{\times}$	Skew-symmetric cross-product matrix
e_{jac}	Error due to the neglected variation of the Jacobian determinant
e_{ma}	Error due to the neglected variation of the convective master parameter coordinate

Numerical Optimization and General Mathematics

$s, \underline{s}, \hat{\underline{s}}, \bar{\underline{s}}$	Slack variables
f	Scalar-valued objective function
$\underline{x}, \underline{y}$	Vectors of unknowns
$\underline{x}^+$	Trial point
$\mathbb{R}$	Real numbers
α	Step length parameter
L	Lipschitz constant
$\underline{\Phi}$	Fix-point function
$\underline{p}$	Search direction
$\underline{\underline{J}}$	Jacobian matrix
$\mathcal{V}$	Definition of a neighborhood around the solution
m_f	Scalar-valued model function for a scalar-valued function f
$\underline{m_{\underline{r}}}$	Vector-valued model function for system of non-linear equations $\underline{r}$
c_1, η_f	Constant of the Armijo rule
β	Step length reduction factor for the line search method
c_2	Constant of the curvature conditions as part of the Wolfe conditions
Δ_{TR}	Trust region radius
ρ_{TR}	Quality measure for the trust region step
$\eta_1^{\mathrm{TR}}, \eta_2^{\mathrm{TR}}, \gamma_1^{\mathrm{TR}}, \gamma_2^{\mathrm{TR}}$	Constants for the trust region adaption
τ	Pseudo-time of the PTC method
δ_τ	Pseudo-time increment of the PTC method
$\mathcal{K}$	Sub-sequence of iterations
$\varnothing$	Empty set

Constrained Optimization

$\underline{g}, g^i, g_i$	(Non-linear) constraints
$\underline{\lambda}$	Lagrange multiplier
n	Total dimension of the primal variable space
m	Number of constraints
$\mathcal{S}$	Index set of all constraints (similar to the slave node set)
$\mathcal{A}_0$	Set of all active constraints at the solution
$\mathcal{A}_+$	Set of all strongly active constraints at the solution
$\mathcal{A}, \mathcal{A}_{c_\mathrm{N}}$	Set of all active constraints, not necessarily at the solution
$\mathcal{I}$	Set of all inactive constraints, not necessarily at the solution
c	Regularization or penalty parameter
$\mathscr{L}$	Lagrangian function
$\mathscr{L}_\mathrm{c}$	Augmented Lagrangian function
$\mathcal{C}$	Cone with first order feasible directions at a feasible point $\underline{x}$
$\mathcal{C}^\star$	Cone of critical directions at the KKT point
$\mathscr{P}_\mathrm{c}$	Penalty function
$\lambda(\underline{x})$	Lagrange multiplier function
$\gamma_1^\lambda, \gamma_2^\lambda$	Constants used for the Lagrange multiplier function
$\underline{p}_x$	General search direction for the primal variables
$\underline{p}_\lambda$	General search direction for the dual variables
$\underline{\lambda}_+$	Trial Lagrange multiplier for the SQP method
$\underline{z}, \underline{z}_\mathrm{p}$	Auxiliary variable
μ_IP	Positive control parameter for the interior point method
$\underline{\underline{\Sigma}}, \underline{\underline{\Lambda}}, \underline{\underline{S}}$	Iteration matrices of the interior point method

Variant of Newton's Method

$\underline{\underline{R}}$	Matrix whose columns span the range space of $\nabla_{\underline{d}} \tilde{\underline{g}}_\mathrm{N}^{\mathcal{A}}$
$\underline{\underline{Z}}$	Matrix whose columns span the null space space of $[\nabla_{\underline{d}} \tilde{\underline{g}}_\mathrm{N}^{\mathcal{A}}]^\mathrm{T}$
$\Delta\underline{d}_\mathrm{R}, \Delta\underline{d}_\mathrm{Z}$	Formal split of the solution displacement vector into a range space and a null space part
$\beta_\varphi^{c_\mathrm{N}}$	Control parameter for the sufficient enclosed angle update routine
$\beta_\Theta^{c_\mathrm{N}}$	Control parameter for the sufficient infeasibility reduction update routine
Θ	Infeasibility measure
m_Θ	Linear model of the infeasibility measure
$\delta_1, \delta_2, \varepsilon, \breve{M}$	Constants defined in Theorem 5.1
$\underline{\Delta}, \underline{l}, \omega$	Auxiliary variables for the proof of Theorem 5.1 defined in Equation (5.42)
$\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)$, $\breve{\underline{\Lambda}}_\mathrm{N}(\underline{\Delta}, \underline{l}, \omega)$	Continuously differentiable functions needed for the proof of Theorem 5.1

$\breve{\mu}$	Upper bound for the inverse modified iteration matrix
$\breve{a}$	Upper bound for the norm of the tributary area and its gradient
$\breve{G}$	Upper bound for the norm of the second order derivative of weighted gap
$c_{\mathrm{a}}, c_{\mathrm{g}}$	Constants defined within Theorem 5.2
$L_{\mathrm{edge}}^{(e)}$	Element edge length of an element e
$\mathcal{B}_{\mathrm{pre}}^{\hat{g}_{\mathrm{N}}}$	User-specified bound for the gap criterion
$\mathcal{B}_{\mathrm{pre}}^{\varphi}$	User-specified bound for the angle criterion
$\mathcal{B}_{\mathrm{pre}}^{\mathrm{res}}$	User-specified bound for the magnitude criterion
γ_{g}	Scaling factor for the minimal detected element edge length
$\mathrm{TOL}_{\mathrm{res}}, \mathrm{TOL}_{\varphi}$	Tolerances for the magnitude and angle criteria
κ_{∞}	Condition number estimate in the infinity norm

Line Search Filter Method

$m_{\mathscr{L}}$	Model equation for the Lagrangian objective function
γ_{Θ}	Scaling factor for the acceptance criterion based on the infeasibility measure
γ_f	Scaling factor for the acceptance criterion based on the Lagrangian objective function
s_f, s_{Θ}	Exponents for the $\mathscr{L}$-type switching condition
ν_{Θ}	Scaling factor for the $\mathscr{L}$-type switching condition
$\mathcal{F}$	Filter set
$\alpha_{(\cdot)}^{\min}$	Minimal step length estimates
TOL_1	Tolerance for the residual norm check (stopping criterion)
TOL_2	Tolerance for the infeasibility norm check (stopping criterion)
TOL_3	Tolerance for the primal increment norm check (stopping criterion)
TOL_4	Tolerance for the Lagrange multiplier increment norm check (stopping criterion)

Correction of the Iteration Matrix (Filter Method)

ω	Correction factor for the upper-left matrix block
$\omega_{\min}, \omega_{\max}$	Upper and lower bound for the matrix correction factor
n_{ω}	Number of successive decrease iterations
N_{ω}	Switch to the unmodified system if n_{ω} reaches this value
n_{bad}	Number of invalid or heavily distorted/changing elements
$r_{\min}, r_{\max}$	Upper and lower bound for the acceptable volume ratio
$v_{\mathrm{curr}}^{(e)}$	Current volume of element e
$v_{\mathrm{ref}}^{(e)}$	Reference volume of element e
s_{curr}	Quadratic ℓ_2-norm of the current primal solution increment

s_{last}	Quadratic ℓ_2-norm of the lastly accepted primal solution increment
$\delta_\omega, \delta_\varepsilon$	Scaling factor for the positive definiteness test
κ_ω^-	Reduction factor for ω
$\kappa_\omega^+, \kappa_\omega^{++}$	Accretion factor for ω

Invalid Element Identification (Filter Method)

$L_{(\cdot)}$	Lagrangian polynomials
$\underline{\underline{J}}^{(e)}$	Element Jacobian matrix
$B_i, \tilde{B}_{ijk}, \tilde{B}_i$	Bézier basis functions
$\tilde{b}^i$	Coefficients of the Bézier basis functions
$j^i, \underline{j}$	Values of the Jacobian determinants
$\underline{\underline{T}}, \underline{\underline{Q}}$	Static algorithmic iteration matrices
$\underline{a}, \underline{b}, \underline{c}$	Auxiliary coordinates for the subdivision algorithm

Further Details on the Globalization Algorithm (Filter Method)

$r_{\text{min}}^{\text{pre}}, r_{\text{max}}^{\text{pre}}$	Upper and lower bound for the acceptable volume ratio within pre-testing
Θ_{min}	The $\mathscr{L}$-type switching condition will be bypassed as long as the infeasibility measure is above this value
Θ_{max}	Upper bound for the scaled constraint violation,
$\gamma_\Theta^{\text{max}}$	Scaling factor to adapt Θ_{max} after a filter reinitialization
$n_{\text{newton}}^{\text{block}}$	Bound for blocking Newton iterations, see also Section 6.7.3
$n_{\text{ls}}^{\text{block}}$	Bound for blocking line search iterations, see also Section 6.7.3
$\beta_{\Theta\text{crit}}^{c_{\text{N}}}$	Critical control parameter for the SIR method to compute a possible reduction of c_{N}
κ_f	Scaling of the first filter coordinate based on the objective function value
κ_Θ	Scaling of the second filter coordinate based on the infeasibility measure
$\tilde{\nu}_\Theta$	Scaled ν_Θ factor for the $\mathscr{L}$-type switching condition

Parallel Redistribution

N_{p}	Number of available processor cores
$t_p^{\{k\}}$	Evaluation time of processor core p in one Newton iteration k
$\mathcal{B}_{\text{proc}}$	Bound for the ratio between maximal and minimal evaluation time per core
γ_{proc}	Important scaling factor for the near/far field decision

Sine-Shaped Membranes (Filter Method)

w	Mid-surface definition of the upper membrane
v	Mid-surface definition of the lower membrane
$\underline{X}_w, \underline{X}_v$	Three-dimensional reference coordinates of the upper and lower membrane mid-surfaces
$\underline{X}_w^{\pm}, \underline{X}_v^{\pm}$	Three-dimensional reference coordinates of the respective lower and upper surfaces
$\underline{\tilde{N}}_w, \underline{\tilde{N}}_v$	Reference non-unit normal field on the respective membrane
$\underline{N}_w, \underline{N}_v$	Reference unit normal field on the respective membrane
t	Thickness of the membranes
o_w, o_v	Initial offset between the two membranes

Abbreviations

FE	Finite element
FEM	Finite element method
VEM	Virtual element method
NTN	Node-to-node
NTS	Node-to-surface, node-to-segment
STS	Segment-to-segment
DD	Domain decomposition
GPTS	Gauss-point-to-surface
X-FEM	Extended finite element method
ILS	Inequality level-set
ANS	Assumed natural strain
DSG	Discrete strain gap
KKT	Karush-Kuhn-Tucker
PEEK	Polyether ether ketone
EAS	Enhanced assumed strain
DOF	Degree(s) of freedom
NCP	Non-linear complementarity function
HEX8	8-node hexahedral element
QUAD4	4-node quadrilateral element
TET4	4-node tetrahedron element
GMRES	Generalized minimal residual
BFGS	Broyden, Fletcher, Goldfarb and Shanno
PTC, ΨTC	Pseudo-transient continuation
SER	Switched evolution relaxation
TTE	Temporal truncation error

LICQ	Linear independence constraint qualification
MFCQ	Mangasarian–Fromovitz constraint qualification
RQP	Recursive quadratic programming
SQP	Sequential quadratic programming, or successive quadratic programming
GQ	Gaussian quadrature
GP	Gauss point
IGA	Isogeometric analysis
NURBS	Non-uniform rational basis spline
SEA	Sufficient enclosed angle
SIR	Sufficient infeasibility reduction
mN	Modified Newton
std	Standard
ILU	Incomplete L(ower) U(pper), see LU-factorization in the literature
SOC	Second order correction

1. Introduction

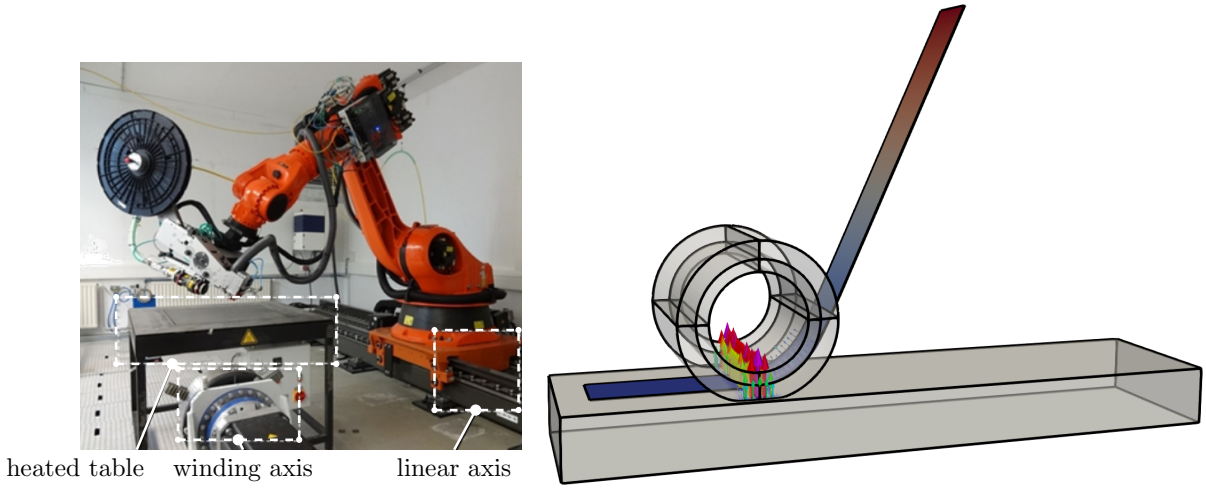
1.1. Motivation

Elastodynamics including contact problems play a major role in our all daily lives. A demonstrative example can be any imaginable connection between two distinct parts via screws, rivets or nails. Quite often these connections can become the most vulnerable spot in large assemblies leading to malfunctions with severe consequences. Such classical mechanical failures require a detailed and on-point mechanical analysis where two main options come into play: The experimental validation and the prediction via numerical simulations. Since the experimental validation can become very expensive, the numerical simulation grows more and more in importance and is now an established tool of the preliminary design and during the final quality check of safety-related components. In the last decades large progress has been made to cover many of the critical scenarios in a consistent and reliable way. For example finite wear problems [86, 255] have quite recently been addressed.

However, one point that seems to become rather less attractive for researchers is the further development of the algorithmic foundation for theses methods. But, exactly there seems to be still a lot of room for improvements. Therefore, the specific application topics shall be put aside throughout this thesis and, instead, the entire attention is on the field of implicit non-linear solution methods for non-linear contact problems. A big motivational aspect is hereby the objective to find a way to improve and simplify the work with contact algorithms and to make them accessible to a greater number of interested users without the demand to become an expert in computational contact mechanics.

In the following, the difficulties encountered in contact simulations are divided into three main topics. The first topic is related to an insufficiently posed problem and can be summarized under the point *problem modeling*. Examples are badly chosen contact boundary conditions, insufficient consideration of dynamic contact aspects, or just a wrong parametrization. Let us look a little bit closer at the last point: issues with the chosen parameters. Some of them might just lead to wrong results. This is typical for all parameters which are directly linked to the used materials as well as to special surface properties which can be tough to model and often ask for a number of conscientiously conducted experiments. Think for example of friction and all the possible variables which can influence the frictional behavior between two contacting bodies, such as temperature, moisture or changing material properties (melting point, changes in the lattice structure, etc.). A demonstrative example of such a real world application is shown in Figure 1.1. The presented structural simulation of an automated fiber placement process in Figure 1.1b has been performed with an early version of the finite deformation contact solution strategy presented later.

On the other hand, there are parameters which have no effect on the simulation outcome but play a crucial role for the behavior of the simulation procedure itself. These parameters are re-



(a) (b)

Figure 1.1.: Example of a complex real-world problem: the automated fiber placement process. Hereby, a robot and a radiative heat source (i.e. a laser) is used for the automated manufacturing of high-performance, near-net shape carbon fiber reinforced plastic parts, see Figure 1.1a. On the other hand, Figure 1.1b shows the non-linear structural contact simulation among multiple bodies, namely the elastic roller and carbon fiber reinforced plastic tape, and a third body, the rigid table.

lated to the second main source of difficulties: the applied *non-linear solution technique*. While the first source, viz. the modeling part, depends strongly on the individual simulation task, it is this second source of errors which can be addressed by a number of very general improvement strategies. These strategies shall be introduced and comprehensively discussed in this thesis and are linked to the proper choice of certain parameters as well as the proper choice of the best solution strategy at the right moment on the way to the solution. However, it is necessary to understand the effects of the different choices first, before an educated guess becomes possible. Therefore, it is a major goal of this thesis to encourage and support the knowledge gain. In addition, a number of simple rules will be introduced which shall help the user towards a meaningful parameter choice and to identify the problem in numerous different, sometimes quite complicated, contact scenarios.

Finally, the third and last source of error shall be addressed: the *inherent imperfections and simplifications* of the applied contact formulations. They can cause inherent instabilities in the respective discretized formulation. These issues are rare in modern formulations and are strongly linked to the individually applied contact scheme or the considered finite element formulation. Since there exists a huge variety of different formulations, it is not possible to discuss and investigate all of them. Instead, the focus of this work lies on mortar-like formulations which enforce the constraints with Lagrange multipliers. A more precious classification will follow in the next section.

1.2. Fundamental Approaches

In this section a brief summary of the fundamental approaches towards a trustworthy and efficient contact simulation together with a robust non-linear solution strategy is made and presented.

1.2.1. Computational Contact Methods

The overall objective is to find a way how the robustness of numerical contact simulations can be drastically improved and the improvement shall be independent of the considered specific contact problem. Therefore, it is required that the applied strategies are designed in such a way that they are easily applicable to many of the established contact algorithms (see [214] for a recent overview). Nevertheless, some restrictions must be made: Firstly, the thesis puts its focus on only one specific discretization method: the finite element method (FEM). However, this can be seen as a rather weak restriction since the FEM is probably the method of choice in the field of elastic contact simulations, either under consideration of small or large deformations.

The triumph of the FEM is based on its beneficial universal applicability and its strong mathematical convergence properties as comprehensively described in Brenner and Scott [35]. This has helped to outperform its early competitors, such as the finite difference method described in Collatz [51], Samarskii [233], in many important research fields. Even though there were certain efforts which should help the finite difference method to regain importance by resolving issues such as the regular mesh requirement [148, 177]. Nevertheless, the success of the FEM could not be stopped in the numerical treatment of non-linear structural problems. Therefore, it is not surprising that there exists a huge number of well-written text books. Examples are Bathe [12], Belytschko et al. [18], Hughes [140], Oden [206], Zienkiewicz and Taylor [297]. A more recent development is the observation that the inherent shape functions of the defining finite elements are replaced by non-uniform rational basis splines (NURBS) instead of the more familiar Lagrange polynomials. This development started with Hughes et al. [141] and, indeed, it brings some appealing properties especially in the context of numerical contact and interface problems discussed here. However, there are also some drawbacks such as the rather intricate construction of proper general-purpose meshes for complex real world geometries. This is a topic which finds exemplarily very recent resonance in the research area of fluid structure interaction [210]. Since the isogeometric analysis (IGA) literature is of rather minor interest for this thesis, the discussion of this topic shall be restricted to only certain specific sub-topics which will be addressed at the respective points, e.g., within the discussion in Sections 4.7.2 and 4.7.3. However, there are many more discretization methods, such as the virtual element method (VEM) allowing arbitrarily shaped elements. A general overview as well as a brief introduction with focus on contact mechanics can be found in Wriggers [284] and the literature therein. For a more comprehensive list of alternative spatial discretization methods the interested reader is kindly referred to Section 2.3.

At this point the attention shall be drawn to the numerical treatment of contact mechanics under consideration of the FEM. Early approaches considered simple *node-to-node* (NTN) approaches. Examples can be found in [100, 143]. Due to its simplicity the NTN method finds also application in a few more recent publications such as [197], however, it stays always restricted to rather simple problems and linear elasticity. The restriction comes from the fact that the nodes on the two opposing contact surfaces must exactly match which can not be guaranteed for ar-

bitrary large deformations. Other early approaches are based on the so-called *node-to-segment* or *node-to-surface* (NTS) formulations [1, 14, 143]. These methods have been extended to large deformations as well as friticonal contact (see e.g. Pietrzak and Curnier [212]) and are still widely spread in many commercial tools, but also in recent research codes (see e.g. Zavarise et al. [294]). The idea is to monitor the penetration of the nodes into the segments formed by the elements of the opposing body. Since the penetration is only tracked for one node set of both surfaces, namely the so-called *slave surface* or *contactor*, it is possible that the nodes of the second surface, the so-called *master* or *target surface*, can still penetrate the segments of the slave body. This is one of the obvious drawbacks which leads to the general advise that the surface with the finer discretization should be considered as the contactor [76]. Some other difficulties in certain geometrical scenarios together with possible solutions are summarized in Zavarise and De Lorenzis [293]. Besides this classical NTS approaches there exist also a number of modified methods. For example the surface smoothing approaches described in Crisfield [58], El-Abbasi et al. [77]. A discussion of other examples for NTS algorithms, such as the two-pass variants [58, 260], can be found in [76]. Therein, also the stability of all these contact algorithms as well as their patch test performances are addressed in detail. In general, NTS formulations are not able to pass the patch test and some of their modified versions might have stability problems due to overconstraining [76]. However, they have also a big advantage compared to the up-coming methods. That is their simplicity. The implementation as well as the necessary computational effort is easier manageable and leads to faster run-times.

The final class of algorithms belong to the so-called *segment-to-segment* (STS) approaches. These methods are mainly considered in this thesis. The later discussed variant is a so-called *mortar-type* or *mortar-like* contact method. The mortar methods have their origin in the research field of non-overlapping domain decomposition (DD) approaches. These methods are used to allow a consistent coupling among different domains at non-conforming discretized interfaces. One of the first articles pointing in this direction is given by Maday et al. [184]. Herein, the mortar method has been discussed in context with the spectral element method. Further publications followed quickly such as Belgacem [15], Bernardi et al. [20], Seshaiyer and Suri [245]. In Belgacem [15] the Lagrange multiplier method is applied to enforce the constraints and in Seshaiyer and Suri [245] a closer look at the superior convergence properties is taken. The success of the mortar method for finite elements is probably due to its underlying idea. It is constructed in such a way that the continuity condition is fulfilled in a weak sense over the domain boundaries and, therefore, it fits perfectly into the context of the FEM. The choice of the correct mortar function spaces can become crucial. However, in these early publications the spaces have often been chosen as a sub-space of the finite element space considered for the interior of the domains. This started to change with Wohlmuth [280, 281] where the so-called *dual-spaces* for the Lagrange multipliers have been introduced. These dual spaces allow the condensation of the additional Lagrange multiplier degrees of freedom from the linear system of equations without worsening the convergence properties of the domain decomposition methods and, indeed, this can be seen as a great advantage for problems with many separate domains or volumetric coupling approaches for multi-physics problems as discussed in Farah and Vuong [84].

All these DD methods are usually concerned with equality constrained problems. Now, for contact problems a reasonable extension to inequality constrained problems becomes necessary. The first necessary steps have been taken by Belgacem et al. [16], Hild [133], McDevitt and Laursen [192]. Afterwards, also large deformation problems have been exemplarily addressed

by Fischer and Wriggers [90], Yang et al. [290]. Both articles are pretty interesting, since Fischer and Wriggers [90] discusses the still relevant *Gauss-point-to-surface* (GPTS) penalty methods (see e.g. Dimitri et al. [71]), while Yang et al. [290] addresses the important topic of objectivity in the context of mortar methods and frictional sliding. Also the very important topic of a meaningful spatial integration scheme is considered by these early publications. The extension to 3-dimensional problems follows then in Puso and Laursen [222], Puso et al. [223]. At this point one advantage of the mortar contact methods compared to the previously mentioned NTS methods must be addressed: As nicely shown in El-Abbasi and Bathe [76], STS methods are able to pass the patch test, however, to do so a suitable integration scheme must be applied which is able to handle possible weak or strong discontinuities. A well-suited so-called *segmented based* integration scheme is introduced and refined by Puso and Laursen [222], Puso et al. [223], Wilking and Bischoff [278], Yang et al. [290]. However, these integration schemes can become numerically pretty expensive. Therefore, also less expensive so-called *element-based* integration schemes have been developed and applied [65, 90] with the obvious drawback that the patch test can not be completely satisfied and the choice of slave and master can become decisive, especially, in the presence of strong discontinuities. This has been pointed out, e.g., by Farah et al. [83]. Therein, a new approach is proposed which uses the segment-based integration only along the boundaries of the active contact zone. In this way the bad influence of strong discontinuities can be completely avoided, while the minor impact of weak discontinuities is tolerated, thus, an overall computational more efficient algorithm is achieved. A number of publications consider also the dual-Lagrange multipliers and transfer this idea to contact mechanics. A certain selection of publications concerning small deformations is given by Brunssen et al. [39], Flemisch and Wohlmuth [91], Hüeber [138]. The extension to the regime of large deformations follows in Gitterle [109], Gitterle et al. [110], Popp et al. [215, 216, 217, 218], Wohlmuth [279]. Also the combination of IGA and mortar methods has been extensively considered in the last decade, e.g., by De Lorenzis et al. [65], Dimitri et al. [71], Seitz et al. [241], Temizer et al. [261]. In summary, the mortar method becomes more and more important and has already achieved to catch up or even pass the NTS methods in the research sector. But also in the field of industrial applications, it starts to gain more attention. An advantage of the mortar method is the naturally quite robust numerical performance and the high quality results for smooth contact scenarios. However, in the presence of sharp edges or corners, a STS approach might not be the right choice and NTS or even a NTN method are probably much better suited. However, also this problem has been recently tackled by a new approach which combines mortar, NTS and NTN methods in a new framework proposed by Farah et al. [85]. In this thesis, however, the focus is on the pure mortar-like contact algorithms.

Another well-known issue of discretized contact formulations is the insufficient representation of the active contact zone boundary. The classical contact methods suffer under the circumstance that the resolution of the a priori unknown boundary of the active contact zone is inherently limited by the finite element mesh used. Therefore, the mesh dependent convergence order for the displacements and Lagrange multiplier values is bound away from the optimal rates expected for pure unconstrained structural problems. This is true even for higher order function spaces, see Wohlmuth [279, Remark 4.13] and the references therein for more information. The work of Graveleau et al. [119] shows a way out of this misery, which builds upon the ideas of Bonfils et al. [33]. The fundamental new idea is to split the contact problem into two parts: The first part is achieved by temporarily fixing the unknown contact domain, thus, only equality constraints

must be considered. In a subsequent step, a so-called *shape optimization* of this contact domain is initiated, which results in an update of the set domain. These two sub-problems are iterated up to convergence. Details for simple linear contact problems can be found in [119]. These steps include level-sets [209] coupled with X-FEM to represent the changing a priori unknown contact zone without the need for repetitive remeshing, which leads to the name *inequality level-set* (ILS) approach. In the master thesis of Hofer [135], first steps towards non-linear contact problems have been made based on the ILS approach and the mortar-like contact formulation presented in this thesis. However, the consideration of X-FEM introduces new complexity, e.g., by additional Lagrange multipliers [199]. Other successful X-FEM methods circumvent this lastly mentioned problem by using Nitsche's method as exemplarily discussed by Schott [237] in the context of complex interface coupled flow problems.

Quite recently a similar approach is followed by [73, 74]. Therefore, a so-called *refined boundary quadrature* method is developed in Duong and Sauer [74]. This method also relies on an accurate localization of the contact boundary via a level-set approach. Subsequently, the gained information is used to apply a special segmentation at the identified contact boundary zone to handle physical discontinuities. The associated contact boundary detection algorithm is described in great detail. Subsequently, a uniform and an adaptive quadrature method are compared to the new approach. This work considers a GPTS approach for large deformations and shows also an accurate post-processing of the contact pressure via enriched nodal values base on a X-FEM representation. In Duong et al. [73], the idea is revisited and extended by an enriched interpolation of the contact pressure which is incorporated into an extended mortar method to address the physical discontinuities. Furthermore, a so-called *two-half-pass* approach is used. This approach is used to achieve an unbiased contact formulation that does no longer contain the mortar coupling term. In this way, the problem with artificial discontinuities due to discontinuities in the normal vectors or of the shape functions and their products can be avoided without the need for an expensive segmentation during the numerical integration inside the contact zone. Therefore, the segmentation is limited to the contact zone boundary. However, in contrast to Bonfils et al. [33], only the contact pressure is enriched while the resolution of the contact zone boundary is handled by the refined boundary quadrature. Again, Duong et al. [73] uses the GPTS method and considers large deformations.

An alternative is the mesh refinement approach without X-FEM, which can be also used to counteract the insufficient localization of the active contact zone boundary. A comparison of different strategies of this kind can be found in Franke et al. [101], for instance. These methods use a number of remeshing and/or relocation steps to obtain the desired goal. A comprehensive introduction to so-called *adaptive mesh refinement* methods for contact problems can be found in Rieger et al. [226]. Therein, the topic of reliable error estimators and indicators is addressed in detail. Despite the fact, that the distortion of the convergence rates and the reduction of approximation errors remain important issues, it must be said that this thesis will put no emphasize on resolving them. Hence, the interested reader is referred to the mentioned literature for more information.

Lastly, the different ways of incorporating the constraints into the governing equations shall be briefly discussed. On the one hand, the Lagrange multiplier method exists, which will also be mainly used within this thesis. This method introduces additional unknowns, the name-giving *Lagrange multipliers*, which can be interpreted as the applied contact pressure. The classical mortar-like contact methods are based on this Lagrange multiplier method and lead to a mixed

discretization, including displacements and stresses as unknowns. These methods fulfill the non-penetration condition not point-wise, but in an integral sense as it is expected for a STS approach. Examples can be found in [90, 130, 131, 222], for instance. However, the Lagrange multiplier as well as the augmented Lagrange multiplier methods are not restricted to STS-methods, but can also be used together with other approaches such as the NTS method, see [1, 212]. Furthermore, there are the already addressed dual-Lagrange methods, which are able to remove the additional unknowns from the linear system of equations by a consistent condensation step [123, 139, 215, 216, 218]. A much more comprehensive introduction into the field of Lagrange multiplier methods will follow in Section 2.2, Section 3.2 and Chapter 4. On the other hand, there exists a second set of methods which does not introduce additional variables. The first, still widely used, representative is the classical penalty method. Its idea is to penalize a violation of the non-penetration condition by adding a large stiffness and a corresponding force to all interface elements which are in active contact, see [90, 290] for a mortar-like formulation or [293, 294] for a NTS approach. The advantage of the penalty method compared to the Lagrange multiplier method is its simplicity, due to the fact that no additional unknowns must be considered. Its drawback is that the contact constraints are generally not exactly fulfilled. Only in the limit case of an infinite value of the penalty parameter, but a large number for the penalty parameter can cause an ill-conditioned system matrix. This point will be reconsidered in Chapter 5. The penalty method can also be combined with the Lagrange multiplier method leading to the already mentioned augmented Lagrange multiplier method, see [1, 65, 131, 212, 222]. Here, it is not necessary to increase the associated penalty parameter to infinite, however, these methods also introduce Lagrange multipliers as additional unknowns. Finally, a last alternative for the constraint enforcement in computational contact mechanics shall be mentioned: *Nitsche's methods*. These methods are designed in such a way that there is no need for additional unknowns just similar to the penalty methods. However, they add additional terms to weakly enforce the constraints, including a *consistent* penalty term, instead. In contrast to the classical penalty methods, Nitsche's methods have the advantage that they converge to the exact solution for sufficient mesh refinement without the demand for an infinite penalty parameter. However, the renunciation of Lagrange multipliers does not come for free. A new complexity can be found in the evaluation of the boundary traction, which is obtained by evaluating the interface-near bulk elements. Another difference is that the consistent penalty parameter asks for the solution of a local Eigenvalue problem. This Eigenvalue problem is necessary to account for the possible stiffness dependence. In case of non-linear elasticity, the stiffness actually depends on the current deformation state, such that the Eigenvalue problem must be updated during the non-linear iteration procedure. Furthermore, so-called *harmonic weights* must be introduced to preserve the high accuracy of Nitsche's method. A more detailed introduction into Nitsche's method for computational contact mechanics can be found in [40, 46, 240, 244]. In this thesis, the focus is mainly on Lagrange multiplier and less on penalty methods. Nitsche's method will not be further considered within this work and was only mentioned for the sake of completeness.

1.2.2. Robust Non-Linear Solution Strategies

Another important ingredient for this thesis is the research towards globally convergent non-linear solution strategies. In many of the previously mentioned publications, the solution strategy is an important part of the applied contact method. This is especially true for large deformations,

but, the strategy is often taken for granted. This was also the case at the beginning of this thesis. For large deformations in computational mechanics usually the Newton Raphson method is applied, see Kelley [154] for an introduction. The reason is the appealing property that the Newton method converges quadratically near the solution. However, this property is only available in a limited region around the desired solution point. If the initial guess for the solution, i.e., the starting point for the simulation, is not inside this neighborhood, the Newton method might diverge. Now, whenever this should happen in a structural simulation, the typical strategy is to manually or automatically halve the last time step and to restart the simulation. Alternatively, it might be also necessary to tweak other parameters. Typical candidates in case of mortar-like contact simulations are the c_{N} or c_{τ} values, see e.g. [87, 213]. Even though this might quickly become a quite frustrating task, it is often possible to find a proper parameter set. However, it can be also impossible to find a solution with a plain Newton method. These examples show often structural instabilities or, mathematically speaking, an indefinite or semi-definite system matrix block. In case of (quasi-)static simulations, the equilibrium path can still be found by applying so-called continuation methods [2]. Typical representatives, such as the classical arc-length method, are described in Crisfield [56, 57], Hellweg and Crisfield [126], Ramm [224]. Some early extensions towards contact mechanics can be also found, e.g., in [163, 251]. However, similar problems can occur in dynamic simulation as well where sudden changes in the displacement field can be cumbersome and might ask for an extremely small time step.

In this thesis, a different direction shall be taken. A look at the mathematical optimization literature on globally convergent algorithms helps to find alternative remedies. Actually, there are several well-written text-books available. If unconstrained such as pure structural problems shall be addressed a look into the text books by Boyd and Vandenberghe [34], Dahmen and Reusken [60], Kelley [153], Schnabel and Frank [236] is advisable or a book completely devoted to continuation methods such as Allgower and Georg [2] might be helpful. However, if more complicated equality and inequality constrained problems are considered, other books must be taken into account. A great general overview of unconstrained and constrained globalization methods is given by Nocedal and Wright [204], while Fletcher [93] provides a number of useful practical tips helping to decide which method might be better suited for certain problems. Other books put their focus on specific sub-topics, e.g., the optimization for non-smooth problems is addressed by Clarke [49], while Bertsekas [23] gives a comprehensive introduction into the optimization with Lagrange multiplier methods.

Since a more detailed introduction into the general field of numerical optimization would go beyond the scope of this introductory section, the interested reader is kindly referred to Chapter 3 and the references therein. Instead, the remainder of this section shall be used to put the focus on the most important methods and strategies for this thesis. The used (augmented) Lagrangian formulation is based on Glad and Polak [111], Rockafellar [230]. While the extension for inequality constraints via slack variables has its origin in Rockafellar [229] and is nicely described in Gill et al. [107]. The advantage of slack variables is that they allow a smooth transition between equality and inequality constrained problems. Under certain restriction this can also be used to formulate the contact problem in a way which is more convenient if the mathematical literature shall be applied. By doing so, the related convergence results become easier accessible as well. A great example is given by Facchinei and Lucidi [79] where the therein described *Low-Cost Newton-Like Methods* fits pretty well with the usually applied solution strategies for (mortar) contact problems [65, 212]. The local solution scheme used here can be obtained either by ap-

plying Newton's method to the underlying Karush-Kuhn-Tucker (KKT) optimality conditions or under direct consideration of the so-called *Sequential Quadratic Programming* (SQP) method. The idea of the SQP method is basically to solve the non-linear optimization problem by solving a sequence of quadratic sub-problems. These quadratic sub-problems are obtained by replacing the non-linear constraints by a linear model and the non-linear objective function by a quadratic model which is additional augmented by second order information of the constraints. The second part is the same result one would obtain when the gradient of the Lagrangian as part of the KKT conditions is linearized, while the linear constraint model follows from the linearization of the (active) constraint equations as part of the KKT conditions. However, the interpretation as a sequential quadratic programming task creates another, often helpful point of view on the original non-linear optimization problem. This SQP idea dates back to Han [125] and remains one of the most popular approaches. Nevertheless, it is not the only local solution method which shall be considered in this thesis. For example the discussion about differentiable penalty functions in Gould [113] will become as well handy, or a certain variant of Newton's method for constrained problems described and briefly discussed in Bertsekas [23]. Furthermore, there are other local solution methods such as the interior point method which are not considered in this thesis but are still worth mentioning. A brief introduction into the interior point method will follow in Section 3.2.3.3.

In addition to these local solution strategies, globalization strategies must be addressed as well. These globalization strategies are the key to gain the property of global convergence. The classical way to achieve a globally convergent solution method has been for many years the construction of a so-called *penalty function*. These penalty functions consist of a simple linear combination of the objective function value and the constraints, where the weighting of the two contributions can be controlled by a penalty parameter. There exists a huge variety of different penalty functions. One often used representative is the augmented Lagrangian merit function Gill et al. [107], Glad and Polak [111], Lucidi [182], but there are many more such as the ℓ_1 or ℓ_2 penalty functions, see also [204, Sec. 15.4] and the references therein. However, the application of penalty functions has the drawback that the penalty parameter must exceed a certain positive scalar such that the penalty function becomes exact, i.e., otherwise the location of the minimum of the penalty function does not coincide with the KKT point. This positive scalar might be defined by the magnitude of the unknown associated Lagrange multiplier value at the solution (as it is the case for the ℓ_1 penalty function). However, since this threshold is unknown as long as the solution is unknown, the algorithms based on these penalty functions must adapt the penalty value whenever it might seem necessary. Furthermore, it is to note that the penalty value can also not be chosen excessively high since this puts too much weight on the constraints resulting in a very poor convergence since the iteration sequence is forced to stay on the (possibly curved) boundary of the feasible region. Therefore, the use of the *watchdog* method for constrained problems, as introduced by Chamberlain et al. [45], gained some attention. This method allows the increase of the penalty function values for a pre-defined finite number of iterations and thus is less restrictive. However, its dependence on a suitable penalty parameter makes the penalty function as general optimality measure less universal and asks for an problem dependent educated guess if the performance shall be improved.

Due to this drawback another method started to become very popular in the last decade. This new approach is based on ideas stemming from multi-objective optimization and avoids the necessity of a correctly chosen penalty parameter. Instead, the two goals of minimization of an

objective function and maintaining feasibility of the solution are considered more or less separately. The fundamental idea for the so-called *filter method* has been proposed in Fletcher and Leyffer [96] and subsequently refined in Fletcher et al. [94], Fletcher and Leyffer [97], Fletcher et al. [98]. All these early attempts have been restricted to *trust region* algorithms [53]. There exist many more publications on this new class of algorithms. For example the interior point implementation by Ulbrich et al. [263] which has been later improved in Silva et al. [246] by replacing the second filter entry with a more reliable optimality measure. Another example is a bundle method for non-smooth optimization by Fletcher and Leyffer [95] or a globalization method for a direct, derivative-free algorithm by Audet and Dennis Jr. [8] to name only a few. Again, all based on the trust region idea. A line search filter method is described in-depth by Wächter and Biegler [270, 271, 272] and this line search approach will also build the foundation for the globalization method used in this thesis. As a side note it is to mention that the filter method is not limited to constrained problems only. The basic idea of a multidimensional filter globalization method has also been successfully extended to unconstrained (Gauss-)Newton-based methods for least-squares problems and non-linear equations [114] and to ℓ_1-optimization problems as proposed by Milzarek and Ulbrich [196]. This underlines once more the great universality of this idea. At the end of this section it shall be mentioned that this thesis is not the first contribution which considers mortar contact formulations and is heading in this direction. There is at least one further author who recently applied the filter method to large deformation contact problems, viz. Youett et al. [291], Youett [292]. These publications put the focus on a trust region implementation in combination with a suitable multigrid linear solver method. Despite the fact that the mortar contact formulation used therein is inspired by the work of Popp [213], Popp et al. [215, 218] which is also the case for this thesis, it shall be emphasized that the mentioned work and this thesis have things in common but have been developed independently. Nevertheless, since both use successfully the filter method as trustworthy globalization strategy, the potential of this method for the research field of large scale contact simulations seems promising.

1.3. Research Objective

The overall objective of this thesis is to develop a robust and efficient simulation toolbox for finite deformation mortar-like contact problems. Therefore, a number of important aspects are addressed, investigated and will be briefly presented in the following. The goal of these investigations is to obtain a better understanding of the underlying connections. The gained scientific insights are then used to improve existing formulations. Thereby, it is an important point that the resulting simulation toolbox consisting of an adapted contact formulation and certain special non-linear solution strategies stays controllable: The number of parameters shall be limited and in the best case a set of default parameters shall be defined which works fine in most of the cases. In the end, the solution of frictionless contact problems with the proposed mortar-type finite element based contact framework shall become less challenging. Furthermore, many of the introduced modifications are described in a very general way such that the adaption to other contact formulations or even completely different equality or inequality constrained (multi-physics) problems should be straight forward.

1.3.1. Specification of Requirements

Since the beginning of computational contact mechanics there has been always the deep wish to develop a universally applicable, robust and efficient simulation framework for large non-linear real world contact problems. However, this has been proven to be such a big obstacle that it is not possible to resolve it at once but it is rather necessary to make small reliable steps towards a better solution strategy. A number of unresolved and, in the opinion of the author, important currently remaining issues on this way will be summarized next.

Non-symmetry of Frictionless Contact Formulations. While it is well-known that frictional contact formulations lead naturally to a non-symmetric system of equations, the opposite can be expected for the frictionless case. In contact scenarios where friction plays only an insignificant role, the modeling without friction is often a reasonable simplification. Now, if friction is neglected, the arising non-linear contact problem can be completely formulated as an optimization problem subject to inequality constraints. In such a case, it is well-known that the evolving KKT system of equations is symmetric. However, there are numerous large deformation contact formulations which violate this assumption (cf. Popp [213], Popp et al. [218], Puso and Laursen [222], Puso et al. [223]). In fact the list can be extended by all large deformation contact formulations considering dual Lagrange multipliers. The reasons for this fact can be traced back to certain assumptions and simplifications in the variational approach. However, the implications of these modifications are rather less well documented. One exemplary obvious consequence is the greater limitation in the choice of suitable linear iterative solvers (see e.g. [275–277]).

Conservation of Angular Momentum. The previously mentioned publications as well as their frictional extensions such as Gitterle [109], Gitterle et al. [110], Yang and Laursen [288], Yang et al. [290] lack additionally the conservation of angular momentum as described in Popp [213], Puso and Laursen [222]. Therefore, the question may rise which steps must be taken to regain this important property. Especially, in case of a ray-tracing projection algorithm, where the smooth normal field is completely defined on the slave side, a detailed investigation is missing.

Unpredictable Behavior of the Applied (Inconsistent) Variational Contact Methods. The next point can be also seen as part of the investigation procedure in this thesis. Before an improvement of existing methods is conceivable, the method itself must be studied and understood. As part of the knowledge gain, it is especially important to detect possible scenarios where the existing methods fail due to inherent imperfections and simplifications. This is also a point which can be hardly found in the literature. In general it is more convenient to show examples where everything is working great instead of presenting failing simulations or settings. However, the simplifications in the variational approach of the mentioned mortar-like contact formulations can have severe consequences under certain circumstances and need to be discussed and documented.

Reliable Handling of Large Initial Penetrations. The treatment of large initial penetrations and suitable start-up procedures have been proposed quite recently by Zavarise et al. [294]. Therein, a NTS penalty formulation is addressed and the necessary switch between a consistent and non-consistent method is explained in detail. However, the proposed strategy asks for an

user provided guess of the expected contact pressure. Furthermore, the presented work is not directly applicable to mortar-like contact formulations using Lagrange multipliers.

Treatment of Structural Instabilities in (Quasi-Static) Contact Simulations. The treatment of structural instabilities is still a difficult task. This is even more true in case of finite deformation contact problems. Furthermore, the proposed arc-length methods (see e.g. Koo and Kwak [163], Simo et al. [251]) can not be transferred to the case of dynamic simulations. There, the problem might be less severe due to the inherent inertia effects, however, it is still possible that a structural instability leads to a sudden unexpected structural response which then needs very small time steps to be resolvable with the common locally convergent approaches. Furthermore, the appropriate treatment of these instabilities can also help to improve the solvability of the related linear systems of equations or it may be used to improve the applicability of certain finite element technologies such as the *enhanced assumed strain* (EAS) formulation.

General Globalization Strategy for a Globally Convergent Simulation Framework. Most of the literature on computational contact mechanics addresses the contact formulations itself or the local solution strategy. It is not so common to address the *global* convergence behavior besides some exceptions (see e.g. [291, 292]). However, the development of a truly globally convergent strategy asks for many different ingredients since already small changes to the formulation can have a huge impact on the non-linear solver performance. It starts with the choice of suitable optimality and feasibility measures, goes over to a meaningful inertia correction scheme for the KKT matrix [246, 272] and ends with things like the influence of special element technologies or the reliable identification of heavily deformed invalid finite elements [150, 151].

1.3.2. Contributions of this Work

In this thesis a novel non-linear solution strategy for large deformation contact problems is comprehensively described and developed which addresses all of the aforementioned requirements. Hereby, the true novelty lies not so much in the different ingredients themselves but in the presented combination of all the small parts to a new reliable simulation tool. The most important new scientific contributions and ideas of the developed non-linear solution techniques for frictionless contact mechanics are summarized in the following:

- development of a truly variationally consistent and symmetric mortar-based contact formulation for finite deformation solid mechanics. Therein, a detailed investigation of the newly added terms is contained and a comparison to an additionally developed slightly inconsistent approach is presented alongside to this investigation, see Hiermeier et al. [131].

- a detailed proof of the conservation of angular momentum for the variationally consistent and slightly inconsistent mortar-based contact formulation including a comparison with existing and well-established formulations, see Hiermeier et al. [131].

- first presentation and description of an inherent instability of certain variationally inconsistent mortar-based STS contact formulations, see Section 4.7.4.

- development of a variant of Newton's method which is suitable to very robustly solve large penetration contact problems. The constraints are enforced with Lagrange multipliers, see Chapter 5.
- presentation of a self-adaptive update scheme for the regularization parameter c_{N}. This novel strategy finally helps to resolve the often cumbersome choice of this parameter. In addition, it is proven that the proposed update scheme together with the variant of Newton's method developed in this thesis generates a bounded sequence of c_{N} parameters, see Sections 5.3.2 and 5.4.
- introduction of a novel switching strategy which allows to easily switch between a robust pre-asymptotic and a fast locally convergent asymptotic strategy, i.e., between the variant of Newton's method of Chapter 5 and the consistently linearized system developed in Chapter 4, see Section 5.5.
- development of a new line search filter method for large deformation mortar-based contact formulations. The novel method combines well-established ideas with important new ingredients which significantly improve the applicability to finite element large scale contact simulations. The new method is capable of handling structural instabilities in quasi-static and dynamic contact simulations, see Chapter 6.
- development of a correction scheme for the linear system of equations. The proposed strategy is self-adaptive and helps to overcome non-positive definite points on the way to the solution. It is designed to be applicable to large scale parallel distributed systems and to be consistent with the proposed filter method, see Section 6.6.
- improvement of the parallel redistribution approach originally proposed by Popp [213]. The novel strategy is superior in case of large initial penetration, see Section 6.9.3.

All the different contributions build on top of each other and complement each other in order to reach the final goal of a globally convergent non-linear solution strategy for computational contact mechanics. Thereby, many more ingredients contribute to the final proposed algorithm such as the correct handling of certain finite element technologies or the detection of invalid elements.

All the presented methods, formulations and algorithms have been implemented in the in-house finite element software package BACI (see Wall and Kronbichler [274]) of the Institute for Computational Mechanics at the Technical University in Munich. The mentioned research code is completely written in C++, well suited for high performance computing (HPC) and takes advantage of third party libraries such as the LAPACK [4] or the Trilinos project (see Heroux et al. [127] for a general overview). Hereby, one package of the Trilinos project must be mentioned more explicitly since it has been used as the foundation for the non-linear solver framework for large-scale contact problems which will be developed throughout this thesis: the *NOX* package. NOX stands for *Object-Oriented Nonlinear Solutions* and is an object-oriented C++ library for the solution of large-scale systems of (unconstrained) non-linear equations. It has been chosen due to its great universality, modularity and HPC qualities. In total, the work presented here is built on top of an already existing multi-physics large scale finite element framework which contains already a mortar-based finite element formulation which has been largely implemented

by Popp [213]. However, the algorithms presented therein have been extended in many ways. Not only the contact formulation but even the underlying structural time integration framework underwent a reimplementation to satisfy the new requirements and improvements.

1.4. Outline of the Thesis

This thesis considers three major pillars of a globally convergent mortar-based contact framework in finite deformation structural dynamics. These three pillars are presented in the Chapters 4 to 6. Large parts of the first pillar, wherein a reliable mortar-based contact formulation is developed, have been already published and are taken from Hiermeier et al. [131]. The other two pillars are planned to be published in the near future. However, before more details are given, let us start from the beginning.

In **Chapter 2** a brief introduction into the field of computational structural mechanics is presented. Therein, a number of well-known relationships will be introduced and discussed. Furthermore, another objective of this chapter is to build the bridge to classical numerical optimization literature. Therefore, the mechanical variational problems are often motivated by the corresponding minimization problems. This is also extended to frictionless and frictional contact problems. Additionally, a brief introduction into the used spatial discretization methods including enhanced locking preventing formulations as well as discrete time integration schemes are addressed.

Next, in **Chapter 3**, a far more detailed mathematical introduction into the field of numerical optimization is presented. Beginning with unconstrained problems and the presentation of a selection of available local and global solution strategies. Afterwards, the attention is drawn to the research area of constrained optimization problems. This chapter is mainly used to discuss many of the used as well as other possible solution approaches in a more general context. However, the focus is always on Lagrange multiplier methods and, to a smaller extent, on penalty methods.

In **Chapter 4** the main part of this thesis is starting with the development of two new reliable mortar-based contact formulations side by side which are derived from Popp [213] and largely inspired by Alart and Curnier [1], De Lorenzis et al. [65], Pietrzak and Curnier [212]. The discussion considers a fully symmetric variationally consistent mortar-based contact formulation as well as a slightly less consistent variational approach. The latter one has much in common with the referenced mortar-based strategies by [87, 109, 110, 213, 218, 222, 223, 288, 290], for instance. The major difference will be that the mortar-based contact formulation presented here satisfies the conservation of linear *and* angular momentum with the obvious drawback that condensation using dual Lagrange multipliers is no longer possible. The opportunity will be taken to compare the symmetric variationally consistent, the new variationally inconsistent and the well-established mortar-based contact formulation [213] with each other. Many parts of this chapter can also be found in Hiermeier et al. [131].

After the knowledge gain in Chapter 4 has been accomplished, the next **Chapter 5** will address a specific variant of Newton's method which is suited for both previously introduced mortar-based contact formulations. The basic idea of the newly introduced solution strategy can be also found in [23] and is also often part of globalization methods to correct the inertia of the linear system of equations, see e.g. [246, 272]. In Chapter 5, however, the method is applied to finite deformation contact problems and is extended to a self-reliant (local) non-linear solution strategy. Therefore, a number of adjustments are made, e.g. the meaningful adaption of the internal regu-

larization parameter c_{N}. In this way, the c_{N} value becomes self-adaptive which is a very beneficial side effect. The newly developed updating rule plays also a decisive role when it comes to the numerical performance of the modified Newton method. Furthermore, a comprehensive mathematical study is made with focus on the local convergence properties. Additionally, a switching scheme is added which allows to regain the quadratic convergence property near the solution. All in all, the presented algorithms in Chapter 5 allow a much larger initial penetration. The effect can be compared with the proposed method in [294], but the method presented here represents the extension to Lagrange multiplier methods and is filled with a reliable mathematical foundation which does not worry about any solution estimates to fit certain parameters.

In the last major **Chapter 6** the step towards a globally convergent non-linear solution strategy is taken. This chapter is mainly devoted to the filter approach and is largely inspired by the great work of [270–272]. The previous chapters addressing the mortar-based contact formulation as well as the modified Newton approach are important corner stones of the finally proposed algorithms. Additionally, the considered computational contact mechanics problems together with the (enhanced) finite element discretization ask for specific modifications of the classical optimization algorithms. For example, it is of severe importance to detect any invalid finite elements, since they are often not sufficiently represented in the filter measures. Another very important point is the reliable modification of the linear system of equations. These modifications play a crucial role when it comes to structural or algorithmic instabilities (e.g. due to hour-glass modes in case of the EAS formulation under pressure [67, 273, 285]).

Finally, a summary as well as a selection listing possible improvements for future work can be found in **Chapter 7**.

2. Computational Mechanics for Large Deformations

The first two chapters are supposed to give an introduction into the mechanical and mathematical foundations necessary for this thesis. Therefore, this chapter is used to start with a brief summary of the finite deformation solid mechanics, followed by a more detailed description of the basic contact mechanics and some further background about the used discretization methods. The ongoing discussion will quickly require a solid mathematical background in numerical optimization. Thus, an introduction into the field of unconstrained and constrained optimization will be given in Chapter 3. The overall objective of this introductory part is to present enough details, thus, everyone with a profound background in mechanical engineering is able to understand the upcoming discussions without problems.

Notation

Before the actual introduction into the mathematical fields can take place, a short explanation of the used mathematical notation is mandatory:

A scalar $c \in \mathbb{R}$ is indicated by non-underlined mostly small letter, a vector $\underline{v} \in \mathbb{R}^n$ is indicated by once underlined mostly small letters and a matrix $\underline{\underline{A}} \in \mathbb{R}^{n \times m}$ is indicated by double underlined mostly capital letters. This is inspired by the typical tensor notation where the number of underlines correlates directly to the dimension of the respective variable. The contravariant components of a vector $\underline{v} \in \mathbb{R}^n$ shall be denoted by v^i, while the contravariant matrix entry i, j is denoted by A^{ij}. Here is i the row index and j the column index. The i-th unit covariant coordinate vector is called $\underline{e}_i$. It is taken advantage of the Einstein notation such that a sum $\sum_i v^i u_i$ can be defined as $v^i u_i$. Accordingly, a vector can be written as $\underline{v} = v^i \underline{e}_i$. The inner product of two contravariant vectors $\underline{v}, \underline{u} \in \mathbb{R}^n$ is denoted by $\langle \underline{v}, \underline{u} \rangle = v^i m_{ij} u^j = v^i u_j$ with the metric m_{ij}, while $\langle \underline{v}, \underline{u} \rangle_{\underline{\underline{A}}} = \underline{v}^T \underline{\underline{A}} \underline{u} = v^i m_{ik} A^{kj} m_{jl} u^l = v^i A_i{}^j u_j$ denotes an inner product scaled by a typically quadratic matrix $\underline{\underline{A}} \in \mathbb{R}^{n \times n}$. The gradient of a contravariant vectorial quantity $\underline{v} \in \mathbb{R}^n$ with respect to a contravariant vectorial quantity $\underline{u} \in \mathbb{R}^m$ is denoted by $\nabla_{\underline{u}} \underline{v} = v^i{}_{,u^j} \underline{e}_i \otimes \underline{e}^j$, $\forall i \in \{1, \dots, n\}$ and $\forall j \in \{1, \dots, m\}$. An equivalent type of notation is used for lower, or higher order contravariant, covariant, or mixed tensors. If not defined differently, a three-dimensional Cartesian coordinate system with base vectors $\{\underline{e}_1, \underline{e}_2, \underline{e}_3\}$ is used and, consequently, the metric m_{ij} usually degenerates to the Kronecker delta $\delta_{ij} = \delta^i{}_j = \delta_i{}^j = \delta^{ij}$.

2.1. Non-Linear Solid Mechanics

Besides the main topic of this thesis, viz. the efficient and robust treatment of finite deformation contact mechanics, also the underlying non-linear solid mechanics must be addressed in some detail. Since actually the combination of both, i.e., complex contact interactions between multiple bodies undergoing large translations and rotations, and the additional occurrence of huge elastic deformations, introduces the final high complexity, which makes this work even necessary. However, not everything what is happening behind the scenes can be considered in detail, such that the presentation is restricted to the basic relationships. But, most of the more sophisticated elastic models such as presented in Bonet and Wood [32], Holzapfel [136], Ogden [208], or plastic models such as described in Bertram [21], Lubarda [180], Lubliner [181] are built upon the same fundamental ideas that are also addressed in the following. Furthermore, the text-book by Marsden et al. [188] must be mentioned in terms of a rigorous tensor notation and comprehensive introduction into the field of non-linear elasticity.

2.1.1. Kinematics

Firstly, a classical Boltzmann continuum model in a three-dimensional Euclidean space shall be considered. The related 2-D equations can be easily derived in a straight-forward manner. However, the direct transition to reduced models embedded in a higher dimensional space such as the non-linear beam or shell theories is not straight forward and is therefore explicitly excluded from the discussion. Furthermore, if not defined differently, a classical Cartesian coordinate system will be assumed. At least two distinct observer frames become necessary for the comprehensive spatial description. These frames are given once for the reference configuration $\Omega_0 \subset \mathbb{R}^3$ at time $t = 0$ and, secondly, for the current configuration $\Omega \subset \mathbb{R}^3$ at time $t \geqslant 0$. The mapping between these two domain configurations is denoted by $\underline{x} = \varphi_t(\underline{X}, t)$, or in the opposite direction by $\underline{X} = \varphi_t^{-1}(\underline{x}, t)$. The absolute displacement field can be deduced directly from these definitions, viz.

$$\underline{u}(\underline{X}, t) = \underline{x}(\underline{X}, t) - \underline{X}. \tag{2.1}$$

Usually, all terms directly linked to pure structural contributions are evaluated in the reference configuration. This strategy is also followed throughout this thesis. This quite convenient approach makes use of the mentioned pull-back operator φ_t^{-1}. This operator is one of the ways the current displacement field comes into play. The difference in the primary field between the reference configuration and the current configuration given in (2.1) defines the set of unknowns in each considered time or load step: the displacement field $\underline{u}$. Thus, it is shown that the important role of the primary solution variables is taken over by the displacements in case of classical elasticity.

Now, to define the mapping between the configurations, the so-called *deformation gradient* holds a crucial role. The deformation gradient is a quantity which puts the current state into direct relation to the reference state. Therefore, the partial derivative of the current spatial position $\underline{x}$ of a given material point with respect to the reference configuration $\underline{X}$ must be computed, yielding

$$\underline{\underline{F}} = \nabla_{\underline{X}} \underline{x}(\underline{X}, t) = F^i{}_J \, \underline{e}_i \otimes \underline{E}^J = \frac{\partial x^i}{\partial X^J} \, \underline{e}_i \otimes \underline{E}^J. \tag{2.2}$$

Bijectivity (i.e., one-to-one correspondence between the two configurations) and sufficient smoothness must be assumed for the map φ_t to formally compute this quantity. Under these assumptions the inverse deformation gradient $\underline{\underline{F}}^{-1}$ is always well-defined. The same holds true for the existence of the pull-back operator φ_t^{-1}. The requirement of these assumptions can be summarized in the demand that the determinant of the deformation gradient $\det(\underline{\underline{F}})$ must stay positive. Keeping this in mind and by taking (2.1) into account, an infinitesimal line segment in the current configuration can be expressed as

$$\begin{aligned} \mathrm{d}\underline{x} &= \mathrm{d}x^i \, \underline{e}_i = a^i{}_I \left[\varphi^{-1}(\mathrm{d}x^i)\right]^I \underline{e}_i + \mathrm{d}u^i \, \underline{e}_i = a^i{}_I \left(F^{-1}\right)^I{}_j \, \mathrm{d}x^j \, \underline{e}_i + \mathrm{d}u^i \, \underline{e}_i \\ &= a^i{}_I \, \mathrm{d}X^I \, \underline{e}_i + \mathrm{d}u^i \, \underline{e}_i, \end{aligned} \tag{2.3}$$

where $\underline{\underline{a}}$ is introduced as a mixed metric living in the current and the reference configuration. However, since the same Cartesian base vectors for the reference and the current configuration are used, $\underline{\underline{m}} = \underline{\underline{a}} = \underline{\underline{\delta}}$ holds and thus the expression can be simplified. The deformation gradient follows directly as

$$F^i{}_J = \delta^i{}_I \frac{\partial X^I}{\partial X^J} \, \underline{e}_i \otimes \underline{E}^J + \frac{\partial u^i}{\partial X^J} \cdot \underline{e}_i \otimes \underline{E}^J = \delta^i{}_J \, \underline{e}_i \otimes \underline{E}^J + \frac{\partial u^i}{\partial X^J} \, \underline{e}_i \otimes \underline{E}^J. \tag{2.4}$$

This derivation demonstrates the actual mathematical meaning of the deformation gradient: It represents the mapping operator for an infinitesimal line segment between the current and the reference configuration via $\mathrm{d}\underline{x} = \underline{\underline{F}} \cdot \mathrm{d}\underline{X}$ known as push-forward operation. While the so-called pull-back operation of this line segment defined by $\mathrm{d}\underline{X} = \underline{\underline{F}}^{-1} \cdot \mathrm{d}\underline{x}$ has been already used in (2.3). By using these definitions, it becomes quite easily possible to define suitable transformations for an infinitesimal volume element by

$$\mathrm{d}v = \det(\underline{\underline{F}}) \, \mathrm{d}V_0. \tag{2.5}$$

Another possible application is the mapping of an infinitesimal area element between its current $\mathrm{d}\underline{a}$ and its reference configuration $\mathrm{d}\underline{A}_0$ under consideration of the so-called Nanson's formula yielding

$$\mathrm{d}\underline{a} = \det(\underline{\underline{F}}) \, \underline{\underline{F}}^{-\mathrm{T}} \cdot \mathrm{d}\underline{A}_0, \tag{2.6}$$

where the area quantities $\mathrm{d}\underline{a}$ and $\mathrm{d}\underline{A}_0$ are defined in terms of a scalar and its corresponding unit normal vector, viz. $\mathrm{d}\underline{a} = \mathrm{d}a \, \underline{n}$ and $\mathrm{d}\underline{A}_0 = \mathrm{d}A_0 \, \underline{N}$. The introduced transpose of the deformation gradient is given as

$$\underline{\underline{F}}^{\mathrm{T}} = (F^i{}_I \, \underline{e}_i \otimes \underline{E}^I)^{\mathrm{T}} = \left(F^{\mathrm{T}}\right)^I{}_i \, \underline{E}_I \otimes \underline{e}^i = F_i{}^I \, \underline{E}_I \otimes \underline{e}^i \tag{2.7}$$

following the comprehensive index notation of Marsden et al. [188]. This at hand and by applying the polar decomposition theorem, it is further possible to state the deformation gradient $\underline{\underline{F}}$ by its decomposition into

$$\underline{\underline{F}} = \underline{\underline{R}} \cdot \underline{\underline{U}} = \underline{\underline{V}} \cdot \underline{\underline{R}}, \tag{2.8}$$

where $\underline{\underline{R}}$ is an orthogonal rotation tensor and $\underline{\underline{U}}$ is the so-called *right stretch tensor*, whereas $\underline{\underline{V}}$ is the *left stretch tensor*. Now, since $\underline{\underline{F}}^{\mathrm{T}} = \underline{\underline{U}}^{\mathrm{T}} \cdot \underline{\underline{R}}^{\mathrm{T}}$ and, further, $\underline{\underline{R}}^{\mathrm{T}} \cdot \underline{\underline{R}} = \underline{\underline{\delta}}$ hold, it is possible to define a tensor describing the material stretch in the reference configuration by

$$\underline{\underline{C}} = \underline{\underline{F}}^{\mathrm{T}} \cdot \underline{\underline{F}} = U_i{}^I U^i{}_J \underline{E}_I \otimes \underline{E}^J. \tag{2.9}$$

It shall be noted that a pure co-variant definition $\underline{\underline{C}}^\flat$ of the *right Cauchy–Green tensor* can readily be obtained by accordingly pre-multiplying (2.9) with the respective metric. The right Cauchy–Green tensor has a number of appealing properties: Compared to the deformation gradient, the right Cauchy–Green tensor lives completely in the reference configuration what simplifies its interpretation a lot. Furthermore, it has the advantage that it is an objective measure, i.e., any superimposed rigid body motion does not change its definition. Lastly, it can be used to map the squares of the infinitesimal line elements including their enclosed angle, viz. $\mathrm{d}\underline{x} \cdot \mathrm{d}\underline{x} = \mathrm{d}\underline{X} \cdot \underline{\underline{C}} \cdot \mathrm{d}\underline{X}$. Now, if a pure rigid body motion is applied and no stretching is involved, the right Cauchy–Green tensor degenerates to the identity tensor, thus the following strain measure, named *Green–Lagrange strain*, is obtained

$$\underline{\underline{E}} = \frac{1}{2}\left(\underline{\underline{C}} - \underline{\underline{I}}\right) = \frac{1}{2}\left(\underline{\underline{F}}^{\mathrm{T}}\underline{\underline{F}} - \underline{\underline{I}}\right) = \frac{1}{2}(U_k{}^I\, U^k{}_J - \delta^I_{\cdot J})\, \underline{E}_I \otimes \underline{E}^J, \tag{2.10}$$

where $\underline{\underline{I}}$ is the mentioned identity tensor. This is, as expected, equal to zero for the undeformed state and, again, by accordingly pre-multiplying the necessary metric a pure co-variant Green–Lagrange strain $\underline{\underline{E}}^\flat$ can be derived.

Remark 2.1. The one-dimensional equivalent to the quite abstract general Green–Lagrange strain definition (2.10) is $E = 1/2(l^2/L_0^2 - 1) = 1/2(\Lambda^2 - 1)$, where L_0 is the initial length, l the current length and Λ the so-called *principal elongation*. This principal elongation correlates linearly with the more convenient technical strain $\varepsilon_{\mathrm{t}} = \Lambda - 1$ known from linear elasticity and, therefore, the principal elongation is not surprisingly often the preferred quantity when it comes to find a strain representative for large deformation experiments, see large deformation stress-strain plots, for instance.

However, the shown Green–Lagrange strain is by far not the only applicable strain measure, instead there are many more, all with benefits for different applications. For example, the typical strain measure for large strains, which is often used in the context of plasticity, is the *logarithmic strain*.

The inherent time dependency of the quantities has stayed unconsidered until now, thus, the necessary steps shall be quickly presented. It is straight forward to define a suitable velocity and

acceleration measure for the primary field variables by simply applying the corresponding time derivatives, viz.

$$\dot{\underline{u}}(\underline{X},t) = \left.\frac{\partial \underline{u}(\underline{X},t)}{\partial t}\right|_{\underline{X}} = \frac{\mathrm{d}\underline{u}(\underline{X},t)}{\mathrm{d}t}, \tag{2.11}$$

$$\ddot{\underline{u}}(\underline{X},t) = \left.\frac{\partial \dot{\underline{u}}(\underline{X},t)}{\partial t}\right|_{\underline{X}} = \frac{\mathrm{d}\dot{\underline{u}}(\underline{X},t)}{\mathrm{d}t} = \frac{\mathrm{d}^2\underline{u}(\underline{X},t)}{\mathrm{d}t^2}. \tag{2.12}$$

Accordingly, it is also possible to define a rate dependent equivalent for the deformation gradient, called the *material velocity gradient* $\underline{\underline{L}} = \dot{\underline{\underline{F}}}$. For more detailed information the reader is kindly referred to the literature mentioned at the beginning of this chapter.

2.1.2. Balance Equations

There are different fundamental balance equations which have to hold for a consistent continuum mechanics approach. In the following a brief overview will be given. This basic governing equations form the foundation of the non-linear structural finite element approach considered throughout this thesis. In particular, the later derived potential formulation will become important in the context of more sophisticated globalization methods. In fact, the potential formulation will help to build the bridge between mechanics and mathematics, since it can be used as a natural representative of a well-posed objective function without the possible drawback of least square estimates (see Byrd et al. [42]).

Conservation of Mass

Under consideration of the already introduced motion $\varphi_t(\underline{X},t)$ applied to the considered reference domain Ω_0, the mass densities $\varrho(\underline{x},t)$ in the current configuration at time t as well as $\varrho_0(\underline{X})$, denoting the mass density of the body in its reference configuration, shall be given. Now, conservation of mass implies that

$$\mathrm{mass}(\Omega) = \mathrm{mass}(\Omega_0) \qquad \Leftrightarrow \qquad \int_{\Omega} \varrho(\underline{x},t)\mathrm{d}v = \int_{\Omega_0} \varrho_0(\underline{X})\mathrm{d}V_0 \tag{2.13}$$

holds for all involved bodies summarized by the mentioned domains Ω_0 and Ω. This is completely plausible for classical elastodynamics, since mass loss during a motion should be always prohibited. Next, the mass density in the reference configuration can be also expressed by its counterpart in the current state, by simply following the introduced volume relationship (2.5). Thus, $\varrho_0 = \varrho \ \det \underline{\underline{F}}$ is obtained. Furthermore, the time derivative of the determinant yields $\mathrm{d}/\mathrm{d}t(\det \underline{\underline{F}}) = \det \underline{\underline{F}} \ \mathrm{div}(\dot{\underline{u}})$, where the divergence of $\dot{\underline{u}}$ is introduced. Next, since $\dot{\varrho}_0(\underline{X}) = 0$ must hold, it is straight forward to derive the equation of continuity by inserting this demand into (2.13) which leads to

$$0 = \int_{\Omega_0} \dot{\varrho}_0(\underline{X})\mathrm{d}V_0 = \int_{\Omega_0} \det \underline{\underline{F}} \ \dot{\varrho} + \varrho \ \det \underline{\underline{F}} \ \mathrm{div}(\dot{\underline{u}})\mathrm{d}V_0 \overset{(2.5)}{=} \int_{\Omega} \dot{\varrho} + \varrho \ \mathrm{div}(\dot{\underline{u}})\mathrm{d}v. \tag{2.14}$$

Finally, the divergence theorem provides the last step to obtain

$$\int_{\Omega} \dot{\varrho} \, \mathrm{d}v = -\int_{\partial\Omega} \varrho \underline{\dot{u}} \cdot \underline{n} \, \mathrm{d}a, \tag{2.15}$$

or in simpler words: No mass is allowed to leave the body over time.

Balance of Momentum

While a body follows its path through time and space, it is typically externally loaded. In general, this load can be split into two contributions, the surface force $\check{\underline{t}}(\underline{x}, \underline{n}, t)$ acting on a current unit surface element at time t across a surface element defined by its unit normal $\underline{n}$, and, secondly, there exists a body force $\check{\underline{b}}$ acting per current unit volume element. Alternatively, it is also possible to define the body force per unit mass element, see e.g. Marsden et al. [188]. Besides these usually prescribed forces, there are additionally the so-called internal forces acting across any possible surface. These at hand, the *balance of linear momentum* is obtained by

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \varrho \underline{\dot{u}} \, \mathrm{d}v = \int_{\partial\Omega} \check{\underline{t}}(\underline{x}, \underline{n}, t) \, \mathrm{d}a + \int_{\Omega} \check{\underline{b}} \, \mathrm{d}v \tag{2.16}$$

as the continuum mechanic's counterpart to Newton's second law. In addition, this equation can be easily extended to the *balance of angular momentum* with respect to the origin of coordinates, viz.

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \underline{x} \times \varrho \underline{\dot{u}} \, \mathrm{d}v = \int_{\partial\Omega} \underline{x} \times \check{\underline{t}}(\underline{x}, \underline{n}, t) \mathrm{d}a + \int_{\Omega} \underline{x} \times \check{\underline{b}} \, \mathrm{d}v. \tag{2.17}$$

Now, if the balance of linear momentum (2.16) holds, it can be concluded under application of Cauchy's theorem [188, ch. 2.1] that the vector field $\check{\underline{t}}$ depends linearly on the current normal vector field $\underline{n}$, thus, $\check{\underline{t}} = \underline{\underline{\sigma}}(\underline{x}, t) \cdot \underline{n}$ follows. Here, σ^{ij} denotes one of the components of the so-called *Cauchy stress tensor* $\underline{\underline{\sigma}} \in \mathbb{R}^{3\times3}$, which is completely defined in the current configuration. By inserting this relation together with $\mathrm{div}(\underline{\underline{\sigma}})^i = \partial\sigma^{ij}/\partial x^j$ into (2.16) and by applying the well-known Reynolds transport theorem to the left hand side of (2.16), which for a given vector field $\underline{f}(\underline{x}, t)$ and a time dependent integration domain Ω generally yields

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \underline{f} \, \mathrm{d}v = \int_{\Omega} \dot{\underline{f}} + \underline{f} \mathrm{div}(\underline{\dot{u}}) \, \mathrm{d}v, \tag{2.18}$$

it directly follows that

$$\int_{\Omega} \varrho \underline{\ddot{u}} \, \mathrm{d}v = \int_{\Omega} \mathrm{div}(\underline{\underline{\sigma}}) \, \mathrm{d}v + \int_{\Omega} \check{\underline{b}} \, \mathrm{d}v. \tag{2.19}$$

Here, the left hand side has been additionally simplified by applying (2.14). This balance demand has to hold for each material point in the current as well as in the reference configuration, thus,

$$\varrho\, \underline{\ddot{u}} = \operatorname{div}(\underline{\underline{\sigma}}) + \underline{\check{b}}, \tag{2.20a}$$

$$\Leftrightarrow \qquad \varrho_0\, \underline{\ddot{u}} = \operatorname{Div}(\underline{\underline{P}}) + \underline{\check{b}}_0, \tag{2.20b}$$

must hold, where (2.20b) denotes exactly the same relationship as (2.20a) just formulated in the reference configuration. The therein newly introduced stress tensor $\underline{\underline{P}}$ is the so-called *first Piola-Kirchhoff stress* which can be derived by $\underline{\underline{P}} = \det(\underline{\underline{F}})\, \underline{\underline{\sigma}} \cdot \underline{\underline{F}}^{-\mathrm{T}}$. The first Piola–Kirchhoff stress tensor is again a mixed tensor, potentially living in two different vector spaces. This brings us to the definition of the so-called *2nd-order Piola–Kirchhoff stress* $\underline{\underline{S}} = \underline{\underline{F}}^{-1} \cdot \underline{\underline{P}}$, which is completely defined in the reference configuration and thus it has also the symmetry property of the Cauchy stresses $\sigma^{ij} = \sigma^{ji}$. This symmetry of the Cauchy stress tensor can be easily shown by repeating the steps which led from (2.16) to (2.19) for the balance of angular momentum (2.17).

2.1.3. Constitutive Laws: Hyperelasticity

The previously defined balance equations are the classical starting point if the Hamiltonian or variational principle shall be studied. However, it is not possible to solve these equations without any knowledge about the stress and strain relationship, i.e., the respective constitutive law. In this section the necessary content shall be presented for pure elastic materials. This allows us to express the first Piola–Kirchhoff stress as a function $\underline{\underline{P}}(\underline{X}, t) = \underline{\underline{\mathfrak{P}}}(\underline{X}, \underline{\underline{F}}(\underline{X}, t))$ depending solely on the material position vector $\underline{X}$ and the deformation gradient with $\det(\underline{\underline{F}}) \geqslant 0$. If additionally a homogeneous material is considered, the local dependence on the position $\underline{X}$ can be dropped as well. In this thesis the theory for finite elasticity shall be restricted to the theory of hyperelasticity such that the existence of a suitable scalar valued *strain energy function* Ψ per unit reference volume can be postulated. This circumstance is very convenient for the later discussed optimization and globalization schemes. Furthermore, the discussion shall be further restricted to perfectly elastic materials, i.e., any internal dissipation effects such as damage, plasticity or viscous mechanisms are excluded. Furthermore, it shall hold that the strain energy function reaches its absolute minimum for the undeformed state and this minimum shall be equal to zero, thus, $\Psi(\underline{\underline{\delta}}) = 0$ and $\Psi(\underline{\underline{F}}) \geqslant 0$. This demand $\Psi(\underline{\underline{\delta}}) = 0$ is equivalent to a stress free reference configuration. Another fundamental assumption is that the strain energy function value tends to infinity if the respective volume unit is expanded to infinity or compressed to zero volume. For further information about the existence of a solution, the reader is kindly referred to the corresponding literature about polyconvexity , e.g., Ball [11] or Marsden et al. [188, Ch. 6.4].

By computing the first order derivative of the strain energy with respect to the deformation gradient, the first Piola–Kirchhoff stress tensor is obtained, viz.

$$\underline{\underline{P}} = \underline{\underline{\mathfrak{P}}}(\underline{\underline{F}}) = \frac{\partial \Psi(\underline{\underline{F}})}{\partial \underline{\underline{F}}} \qquad \Leftrightarrow \qquad P_i{}^J = \frac{\partial \Psi}{\partial F^i{}_J}. \tag{2.21}$$

Now, since objectivity or also known as frame indifference must hold, the following relationships are revealed

$$\Psi(\underline{\underline{F}}) = \Psi(\underline{\underline{U}}) = \Psi(\underline{\underline{C}}) = \Psi(\underline{\underline{E}}), \tag{2.22}$$

i.e., the strain energy can be equivalently expressed in terms of the deformation gradient (2.2), the right stretch tensor (2.8), the right Cauchy Green tensor (2.9), or the Green–Lagrange strain (2.10). Finally, the derivative of the strain energy function with respect to the deformation gradient (2.21) leading to first Piola–Kirchhoff stress definition can also be stated as

$$\underline{\underline{P}} = 2\underline{\underline{F}}\frac{\partial\Psi(\underline{\underline{C}})}{\partial\underline{\underline{C}}}, \qquad \underline{\underline{S}} = 2\frac{\partial\Psi(\underline{\underline{C}})}{\partial\underline{\underline{C}}} = \frac{\partial\Psi(\underline{\underline{E}})}{\partial\underline{\underline{E}}}. \tag{2.23}$$

Now, if the description is further limited to the special case of isotropy such that the material behavior is supposed to be identical in any material direction, the strain density function becomes independent of the defined material axes and is only a function of the scalar-valued invariants of $\underline{\underline{C}}$, which are defined as

$$I_1 = \mathrm{tr}(\underline{\underline{C}}), \qquad I_2 = \frac{1}{2}\left[(\mathrm{tr}\underline{\underline{C}})^2 - \mathrm{tr}(\underline{\underline{C}}^2)\right], \qquad I_3 = \det(\underline{\underline{C}}) = (\det(\underline{\underline{F}}))^2. \tag{2.24}$$

Thus, the second Piola–Kirchhoff stress can also be expressed as

$$\underline{\underline{S}} = 2\frac{\partial\Psi}{\partial\underline{\underline{C}}} = 2\sum_{i=1}^{3}\frac{\partial\Psi}{\partial I_i}\frac{\partial I_i}{\partial\underline{\underline{C}}} \tag{2.25}$$

under these assumptions. The necessary derivatives of the presented invariants can be found in the related literature, e.g., Bonet and Wood [32], Holzapfel [136]. Now, under consideration of this very basic hyperelasicity framework for homogeneous, isotropic constitutive laws, two examples shall be presented which are widely used during the simulations in this thesis. Both of them can be summarized as compressible neo-Hookean materials. The big advantage of these materials is that their free parameters can easily be identified by the material parameters known from linear elasticity, such as the Young's modulus E, the Poisson's ratio ν, or the directly related shear modulus μ. The first one is the strain density function for the coupled neo-Hookean material given in Holzapfel [136, p. 247]

$$\Psi_{\mathrm{nH}} = \frac{\mu_{\mathrm{nH}}}{2\beta_{\mathrm{nH}}}(I_3^{-\beta_{\mathrm{nH}}} - 1) + \frac{\mu_{\mathrm{nH}}}{2}(I_1 - 3), \qquad \beta = \frac{\nu}{1-2\nu}. \tag{2.26}$$

The second neo-Hookean material law follows the closely related definition provided in Bonet and Wood [32, p. 162], viz.

$$\Psi_{\mathrm{nHlog}} = \frac{\mu_{\mathrm{nH}}}{2}(I_1 - 3) - \frac{\mu_{\mathrm{nH}}}{2}\ln(I_3) + \frac{\lambda_{\mathrm{nH}}}{8}[\ln(I_3)]^2, \qquad \lambda_{\mathrm{nH}} = \frac{\nu E}{(1+\nu)(1-2\nu)} \tag{2.27}$$

here denoted as logarithmic neo-Hookean. However, the material choice is not limited to these two examples, instead, any hyperelastic material can be chosen as long as it is polyconvex and capable of large deformations. An example for an ill-suited material law might be the very simple St. Venant-Kirchhoff material which exhibits serious difficulties if it is used for large strain problems. In fact, it is unable to predict a realistic deformation pattern under such circumstances, see Bonet and Burton [31] for a demonstrative example.

2.1.4. Weak Form

In the preliminary sections the fundamental relations of classical non-linear continuum mechanics have been presented. However, if the goal is to solve complex elastodynamical problems it is often not possible to find a closed analytical solution for these equations. Instead, numerical methods start to become important and propose a way to find an approximate solution. Now, in this section two possible ways to the so-called *weak form* shall be presented. The weak form represents the reformulation which is necessary such that the proposed general continuum's mechanics equations become applicable to numerical schemes.

Based on the Strong Form

First, the classical way is shown. Starting with the initial boundary value problem for the underlying partial differential equations, it is quite simple to obtain the necessary reformulation. In elastodynamics the strong form follows as

$$\varrho_0 \ddot{\underline{u}} = \mathrm{Div}\underline{\underline{P}} + \check{\underline{b}}_0 \qquad \text{in } \Omega_0 \times [0,T], \tag{2.28a}$$

$$\underline{u} = \check{\underline{u}} \qquad \text{on } \Gamma_u \times [0,T], \tag{2.28b}$$

$$\underline{\underline{P}} \cdot \underline{N} = \check{\underline{t}}_0 \qquad \text{on } \Gamma_\sigma \times [0,T], \tag{2.28c}$$

with the inital conditions

$$\underline{u}(\underline{X},0) = \check{\underline{u}} \qquad \text{in } \Omega, \tag{2.28d}$$

$$\dot{\underline{u}}(\underline{X},0) = \check{\dot{\underline{u}}} \qquad \text{in } \Omega. \tag{2.28e}$$

The stress tensor $\underline{\underline{P}}$ still denotes the first Piola–Kirchhoff stress (2.21), the vector $\check{\underline{b}}_0$ summarizes the acting volume forces, and the vector $\check{\underline{t}}_0$ all applied Neumann loads on the Neumann boundary Γ_σ. These last two quantities have already been introduced in Section 2.1.2. In addition, $\check{\underline{u}}$ denotes prescribed values of the primary field variables. Namely, the displacement field on the Dirichlet boundary Γ_u. If a dynamic problem is considered, additional values for the initial displacements $\check{\underline{u}}$ and velocities $\check{\dot{\underline{u}}}$ must be provided in the entire domain Ω_0 at $t = 0$. Next, an arbitrary weighting function $\delta\underline{u}$ is introduced, such that the given strong form directly leads to

$$0 = \int_{\Omega_0} (\varrho_0 \ddot{\underline{u}} - \mathrm{Div}\underline{\underline{P}} - \check{\underline{b}}_0) \cdot \delta\underline{u} \, \mathrm{d}V_0 + \int_{\Gamma_\sigma} (\underline{\underline{P}} \cdot \underline{N} - \check{\underline{t}}_0) \cdot \delta\underline{u} \, \mathrm{d}A_0. \tag{2.29}$$

In the given structural mechanics context the introduced weighting function can be interpreted as a virtual displacement vector with

$$\delta\underline{u} \in \mathcal{W} = \{\delta\underline{u} \in \mathcal{H}^1(\Omega_0) : \ \delta\underline{u} = \underline{0} \text{ on } \Gamma_u\}, \tag{2.30}$$

where $\mathcal{W}$ is the weighting function space. In addition, $\mathcal{H}^1(\Omega_0)$ is the so-called Sobolev function space following

$$\mathcal{H}^1(\Omega_0) = \{f \in L_2(\Omega_0) : \int_{\Omega_0} \|\nabla f\|^2 \mathrm{d}\Omega < \infty\}, \quad \text{where } L_2(\Omega_0) = \{f : \int_{\Omega_0} f^2 \mathrm{d}\Omega < \infty\}. \tag{2.31}$$

Therefore, the meaning as well as the solution have not changed from (2.28) to (2.29). Next, the divergence theorem is applied to the boundary integral in (2.29) yielding

$$\int_{\Gamma_\sigma} (\underline{\underline{P}} \cdot \underline{N} - \check{\underline{t}}_0) \cdot \delta\underline{u} \, \mathrm{d}A_0 = \int_{\Omega_0} \underline{\underline{P}} : \nabla_{\underline{X}}(\delta\underline{u}) \mathrm{d}V_0 + \int_{\Omega_0} \delta\underline{u} \cdot \mathrm{Div}(\underline{\underline{P}}) \mathrm{d}V_0 - \int_{\Gamma_\sigma} \check{\underline{t}}_0 \cdot \delta\underline{u} \, \mathrm{d}A_0. \tag{2.32}$$

This is inserted into (2.29) together with the identities $\nabla_{\underline{X}}(\delta\underline{u}) = \partial(\delta u)^i / \partial X^J = \delta F^i{}_J$ and $\underline{\underline{P}} = \underline{\underline{F}} \cdot \underline{\underline{S}}$ such that

$$0 = \int_{\Omega_0} (\varrho_0 \ddot{\underline{u}} - \check{\underline{b}}_0) \cdot \delta\underline{u} \, \mathrm{d}V_0 - \int_{\Gamma_\sigma} \check{\underline{t}}_0 \cdot \delta\underline{u} \, \mathrm{d}A_0 + \int_{\Omega_0} (\underline{\underline{F}} \cdot \underline{\underline{S}}) : \delta\underline{\underline{F}} \, \mathrm{d}V_0, \tag{2.33}$$

where the last term can be reformulated as

$$\begin{aligned} &\int_{\Omega_0} (\underline{\underline{F}} \cdot \underline{\underline{S}}) : \delta\underline{\underline{F}} \, \mathrm{d}V_0 = \int_{\Omega_0} F^a{}_A \, S^{AB} \, \delta F_{aB} \, \mathrm{d}V_0 \\ &= \int_{\Omega_0} \frac{1}{2} (F^a{}_A \, \delta F_{aB} \, S^{AB} + F^a{}_A \, \delta F_{aB} \, S^{BA}) \, \mathrm{d}V_0 \\ &= \int_{\Omega_0} \frac{1}{2} (\delta\underline{\underline{F}}^{\mathrm{T}} \cdot \underline{\underline{F}} + \underline{\underline{F}}^{\mathrm{T}} \cdot \delta\underline{\underline{F}}) : \underline{\underline{S}} \, \mathrm{d}V_0 = \int_{\Omega_0} \delta\underline{\underline{E}} : \underline{\underline{S}} \, \mathrm{d}V_0, \end{aligned} \tag{2.34}$$

In (2.34) the symmetry of the second Piola–Kirchhoff stress tensor has been used explicitly. Thus, the final weak form is obtained by

$$\delta\mathscr{U} = \int_{\Omega_0} \varrho_0 \ddot{\underline{u}} \cdot \delta\underline{u} \, \mathrm{d}V_0 + \int_{\Omega_0} \underline{\underline{S}} : \delta\underline{\underline{E}} \, \mathrm{d}V_0 - \int_{\Omega_0} \check{\underline{b}}_0 \cdot \delta\underline{u} \, \mathrm{d}V_0 - \int_{\Gamma_\sigma} \check{\underline{t}}_0 \cdot \delta\underline{u} \, \mathrm{d}A_0 = 0. \tag{2.35}$$

This is also known as the demand that the virtual work $\delta\mathscr{U}$ vanishes at the solution. The notation $\delta\mathscr{U}$ already implies the second possibility to derive this result: as variation of a scalar valued potential function $\mathscr{U}$. However, it must be highly emphasized that the principle of virtual work derived here does not need an underlying potential and is therefore also applicable to problems involving non-conservative contributions such as frictional contact, plasticity, viscoelasticity, or air resistance to name only a few. Consequently, it does neither need the existence of a strain energy function (see Section 2.1.3 for a brief introduction into this topic). Thus, the presented principle can be seen as a much more general starting point for the derivation of the structural finite element method in contrast to the following derivation from a potential formulation.

Based on the Balance of Mechanical Energy

Even though the next derivation is less general, it brings another advantage: It introduces a scalar-valued mechanical energy function which can be used as an objective function later on. The existence of such a function brings a lot of useful properties, e.g., many publications about sophisticated optimization algorithms need some kind of merit function (see Nocedal and Wright [204] for an overview). In addition, the consistent variation starting from an objective function preserves a symmetric Hessian matrix. This simplifies on the one hand many theoretical considerations since a clear definition for positive definiteness is available and, on the other hand, it opens a much wider field for iterative linear solvers.

In this section as well as in the entire thesis only mechanical sources of energy shall be considered, i.e. any other chemical, thermal, electric or magnetic source is excluded from the discussion. Furthermore, the balance of mechanical energy shown here does not contribute with a new balance equation to the discussion in Section 2.1.2, but instead the balance of mechanical energy can be derived from the balance of linear momentum (2.16).

First, the *kinetic energy* $\mathscr{K}$ of the considered continuum problem shall be investigated. Therefore, the well-known relationship of Newton's mechanics is generalized for the continuum mechanics approach proposed here leading to

$$\mathscr{K}(t) = \int_{\Omega_0} \frac{1}{2} \varrho_0 \underline{\dot{u}} \cdot \underline{\dot{u}} \mathrm{d}V_0 = \frac{1}{2} \int_{\Omega} \varrho \, \underline{\dot{u}} \cdot \underline{\dot{u}} \mathrm{d}v. \tag{2.36}$$

Next, the *rate of external mechanical work* $\mathcal{P}_{\mathrm{ext}}$ shall be introduced by

$$\mathcal{P}_{\mathrm{ext}}(t) = \int_{\Omega_0} \check{\underline{b}}_0 \cdot \underline{\dot{u}} \, \mathrm{d}V_0 + \int_{\partial\Omega_0} \check{\underline{t}}_0 \cdot \underline{\dot{u}} \, \mathrm{d}A_0 = \int_{\Omega} \check{\underline{b}} \cdot \underline{\dot{u}} \, \mathrm{d}v + \int_{\partial\Omega} \check{\underline{t}} \cdot \underline{\dot{u}} \, \mathrm{d}a \tag{2.37}$$

This is the rate form of the introduced energy by the externally applied loads. Now, only the *rate of internal mechanical work* $\mathcal{P}_{\mathrm{int}}$ is missing. This term represents the mechanical response, i.e., the rate of change of internally stored energy due to deformation, also called *stress power*, and is generally obtained by

$$\mathcal{P}_{\mathrm{int}}(t) = \int_{\Omega_0} \underline{\underline{P}} : \dot{\underline{\underline{F}}} \, \mathrm{d}V_0 = \int_{\Omega_0} \underline{\underline{S}} : \dot{\underline{\underline{E}}} \, \mathrm{d}V_0 = \int_{\Omega} \underline{\underline{\sigma}} : (\dot{\underline{\underline{F}}} \cdot \underline{\underline{F}}^{-1}) \, \mathrm{d}v. \tag{2.38}$$

Finally, the often called *balance of mechanical energy*, or more precisely the *theorem of power expended* can be stated as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{K}(t) + \mathcal{P}_{\mathrm{int}}(t) = \mathcal{P}_{\mathrm{ext}}(t). \tag{2.39}$$

Hence, the rate of external mechanical work is balanced by the rate of internal work and kinetic work. In other words: The rate of the kinetic energy contains contributions from the external as well as from the internal work and is consequently not generally conserved. A closer look at (2.39) reveals some well-known simplifications of this balance equation: First, if the rate of external work $\mathcal{P}_{\mathrm{ext}}$ vanishes, a problem of free vibrations is obtained. Secondly, if the rate of change with respect to the kinetic energy $\mathrm{d}/\mathrm{d}t(\mathscr{K}(t))$ vanishes a quasi-static problem becomes present. However, the latter one still allows time dependence of other terms. An example would be a quasi-static creep deformation modeled by a time dependent material law. The proof that the left and right hand side of (2.39) are indeed equal can be found in Holzapfel [136, p. 154], for instance. This formulation in terms of rates rather than conserved absolute quantities is generally holding. However, it shall be taken one step further: The consideration of *conservative systems*. The existence of a conservative system implies that a scalar-valued *total potential energy*

$$\mathscr{U}_{\mathrm{tot}}(t) = \mathscr{U}_{\mathrm{ext}}(t) + \mathscr{U}_{\mathrm{int}}(t) \tag{2.40}$$

can be formulated, such that the *rate of external mechanical work* (2.37) can be deduced from the *potential energy of the external loading* by

$$\mathcal{P}_{\mathrm{ext}}(t) = -\frac{\mathrm{d}\mathscr{U}_{\mathrm{ext}}(t)}{\mathrm{d}t} = -\dot{\mathscr{U}}_{\mathrm{ext}}(t), \tag{2.41}$$

and the *stress power* (2.38) from the *total strain energy*

$$\mathcal{P}_{\mathrm{int}}(t) = \frac{\mathrm{d}\mathscr{U}_{\mathrm{int}}(t)}{\mathrm{d}t} = \dot{\mathscr{U}}(t), \qquad \text{with } \mathscr{U}_{\mathrm{int}}(t) = \int_{\Omega_0} \Psi \, \mathrm{d}V_0. \tag{2.42}$$

The function Ψ in (2.42) represents the strain energy function introduced in Section 2.1.3. Next, with all this at hand, the explicit form of the *total potential energy* in elastostatics can be given by

$$\mathscr{U}_{\mathrm{tot}}(\underline{u}) = \int_{\Omega_0} \Psi(\underline{X}, \underline{\underline{F}}) - \check{\underline{b}}_0(\underline{X}) \cdot \underline{u}(\underline{X}) \, \mathrm{d}V_0 - \int_{\Gamma_\sigma} \check{\underline{t}}_0(\underline{X}) \cdot \underline{u}(\underline{X}) \, \mathrm{d}A_0. \tag{2.43}$$

Note that it is assumed in the following, that all considered loads are dead (see e.g. Marsden et al. [188, p. 212]), i.e., the prescribed traction $\check{\underline{t}}(\underline{X}) = \underline{\underline{P}} \cdot \underline{N}$ and the volume load $\check{\underline{b}}$ shall be constant during the considered motion. Thus, the directional derivative of (2.43) reveals

$$\begin{aligned} \mathrm{D}_{\delta\underline{u}}(\mathscr{U}_{\mathrm{tot}}(\underline{u})) &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathscr{U}_{\mathrm{tot}}(\underline{u}+\varepsilon\delta\underline{u})\Big|_{\varepsilon=0} \\ &= \int_{\Omega_0} \frac{\partial\Psi}{\partial\underline{\underline{F}}}\cdot\nabla_{\underline{X}}(\delta\underline{u}) - \check{\underline{b}}_0\cdot\delta\underline{u}\,\mathrm{d}V_0 - \int_{\partial\Omega}\check{\underline{t}}_0\cdot\delta\underline{u}\,\mathrm{d}A_0 \end{aligned} \tag{2.44}$$

and this coincides with the quasi-static part of $\delta\mathscr{U}$ in (2.35). Additionally, if others than dead loads are considered and, consequently, the load depends on the current point position $\underline{x}$, it must be ensured that these loads are conservative, i.e., derivable from a potential such that $\mathrm{D}_{\delta\underline{u}}(\mathscr{V}_{\check{b}}(\underline{u})) = -\check{\underline{b}}\cdot\delta\underline{u}$ and $\mathrm{D}_{\delta\underline{u}}(\mathscr{V}_{\check{t}}(\underline{u})) = -\check{\underline{t}}\cdot\delta\underline{u}$ hold.

The existence of such a total energy formulation can be mathematically very precisely formulated by the demand that

$$\mathrm{D}_{\underline{w}}(\mathrm{D}_{\underline{v}}(\mathscr{U}_{\mathrm{tot}}(\underline{u}))) = \mathrm{D}_{\underline{v}}(\mathrm{D}_{\underline{w}}(\mathscr{U}_{\mathrm{tot}}(\underline{u}))) \tag{2.45}$$

holds for all $\underline{v},\underline{w}\in\mathcal{W}$ defined in (2.30). This is equivalent to the demand, that the Jacobian matrix $\mathrm{D}^2_{\delta\underline{u}}(\mathscr{U}_{\mathrm{tot}}(\underline{u}))$ is symmetric. The strain energy fulfills this demand if the fourth order elasticity tensor $\mathbb{C}$ has the symmetry $C_a{}^A{}_b{}^B = C_b{}^B{}_a{}^A$ which can be directly concluded if a strain energy function is used. For more information the reader is kindly referred to Marsden et al. [188, Ch. 5.1].

The extension to elastodynamics can be achieved in different ways. One possibility is to add up (2.36) and (2.40) and formulate the demand

$$\mathscr{K}(t) + \mathscr{U}_{\mathrm{tot}}(t) = \mathrm{const.} \tag{2.46}$$

Remark 2.2. For the sake of completeness it should be mentioned that the total potential (2.43) can be extended by an additional additive term representing the energy input into the system by the prescribed displacements. This term lives completely on the Dirichlet boundary Γ_u and is defined as

$$\int_{\Gamma_u}\underline{t}_0(\underline{X})\cdot(\check{\underline{u}}-\underline{u}(\underline{X}))\,\mathrm{d}A_0.$$

However, it is not relevant for the upcoming discussion and, therefore, it is neglected.

Based on the Lagrangian Field Theory

Another way is based on the more general Lagrangian field theory which shall be applied here. This theory provides a smooth transition from elastostatics to the elastodynamics. For instance, a comprehensive derivation can be found in Marsden et al. [188, Ch. 5.4, p. 275]. Here, only the final result shall be presented:

Under consideration of the *Lagrangian density function*

$$\mathcal{L}(\underline{X},\underline{u},\dot{\underline{u}},\underline{\underline{F}}) \qquad\qquad \mathcal{L}(X^I,u^i,\dot{u}^i,F^i_I), \tag{2.47}$$

the structural Lagrangian can be defined by

$$\mathscr{U}(\underline{u}, \dot{\underline{u}}) = \int_{\Omega_0} \mathcal{L}(\underline{X}, \underline{u}(\underline{X}), \dot{\underline{u}}(\underline{X}), \underline{\underline{F}}(\underline{X}))\mathrm{d}V_0 - \int_{\partial\Omega_0} \mathscr{V}_{\check{t}}(\underline{u}(\underline{X}))\mathrm{d}A_0, \tag{2.48}$$

where as previously introduced $\nabla_{\underline{u}} \mathscr{V}_{\check{t}}(\underline{u}(\underline{X}) = -\underline{\check{t}}$ holds, and, therefore, allows the consideration of conservative Neumann loads. The Piola–Kirchhoff stress tensor can be derived by

$$\underline{\underline{P}} = -\frac{\partial \mathcal{L}}{\partial \underline{\underline{F}}} \qquad\qquad P_i{}^I = -\frac{\partial \mathcal{L}}{\partial F^i{}_I}. \tag{2.49}$$

The weak form of the Lagrange density equations follows then as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega_0} \frac{\partial}{\partial \dot{\underline{u}}}[\mathcal{L}(\underline{u}, \dot{\underline{u}}, \underline{\underline{F}})] \cdot \delta\underline{u}\, \mathrm{d}V_0 = \int_{\Omega_0} \frac{\partial}{\partial \underline{u}}[\mathcal{L}(\underline{u}, \dot{\underline{u}}, \underline{\underline{F}})] \cdot \delta\underline{u}\, \mathrm{d}V_0 + \int_{\Omega_0} \frac{\partial}{\partial \underline{\underline{F}}}[\mathcal{L}(\underline{u}, \dot{\underline{u}}, \underline{\underline{F}})] : \delta\underline{\underline{F}}\, \mathrm{d}V_0 \\ + \int_{\partial\Omega_0} \underline{\check{t}}_0 \cdot \delta\underline{u}\, \mathrm{d}A_0.\end{aligned} \tag{2.50}$$

Additionally, the strong form is obtained by

$$\frac{\partial}{\partial t}\frac{\partial \mathcal{L}}{\partial \dot{\underline{u}}} = \frac{\partial \mathcal{L}}{\partial \underline{u}} - \mathrm{Div}\frac{\partial \mathcal{L}}{\partial \underline{\underline{F}}} \qquad\qquad \underline{\underline{P}} \cdot \underline{N} = \underline{\check{t}}_0 \text{ on } \Gamma_\sigma. \tag{2.51}$$

Finally, in the case of hyperelasticity the Lagrange density function can be directly identified by

$$\mathcal{L}(\underline{u}, \dot{\underline{u}}, \underline{\underline{F}}) = \frac{1}{2}\varrho_0 \dot{\underline{u}} \cdot \dot{\underline{u}} - \Psi(\underline{\underline{F}}) - \mathscr{V}_{\check{b}}(\underline{u}). \tag{2.52}$$

These functions describe precisely the problem stated in the strong (2.28) and weak form (2.35), however, under stricter assumptions due to the dependence on a scalar-valued Lagrange density function (2.52) and the demand for a potential formulation to include the external loads, which together form the Lagrangian in (2.48).

Remark 2.3. The request for conservative systems stated here can actually be weakened for the Lagrange equations by consideration of so-called *generialized* or *velocity-dependent potentials*. A interesting discussion on this topic can be exemplarily found in Goldstein et al. [112, Ch. 1.5]. In the cited chapter two different velocity dependent forces are discussed, namely, the electromagnetic forces on moving charges and, secondly, friction forces following the so-called Rayleigh's dissipation function. Furthermore, in terms of frictional contact problems in continuum mechanics the *quasi-augmented Lagrangian* function in Alart and Curnier [1] or the extension to large deformations in Pietrzak and Curnier [212] is worth mentioning. A more detailed description of the latter one will follow in Section 2.2.3.

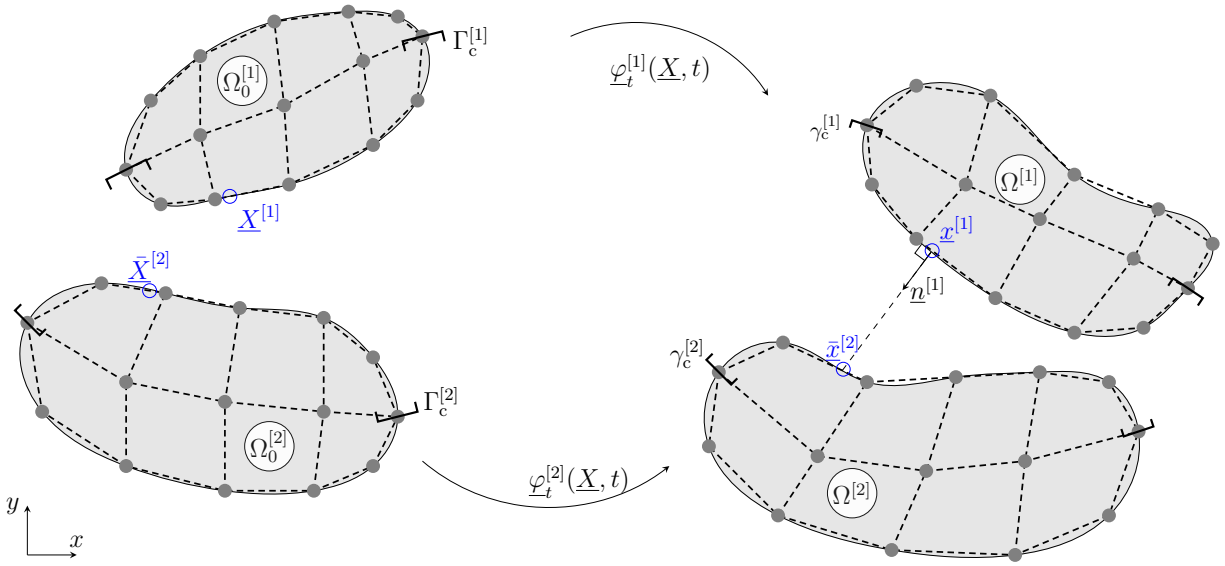


Figure 2.1.: Basic notations and kinematics for a unilateral large deformation contact problem in 2-D. The potential contact boundary zone is enclosed by square brackets in the reference and current configuration, respectively.

2.2. Contact Mechanics

Next, the attention is drawn to the field of contact mechanics. First the basic kinematic relationships between two contacting bodies shall be presented where the focus is put on frictionless contact scenarios. However, the interested reader is referred to the comprehensive literature on contact mechanics such as Kikuchi and Oden [157], Laursen [170], Wriggers [283] for a much deeper insight into the basic relationships. After the kinematics have been discussed, the fundamental contact problem will be introduced. The used notation is chosen in such a way that the transition to the following constrained numerical optimization chapter is as easy as possible. For convenience, the up-coming introduction is restricted to unilateral contact between two elastic bodies. However, the extension to more involved problems such as self-contact or multi-body contact is possible and should be straight forward. Finally, the possible extension to frictional contact is briefly presented as well.

2.2.1. Contact Kinematics

At the beginning the introduction shall be started with a brief overview of the underlying kinematic equations where a classical Boltzmann continuum model in a three-dimensional Euclidean space is considered. Since this is the most general case, the related 2-D equations can be easily derived. Furthermore, a Cartesian coordinate system will be used, if not explicitly defined differently. For the spatial description, two distinct observer frames are defined: One for the reference configuration $\Omega_0 \subset \mathbb{R}^3$ at $t = 0$ and one for the current configuration $\Omega \subset \mathbb{R}^3$ at time $t > 0$. The mapping between these two configurations is - as already previously defined - given by $\underline{x} = \varphi_t(\underline{X}, t)$ or $\underline{X} = \varphi_t^{-1}(\underline{x}, t)$. The absolute displacement of a material point is then obtained by $\underline{u}(\underline{X}, t) = \underline{x}(\underline{X}, t) - \underline{X}$. For a more detailed description of non-linear continuum mechanics aspects the reader is referred to the first part of this chapter.

As mentioned in the introduction to this section, the description is restricted to the unilateral contact case between two elastic bodies. The current spatial configuration of each body is denoted by $\underline{x}^{[b]} = \underline{X}^{[b]} + \underline{u}^{[b]}$, where $b = 1$ identifies the so-called *slave* or *non-mortar* and $b = 2$ the so-called *master* or *mortar* body. The potential contact interfaces in the reference configuration are given by $\Gamma_{\mathrm{c}}^{[b]} \in \mathbb{R}^2 \subset \Omega_0^{[b]}$ for each body $b \in \{1, 2\}$. Their counterparts in the current configuration are represented by $\gamma_{\mathrm{c}}^{[1]}$, $\gamma_{\mathrm{c}}^{[2]}$, respectively.

Before the contact details are presented, a more general convective coordinate system in the current configuration on the respective contact surface shall be defined by the covariant base vectors $\underline{\tau}^{[b]}{}_i$, $i \in \{1, 2\}$. The associated contravariant coordinates are given by $\{\zeta^{[b]i}\}_{i\in\{1,2\}}$, such that the definition of the covariant base vectors follows as $\underline{\tau}^{[b]}{}_i = \underline{x}^{[b]}{}_{,\zeta^{[b]i}} = \partial \underline{x}^{[b]} / \partial \zeta^{[b]i} = \partial \varphi_t^{[b]}(\underline{X}, t) / \partial \zeta^{[b]i}$. Thus, a point $\underline{x}^{[1]}$ on the slave surface can be expressed as a function of the convective coordinates via $\underline{x}^{[1]} = \underline{x}(\zeta^{[1]1}, \zeta^{[1]2}) = \underline{x}(\zeta^{[1]i})$. Furthermore, the outward pointing unit surface normal on the slave side is easily obtained by

$$\underline{n}^{[1]} = \underline{n}(\underline{x}(\zeta^{[1]i})) = \frac{\hat{\underline{n}}(\underline{x}(\zeta^{[1]i}))}{\|\hat{\underline{n}}(\underline{x}(\zeta^{[1]i}))\|} = \frac{\underline{\tau}^{[1]}{}_1 \times \underline{\tau}^{[1]}{}_2}{\|\underline{\tau}^{[1]}{}_1 \times \underline{\tau}^{[1]}{}_2\|}. \tag{2.53}$$

Equivalent definitions can be derived for the master side.

A key ingredient of any frictionless contact formulation is the normal gap definition between the two bodies. There are several options how to define this function. They mostly differ in the definition of the involved normal vector, which is either defined on the slave as in Popp et al. [215], Yang et al. [290] or on the master surface, viz. Laursen [170], Wriggers [283]. The latter variant is closely related to the so-called closest point projection. Nonetheless, whichever definition is in use, the normals are supposed to coincide in the converged active contact zone.

In this thesis a normal field defined on the slave side as given in Popp et al. [215], Yang et al. [290] is used. With this definition it is possible to define the projection of a position $\underline{x}^{[1]}$ on $\gamma_{\mathrm{c}}^{[1]}$ onto an arbitrary point located at $\underline{x}^{[2]} = \underline{x}^{[2]}(\zeta^{[2]i})$ on $\gamma_{\mathrm{c}}^{[2]}$ as the root of

$$\underline{\chi}(\hat{\underline{x}}, \zeta^{[2]i}, \alpha_\chi)\big|_{\hat{\underline{x}}=\underline{x}^{[1]}} = \left\{\underline{x}^{[2]}(\zeta^{[2]i}) - \hat{\underline{x}} - \alpha_\chi \underline{n}(\hat{\underline{x}})\right\}\big|_{\hat{\underline{x}}=\underline{x}^{[1]}} \stackrel{!}{=} \underline{0}, \tag{2.54}$$

where $\underline{n}^{[1]}$ is the continuous normal field defined in (2.53) and α_χ denotes an auxiliary distance factor. It is obvious that there is more than one possible definition of this projection procedure. See for example Popp et al. [215] for an alternative definition based on the cross-product in the two-dimensional case. Nevertheless, one advantage of (2.54) is that the auxiliary distance factor can be helpful to define a unique projection point on the master side if more than one candidate was found. In such a case, the point with the smallest $\bar{\alpha}_\chi$ value and the correct orientation relative to the employed normal is most likely the correct choice (see Figure 2.2 for an illustrative example). All projection rules have in common that, in general, a non-linear system of equations has to be solved. Here, a local Newton scheme is used for this task. The solution of the projection is often denoted by variables with a bar symbol such as $\bar{\underline{x}}^{[2]} = \underline{x}^{[2]}(\bar{\zeta}^{[2]i})$ and $\bar{\alpha}_\chi$. Throughout this thesis this rule will be followed in conjunction with contact related terms, but it is still stated here that all variables $\underline{x}^{[2]}$ are meant to be the projected counterparts of some point $\underline{x}^{[1]}$, if not explicitly defined differently.

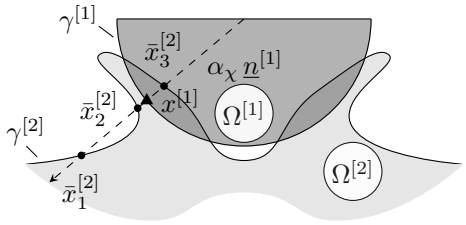


Figure 2.2.: Illustration of a difficult projection scenario during a non-linear solution procedure with three possible master point candidates (•) for the slave point (▲). While $\underline{\bar{x}}_2^{[2]}$ can be excluded due to the enclosed angle of the master and slave normals, the point $\underline{\bar{x}}_1^{[2]}$ disqualifies due to the corresponding positive auxiliary distance factor, such that $\underline{\bar{x}}_3^{[2]}$ is the only remaining unique projection point.

Finally, the continuous normal gap function definition for contact follows as

$$g_{\mathrm{N}}(\underline{x}(\zeta^{[1]i}), \underline{x}(\bar{\zeta}^{[2]j})) = \langle \underline{n}^{[1]}(\underline{x}^{[1]}),\ \underline{\bar{x}}^{[2]} - \underline{x}^{[1]} \rangle, \tag{2.55}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vector quantities. Note that a positive value of $g_{\mathrm{N}}(\underline{x}) : \mathbb{R}^2 \to \mathbb{R}$ indicates an open gap between the two bodies, while a negative value corresponds to some kind of overlap.

2.2.2. General Frictionless Contact Problem

There are different ways to derive the general contact problem. In Hiermeier et al. [131] first the different continuously formulated contributions have been discretized and, afterwards, inserted into a Lagrangian functional. Alternatively, it is possible to start at the strong form, given by the initial boundary value problem of elastodynamics (2.28) for each subdomain $\Omega_0^{[b]}$. This at hand, the corresponding constraints have to be defined to prevent a non-physical overlap of the considered bodies. Therefore, the so-called *Karush–Kuhn–Tucker* conditions or in the context of contact problems also known as *Hertz–Signorini–Moreau* conditions are introduced

$$g_{\mathrm{N}}(\underline{x}(\underline{X}, t)) \geqslant \underline{0}, \qquad p_{\mathrm{N}}(\underline{x}(\underline{X}, t)) \leqslant 0, \qquad p_{\mathrm{N}}(\underline{x}(\underline{X}, t)\, g_{\mathrm{N}}(\underline{x}(\underline{X}, t)) = 0, \tag{2.56}$$

on $\gamma_{\mathrm{c}}^{[1]} \times [0, T]$, where p_{N} represents the contact pressure acting on the potential contact boundary $\gamma_{\mathrm{c}}^{[b]}$ in the current configuration. This boundary zone is the counterpart to $\Gamma_{\mathrm{c}}^{[b]}$ in the reference configuration which represents one sub-set of the entire surface of each body $\partial\Omega_0^{[b]}$. The remaining surface can be subdivided into the already defined Dirichlet boundary $\Gamma_u^{[b]}$ and the Neumann boundary zones $\Gamma_\sigma^{[b]}$, respectively. The last equality in (2.56) is the so-called *complementarity condition* which ensures that the gap is always equal to zero as long as the contact pressure is unequal to zero and the other way around. To put it differently: There can only exist a contact pressure as long as the bodies are in contact ($g_{\mathrm{N}} = 0$). Furthermore, the first inequality in (2.56) claims that the bodies are not allowed to overlap, while the second inequality has the task to avoid tensile forces between the contacting bodies. This is totally meaningful as long as the modeling of adhesion effects is not part of the formulation. For an extension in this direction the reader is referred to Sauer and De Lorenzis [234], Sauer and Wriggers [235].

Using the presented strong form (2.28), the well-known principle of virtual work can be directly applied, thus,

$$0 = \sum_{b=1}^{2} \{ \int_{\Omega_0^{[b]}} (\varrho_0^{[b]} \ddot{\underline{u}}^{[b]} - \mathrm{Div} \underline{\underline{P}}^{[b]} - \check{\underline{b}}_0^{[b]}) \cdot \delta \underline{u}^{[b]} \, \mathrm{d}V_0 + \int_{\Gamma_\sigma^{[b]}} (\underline{\underline{P}}^{[b]} \cdot \underline{N}^{[b]} - \check{\underline{t}}_0^{[b]}) \cdot \delta \underline{u}^{[b]} \, \mathrm{d}A_0 \}$$
$$- \int_{\gamma_c^{[1]}} \lambda_\mathrm{N} \, (\delta g_\mathrm{N} - \delta s) \; + \delta \lambda_\mathrm{N} \, (g_\mathrm{N} - s) \, \mathrm{d}a \tag{2.57}$$

is obtained, where the new primary variables λ_N and s have been introduced. While the first one represents the Lagrange multiplier for the non-penetration constraints and can be directly identified as $\lambda_\mathrm{N} = -p_\mathrm{N}$, the second one is a so-called *slack variable* s which allows to replace the inequality constraint (2.56) by an equality constraint as explained in Gill et al. [107], Rockafellar [229]. The optimal choice for s can be easily derived by $s^\star = \max\{0, g_\mathrm{N} - \lambda_\mathrm{N}/c_\mathrm{N}\}$, where c_N is a so-called *complementarity* or *regularization* parameter. A detailed introduction will follow at the beginning of Section 3.2. This allows the integral over the current potential contact zone $\gamma_c^{[1]}$ to be divided into two parts: The first, so-called *inactive*, part is defined by $g_\mathrm{N} > \lambda_\mathrm{N}/c_\mathrm{N}$ and denoted by $\gamma_c^{[1]\mathcal{I}}$. The complementary, so-called *active*, sub-domain is denoted by $\gamma_c^{[1]\mathcal{A}}$. After all, it can be concluded with the final weak form

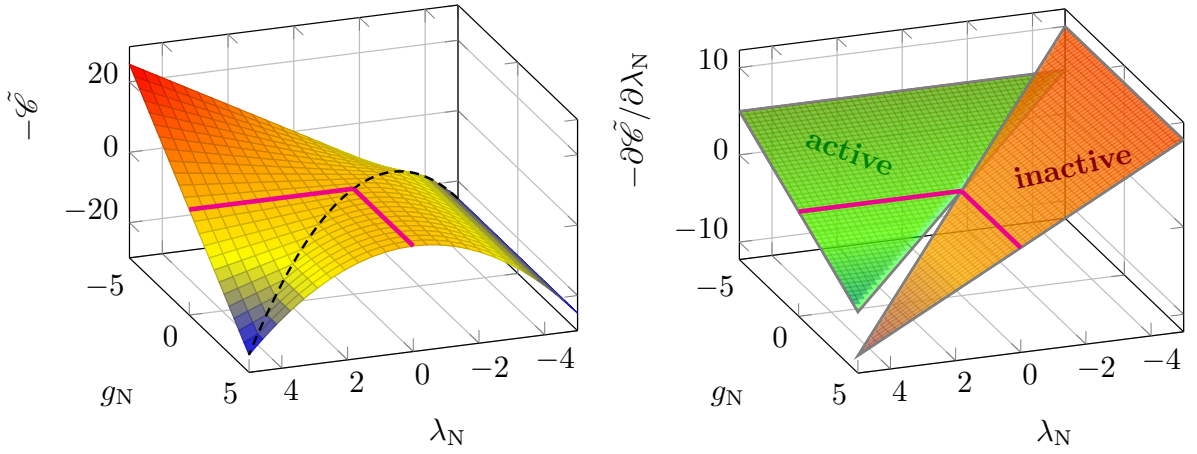
$$0 = \sum_{b=1}^{2} \{ \int_{\Omega_0^{[b]}} \rho_0^{[b]} \ddot{\underline{u}}^{[b]} \cdot \delta \underline{u}^{[b]} + \underline{\underline{S}}^{[b]} : \delta \underline{\underline{E}}^{[b]} - \check{\underline{b}}_0^{[b]} \cdot \delta \underline{u}^{[b]} \mathrm{d}V_0 - \int_{\Gamma_\sigma^{[b]}} \check{\underline{t}}_0^{[b]} \cdot \delta \underline{u}^{[b]} \, \mathrm{d}A_0 \}$$
$$- \int_{\gamma_c^{[1]\mathcal{A}}} \lambda_\mathrm{N} \, \delta g_\mathrm{N} \; + \delta \lambda_\mathrm{N} \, g_\mathrm{N} \, \mathrm{d}a - \int_{\gamma_c^{[1]\mathcal{I}}} \frac{2}{c_\mathrm{N}} \lambda_\mathrm{N} \delta \lambda_\mathrm{N} \, \mathrm{d}a. \tag{2.58}$$

The reformulation of the structural part is comprehensively described in Section 2.1.4. The first line in (2.58) is identified as the variation of the structural potential $\delta \mathscr{U}$ (under the prerequisite of potential based loads) and the second line as the variation of the contact potential $\delta \mathscr{C}$ yielding $\delta \mathscr{L} = \delta \mathscr{U} - \delta \mathscr{C} = 0$, where $\mathscr{L}(\underline{u}, \lambda_\mathrm{N}) : \; \mathbb{R}^{d+1} \to \mathbb{R}$ with the spatial dimension $d \in \{2, 3\}$ denotes the so-called Lagrangian functional

$$\mathscr{L}(\underline{u}, \lambda_\mathrm{N}) = \mathscr{U}(\underline{u}) - \int_{\gamma_c^{[1]}} \lambda_\mathrm{N} \, [g_\mathrm{N}(\underline{u}) - \max\{0, g_\mathrm{N}(\underline{u}) - \frac{\lambda_\mathrm{N}}{c_\mathrm{N}}\}] \, \mathrm{d}a \tag{2.59a}$$
$$= \mathscr{U}(\underline{u}) - \int_{\gamma_c^{[1]}} \tilde{\mathscr{C}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}, \lambda_\mathrm{N}) \, \mathrm{d}a. \tag{2.59b}$$

Next, a brief discussion of the frictionless contact energy density term

$$\tilde{\mathscr{C}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}, \lambda_\mathrm{N}) = \lambda_\mathrm{N} \, [g_\mathrm{N}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}) - \max\{0, g_\mathrm{N}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}) - \frac{\lambda_\mathrm{N}}{c_\mathrm{N}}\}] \tag{2.60}$$



(a) (b)

Figure 2.3.: Visualization of the frictionless contact energy density function for $c_N = 1$. Figure 2.3b shows the first order derivative of $\tilde{\mathscr{C}}$ with respect to λ_N which reveals the C^0-smoothness of the standard variant.

follows. An interesting observation is that the contribution to the Lagrangian disappears at the solution, i.e. $\mathscr{L}^\star = \mathscr{U}^\star$. This follows directly from the complementarity condition in (2.56). Furthermore, by taking a look at an illustrative visualization for $c_N = 1$ in Figure 2.3, it is obvious that this function is only C^0-continuous.

However, there exists an alternative formulation which is C^1-continuous. Therefore, the frictionless contact energy density $\tilde{\mathscr{C}}$ is replaced by its augmented counterpart $\tilde{\mathscr{C}}_{c_N}$ compactly defined by

$$\tilde{\mathscr{C}}_{c_N}(\underline{x}^{[1]}, \underline{\bar{x}}^{[2]}, \lambda_N) = \frac{1}{2c_N}(\lambda_N^2 - [\max\{0, \lambda_N - c_N g_N\}]^2). \tag{2.61}$$

This function as well as the associated derivative with respect to the frictionless Lagrange multiplier are shown in Figure 2.4. It is C^1-continuous since the kink along the active-inactive decision can be now found in the first order derivative. A similar result could have been also obtained if the derivative with respect to the displacements was visualized, instead.

Remark 2.4. It is important to note that the first order derivative of the *augmented* frictionless contact energy density function $\tilde{\mathscr{C}}_{c_N}$ is equivalent to the so-called *non-linear complementarity function* (NCP) which is often mentioned in the context of primal-dual active set strategies [134, 257] and plays also a major role in case of the variational inequality approach applied to mortar contact problems such as followed by [87, 138, 213, 215, 218] and the references therein. Straight forward calculation reveals

$$\frac{\partial}{\partial\lambda_N}\tilde{\mathscr{C}}_{c_N}(\underline{x}^{[1]}, \underline{\bar{x}}^{[2]}, \lambda_N) = \frac{1}{c_N}[\lambda_N - \max\{0, \lambda_N - c_N g_N\}] \stackrel{!}{=} 0. \tag{2.62}$$

Thus, only the augmented formulation contains the NCP as a natural part while the standard Lagrangian functional has a kink and is therefore less smooth but has still the same solution.

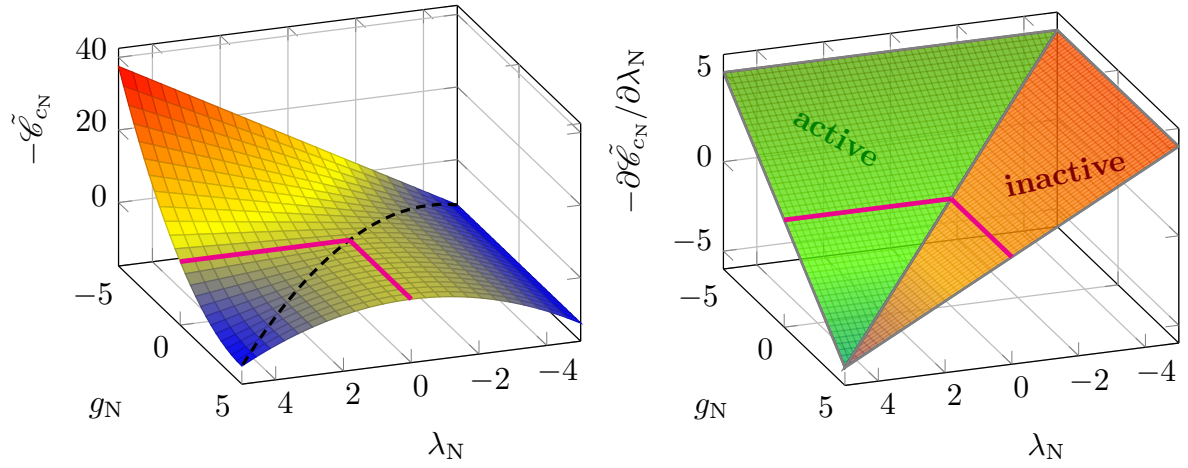


Figure 2.4.: Visualization of the augmented frictionless contact energy density. Figure 2.3b shows the first order derivative of $\tilde{\mathscr{C}}_{c_N}$ with respect to λ_N which indicates the C^1 smoothness of the augmented variant.

This underlines that the augmented formulation is to some extent the more consistent formulation, as it fits better the overall pre-asymptotic active set decision. However, in this thesis both approaches will be followed, i.e., the augmented as well as the standard contact Lagrangian are taken into account and will be compared to each other in Chapter 4.

For a much more comprehensive introduction into the general treatment of inequality constraints as well as for the derivation of the optimal slacks and the augmented formulation, the reader is kindly referred to Section 3.2 and the references therein.

2.2.3. Extension to Frictional Contact Problems

This section follows closely the publications of Alart and Curnier [1], Pietrzak and Curnier [212], wherein a general framework for the treatment of frictional contact problems is presented. This framework can be also applied to the formulation used in this thesis. Friction is a very important topic which increases the complexity severely. However, since the focus of this thesis is on frictionless contact problems, a detailed description would go far beyond the scope of this work and the topic shall therefore be only briefly addressed. The following investigations are given as basis for potential future work which eventually extends the later proposed methods to frictional contact.

First, one additional important kinematic quantity must be addressed: the *relative velocity* of a slave point $\underline{X}^{[1]} \in \Gamma_c^{[1]}$ with its current position $\underline{x}^{[1]} \in \gamma_c^{[1]}$ defined with respect to a master point $\underline{\bar{X}}^{[2]} \in \Gamma_c^{[2]}$ located at $\underline{\bar{x}}^{[2]}(\underline{\bar{X}}^{[2]}(\underline{X}^{[1]}, t), t) \in \gamma_c^{[2]}$ at time t. This quantity is typically given by

$$\underline{v}_{\mathrm{rel}} = \underline{\dot{x}}^{[1]}(\underline{X}, t) - \partial_t \underline{\bar{x}}^{[2]}(\underline{\bar{X}}^{[2]}(\underline{X}^{[1]}, t), t) = v_{\mathrm{N,rel}}\, \underline{n}^{[1]} + \underline{v}_{\tau,\mathrm{rel}}, \tag{2.63}$$

where $\partial_t \underline{\bar{x}}^{[2]}(\underline{\bar{X}}^{[2]}(\underline{X}^{[1]}, t), t)$ denotes the current velocity at a fixed material point $\underline{\bar{X}}^{[2]}$, i.e., at the material point associated to $\underline{X}^{[1]}$ via the applied projection at time t. The components in the tangential directions are defined by

$$\underline{v}_{\tau,\mathrm{rel}} = (\underline{\underline{I}} - \underline{n}^{[1]} \otimes \underline{n}^{[1]})\underline{v}_{\mathrm{rel}}. \tag{2.64}$$

As discussed in Laursen [170], Yang et al. [290] this quantity is only *frame indifferent* (also called *objective*) as long as $g_{\mathrm{N}} = 0$ holds at the location where $\underline{v}_{\tau,\mathrm{rel}}$ is evaluated. In other words: The relative velocity is not well-defined when the contacting bodies are not effectively in contact such as in case of a slight separation or penetration. However, it would be advantageous that the objectivity holds also in these cases, since it allows the formulation of the contact conditions as a tribological law instead of a boundary condition. This is beneficial when it comes to numerical solution procedures where a small local gap at an evaluation point, such as a Gauss point, might be unavoidable. From a continuum mechanical perspective an excellent explanation of the issue is provided by Curnier et al. [59], where the interested reader is referred to for more information. Possible modifications for mortar contact implementations can be found in Gitterle [109], Yang et al. [290], for instance. However, to what extent these modifications are also suitable for the later introduced mortar-like contact formulation is a question which needs further investigations. An alternative is the formulation in so-called *slip advected bases* which is often applied in the context of node-to-segment contact formulations considering a closest point projection, see Laursen [170], Pietrzak and Curnier [212] for further details. This approach is discussed in Curnier et al. [59] as well. Basically, the idea is to augment (2.64) by an additive term which vanishes in the case of perfect sliding, i.e. $g_{\mathrm{N}} = 0$.

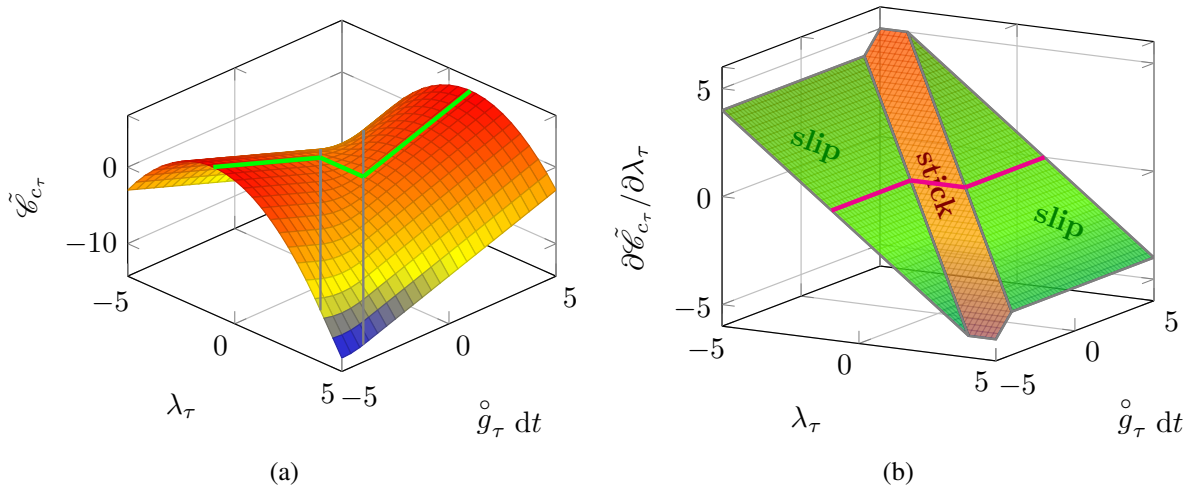
Objectivity plays a major role for large deformations since it guarantees that a change of reference frame, consisting of a translation and/or rotation, does not change the considered quantity. It is obvious that this prerequisite must be satisfied. In the following it shall be assumed that a suitable objective representation of the relative velocity is available (at least in its discrete form), which shall be denoted by $\overset{\circ}{\underline{g}}_{\tau}$ and is called the *tangential slip rate*. The quantity can either coincide with (2.64) and is only modified in the discrete form to become objective in case of slight penetrations or positive gaps following [109, 290] or it is modified in such a way that it becomes objective also in its continuum representation as suggested by [59, 170, 212]. Under consideration of this tangential slip rate and the *friction shear* $\underline{p}_{\tau}(\underline{X}^{[1]}, t)$, which implicitly relies on the normal pressure p_{N} introduced in (2.56), the tangential Coulomb friction law can be defined by three conditions:

- *slip rule*: The first condition claims that the friction shear $\underline{p}_{\tau}$ is collinear with the tangential relative velocity vector $\overset{\circ}{\underline{g}}_{\tau}$. This demand is formulated by

$$\| \overset{\circ}{\underline{g}}_{\tau}(\underline{X}^{[1]}, \bar{\underline{X}}^{[2]}, t)\| \, \underline{p}_{\tau}(\underline{X}^{[1]}, t) = \|\underline{p}_{\tau}(\underline{X}^{[1]}, t)\| \, \overset{\circ}{\underline{g}}_{\tau}(\underline{X}^{[1]}, \bar{\underline{X}}^{[2]}, t). \tag{2.65}$$

- *friction (Coulomb) criterion*: The second condition is used to limit the magnitude of the friction shear via

$$\|\underline{p}_{\tau}(\underline{X}^{[1]}, t)\| + \mathfrak{F}\, p_{\mathrm{N}}(\underline{X}^{[1]}, t) \leqslant 0, \tag{2.66}$$



(a) (b)

Figure 2.5.: Illustration of the C^1-continuous incremental frictional contact energy density function for $c_\tau = 1$ and $\hat{p}_\mathrm{N} = -1$. Presented is the dimensional reduced case, i.e., only one sliding direction is considered. In (2.5b) the first order derivative with respect to the tangential Lagrange multiplier is shown. This function is also known as the frictional nonlinear complementarity function.

where $p_\mathrm{N}(\underline{X}^{[1]}, t) \leqslant 0$ holds as defined in (2.56) and the scalar $\mathfrak{F} \geqslant 0$ is the so-called *friction coefficient*. The inequality in (2.66) can be illustrated as *Coulomb's cone*. A slice through this cone for a fixed value of the normal pressure p_N is a convex disk $C_\mathfrak{F}(p_\mathrm{N}) = \{\underline{p}_\tau \mid \|\underline{p}_\tau\| \leqslant -\mathfrak{F}\, p_\mathrm{N}\}$ where the disk radius is given by $-\mathfrak{F}\, p_\mathrm{N}$. The boundary of the disk defines the upper bound of the friction shear magnitude at a certain level of normal pressure.

- *complementarity condition*: There also exists a complementarity condition for the frictional conditions, just similar to the frictionless case. This condition states that either the magnitude of the friction shear is equal to $-\mathfrak{F}\, p_\mathrm{N}(\underline{X}^{[1]}, t)$ allowing a finite slip rate $\overset{\circ}{\underline{g}}_\tau$ *or* the magnitude of the friction shear is smaller than this upper bound and the slip rate is consequently equal to zero. This can be formulated as

$$\| \overset{\circ}{\underline{g}}_\tau\, (\underline{X}^{[1]}, \bar{\underline{X}}^{[2]}, t)\| \{ \|\underline{p}_\tau(\underline{X}^{[1]}, t)\| + \mathfrak{F}\, p_\mathrm{N}(\underline{X}^{[1]}, t) \} = 0. \tag{2.67}$$

These three conditions (2.65), (2.66), and (2.67) can be summarized in two possible friction states: The first one is called *stick* and is present as long as the ℓ_2-norm of the friction shear is smaller or equal to the current disk radius $-\mathfrak{F}\, p_\mathrm{N}$. It is characterized by a slip rate equal to zero. The second state is called *slip* which in turn is characterized by a slip rate unequal to zero and makes it simultaneously necessary that the magnitude of the friction shear is exactly equal to the disk radius $-\mathfrak{F}\, p_\mathrm{N}$.

The discussion already starts to reveal the cumbersome part of the Coulomb friction law in contrast to a frictionless formulation: The friction conditions depend explicitly on the normal pressure and this normal pressure is the Lagrange multiplier of the constrained optimization problem. This direct coupling makes it far more complicated to find a suitable Lagrangian. In

Pietrzak and Curnier [212], however, an elegant suggestion is made by a so-called *incremental augmented Lagrangian* formulation. This idea shall be briefly presented in the following: The first part is represented by the augmented Lagrangian formulation of the frictionless contact formulation (2.61), while the second part follows from the aforementioned friction conditions. The finally proposed augmented *frictional contact Lagrangian* is defined as

$$\mathscr{L}_{c_\tau}[\underline{u}^{[b]}, \lambda_{\mathrm{N}}, \underline{\lambda}_\tau; \hat{p}_{\mathrm{N}}] = \sum_{b=1}^{2} \mathscr{U}^{[b]} + \int_{\gamma_{\mathrm{c}}^{[1]}} \frac{1}{2c_{\mathrm{N}}} \{[\max\{0, \lambda_{\mathrm{N}} - c_{\mathrm{N}} g_{\mathrm{N}}\}]^2 - \lambda_{\mathrm{N}}^2\} \mathrm{d}a \tag{2.68a}$$

$$+ \int_{\gamma_{\mathrm{c}}^{[1]}} \langle \underline{\lambda}_\tau, \overset{\circ}{\underline{g}}_\tau \, \mathrm{d}t \rangle + \frac{c_\tau}{2} \| \overset{\circ}{\underline{g}}_\tau \, \mathrm{d}t \|^2 \mathrm{d}a \tag{2.68b}$$

$$- \int_{\gamma_{\mathrm{c}}^{[1]}} \frac{1}{2c_\tau} [\max\{0, \| \hat{\underline{\lambda}}_\tau \| + \mathfrak{F} \hat{p}_{\mathrm{N}}\}]^2 \mathrm{d}a, \tag{2.68c}$$

where $\mathrm{d}t$ denotes an infinitesimal time increment and, consequently, $\overset{\circ}{\underline{g}}_\tau \, \mathrm{d}t$ is an infinitesimal *slip increment*. Furthermore, the abbreviation

$$\hat{\underline{\lambda}}_\tau = \underline{\lambda}_\tau + c_\tau \, \overset{\circ}{\underline{g}}_\tau \, \mathrm{d}t \tag{2.69}$$

is introduced. This shortcut is often called *augmented tangential Lagrange multiplier*. The augmented contact pressure $\hat{p}_{\mathrm{N}}$ is also introduced. Its value is defined by $\hat{p}_{\mathrm{N}} = p_{\mathrm{N}} + c_{\mathrm{N}} g_{\mathrm{N}}$. Alternatively, it can be also identified by $\hat{p}_{\mathrm{N}} = -(\lambda_{\mathrm{N}} - c_{\mathrm{N}} g_{\mathrm{N}})$ under consideration of the relationship between pressure and frictionless Lagrange multiplier. However, the important point is that $\hat{p}_{\mathrm{N}}$ enters (2.68) as an independent variable which is kept constant during the variation and linearization, see [212].

Furthermore, just similar to the frictionless case, it is possible to extract a frictional contact energy density function from (2.68) given by

$$\tilde{\mathscr{C}}_{c_\tau}(\underline{u}^{[b]}, \lambda_{\mathrm{N}}, \underline{\lambda}_\tau; \hat{p}_{\mathrm{N}}) = \langle \underline{\lambda}_\tau, \overset{\circ}{\underline{g}}_\tau \, \mathrm{d}t \rangle + \frac{c_\tau}{2} \| \overset{\circ}{\underline{g}}_\tau \, \mathrm{d}t \|^2 - \frac{1}{2c_\tau} [\max\{0, \| \hat{\underline{\lambda}}_\tau \| + \mathfrak{F} \hat{p}_{\mathrm{N}}\}]^2. \tag{2.70}$$

A demonstrative illustration is presented in Figure 2.5. By taking a closer look numerous facts become immediately aware: The contributions of the friction conditions to $\mathscr{L}_{c_\tau}^\star$ at the solution (denoted by $(\cdot)^\star$) vanish only in case of stick, i.e. $\tilde{\mathscr{C}}_{c_\tau,\mathrm{stick}}^\star = 0$. However, in the case of slip the frictional contribution to the incremental augmented Lagrangian value is independent of the regularization parameter c_τ and, consequently, c_τ is really only a regularization parameter which is important for the pre-asymptotic stick/slip decision. Similar to c_{N} in case of frictionless contact, c_τ has no direct influence on the solution. Finally, the remaining frictional contribution to the total value of the incremental augmented Lagrangian at the solution is given by

$$\tilde{\mathscr{C}}_{c_\tau,\mathrm{slip}}^\star = |p_{\mathrm{N}}^\star| \, \| \overset{\circ}{\underline{g}}_\tau^\star \, \mathrm{d}t \|. \tag{2.71}$$

Another interesting point is that the first order derivative of $\tilde{\mathscr{C}}_{c_\tau}$ with respect to $\underline{\lambda}_\tau$ reveals once more the associated frictional *nonlinear complementarity function* as shown in Figure 2.5b. For further information on the frictional NCP the reader is kindly referred to Farah [87], Gitterle [109], Hüeber [138], Seitz et al. [243], Sitzmann [252] and the references therein. In addition, it shall be briefly mentioned that the formulation as NCP can be also found in other research fields. One example is plasticity which has much in common with Coulomb friction [124, 242]. At this point the discussion of the frictional contact problem shall be stopped. However, the presented incremental augmented Lagrangian formulation (2.68) can be considered as a starting point for a possible augmentation of the up-coming modifications from the frictionless case to frictional problems. Furthermore, the interested reader is referred to Pietrzak and Curnier [212] for further details concerning the variation and linearization of the frictional approach presented here.

2.3. Spatial Discretization

The discretization method for the spatial domain is another important topic. There exists a huge variety of applicable methods for this task, just similar to the later discussed numerical time integration. In the early days the *finite difference method* was quite popular as described in Collatz [51], Samarskii [233], even though it had initially shown a number of unfavorable drawbacks such as the restriction to regular meshes. However, this could be resolved in later publications by Jensen [148], Liszka and Orkisz [177]. Nevertheless, in structural dynamics the finite difference method has been replaced almost entirely by the now very popular *finite element method* [206, 297] due to many advantages of the latter one. Examples are its universality and strong mathematical convergence properties [35]. Nowadays, there can be found numerous well-written text books on this topic, e.g., Bathe [12], Belytschko et al. [18], Hughes [140]. The finite element method will be also the method of choice throughout this thesis. However, before the discussion goes deeper into this, it shall be mentioned that there exists another discretization class which has gained increasing attention during the last couple of decades. That is the class of *mesh-free methods*. While its popularity in fluid dynamics starts to rise more and more [179], there are also interesting publications on mesh-free methods in the field of structural dynamics. On the one hand, there are the so-called *discrete element methods* (DEM) which mainly deal with the interaction between freely moving particles such as in granular flow [121] or, as a more specific example, in the application during powder bed metal additive manufacturing [193]. But, there are many more applications of DEM such as fracture mechanics [200], where it is used in combination with FEM. Additionally, the class of meshfree methods which are more similar to the classical finite elements shall be briefly addressed: Examples are *(generalized) moving least squares* or *element free Galerkin methods*. For an introduction into this topic the interested reader is referred to review publications such as [17, 102]. Some of the therein discussed methods have very promising properties but make it difficult to impose Dirichlet boundary conditions (DBC) [88] or introduce a new issue with respect to the numerical integration as exemplarily discussed in [64, 72]. More recently, however, the so-called *local maximum entropy method* has been proposed by Arroyo and Ortiz [7], Sukumar and Wright [256], which resolves at least the DBC problem by its convex hull property. Its convergence properties [30] as well as its smooth transition to the finite element method [7] makes it quite appealing for structural problems and thus it is not surprising that the maximum entropy method has been applied successfully to the

thin shell analysis [194, 195]. In the author's opinion this new type of methods are well-suited for certain research fields which naturally come along with large structural deformations such as cell migration in the biomedical sector [225].

2.3.1. Finite Element Discretization

The fundamental idea of the finite element method or even any discretization method is to reduce the complexity of the original problem by limiting the number of degrees of freedom. In case of the finite element method, this is achieved by subdividing the continuous body into a finite number of small elements $\mathrm{N_e}$, where each element occupies a closed sub-domain of the entire structural body $\Omega_0^{(e)} \subseteq \Omega_{0,\mathrm{h}}$. Now, the disjoint union of all element domains denotes the domain estimate of the respective body. Under the assumption of a regular mesh refinement as well as certain smoothness assumptions with respect to the solution, it can be expected that the approximation error decreases with an increasing number of finite elements. This is called *h*-convergence. However, the definition of the discrete degrees of freedom is still missing. Therefore, the attention is drawn to one arbitrary element $e \in \mathcal{E} = \{1, \dots, \mathrm{N_e}\}$. Such an element contains a number of defining nodes (or control points) $\mathrm{N_n}$ and at these nodes a certain number of degrees of freedom (DOF) can be specified. For instance, in case of a 3-dimensional discretized pure structural problem these degrees of freedom are given by the three coordinates of a displacement vector associated to each of the element nodes. Having said this, the value of any arbitrary nodal quantity inside an element e can be computed by interpolating between its nodal values. A simple example would be the current position vector $\underline{x}(\underline{X}, t)$ at $(\underline{X}, t) \in \Omega_0 \times [0, T]$ yielding

$$\underline{x}(\xi^{(e)k}) = N^{(e)}{}_j \, \underline{x}^{(e)j} = N^{(e)}{}_j \, (\underline{X}^{(e)j} + \underline{d}^{(e)j}), \qquad \forall j \in \{1, \dots, \mathrm{N_n^{(e)}}\}, \tag{2.72}$$

where $\mathrm{N_n^{(e)}}$ is the number of nodes per element and $\{N^{(e)}{}_j\}$ is the set of shape functions, where each shape functon is associated to one node j of element e. In addition, $\{\underline{X}^j\}$ is the set of the nodal reference position vectors. The finally remaining unknown quantities are summarized in the set of nodal displacement vectors $\{\underline{d}^j\}$. One reason for the great success of the finite element method is that this element-wise view can be utilized to design computer codes which are suitable for a large variety of different problems. This is achieved by splitting the global problem into smaller parts related to each element. Subsequently, a assembly strategy is applied which brings the distinct parts back together by simple summation. Therefore, a special numbering scheme is invoked which is known as the *global-to-local* index [35]. This indexing scheme given by $i(e, j)$ is used to assign a local node number j of a specific element e to its global position index i in the considered system. Therefore, the finite element method offers a very systematic way to solve complex problems. Another beneficial effect is the locality of the used shape functions which lead to a so-called *sparsity pattern* of the system matrices. These *sparse matrices* need severely less memory than their dense counterparts and, therefore, the related system of equations can be efficiently solved, e.g., under consideration of iterative linear solvers such as the GMRES based methods [153, 232].

2.3.2. Important Requirements for Convergence

There is a number of very important requirements with respect to the finite element formulation in the context of structural problems. Without these requirements monotonic convergence can not be expected, i.e., the accuracy of the solution might not improve by successive mesh refinement. The first requirement asks for *completeness* of the considered function space. This means that the used shape functions of the elements must be able to represent all rigid body modes and constant strain states [12, Sec. 4.3.2]. Rigid body modes are deformation modes which induce no strain (or stress). For example a pure translation or rotation is not allowed to cause any stresses in a finite element. An interesting side note is that the number of straining or natural modes of a element can be easily obtained by subtracting the number of rigid body modes from the number of degrees of freedom (DOF) of the respective element. Thus, a 2-dimensional quadrilateral element with four nodes (and, therefore, linear shape functions) has $2 \cdot 4 = 8$ DOF and there exist three rigid body modes: Translation in x- and y-direction as well as rotation around the outward pointing normal. Therefore, it can represent $8 - 3 = 5$ natural modes. This result can also be obtained by solving the eigenvalue problem of the element stiffness matrix. Three of the eigenvalues will be zero and the associated eigenvectors are the rigid body modes. The remaining eigenvalues are unequal zero and represent the stiffness for the corresponding mode given by the eigenvector [12]. These straining modes will become also important with respect to the undesired phenomenon of *locking* which is discussed in Section 2.3.3.

The second requirement asks for *compatibility*. In other words, the displacements in the interior of an element and across its boundaries is supposed to be continuous, i.e., non-physical gaps inside the domain shall be avoided. For elements with displacement DOF only, this can be achieved by claiming a continuity of the nodal displacement DOF. However, in case of reduced formulations such as shells or beams embedded into a higher dimensional space, rotational DOF are often additionally introduced. In such a case the enforcement of compatibility with respect to the kinematic assumptions can become much more involved (see for example Bischoff [27], Gee [104]).

These are two important requirements concerning the finite element function space. However, there are even more criteria which must be considered. For example, the *stability* of the discrete solution must be maintained. This stability issue plays a role in different areas: On the one hand, stability problems can arise in form of so-called *zero-energy modes*. Zero-energy modes are related to further zero eigenvalues of the element stiffness matrix and are consequently mathematically spoken equivalent to a rank deficiency of this matrix, i.e., there are more zero eigenvalues than rigid body modes. Different types of zero energy modes exist: Some of them are *non-communicable* and it is therefore less likely that they occur in a mesh consisting of multiple elements. However, there are also so-called *communicable* zero-energy modes which can become a severe problem since they are not blocked by their surrounding neighbors. These zero-energy modes can lead to a locally singular system matrix and might prevent convergence. It is a phenomenon which sometimes occurs during the application of so-called *selective reduced integration* which has been firstly proposed by Zienkiewicz et al. [298] to avoid certain kinds of locking. An excellent overview of the zero-energy problem during reduced integration can be found in Koschnick [164]. However, selective reduced integration is not the only instance where zero-energy modes can occur. It can also occur in the field of finite elasticity when a consistent numerical integration scheme is considered but certain *enhanced assumed strain* (EAS) formu-

lations are added. This has been detected and discussed in de Souza Neto et al. [67], Wriggers and Reese [285]. Since the latter case can be of crucial importance for the application of EAS during contact simulations, it will be addressed again in Section 2.3.4.

Another origin of instabilities is given by the inproper choice of the Lagrange multiplier function spaces. The correct choice of these function spaces becomes not only important with respect to the contact algorithms discussed here but plays also an important role in conjunction with the modeling of totally incompressible conditions, i.e., for materials where Poisson's ratio tends to 0.5. In such a case, the pressure becomes an independent variable leading to a saddle-point problem, thus, the pressure can be interpreted as a Lagrange multiplier. Other examples are given by multi-field functionals such as the three-field Hu-Waishizu functional or the two-field Hellinger Reissner principle, see Koschnick [164] for more information. Furthermore, also the pressure degree of freedom in case of fluid simulations is of this kind. Stability for those mixed formulation is obtained, if the *ellipticity* as well as the *inf-sup* conditions are satisfied [9, 36].

The attention is now on the treatment of structural contact problems with the finite element method and the inherent Lagrange multiplier as introduced in Section 2.2.2. The so-called *ellipticity condition* is independent of the contact conditions and is fulfilled as soon as appropriate finite elements and boundary conditions are applied, see El-Abbasi and Bathe [76]. The second important condition is the *inf-sup condition*, which is a topic on its own. A nice introduction for contact problems can be found in Bathe and Brezzi [13]. Furthermore, a comprehensive introduction into the field of (dual) mortar contact methods is given in Wohlmuth [279]. Under reconsideration of the weighting function space (2.30) which is chosen as a subspace of the solution space defined by

$$\underline{u} \in \mathcal{U} = \{\underline{w} \in \mathcal{H}^1(\Omega_0) : \ \underline{w} = \breve{\underline{u}} \text{ on } \Gamma_u\}, \tag{2.73}$$

the function space for the Lagrange multiplier λ_N follows as

$$\lambda_\mathrm{N} \in \mathcal{M}_+ = \{w_\mathrm{N} \in H^{-1/2}(\gamma_\mathrm{c}^{[1]}) : w_\mathrm{N} \geqslant 0\}. \tag{2.74}$$

Note that the function space $\mathcal{M}_+$ does not implicitly depend on the Lagrange multiplier value since only frictionless problems shall be considered. A very brief introduction into the notation of functional analysis becomes necessary to clarify (2.74): $H^{1/2}(\gamma_\mathrm{c}^{[1]})$ is known as the trace space of $\mathcal{W}$ and $H^{-1/2}(\gamma_\mathrm{c}^{[1]})$ denotes the dual space of $H^{1/2}(\gamma_\mathrm{c}^{[1]})$. For more detailed information the reader is kindly referred to Brezzi and Fortin [37, Ch. 3]. Throughout this thesis, the discrete Lagrange multiplier shape functions are chosen equal to the discrete solution space functions in the interior domain by restricting them to the contact boundary. This choice is inf-sup stable as stated in Bathe and Brezzi [13, Theorem 3] or numerically shown in El-Abbasi and Bathe [76]. However, there are choices which are not inf-sup stable, e.g., a linear interpolation of the solution but a constant interpolation of the Lagrange multiplier field would not be inf-sup stable. Possible remedies are suggested in Wohlmuth [279, Remark 3.2]. Another interesting fact is that the inf-sup stability result for the contact mortar methods discussed in [279] is independent of the degrees of freedom on the master side. Consequently, the inf-sup result is completely independent of the ratio of the element sizes on the slave and master sides which indicates that the non-matching character of the meshes does not affect the stability.

2.3.3. Locking of Structural Elements

In the previous section the most important requirements have been discussed which are necessary to maintain convergence of the finite element solution for consecutive mesh refinement. Even though, a detailed analysis of the expected convergence speed shall not be topic of this thesis (see e.g. Bathe [12], Brenner and Scott [35] for a detailed description), there are certain influences which can vastly decrease the convergence rate and thus badly affect the expected solution quality. Nowadays, there is a high interest in understanding these effects and developing remedies. The first important work in this direction has been given by Babuška and Suri [10]. First as a technical report and later published in the cited journal.

The convergence rate can be affected by many different parameters. However, in this section the discussion is restricted to the so-called *locking effect* of finite element formulations. There is not really one unique definition for locking, but all of them have in common that they can be characterized in dependence on one critical parameter. With respect to this critical parameter, locking can be further subdivided into artificial stiffening effects based on material or geometrical parameters. However, in both cases the main reason for the bad performance is often given by the fact that the used shape function space is not sufficiently well-suited to represent certain element deflections which would be necessary to represent the deformation without introducing unwanted strains and stresses. For a really comprehensive overview and discussion the reader is kindly referred to Koschnick [164]. Some locking effects are mainly related to structural finite element formulations, i.e., special elements which use some kind of kinematic assumptions to reduce their spatial dimensions. Examples are beam, shell and plate elements. The focus of this thesis is clearly on classical continua elements. The main locking effects are here:

- *Shear locking:* Shear locking affects 2-dimensional as well as 3-dimensional continua elements and has a severe impact for shape functions based on low order polynomials. The classical example are four-node quadrilateral elements (QUAD4) and eight-node hexahedral elements (HEX8) for 2- and 3-dimensional problems, respectively. The influence of the parasitic linear shear stresses becomes obvious in case of bending deformations. However, the critical parameter is the aspect ratio of the element edges, e.g., the ratio of element width to height in case of QUAD4 elements, for instance.

- *Trapezoidal locking:* This is a locking phenomenon which can be observed for curved structures discretized by continua elements. Note that the definition for small deformations is restricted to initially curved structures to avoid a mix-up with the influences related to mesh distortion. However, in case of large deformations this effect can also become apparent for initially straight structures which become curved through the applied deformation. Once more a bending deformation shall be assumed. A trapezoidal shaped QUAD4 element would imply a number of parasitic strains in dependence on the degree of curvature. First, it would again show a linear shear strain which vanishes at the element center, but also the expected strain in parametric ξ_1-direction would be slightly affected. Furthermore, a quadratic artificial strain in the parametric ξ_2- direction would occur as well which vanishes only at the element boundaries for $\xi_1 = \pm 1$. For a fixed curvature the element aspect ratio can be identified as the critical parameter.

- *Volumetric locking:* While the first two locking phenomena are purely influenced by geometrical parameters, the volume locking is one representative of a parasitic effect due to a material parameter. The critical parameter is the Poisson's ratio ν. The effect disappears for $\nu = 0$ and will be the worst for $\nu \to 0.5$, i.e., for the limit case of incompressibility. The underlying constraint can be formulated as $\mathrm{div}(\underline{u}) = 0$. This effect is apparent in 2-dimensional and 3-dimensional solid elements and can be explained by the inability of the simple elements to represent this constraint point-wise. The result is a stiffening effect due to very high artificial normal stresses.

More information can be found in [164, Chapter 5]. Possible remedies for the discussed unwanted effects follow in the next Section 2.3.4.

2.3.4. Enhanced Assumed Strain Formulation

The main focus of this section will lie on the *enhanced assumed strain* (EAS) formulation since it will also be considered throughout the later discussed globalization techniques for large deformation contact problems in Chapter 6. However, this is not the only possibility. As already mentioned there is for instance the selective reduced integration [298] and even though it seems unlikely it is indeed possible to show equivalence to mixed finite element methods for this simple idea as demonstrated in Malkus and Hughes [185]. Another well-suited anti-locking technique is the *assumed natural strain* (ANS) method which can also be derived from variational principles as described in Simo and Hughes [248]. The basic idea is to specify so-called collocation or sampling points in the element which are then used to interpolate and evaluate the correct strains for critical deformation states [164]. A great example for large deformations is given in Vu-Quoc and Tan [269]. Therein, a big advantage of the ANS formulation compared to EAS is given as well: That is the much better handling of transverse shear strains. In Vu-Quoc and Tan [269] is shown that even a 30 parameter EAS formulation as proposed by Klinkel and Wagner [158] is not able to fulfill the out-of-plane bending patch test without an additional ANS modification. The drawback of the ANS methods is that the location, the number as well as suitable interpolation functions must be newly defined for each considered element and stiffening effect [164]. This drawback is supposed to be resolved by the proposed *discrete strain gap* (DSG) method which is comprehensively described and discussed in Koschnick [164]. For small deformations the ANS as well as the DSG method can be summarized in the so-called *B-bar* methods as proposed by [248], while for large deformations rather the deformation gradient instead of the B-operator has to be considered. A drawback of these B-bar methods, i.e. ANS as well as DSG, is the fact that they are not capable of avoiding volumetric locking. However, it must be noted that there exist more methods than the EAS method to avoid volumetric locking completely, even in presence of large deformations. For instance, one method has been proposed by De Souza Neto et al. [66] and uses a multiplicative split of the deformation gradient into a deviatoric and volumetric part. The volumetric part is then only evaluated at the center point of the element and subsequently, the two parts are recombined resulting in the F-bar deformation gradient. The drawback is that this method helps only against volumetric but not against shear locking or any other geometrical locking effect.

This brings us finally to the EAS formulation which is able to avoid shear and volumetric locking, but fails to avoid transversal shear-locking (see e.g. solid shell elements [269]). The up-

coming brief description of the EAS method for large deformations follows Simo and Armero [247], Simo and Rifai [249] and Wall et al. [273]. Let us start with the variational view on the three-field Hu-Washizu functional which yields

$$\mathscr{U}(\underline{u}, \underline{\underline{E}}, \underline{\underline{S}}) = \int_{\Omega_0} \Psi(\underline{\underline{E}}) \mathrm{d}V_0 + \int_{\Omega_0} \underline{\underline{S}} : (\frac{1}{2}(\underline{\underline{F}}(\underline{u})^{\mathrm{T}} \underline{\underline{F}}(\underline{u}) - \underline{\underline{I}}) - \underline{\underline{E}}) \mathrm{d}V_0 - \int_{\Omega_0} \check{\underline{b}}_0 \cdot \underline{u} \mathrm{d}V_0 - \int_{\Gamma_\sigma} \check{\underline{t}} \cdot \underline{u} \, \mathrm{d}A_0, \tag{2.75}$$

where three different unknowns are specified: the displacement vector $\underline{u}$, the Green–Lagrange strain tensor $\underline{\underline{E}}$ and the second Piola-Kirchhoff stress tensor $\underline{\underline{S}}$. Next, the EAS method introduces a reparametrization of the free Green Lagrange strain tensor by an additive split

$$\underline{\underline{E}} = \underline{\underline{E}}^{\mathrm{u}} + \tilde{\underline{\underline{E}}}, \qquad \text{with } \underline{\underline{E}}^{\mathrm{u}} = \frac{1}{2}(\underline{\underline{F}}(\underline{u})^{\mathrm{T}} \underline{\underline{F}}(\underline{u}) - \underline{\underline{I}}). \tag{2.76}$$

Now, $\tilde{\underline{\underline{E}}}$ denotes the additive enhancement to the deformation dependent strain part $\underline{\underline{E}}^{\mathrm{u}}$. This inserted into (2.75) results in

$$\mathscr{U}(\underline{u}, \underline{\underline{E}}, \underline{\underline{S}}) = \int_{\Omega_0} \Psi(\underline{\underline{E}}^{\mathrm{u}} + \tilde{\underline{\underline{E}}}) \mathrm{d}V_0 + \int_{\Omega_0} \underline{\underline{S}} : \tilde{\underline{\underline{E}}} \mathrm{d}V_0 - \int_{\Omega_0} \check{\underline{b}}_0 \cdot \underline{u} \mathrm{d}V_0 - \int_{\Gamma_\sigma} \check{\underline{t}} \cdot \underline{u} \, \mathrm{d}A_0. \tag{2.77}$$

This functional will be discretized by a suitable interpolation scheme which is given by

$$\underline{x} = \underline{\underline{N}} \, (\underline{X} + \underline{d}), \qquad \tilde{\underline{\underline{E}}}^{\mathrm{h}} = \mathcal{Q}_1 \, \underline{\underline{\alpha}}_{\mathrm{eas}}, \qquad \underline{\underline{S}}^{\mathrm{h}} = \mathcal{Q}_2 \, \underline{\underline{\beta}}_{\mathrm{eas}}, \tag{2.78}$$

where $\underline{\underline{N}}$, $\mathcal{Q}_1$ and $\mathcal{Q}_2$ denote appropriate mapping operators based on the respective shape functions. Furthermore, $\underline{\underline{\alpha}}_{\mathrm{eas}}$ and $\underline{\underline{\beta}}_{\mathrm{eas}}$ represent the related unknown coefficients of the strains and stresses, while $\underline{d}$ denotes the unknown nodal displacements. Please note that it is tried to minimize the explicit marking of individual variables as discretized by the superscript h. Therefore, it is only introduced where it actively supports the understanding. Next, Simo and Rifai [249] suggest to choose the interpolation of the stresses and strains orthogonal to each other such that

$$\int_{\Omega_0} \underline{\underline{S}}^{\mathrm{h}} : \tilde{\underline{\underline{E}}}^{\mathrm{h}} \mathrm{d}V_0 = 0 \tag{2.79}$$

holds. Another prerequisite is that the enhanced strains are not linearly dependent of the compatible strains to avoid any unwanted rank deficiency of the related stiffness matrices. By inserting (2.79) into (2.77), the new discretized functional follows as

$$\mathscr{U}(\underline{d}, \underline{\underline{\alpha}}_{\mathrm{eas}}) = \int_{\Omega_0} \Psi[\underline{\underline{E}}^{\mathrm{u}}(\underline{d}) + \tilde{\underline{\underline{E}}}^{\mathrm{h}}(\underline{\underline{\alpha}}_{\mathrm{eas}})] \, \mathrm{d}V_0 - \underline{d}^{\mathrm{T}} \underline{f}_{\mathrm{ext}}, \tag{2.80}$$

where $\underline{f}_{\text{ext}}$ is the nodal external force vector. Next, the variation and linearization of this function with respect to the displacements and the EAS strain parameters follow. Details can be found for instance in Wall et al. [273]. The important part for this thesis are the contributions of each element to the global system of equations and the global right hand side vector which look like

$$\begin{pmatrix} \underline{\underline{K}}^{(e)} & \underline{\underline{\tilde{L}}}^{(e)}_{\text{eas}} \\ [\underline{\underline{\tilde{L}}}^{(e)}_{\text{eas}}]^{\text{T}} & \underline{\underline{\tilde{D}}}^{(e)}_{\text{eas}} \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} \underline{f}^{(e)}_{\text{ext}} - \underline{f}^{(e)}_{\text{int}} \\ -\underline{\tilde{r}}^{(e)}_{\text{eas}} \end{pmatrix} \tag{2.81}$$

for each element e. See again [273] for the detailed definition of the matrices and right hand side contributions. Finally, since the enhanced strains do not need to be compatible over the element boundaries, it is possible to condensate the additional degrees of freedom on element level, i.e., before the global system matrix is assembled, resulting in the already condensed contributions

$$\underline{\underline{K}}^{(e)} - \underline{\underline{\tilde{L}}}^{(e)}_{\text{eas}} \, [\underline{\underline{\tilde{D}}}^{(e)}_{\text{eas}}]^{-1} \, [\underline{\underline{\tilde{L}}}^{(e)}_{\text{eas}}]^{\text{T}}, \qquad \text{and} \qquad \underline{f}^{(e)}_{\text{ext}} - \underline{f}^{(e)}_{\text{int}} + \underline{\underline{\tilde{L}}}^{(e)}_{\text{eas}} \, [\underline{\underline{\tilde{D}}}^{(e)}_{\text{eas}}]^{-1} \, \underline{\tilde{r}}^{(e)}_{\text{eas}}. \tag{2.82}$$

The enhanced strain increment can be computed in a post-processing step for each element:

$$\Delta \underline{\tilde{\alpha}}^{(e)}_{\text{eas}} = -[\underline{\underline{\tilde{D}}}^{(e)}_{\text{eas}}]^{-1} \, \{\underline{\tilde{r}}^{(e)}_{\text{eas}} + [\underline{\underline{\tilde{L}}}^{(e)}_{\text{eas}}]^{\text{T}}] \Delta \underline{d}^{(e)}\}, \tag{2.83}$$

where $\Delta \underline{d}^{(e)}$ is the part of the global displacement increment related to the DOF of element e and $\Delta \underline{\tilde{\alpha}}^{(e)}_{\text{eas}}$ the internal enhanced strain increment in Voigt notation. The changed notation is indicated by the tilde superscript. Therefore, compared to other methods such as ANS and DSG, the EAS method includes an additional effort in form of evaluating these extra matrices, the condensation and the necessary inversion of the matrix $\underline{\underline{\tilde{D}}}^{(e)}_{\text{eas}}$. The dimension of this matrix depends on the number of used enhanced strain modes per element. In this thesis 21 additional modes for linear hexahedron elements will be considered, thus, $\underline{\underline{\tilde{D}}}^{(e)}_{\text{eas}} \in \mathbb{R}^{21 \times 21}$. Nevertheless, the EAS formulation still shares the advantage that the size of the global system of equations remains unchanged and that the evaluation of the internal energy contributions stays also unchanged from a global point of view. This makes the application of the globalization techniques later introduced in Section 6.8.2 much easier.

Remark 2.5. The section is closed by an important remark concerning the EAS formulation. The enhanced assumed strain formulation is *not* completely stable for large deformations. This is something which is already known for quite some time since it had been demonstrated by de Souza Neto et al. [67] and Wriggers and Reese [285]. However, most of the EAS literature is written with respect to small deformations and therein this phenomenon is hardly mentioned since it does not appear. This leads to the situation that not everyone is aware of this problem. Responsible for the instabilities are the additional modes addressing shear locking and, as reported in Wall et al. [273], these shear modes can lead to hour-glassing (i.e. zero energy modes). However, in contrast to the selective reduced integration method where these effects are somewhat expected due to the insufficient integration, they will occur also for an analytical exact integration. The instability will occur for large compression strains and also for large tensile strains. In combination with contact the compression is more likely, consequently, it is not surprising that

the hour-glassing might become apparent during large deformation contact simulation considering EAS. For the simple 2-dimensional example discussed in [273] the critical state is reached for a height reduction of the QUAD4 element by a factor of $1/\sqrt{3}$. It is also stated that the instability can be alternatively detected by an eigenvalue analysis of the element stiffness matrix (see also the analytical derivations in [285]). This entire EAS instability problem is mainly mentioned in this thesis, since it probably explains some of the necessary matrix modifications in Section 6.10. The there used algorithm will naturally counteract any negative or zero eigenvalues in the global system matrix. However, in the author's opinion it might be a better idea to tackle the problem at its origin by the suggested stabilization approach in Wall et al. [273]. Therefore, it is recommended to implement the stabilization in the future.

2.4. Discrete Time Integration

The discrete treatment of time integrals and time dependency is a topic on its own. Therefore, the discussion in this thesis is restricted to only one representative which is quite popular in structural dynamics. The so-called *Generalized-α* method, firstly introduced by Chung and Hulbert [48]. One of the big advantages of this method is that it allows a controlled numerical dissipation which acts in such a way that mainly contributions stemming from high-frequencies in the structural energy are damped while the more important low frequency response stays almost unbiased. This is an important property for contact problems, where the discrete resolution of the impact from one body into the other leads almost always to a non-continuous scenario coming along with some artificial high-frequency solution artifacts. However, it is by far not the only possible time integration scheme. Actually, there exists a huge variety of different methods. For instance, the for the spatial discretization preferred finite element method (FEM) can also be applied to the time domain resulting in so-called *space-time finite element methods*. These methods have also promising properties. The interested reader is referred to Hughes and Hulbert [142], Hulbert [144] for an early example in the field of elastodynamics. However, also the application to other research fields is possible such as the fluid-structure interaction, see also Hübner et al. [137] and the references therein.

Before the attention is drawn to a detailed description of the Generalized-α method considered here, the very basic idea of almost any time integration scheme shall be explained. First of all, the goal is to find a numerical scheme which is able to reproduce a transient solution path while maintaining important properties of the solution such as its magnitude or its phase angle. Think for example of a harmonic oscillation. To achieve this it becomes necessary to split the desired time interval $[0, T]$ into discrete smaller sections. The size of these sections is given by the user-defined time step Δt. Now, concerning this time step two things should not be mixed up:

- Firstly, a too large chosen time step size can lead, under certain circumstances, to a blow-up of the simulation due to inherent *instabilities* of the discrete time integration scheme. One well-known example is the explicit forward-Euler scheme. For example, this simple numerical scheme cannot be applied to a harmonic oscillation since it will always lead to an increasing amplitude even for an extremely small time step size. However, there are ways to avoid these problems, e.g., by considering appropriate implicit time integration schemes. But, nothing comes for free and thus implicit schemes have the drawback that they ask for a much higher

computational effort. Nevertheless, if large time steps are sufficient to accurately represent the solution, the implementation and computational effort might pay off by in the end much smaller total simulation times (even though one single time step might take much longer compared to an explicit scheme).

- This brings us to the second distinct point: The *solution accuracy*. While implicit schemes might allow an artificially large time step, the solution quality is not necessarily superior compared to an explicit scheme and consequently the time step shows a second upper bound which relies on the demanded solution quality. In fact, the error order depends heavily on the definition of the different schemes, e.g., it plays a decisive role which information is used for the construction: Only the solution of the previous time step at t_n, e.g. explicit Euler, the information of the current time step t_{n+1} leading to implicit schemes, or shall also information from previous steps be taken into account $\{t_{n-1}, t_{n-2}, \dots\}$ such as for BDF, Adams–Bashforth or Adams–Moulton methods. Furthermore, it is also possible to use information from artificial points placed at locations in time between t_n and t_{n+1}, see the Runge–Kutta schemes. As a rule of thumb one can say that with rising complexity by considering more and more solution points in time also the solution accuracy can be expected to be improved. But again: A high error-order, i.e. small errors, does not automatically mean that the chosen solution step is free of instabilities. These are two very different issues.

Now, with the previously said in mind, the application field for implicit methods are for example stiff differential equations. One might think of a system which contains a high-frequency response which naturally fades out very quickly and a second overlaid response which is of a low-frequency harmonic character. In this case an implicit scheme in conjunction with an adaptive time step can be much faster than an explicit scheme with the same error order, since it allows to switch to a larger time step as soon as the high frequency response disappears. In contrast: An explicit scheme would be bound to the small time step just because of the instability issue. On the other hand, if the solution is mainly defined by high-frequency contributions over the entire time interval of interest, it might be much more efficient to use an explicit scheme since the implicit scheme would be also bound to a small time step, which is computationally much more expensive. For further information the reader is referred to the literature on the treatment of numerical initial value problems such as Gear [103].

2.4.1. Generalized-α Method

After this short introduction into numerical time integration the remaining part of this section is solely devoted to the *Generalized-α* method. It is an implicit one-step, three-stage time integration scheme, i.e., it needs only information from t_n and t_{n+1} to be constructed and is described by the three stages corresponding to the displacement $\underline{d}^{\{n\}}$, velocity $\underline{v}^{\{n\}}$ and acceleration $\underline{a}^{\{n\}}$ fields. Under these circumstances the accuracy can become maximally second order and, indeed, the *Generalized-α* method is second order accurate. The construction of the three stage vectors relies on the Newmark-β methods which provide simple update rules for the discrete approximate velocity $\underline{v}^{\{n+1\}} \approx \dot{\underline{d}}(t_{n+1})$ and acceleration $\underline{a}^{\{n+1\}} \approx \ddot{\underline{d}}(t_{n+1})$ fields depending on the three old stage vectors and the current displacement field. These updating formulas are namely

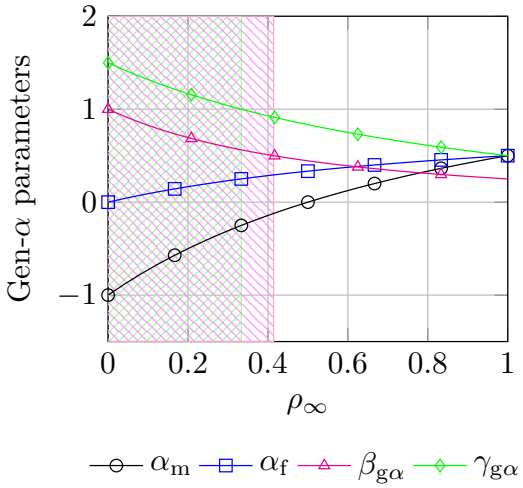


Figure 2.6.: Generalized-α parameters in dependency on the spectral radius ρ_∞. The hatched areas highlight the domains where $\beta_{\mathrm{g}\alpha} \in [0, 0.5]$ and $\gamma_{\mathrm{g}\alpha} \in [0, 1.0]$ exceed their bounds, respectively.

$$\underline{v}^{\{n+1\}}(\underline{d}^{\{n+1\}}) = \frac{\gamma_{\mathrm{g}\alpha}}{\beta_{\mathrm{g}\alpha}\,\Delta t}(\underline{d}^{\{n+1\}} - \underline{d}^{\{n\}}) - \frac{\gamma_{\mathrm{g}\alpha} - \beta_{\mathrm{g}\alpha}}{\beta_{\mathrm{g}\alpha}}\underline{v}^{\{n\}} - \frac{\gamma_{\mathrm{g}\alpha} - 2\beta_{\mathrm{g}\alpha}}{2\beta_{\mathrm{g}\alpha}}\Delta t\underline{a}^{\{n\}}, \tag{2.84a}$$

$$\underline{a}^{\{n+1\}}(\underline{d}^{\{n+1\}}) = \frac{1}{\beta_{\mathrm{g}\alpha}\,(\Delta t)^2}(\underline{d}^{\{n+1\}} - \underline{d}^{\{n\}}) - \frac{1}{\beta_{\mathrm{g}\alpha}\,\Delta t}\underline{v}^{\{n\}} - \frac{1 - 2\beta_{\mathrm{g}\alpha}}{2\beta_{\mathrm{g}\alpha}}\underline{a}^{\{n\}}, \tag{2.84b}$$

where $\gamma_{\mathrm{g}\alpha} \in [0, 1]$ and $\beta_{\mathrm{g}\alpha} \in [0,\, 0.5]$ are two important parameters of the Newmark-β methods which significantly control the character and accuracy of these updating rules. Now, the Generalized-α method makes use of these equations, but introduces additional special mid-points based on the parameters $\alpha_\mathrm{f} \in [0.0,\, 0.5]$ and $\alpha_\mathrm{m} \leqslant \alpha_\mathrm{f}$. Thus, the evaluation is shifted to a time-point $t_{n+1-\alpha_\mathrm{m}}$ and $t_{n+1-\alpha_\mathrm{f}}$. In accordance to these mid-points the linear momentum balance equations can be stated in matrix form, viz.

$$\underline{\underline{M}}\,\underline{a}^{\{n+1-\alpha_\mathrm{m}\}} + \underline{\underline{C}}\,\underline{v}^{\{n+1-\alpha_\mathrm{f}\}} + \underline{f}_{\mathrm{int}}^{\{n+1-\alpha_\mathrm{f}\}} - \underline{f}_{\mathrm{ext}}^{\{n+1-\alpha_\mathrm{f}\}} = \underline{0}, \tag{2.85}$$

where $\underline{\underline{M}}$ is the so-called *mass matrix* and $\underline{\underline{C}}$ the so-called *damping matrix*. In the framework discussed here, $\underline{f}_{\mathrm{int}}$ can be identified by the gradient of the internal energy with respect to the displacements evaluated at the mid-point $t_{n+1-\alpha_\mathrm{f}}$. In a similar way, the external force vector $\underline{f}_{\mathrm{ext}}^{n+1-\alpha_\mathrm{f}}$ can be interpreted as the gradient of a conservative external potential which considers tractions and/or volume forces. However, the balance equations stay valid also in a more general context in which these scalar-valued potentials might not exist. Now, to be able to apply these equations, the corresponding mid-point state vectors must be defined. Typically a simple linear interpolation is used, yielding

$$\underline{d}^{\{n+1-\alpha_\mathrm{f}\}} = (1 - \alpha_\mathrm{f})\underline{d}^{\{n+1\}} + \alpha_\mathrm{f}\underline{d}^{\{n\}}, \tag{2.86}$$

$$\underline{v}^{\{n+1-\alpha_\mathrm{f}\}} = (1 - \alpha_\mathrm{f})\underline{v}^{\{n+1\}} + \alpha_\mathrm{f}\underline{v}^{\{n\}}, \tag{2.87}$$

$$\underline{a}^{\{n+1-\alpha_\mathrm{m}\}} = (1 - \alpha_\mathrm{m})\underline{a}^{\{n+1\}} + \alpha_\mathrm{m}\underline{a}^{\{n\}}. \tag{2.88}$$

The original paper by Chung and Hulbert [48] proposes an elegant and meaningful way to define the free parameters $\alpha_\mathrm{f}, \alpha_\mathrm{m}, \gamma_{\mathrm{g}\alpha}$ and $\beta_{\mathrm{g}\alpha}$ all in dependency to the desired high-frequency dissipation. This is achieved by a new scalar parameter, the so-called *spectral radius* ρ_∞. The remaining parameters can be defined with respect to $\rho_\infty \leqslant 1$ such that the algorithm reaches an optimal and controllable combination of high-frequency and low-frequency dissipation, while additional conditions ensure second order accuracy and an unconditionally stable behavior of the algorithm. The parameters are finally given by

$$\alpha_\mathrm{m} = \frac{2\rho_\infty - 1}{\rho_\infty + 1}, \quad \alpha_\mathrm{f} = \frac{\rho_\infty}{\rho_\infty + 1}, \quad \beta_{\mathrm{g}\alpha} = \frac{1}{4}(1 - \alpha_\mathrm{m} + \alpha_\mathrm{f})^2, \quad \gamma_{\mathrm{g}\alpha} = \frac{1}{2} - \alpha_\mathrm{m} + \alpha_\mathrm{f}. \tag{2.89}$$

One special case is $\rho_\infty = 1$. Then, no numerical dissipation is activated. However, the domains for $\beta_{\mathrm{g}\alpha} \in [0, 0.5]$ and $\gamma_{\mathrm{g}\alpha} \in [0, 1.0]$ introduce additional lower bounds for ρ_∞ (see Figure 2.6 for an illustration). Thus, it follows that $(\sqrt{2} - 1) \leqslant \rho_\infty \leqslant 1.0$ must hold.

One further open question is how to evaluate $\underline{f}_\mathrm{int}^{\{n+1-\alpha_\mathrm{f}\}}$ in (2.85). There exist two usual variants. The first one is the so-called *implicit mid-point rule* resulting in

$$\underline{f}_\mathrm{int}^{\{n+1-\alpha_\mathrm{f}\}} = \nabla_{\underline{d}} \mathscr{U}\big|_{\underline{d}^{\{n+1-\alpha_\mathrm{f}\}}}, \tag{2.90}$$

where the gradient of the internal energy is evaluated with respect to the linearily interpolated displacement field defined in (2.86). The second option, which will also be followed throughout this thesis, is based on a *trapezoidal rule*, i.e.,

$$\underline{f}_\mathrm{int}^{\{n+1-\alpha_\mathrm{f}\}} = (1 - \alpha_\mathrm{f})\, \nabla_{\underline{d}} \mathscr{U}\big|_{\underline{d}^{\{n+1\}}} + \alpha_\mathrm{f}\, \nabla_{\underline{d}} \mathscr{U}\big|_{\underline{d}^{\{n\}}}. \tag{2.91}$$

There is no doubt that both formulations (2.90) and (2.91) do coincide in the special case of small deformations and linear elasticity. However, the difference becomes also generally smaller and smaller in case of a decreasing time step size. Furthermore, in case of a general non-linear material behavior it is hard to guess which one is favorable. The same holds true for the external forces introduced in $\underline{f}_\mathrm{ext}^{n+1-\alpha_\mathrm{f}}$. The only difference is that the simple underlying external potential law often naturally leads to a result similar to (2.91).

2.4.2. Linearization of the Generalized-α Method

In the following brief discussion the basic steps of the fundamental linearization approach are demonstrated. Even though this approach is by no means restricted to the Generalized-α method, it shall be introduced in this context, since (2.85) represents one of the more advanced structural balance equations in terms of all the considered contributions. Furthermore, it contains the quasi-static case as special case. However, the discussion in this section is solely restricted to a pure structural dynamic simulation, i.e., contact contributions are not considered, yet. The necessary extensions will follow in Sections 4.6 and 6.8.1.

For now the in (2.85) stated balance equations in matrix-vector form under consideration of the trapezoidal rule shall be considered as the underlying residual and the task will be to find a solution which fulfills these non-linear equations such that $\underline{r}_{\mathrm{g}\alpha}(\underline{d}_{\{k\}}^{n+1}) = \underline{0}$ with

$$\underline{r}_{\mathrm{g}\alpha}(\underline{d}_{\{k\}}^{\{n+1\}}) = \underline{\underline{M}}\,((1-\alpha_{\mathrm{m}})\,\underline{a}_{\{k\}}^{\{n+1\}} + \alpha_{\mathrm{m}}\,\underline{a}^{\{n\}}) \tag{2.92a}$$

$$+\,\underline{\underline{C}}\,((1+\alpha_{\mathrm{f}})\underline{v}_{\{k\}}^{\{n+1\}} + \alpha_{\mathrm{f}}\underline{v}^{\{n\}}) \tag{2.92b}$$

$$+\,((1-\alpha_{\mathrm{f}})\,\nabla_{\underline{d}}\mathscr{U}|_{\underline{d}_{\{k\}}^{\{n+1\}}} + \alpha_{\mathrm{f}}\,\nabla_{\underline{d}}\mathscr{U}|_{\underline{d}^{\{n\}}}) \tag{2.92c}$$

$$-\,((1-\alpha_{\mathrm{f}})\,\nabla_{\underline{d}}\mathscr{V}_{\mathrm{ext}}|_{\underline{d}_{\{k\}}^{\{n+1\}}} + \alpha_{\mathrm{f}}\,\nabla_{\underline{d}}\mathscr{V}_{\mathrm{ext}}|_{\underline{d}^{\{n\}}}) \stackrel{!}{=} 0. \tag{2.92d}$$

Again, as already mentioned multiple times, the gradient of an external potential can also directly be replaced by a suitable external force vector. The balance equations stay valid. The advantage of the notation as an (auxiliary) external potential will become obvious later in Section 6.8.1.

With (2.92) at hand the further steps follow closely the classical Newton-Raphson method for a set of non-linear equations. Here, only the basic idea and the related system of equations will be presented, while a much more general discussion follows in Section 3.1.1. Now, to find the desired solution the demand is formulated that the linear model of Equation (2.92) with respect to the unknown $\underline{d}^{n+1}$ vector shall be equal to zero, thus,

$$\underline{r}_{\mathrm{g}\alpha}(\underline{d}_{\{k\}}^{\{n+1\}}) + (\underline{d}_{+}^{\{n+1\}} - \underline{d}_{\{k\}}^{\{n+1\}})^{\mathrm{T}}\,\nabla_{\underline{d}}\underline{r}_{\mathrm{g}\alpha}\big|_{\underline{d}^{\{n\}}} \stackrel{!}{=} \underline{0}. \tag{2.93}$$

Obviously, this represents only an approximation. However, under certain circumstances it can be expected that the obtained sequence $\underline{d}_{\{k+1\}}^{\{n+1\}} = \underline{d}_{+}^{\{n+1\}}$ with $k = \{0, 1, \dots\}$ will converge to the solution vector $\underline{d}_{\star}^{\{n+1\}}$ and as soon as $\underline{d}_{\{k\}}^{\{n+1\}}$ enters a region close enough to the solution, the sequence will stay in the neighborhood and will converge q-quadratically. For more information the reader is kindly referred to Chapter 3 and the information therein. Typically, (2.93) is written as

$$\underline{\underline{K}}_{\mathrm{g}\alpha}(\underline{d}^{n})\,\Delta\underline{d}_{+}^{\{n+1\}} = -\underline{r}_{\mathrm{g}\alpha}(\underline{d}_{\{k\}}^{\{n+1\}}), \tag{2.94}$$

where $\underline{\underline{K}}_{\mathrm{g}\alpha}$ is the so-called *Jacobian* matrix which is defined as the transpose of $\nabla_{\underline{d}}\underline{r}_{\mathrm{g}\alpha}^{\mathrm{T}}$. In addition, the incremental vector $\Delta\underline{d}_{+}^{\{n+1\}}$ is used as an abbreviation for $\underline{d}_{+}^{\{n+1\}} - \underline{d}_{\{k\}}^{\{n+1\}}$ leading to the simple update formula

$$\underline{d}_{\{k+1\}}^{\{n+1\}} = \underline{d}_{+}^{\{n+1\}} = \underline{d}_{\{k\}}^{\{n+1\}} + \Delta\underline{d}_{+}^{\{n+1\}}. \tag{2.95}$$

Therefore, two questions remain: What is the actual definition of the Jacobian matrix and, secondly, how to solve (2.94)? For a general answer to the latter question the reader is kindly referred to the comprehensive literature on linear systems of equations. However, during this thesis special techniques will be proposed which not only improve the non-linear but also the linear solvability of the related systems. The thereby improved conditioning of the corresponding system of equations can become very important when large problems are considered and, consequently, linear iterative solvers must be applied. For more information see Chapter 5 and, especially, Section 6.6.

The answer to the first question is briefly addressed in the following:

$$\underline{\underline{K}}_{\mathrm{g}\alpha} = \frac{(1-\alpha_{\mathrm{m}})}{\beta_{\mathrm{g}\alpha}(\Delta t)^2}\underline{\underline{M}} + \frac{(1-\alpha_{\mathrm{f}})\,\gamma_{\mathrm{g}\alpha}}{\beta_{\mathrm{g}\alpha}\,\Delta t}\underline{\underline{C}} + (1-\alpha_{\mathrm{f}})\;\; \nabla^2_{\underline{d}\,\underline{d}}\mathscr{U}\big|^{\mathrm{T}}_{\underline{d}^{\{n+1\}}_{\{k\}}}\,, \tag{2.96}$$

where the first two additive terms can be easily derived under consideration of (2.84). The last term is the so-called tangential stiffness matrix which is also of major importance for the static case. Theoretically, it is also possible that the derivative of the external forces with respect to the currently unknown displacement field must be taken into account, if the loads depend on the current deformation. However, this case shall be excluded during this thesis.

In summary, the solution approach via the Newton Raphson scheme asks for two expensive operations: The first one is the evaluation of the tangential stiffness matrix which will get increasingly expensive as soon as the later discussed contact contributions have to be considered as well. Secondly, the solution of the linear system of equations can be a time consuming point. However, throughout this thesis a preconditioned linear solver method based on the *Generalized Minimal Residual* (GMRES) approach will be used as soon as a direct solver becomes too expensive or non-applicable due to a too high memory consumption. The GMRES based methods proved to be quite efficient for large linear systems. Further information about this linear solver strategy can be found in Saad and Schultz [232], and Kelley [153, Chapter 3], for instance.

3. Numerical Optimization

The second part of the mathematical fundamentals is covered by the field of numerical optimization. This field is extremely wide and can be separated into a large number of different sub-topics. Actually, even the classification is not unique, since it can be done with respect to very different aspects. In this thesis it is all about the development of a robust and efficient solution method for frictionless contact problems. Thus, the following introduction starts with the solution of unconstrained problems to then focus on the class of inequality constrained problems.

3.1. Unconstrained Optimization

The basic objective of any non-linear solution method is to find a root of a generally non-linear set of equations $\underline{r}(\underline{x}) = \underline{0}$, where $\underline{r}(\underline{x}) : \mathbb{R}^m \rightarrow \mathbb{R}^n$. The case $m < n$ is valid and aims for so-called least-squares problems (see Marquardt [187] for a classical example, or Conn et al. [53, Ch. 16] for a great introduction). However, at this point the discussion shall be restricted to the special case $m = n$. If further $\nabla_{\underline{x}}\underline{r}(\underline{x}) = [\nabla_{\underline{x}}\underline{r}(\underline{x})]^{\mathrm{T}}$ holds, i.e., the associated Jacobian matrix is symmetric, then it is possible to directly obtain a scalar-valued objective function $f(\underline{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ by

$$f(\underline{x}) = \int_0^1 \langle \underline{r}(t\underline{x}),\ \underline{x} \rangle \, \mathrm{d}t, \tag{3.1}$$

such that $\nabla_{\underline{x}} f(\underline{x}) = \underline{r}(\underline{x})$ and $\nabla^2_{\underline{x}\,\underline{x}} f(\underline{x}) = \nabla_{\underline{x}}\underline{r}(\underline{x})$ hold. The derived objective function is not unique since it can be manipulated by multiplication or summation with some arbitrary scalar without changing the root of $\underline{r}$ or $\nabla_{\underline{x}} f$. Furthermore, whenever the derivation of an objective function is possible, the Jacobian matrix becomes the so-called Hessian matrix and the root finding objective can be reformulated as a search for an extremum of f. This is probably an unusual approach at first glance, since typically not the gradient is used to start the discussion but the objective function. However, it is done in this order to close the cycle with (2.45). In the finite element community it is typically more convenient to start with the strong (2.28) and weak forms (2.35) rather than with a scalar-valued objective, or potential function. Especially, the consideration of non-conservative (constraint) forces like friction will automatically lead to a non-symmetric Hessian such that (3.1) is no longer valid. Nevertheless, in case of a pure structural problem and under consideration of conservative forces (e.g. dead loads), the derivation of a suitable objective function is meaningful (see Section 2.1.4). Additionally, if the reconstruction of a suitable objective function is no option, there is still the possibility to consider an "artificial" merit function, e.g., leading to the already mentioned least-squares problem

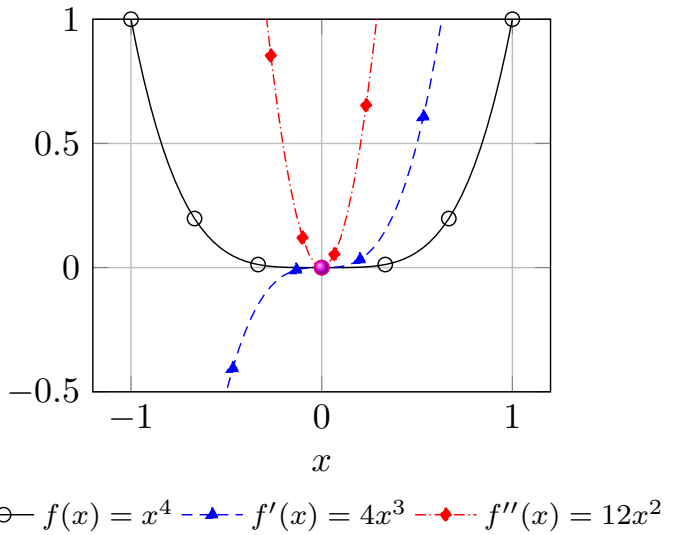


Figure 3.1.: A simple example for a one-dimensional optimization problem is presented. The goal is to find the (global) minimizer of the objective function $f(x) = x^4$. The obvious solution is $x^\star = 0$. However, the function is special with respect to Theorem 3.3: Despite the fact that the solution point is a global minimizer, the mentioned theorem is not satisfied since the second order derivative f'' is zero at $x^\star$. This underlines the important difference between *sufficient* and *necessary* second order conditions.

$$\min_{\underline{x} \in \mathbb{R}^n} \frac{1}{2} \|\underline{r}(\underline{x})\|^2. \tag{3.2}$$

However, such a formulation introduces other drawbacks which will also be briefly addressed within this chapter. For now, it shall be assumed that a suitable scalar-valued objective function f is available, such that the *necessary first order conditions for optimality* are given by:

Theorem 3.1. *First-Order Necessary Conditions for Unconstrained Optimization.* If $\underline{x}^\star$ is a local minimizer and the objective function f is continuously differentiable in an open neighborhood of the solution point $\underline{x}^\star$, then $\nabla f(\underline{x}^\star) = \underline{0}$.

This very basic theorem is taken from Nocedal and Wright [204], where the proof can be found as well. Now, these necessary conditions make a point $\underline{x}$ to a stationary point, but not necessarily to a local minimizer of the considered problem, at least not if the problem is not strictly convex as discussed in Boyd and Vandenberghe [34]. Luckily, convexity can be often taken as granted which significantly simplifies the search for a (global) minimizer. Typical convex problems are classical pure structural problems (without instabilities). A general problem is called convex iff

$$f(\alpha \underline{x} + (1-\alpha)\underline{y}) \leqslant \alpha f(\underline{x}) + (1-\alpha) f(\underline{y}), \qquad \forall \underline{x}, \underline{y} \in E, \alpha \in [0,1], \tag{3.3}$$

where E defines the domain of convexity. Strict convexity means that the inequality (3.3) becomes strict for any $\underline{x} \neq \underline{y} \in E$ and $\alpha \in (0,1)$. A function $f : E \to \mathbb{R}$ is called concave if $-f$ is convex. However, if strict convexity does not hold, the given conditions in Theorem 3.1 must be extended to obtain sufficient conditions for the identification of a local minimizer of f.

Therefore, first, the so-called *second-order necessary conditions for unconstrained optimization* are introduced.

Theorem 3.2. *Second-Order Necessary Conditions for Unconstrained Optimization.* If $\underline{x}^\star$ is a local minimizer of f and the Hessian $\nabla^2_{\underline{x}\,\underline{x}} f$ exists and is continuous in an open neighborhood around $\underline{x}^\star$, then $\nabla_{\underline{x}} f(\underline{x}^\star) = \underline{0}$ and $\langle \underline{v}, \nabla^2_{\underline{x}\,\underline{x}} f(\underline{x}^\star)\underline{v}\rangle \geqslant 0$ hold for all $\underline{v} \neq \underline{0}$, i.e., the Hessian is at least positive semi-definite.

If one of the two theorems, i.e., Theorem 3.1 or 3.2, should be violated, the point $\underline{x}^\star$ cannot be a local minimizer. However, the stated necessary conditions are still not sufficient for a strict local minimizer. Therefore, it is possible that both theorems hold and the considered point is still just a stationary point. To obtain sufficient conditions which guarantee that the point $\underline{x}^\star$ is a local minimizer, the so-called *second order sufficient conditions* must hold.

Theorem 3.3. *Second Order Sufficient Conditions for Unconstrained Optimization.* The point $\underline{x}^\star$ is a strict local minimizer of f if the Hessian $\nabla^2_{\underline{x}\,\underline{x}} f$ exists and is continuous in an open neighborhood around $\underline{x}^\star$, and if $\nabla_{\underline{x}} f(\underline{x}^\star) = \underline{0}$ and $\langle \underline{v}, \nabla^2_{\underline{x}\,\underline{x}} f(\underline{x}^\star)\underline{v}\rangle > 0$ hold for all $\underline{v} \neq \underline{0}$, i.e. the Hessian must be positive definite.

This is a slightly stricter definition as the stated second order necessary conditions in Theorem 3.2, since now the considered Hessian must be strictly positive definite. However, even though this guarantees a local minimizer, it is not said that any strict local minimizer must fulfill these stronger assumptions. A easy counter-example is the function $f(x) = x^4$ which has a strict minimum at $x = 0$, although its second order derivative is zero at this point (see Figure 3.1 for a visualization).

3.1.1. Local Iterative Solution Methods

After this brief introduction into unconstrained optimization where some of the most important basic theorems have already been given, the attention is drawn to the numerical treatment of such optimization problems. In its most general form, the problem

$$\underset{\underline{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\underline{x}) \tag{3.4}$$

shall be solved. The following text gives a short overview of different methods which are able to solve these problems. As a starting point an iterate $\underline{x}^{\{k\}}$ shall be considered, and, for the moment, only the coordinates of this point and its associated function value $f^{\{k\}} = f(\underline{x}^{\{k\}})$ shall be taken into account. The overall objective is to find a sequence of iterates $\{\underline{x}^{\{k\}}\}$ which leads to smaller function values, such that at least for some $\bar{k} > k$ the new function value $f^{\{\bar{k}\}}$ shows a sufficient reduction compared to the current value $f^{\{k\}}$.

3.1.1.1. Direct Methods

Direct, or derivative-free, algorithms aim for applications where it is not possible, or too expensive, to compute a reliable gradient information. Examples are manifold and can be found in

many different research fields such as medical, engineering or location problems. Furthermore, based on the early work of Nelder and Mead [201], Spendley et al. [253], a sub-set of the most recent algorithms are not only able to find local minimizers, but are also designed in such a way that they are capable of searching for global minimizers in non-convex problems. That is something which goes beyond the scope of this thesis, but is very important in other fields. Furthermore, since no derivatives of the objective function must be computed, it becomes possible to consider non-smooth problems to a much greater extent. The treatment of constrained problems is an ongoing research field. For a comprehensive introduction and review on these interesting methods the reader is referred to Conn et al. [54], Kolda et al. [162], Rios and Sahinidis [227]. Especially, Rios and Sahinidis [227] provide a recent insight by performing some extensive studies. Finally, it is to mention that not all of these methods are entirely restricted to pure objective function evaluations. A group of direct methods uses gradient information as well, but not of the actual objective function but instead for a so-called *surrogate function*. Basically, surrogate functions are model functions which are supposed to be as simple as possible while still maintaining the main features of the actual objective function. The consideration of such surrogate functions can even reopen—under certain assumptions—the field of the later discussed globalization methods.

3.1.1.2. Fixed-Point Methods

Before discussing the classical gradient-based methods, the idea of fixed-point iterations shall be presented. Fixed-point methods are not restricted to problems where a scalar-valued objective function is readily available. Instead, they represent classical root-finding methods which can directly be applied to the initial problem $\underline{r}(\underline{x}) \stackrel{!}{=} \underline{0}$, where $\underline{r} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ denotes the previously introduced system of non-linear equations. The desired sequence of iterates $\{\underline{x}^{\{k\}}\}$ can be obtained by

$$\underline{x}^{\{k+1\}} = \underline{x}^{\{k\}} + \Delta\underline{x}^{\{k\}} = \underline{x}^{\{k\}} - \underline{\underline{B}}^{-1}(\underline{x}^{\{k\}})\, \underline{r}(\underline{x}^{\{k\}}) = \underline{\Phi}(\underline{x}^{\{k\}}), \tag{3.5}$$

where the iteration matrix $\underline{\underline{B}}(\underline{x})$ is chosen in such a way that it is invertible in a neighborhood around the solution $\underline{x}^\star$ and $\underline{\Phi} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a so-called *fix-point mapping* or *fixed-point function*. Note that at the solution $\underline{x}^\star = \underline{x}^\star - [\underline{\underline{B}}^\star]^{-1}\underline{r}^\star = \underline{x}^\star$ follows, since $\underline{r}(\underline{x}^\star) = \underline{0}$ holds. The generated sequence $\{\underline{x}^{\{k\}}\}$ will converge to a fixed point $\underline{x}^\star$ if $\underline{\Phi}(\underline{x})$ is Lipschitz continuous on the considered domain $\Omega \subset \mathbb{R}^n$ with a Lipschitz constant $L < 1$. In this case $\underline{\Phi}$ is a so-called contraction mapping such that $\Omega \rightarrow \Omega$ holds. The classical result is known as the Banach Fixed-Point Theorem as exemplarily presented in Dahmen and Reusken [60], Kelley [153]. Further note that (3.5) is the non-linear variant of the so-called Richardson iteration which provides the foundation for many linear iterative methods [153]. Furthermore, there has been an effort to increase the convergence speed of the generated vector sequence, since the convergence speed of the classical fixed-point method is closely related to the Lipschitz constant where a constant close to 1 can lead to a very poor performance. For more information on this topic the reader is referred to the literature on the so-called *Aitken relaxation* (see e.g. Irons and Tuck [146], Křížek et al. [165], Macleod [183]). These ideas have, for instance, successfully been applied to fluid structure interaction problems as demonstrated in Küttler [168], Küttler and Wall [169].

3.1.1.3. Gradient-Based Methods

If the gradient information is available, another large variety of different methods becomes accessible. First, it shall be assumed that besides the function value at the current iterate $f^{\{k\}}$ only the gradient information $\nabla_{\underline{x}} f^{\{k\}} = \nabla_{\underline{x}} f(\underline{x}^{\{k\}})$ is additionally available. In such a case, e.g., the *steepest descent* method can be used. Variants of this method are all based on the idea that a sufficient reduction of the objective function value must be achievable if the direction with the locally fastest decrease is chosen, i.e., the direction orthogonal to the current function value isolines. Therefore, the local rate of change in f must be identified. This is possible by considering the mean value theorem following from the Taylor series expansion

$$f(\underline{x}^{\{k\}} + \alpha\, \underline{p}) = f^{\{k\}} + \alpha \langle \underline{p}, \nabla_{\underline{x}} f^{\{k\}} \rangle + \frac{\alpha^2}{2} \langle \underline{p}, \nabla^2_{\underline{x}\,\underline{x}} f(\underline{x}^{\{k\}} + t\, \underline{p})\, \underline{p} \rangle, \tag{3.6}$$

with $t \in (0, \alpha)$. Here, the term scaled by the single step length α denotes the considered rate of change. The desired direction is obtained by $\underline{p}^{\{k\}} = -\nabla_{\underline{x}} f^{\{k\}}$. This direction is by construction perpendicular to the contour lines. However, a closer look at this direction reveals that the convergence rate can become arbitrarily slow if this search direction is used in an iterative method, i.e., by applying the update rule $\underline{x}^{\{k+1\}} = \underline{x}^{\{k\}} - \alpha \nabla_{\underline{x}} f^{\{k\}}$. Additionally, a suitable step length α must be obtained, since the gradient alone holds no length information for a meaningful step. How this step length can be obtained will be explained in Section 3.1.2. A detailed discussion of the steepest descent method can be found in almost all of the text books on this topic, e.g. Deuflhard [69], Fletcher [93], Nocedal and Wright [204]. Finally, it is to mention that the presented update routine is a special case of (3.5), where $\underline{\underline{B}}$ is set to $\alpha\, \underline{\underline{I}}$ with the identity matrix $\underline{\underline{I}}$.

Besides the basic steepest descent method, there are also better options which are still based on the pure gradient information but choose a more sophisticated search direction. One example are the non-linear conjugate direction methods (see Fletcher [93, Ch. 4] or Nocedal and Wright [204, Ch. 5]).

3.1.1.4. Newton-Like Methods

Next, it is assumed that not only the gradient information, i.e., information about the local slope of the currently considered objective function, but also information about its curvature at the current iterate $\underline{x}^{\{k\}}$ are available.

Newton-Raphson Method

If the second order derivative of the underlying objective function can be evaluated, the well-known Newton-Raphson method is obtained. Furthermore, the Newton method is not limited to the case of a scalar valued objective function, but can also be applied if a set of non-linear equations is considered as described in Dennis Jr. and Schnabel [68], Kelley [153, 154]. The fundamental update rule for the iterates follows again from (3.5), where the iteration matrix $\underline{\underline{B}}$ is now set to the Jacobian $\underline{\underline{J}} = \nabla_{\underline{x}} \underline{r}^{\mathrm{T}}$, i.e., the transposed gradient of the residual vector $\underline{r}$ with respect to the solution variables. Thus, if the following *standard assumptions* hold,

AS 3.1. The considered problem $\underline{r}(\underline{x}) = \underline{0}$ has a solution $\underline{x}^{\star}$;

AS 3.2. The matrix and the corresponding mapping $\underline{\underline{J}}: \Omega \to \mathbb{R}^{n \times n}$ is Lipschitz continuous with a Lipschitz constant L;

AS 3.3. The Jacobian $\underline{\underline{J}}(\underline{x}^\star)$ is non-singular at the solution;

it can be shown that the generated sequence $\{\underline{x}^{\{k\}}\}$ converges q-quadratically as long as the initial point $\underline{x}^{\{k\}}$ is inside a ball around the solution point $\underline{x}^\star$, i.e., $x^{\{k\}} \in \mathcal{V}(\delta) = \{\underline{x} \in \mathbb{R}^n : \|\underline{x} - \underline{x}^\star\| < \delta\}$. The proof can be found in the mentioned text books. The term q-quadratic means that

$$\|\underline{x}^{\{k+1\}} - \underline{x}^\star\| \leqslant C\|\underline{x}^{\{k\}} - \underline{x}^\star\|^2, \qquad \forall \underline{x}^{\{k\}} \in \mathcal{V}(\delta) \tag{3.7}$$

holds, where the "q" stands for quotient. In other words: Close to the solution, each newly computed iterate coincides by roughly twice as many digits with the desired result vector $\underline{x}^\star$. Consequently, if the quadratic convergence is activated by $\underline{x}^{\{k\}}$ entering $\mathcal{V}(\delta)$, a very fast convergence can be expected. However, an actual implementation will suffer from rounding errors at some point, thus convergence to the analytical result can in general not be guaranteed. For details on these numerical issues the reader is kindly referred to Higham [132]. The computed Newton direction is well-suited for some of the following discussed globalization methods as long as the *descent property*

$$\langle \Delta\underline{x}, \nabla_{\underline{x}} f \rangle = -\langle \nabla_{\underline{x}} f,\ \underline{\underline{B}}^{-1} \nabla_{\underline{x}} f \rangle < 0 \tag{3.8}$$

holds. This criterion ensures that the enclosed angle between the introduced steepest descent direction and the search direction is smaller than $\pi/2$ and, therefore, a decrease of the objective function can be (at least locally) expected. However, the descent condition (3.8) is naturally fulfilled when the matrix $\underline{\underline{B}}$ is positive definite. Otherwise, an appropriate adaption must be considered. Examples can be found in Conn et al. [53], Nocedal and Wright [204], Wächter and Biegler [272] and will also be discussed in Section 6.6. Another point which must be mentioned is that the Newton method contains a meaningful inherent step length information. This is in contrast to the pure gradient methods, such as the conjugate gradient or steepest descent method, and also in contrast to approximate methods like the up-coming quasi-Newton methods. This property of the Newton method becomes obvious by the following consideration: Starting with the case where a scalar-valued objective function is readily available, a quadratic model equation can be derived and follows directly from the Taylor series expansion similar to (3.6). This model yields

$$m_f^{\{k\}}(\underline{x}) = f(\underline{x}^{\{k\}}) + \langle \underline{x} - \underline{x}^{\{k\}}, \nabla_{\underline{x}} f(\underline{x}^{\{k\}}) \rangle + \frac{1}{2} \langle \underline{x} - \underline{x}^{\{k\}}, \nabla^2_{\underline{x}\,\underline{x}} f(\underline{x}^{\{k\}})(\underline{x} - \underline{x}^{\{k\}}) \rangle. \tag{3.9}$$

Now, the search direction of the classical full Newton method, where "full" means that a full step length of $\alpha = 1$ is considered, with $\underline{\underline{B}}^{\{k\}} = \nabla^2_{\underline{x}\,\underline{x}} f^{\{k\}}$ minimizes this quadratic model, since

$$\begin{aligned}
& \underset{\underline{x} \in \mathbb{R}^n}{\text{minimize}} \quad m_f^{\{k\}}(\underline{x}) \\
\Rightarrow \quad & \nabla_{\underline{x}} m_f^{\{k\}}(\underline{x}) = \nabla_{\underline{x}} f(\underline{x}^{\{k\}}) + \nabla^2_{\underline{x}\,\underline{x}} f(\underline{x}^{\{k\}})(\underline{x} - \underline{x}^{\{k\}}) \stackrel{!}{=} \underline{0} \\
\Rightarrow \quad & \underline{x}^\star = \underline{x}^{\{k\}} - \nabla^2_{\underline{x}\,\underline{x}} f(\underline{x}^{\{k\}})^{-1} \nabla_{\underline{x}} f(\underline{x}^{\{k\}}).
\end{aligned} \tag{3.10}$$

Under the assumption that the quadratic model (3.9) represents a good approximation of the underlying objective function, it makes sense to stick to the unmodified computed step and exploit the possible quadratic convergence. Obviously that is only a good idea close to the solution, i.e., as soon as the influence of the higher order terms neglected in (3.9) starts to vanish.

In contrast, the same interpretation gets a little bit more involved for a given system of non-linear equations. In such a case there are different possibilities how to construct a suitable objective function. Probably the most common one is the so-called *Gauss-Newton model*, where the initial step is a linear model for the set of non-linear equations rather than a quadratic model for the least-squares problem, viz.

$$\underline{m}_r^{\{k\}}(\underline{x}) = \underline{r}(\underline{x}^{\{k\}}) + \underline{\underline{J}}(\underline{x}^{\{k\}})(\underline{x} - \underline{x}^{\{k\}}), \tag{3.11}$$

where $\underline{\underline{J}}$ is again the associated Jacobian matrix. This at hand, the (quasi-)quadratic model can easily be formed by

$$\begin{aligned} m_f^{\{k\}}(\underline{x}) =& \frac{1}{2}\|\underline{m}_r^{\{k\}}(\underline{x})\|^2 = \frac{1}{2}\|\underline{r}^{\{k\}} + \underline{\underline{J}}^{\{k\}}\ (\underline{x} - \underline{x}^{\{k\}})\|^2 \\ =& \frac{1}{2}\|\underline{r}^{\{k\}}\|^2 + \langle \underline{r}^{\{k\}},\ \underline{\underline{J}}^{\{k\}}\ (\underline{x} - \underline{x}^{\{k\}})\rangle + \frac{1}{2}\|\underline{\underline{J}}^{\{k\}}(\underline{x} - \underline{x}^{\{k\}})\|^2. \end{aligned} \tag{3.12}$$

The first-order optimality conditions for a minimization problem based on the model (3.12) deliver

$$\begin{aligned} & \nabla_{\underline{x}} m_f^{\{k\}}(\underline{x}) = [\underline{\underline{J}}^{\{k\}}]^{\mathrm{T}}\ \underline{r}^{\{k\}} + [\underline{\underline{J}}^{\{k\}}]^{\mathrm{T}}\ \underline{\underline{J}}^{\{k\}}(\underline{x} - \underline{x}^{\{k\}}) \stackrel{!}{=} \underline{0}, \\ \Rightarrow \qquad & \underline{x}^{\star} = \underline{x}^{\{k\}} - \{[\underline{\underline{J}}^{\{k\}}]^{\mathrm{T}}\ \underline{\underline{J}}^{\{k\}}\}^{-1}\ [\underline{\underline{J}}^{\{k\}}]^{\mathrm{T}}\ \underline{r}^{\{k\}} \stackrel{(\dagger)}{=} \underline{x}^{\{k\}} - [\underline{\underline{J}}^{\{k\}}]^{-1}\underline{r}^{\{k\}}, \end{aligned} \tag{3.13}$$

where the last step $(\dagger)$ becomes possible, since it is assumed that the Jacobian matrix $\underline{\underline{J}}$ is square. This concludes the initial statement that the Newton method for a system of non-linear equations can also be interpreted as the minimization of a scalar valued (model) function. Furthermore, it is also possible to use other norms than the presented ℓ_2-norm. The ℓ_2-norm has the advantage that it minimizes the variance and, therefore, has some statistical relevance. However, there are also circumstances where other norms are preferable as explained in Conn et al. [54, Sec. 16.2].

Remark 3.1. The Gauss-Newton model is used quite often in the optimization literature, even though it has an obvious drawback: The pure convexity of the model does not necessarily remodel the true second order derivative, which would be given by

$$\nabla^2_{\underline{x}\,\underline{x}}[\frac{1}{2}\|\underline{r}(\underline{x})\|^2] = \underline{\underline{J}}^{\mathrm{T}}\underline{\underline{J}} + \sum_{i=1}^{m} r_i(\underline{x})\ \nabla^2_{\underline{x}\,\underline{x}} r_i(\underline{x}), \tag{3.14}$$

and can consequently lead to convergence to stationary but not second-order optimal points (saddle-points). However, this drawback can be resolved by considering the actual second-order model, known as the *Newton model*, of the least squares problem (3.2), or even more sophisticated models, such as *tensor methods* as discussed in Schnabel and Frank [236]. The reader is

kindly referred to Conn et al. [54, Sec. 6.5 and Sec. 16.1] for a detailed explanation of this issue in the context of Trust Region methods. Otherwise, if pure line search methods are used and the Jacobian matrix shows some degree of rank deficiency, it is possible that the solution procedure is attracted by non-stationary points as presented in Byrd et al. [42]. This can even happen for regularized Newton methods [68, Sec. 6.5].

Quasi-Newton Methods

The next short paragraph is dedicated to the so-called *quasi-Newton methods*. The idea of quasi-Newton methods is to replace the calculation of the second order derivatives, i.e., the Jacobian or Hessian matrix, by a well-defined estimate in order to reduce the high associated costs. Alternatively, it is also possible to compute an estimate for the inverse of these matrices, which has the additional advantage that not only the expensive evaluation but also the solution of the linear system of equations becomes much cheaper. In the case of a symmetric Hessian matrix, suitable methods can be found under the key-word *BFGS* (abbreviation for Broyden, Fletcher, Goldfarb and Shanno), which is the dual spin-off of the previously published *Davidon-Fletcher-Powell algorithm* [61]. These methods are based on the idea that the proposed estimate is symmetric positive-definite, fulfills the secant equation and the deviation from the previous estimate is expected to be minimized. This can be summarized for the BFGS methods by

$$\begin{aligned} &\underset{\underline{\underline{B}}^{-1}}{\text{minimize}} \quad \|\underline{\underline{B}}^{-1} - [\underline{\underline{B}}^{\{k\}}]^{-1}\|, \\ \text{s. t.} \quad &\underline{\underline{B}}^{-1} = \underline{\underline{B}}^{-\mathrm{T}} \text{ and } \underline{\underline{B}}^{-1}(\nabla_{\underline{x}} f^{\{k+1\}} - \nabla_{\underline{x}} f^{\{k\}}) = \underline{x}^{\{k+1\}} - \underline{x}^{\{k\}}. \end{aligned} \tag{3.15}$$

However, the BFGS method is not considered in this thesis, since it asks for a symmetric Jacobian matrix and this prerequisite is not always satisfied due to the later applied and discussed modifications to the variational form of the frictionless contact problem (see Chapter 4). Nevertheless, it is an interesting topic. Especially the fact that the algorithm maintains a positive-definite system matrix can be very helpful if the minimization of non-convex problems is considered. For more information, especially concerning the low-memory variant (L-BFGS), the reader is kindly referred to Erway and Marcia [78], Li and Fukushima [176], Liu and Nocedal [178]. Additionally, there are also more recent publications concerning non-smooth problems, e.g. Lewis and Overton [175].

Besides the well-known BFGS method, another class of inexact Newton methods is applicable to non-symmetric system matrices as well. These methods are called *Broyden methods* named after Broyden [38]. A summary of different variants for this family of methods can be found in Martinez [189]. Furthermore, the important treatment of sparse matrices is addressed in Bogle and Perkins [28], Martinez and Zambaldi [190], Marwil [191], Schubert [238]. This sub-class is often called *Schubert's method*. However, all of the cited methods concerning sparse matrices expect a constant sparsity pattern which is generally not available for contact problems. Nevertheless, the idea to approximate the system matrix or a part of the system matrix by these methods can be an interesting future research topic.

3.1.2. Globalization Techniques

This thesis is mainly about improved robustness in the field of non-linear solution techniques. Therefore, after a short introduction into the local solution schemes in Section 3.1.1, the attention is on globalization strategies. For the time being, the focus is still on unconstrained optimization. First of all, there are two major strategies to construct globally convergent methods: Once, the so-called *line search* methods which will be used throughout this thesis and, secondly, the *trust region*, or *Levenberg-Marquardt methods*. Even though both of them follow the same objective and provide a meaningful safe-guarding strategy throughout the non-linear solution procedure, they are still very different in their properties and fundamental idea. To highlight the main differences, both approaches shall be discussed in the following.

3.1.2.1. Line Search Methods

The fundamental idea of line search methods is that as long as the *descent property* (3.8) is fulfilled, it must be possible to find a new iterate $\underline{x}^{\{k+1\}} = \underline{x}^{\{k\}} + \Delta\underline{x}^{\{k\}}$ with $\Delta\underline{x}^{\{k\}} = \alpha^{\{k\}}\underline{p}^{\{k\}}$ along the given search direction $\underline{p}^{\{k\}}$ such that the objective function (or a suitable merit function) is sufficiently reduced. Sufficiently reduced means that the step length $\alpha^{\{k\}} > 0$ is chosen such that

$$f(\underline{x}^{\{k\}} + \alpha^{\{k\}}\underline{p}^{\{k\}}) \leqslant f_{\text{ref}}^{\{k\}} + c_1\alpha^{\{k\}}\langle\nabla_{\underline{x}} f^{\{k\}},\ \underline{p}^{\{k\}}\rangle, \tag{3.16}$$

where $c_1 \in (0,\ 0.5)$ is a constant scalar. The reference function value $f_{\text{ref}}^{\{k\}}$ is classically chosen as the previously accepted function value $f_{\text{ref}}^{\{k\}} = f(\underline{x}^{\{k\}})$. However, other choices are possible as well. The explanation for the fact that the constant c_1 must be smaller than 0.5 can exemplarily be found in Boyd and Vandenberghe [34, Sec. 9.2 and 9.5]. The demand behind the inequality (3.16) can be summarized in words as follows: The objective function value must be reduced proportional to the step length. The inequality (3.16) is called the *Armijo rule*. Theoretically, there is of course the option to perform an exact line search, i.e.,

$$\alpha^{\{k\}} = \arg\min_{t \geqslant 0} f(\underline{x}^{\{k\}} + t\underline{p}^{\{k\}}). \tag{3.17}$$

In general, (3.17) asks for a one-dimensional optimization problem on its own. Even if a local, instead of a global, minimizer in search direction is sufficient, the solution of this problem will come along with a number of additional, probably very expensive, evaluate calls. Fortunately, another less expensive inexact, so-called *backtracking*, approach is available as well and is presented in Algorithm 3.1.

Extensive studies (see e.g. Boyd and Vandenberghe [34, Sec. 9.5.4]) reveal that exact line search shows typically only a small improvement in terms of the total necessary iteration counts of the non-linear solution procedure compared to the backtracking line search approach. Therefore, and due to the smaller implementation burden, the backtracking line search scheme is used throughout this thesis.

The formulated demand in (3.17) asks in its usual variant for a monotone solution path, where the additional reduction proportional to the step length is introduced to prevent the acceptance of

Algorithm 3.1 BACKTRACKING LINE-SEARCH

Given. Assume that a constant $c_1 \in (0,\ \frac{1}{2})$ and a second constant $\beta \in (0,\ 1)$ are given. Furthermore, a reference merit function value $f_{\text{ref}}^{\{k\}}$ and the evaluated gradient at the last accepted iterate, i.e., $\nabla_{\underline{x}} f(\underline{x}^{\{k\}})$, as well as the current search direction $\underline{p}^{\{k\}}$ are known.

0. *Initialize.* Set the line search iteration counter $l = 0$, initialize the step-length parameter $\alpha^{\{k,\,l\}} = 1.0$ and compute the directional derivative $\mathrm{D}_{\underline{p}^{\{k\}}}(f^{\{k\}}) = \langle \nabla_{\underline{x}} f(\underline{x}^{\{k\}}), \underline{p}^{\{k\}} \rangle$.
1. *Evaluate.* Compute the new trial point $\underline{x}^{+} = \underline{x}^{\{k\}} + \alpha^{\{k,\,l\}} \underline{p}^{\{k\}}$ and evaluate $f(\underline{x}^{+})$.
2. *Check the Armijo rule.* Check (3.16).
   2.1. If the inequality is not fulfilled, set $\alpha^{\{k,l+1\}} = \beta \alpha^{\{k,l\}}$, increase the line-search iteration counter $l \leftarrow l + 1$ and go to Step 1.
   2.2. Otherwise, accept the current trial point and set $\underline{x}^{\{k+1\}} = \underline{x}^{+}$.

steps which are only infinitesimal better than the current iterate even though the local information based on the gradient signalize a possible much better progress. A good explanation can be found in Nocedal and Wright [204, Sec. 3.1]: Imagine an objective function with a minimal function value $f^{\star}$ equal to -1 and, now, imagine at the same time that the chosen algorithm generates a sequence of successively smaller function values in form of $f^{\{k\}} = 5/k$, for $k = 1,\ 2, \ldots$. The given sequence is monotonically decreasing, but would converge to the wrong, non-optimal function value of zero. To avoid such an unfavorable behavior, the so-called *sufficient decrease* condition in (3.16) is necessary. There are also circumstances where such a strict demand for a monotone sequence can be troublesome since it might lead to a much greater number of iterations, just because the algorithm is not allowed to move freely. This observation leads to a number of so-called *non-monotone line search* algorithms. One of the first publications is given by Grippo et al. [120]. The idea therein is easily explained by a slight modification of (3.16). Instead of setting the reference function value equal to the last accepted function value, the expression

$$f_{\text{ref}}^{\{k\}} = \max_{0 \leqslant j \leqslant m(k)} \{f^{\{k-j\}}\} \tag{3.18}$$

is used, where $m(k)$ is initialized by $m(0) = 0$ in Step 0 of Algorithm 3.1. Subsequently, if the iterate is accepted in Step 2, the variable is updated by $m(k+1) = \min\{m(k) + 1, M\}$. Thus, only the largest value of the last M accepted function values is considered for the modified Armijo rule. That is one way how a controllable non-monotone behavior can be added to the algorithm. However, there are more possibilities. See for example Zhang and Hager [295], where an averaging approach of previously accepted function values is followed and, then, the demand is formulated with respect to the averaged reference value. Furthermore, a review over different non-monotone algorithms, including also a derivative-free variant, is given in Eisenträger [75].

In case of a Newton direction the Armijo rule is usually sufficient to achieve global convergence. This is mainly due to the fact that the Newton method already contains a inherent step length information. On the other hand, if a pure gradient or a quasi-Newton method is used, it is possible that the step is not too long, but can be also too short to reach sufficient progress. In such a case, the Armijo rule can be extended to the so-called *(strong) Wolfe conditions* by additionally considering one variant of the following *curvature conditions*

$$\langle \nabla_{\underline{x}} f(\underline{x}^{\{k\}} + \alpha^{\{k\}} \underline{p}^{\{k\}}),\ \underline{p}^{\{k\}} \rangle \geqslant c_2 \langle \nabla_{\underline{x}} f(\underline{x}^{\{k\}}),\ \underline{p}^{\{k\}} \rangle \tag{3.19a}$$

$$\text{or} \qquad |\langle \nabla_{\underline{x}} f(\underline{x}^{\{k\}} + \alpha^{\{k\}} \underline{p}^{\{k\}}),\ \underline{p}^{\{k\}} \rangle| \leqslant c_2 |\langle \nabla_{\underline{x}} f(\underline{x}^{\{k\}}),\ \underline{p}^{\{k\}} \rangle|, \tag{3.19b}$$

where $c_2 \in (0,\ 1)$. The inequality (3.19a) is the default variant which avoids too small steps by the demand that the slope at the new trial point must be less negative or even positive compared to the reference point. This condition can be strengthened by (3.19b): The so-called *strong Wolfe* conditions avoid additionally too long steps, i.e., the positive slope at the new trial point is also not allowed to be too steep. Note that the satisfaction of the Wolfe conditions asks for the fulfillment of the presented curvature conditions (3.19) in combination with the Armijo rule (3.16). The Wolfe conditions also play a decisive role for the BFGS method where the curvature condition enforces the positive definiteness of the matrix estimate as explained in Nocedal and Wright [204].

Finally, it is worth to mention that there are a variety of possible modifications for the case that the assumption of a descent property is violated, the considered Jacobian matrix is (nearly) singular, or the computed Newton direction is in some way too small or too large [120]. See also Nocedal and Wright [204, Sec. 3.4], Wächter and Biegler [272] and Section 6.6 for suggestions on a possible modification of the system matrix or Remark 3.1 for tips concerning a singular Jacobian matrix stemming from a non-linear set of equations.

3.1.2.2. Trust Region Methods

Trust region methods represent the second big family of globalization methods. However, since a line search approach is followed in this thesis, only the basic idea of trust region methods shall be explained here. A much more comprehensive overview of the trust region method is given in the excellent text-book by Conn et al. [54]. However, here the discussion begins by the assumption that a meaningful initial trust region radius $\Delta_{\mathrm{TR}}^{\{0\}}$ is given. Information on how this initial radius can be obtained is also given in Conn et al. [54, Sec. 17.2]. With the initial radius at hand, the following spherical trust region can be defined

$$\mathcal{V}(\Delta_{\mathrm{TR}}^{\{k\}}) = \{\underline{x} \in \mathbb{R}^n \mid \|\underline{x} - \underline{x}^{\{k\}}\| \leqslant \Delta_{\mathrm{TR}}^{\{k\}}\}. \tag{3.20}$$

The region can also have a non-spherical shape by inserting a different norm. However, the discussion shall be restricted to the ℓ_2-norm. Next, a suitable model function is considered, e.g., (3.9) or (3.12), which is supposed to be a valid approximation for the underlying objective function, as long as $\underline{x}^{\{k\}} \in \mathcal{V}(\Delta_{\mathrm{TR}}^{\{k\}})$. Then, the step to the minimum of the model in the defined trust region is computed. This step can lead to a solution on the boundary of the trust region, such that $\|\underline{p}^{\{k\}}\| = \Delta_{\mathrm{TR}}^{\{k\}}$, or, under consideration of the associated full Newton step, it is also possible that the step ends inside the trust region, such that $\|\underline{p}^{\{k\}}\| < \Delta_{\mathrm{TR}}^{\{k\}}$. Now, the quality of the computed step is checked by the quotient

$$\rho_{\mathrm{TR}}^{\{k\}} = \frac{f^{\{k\}} - f(\underline{x}^{\{k\}} + \underline{p}^{\{k\}})}{m^{\{k\}}(\underline{x}^{\{k\}}) - m^{\{k\}}(\underline{x}^{\{k\}} + \underline{p}^{\{k\}})}. \tag{3.21}$$

The step $\underline{p}^{\{k\}}$ is accepted if $\rho_{\mathrm{TR}}^{\{k\}} > \eta_1^{\mathrm{TR}} > 0$. Afterwards, the trust region radius is updated by

$$\Delta_{\mathrm{TR}}^{\{k+1\}} \in \begin{cases} [\Delta_{\mathrm{TR}}^{\{k\}}, \infty) & \text{if } \rho_{\mathrm{TR}}^{\{k\}} \geqslant \eta_2^{\mathrm{TR}}, \\ [\gamma_2^{\mathrm{TR}} \Delta_{\mathrm{TR}}^{\{k\}},\ \Delta_{\mathrm{TR}}^{\{k\}}] & \text{if } \rho_{\mathrm{TR}}^{\{k\}} \in [\eta_1^{\mathrm{TR}},\ \eta_2^{\mathrm{TR}}), \\ [\gamma_1^{\mathrm{TR}} \Delta_{\mathrm{TR}}^{\{k\}},\ \gamma_2^{\mathrm{TR}} \Delta_{\mathrm{TR}}^{\{k\}}] & \text{if } \rho_{\mathrm{TR}}^{\{k\}} < \eta_1^{\mathrm{TR}}, \end{cases} \tag{3.22}$$

and the iteration counter is increased by one, i.e. $k \leftarrow k+1$. Typical values for the introduced constants are $\eta_1^{\mathrm{TR}} = 0.01$, $\eta_2^{\mathrm{TR}} = 0.9$, $\gamma_1^{\mathrm{TR}} = \gamma_2^{\mathrm{TR}} = 0.5$. The trust region algorithm contains the solution of an inequality constrained minimization problem as a sub-step, namely,

$$\underset{\underline{p} \in \mathbb{R}^n}{\text{minimize}} \qquad m^{\{k\}}(\underline{p}), \tag{3.23a}$$

$$\text{s. t.} \qquad \|\underline{p}\| \leqslant \Delta_{\mathrm{TR}}^{\{k\}}. \tag{3.23b}$$

Possible solution procedures for this problem (3.23) can be found in Conn et al. [54, Ch. 7] and lead, in case of the ℓ_2-norm, to a system of the form

$$(\underline{\underline{H}}^{\{k\}} + \lambda_{\mathrm{TR}}^{\star} \underline{\underline{I}})\ \underline{p}^{\{k\}} = -\nabla_{\underline{x}} f^{\{k\}}, \tag{3.24a}$$

$$\|\underline{p}^{\{k\}}\|^2 - (\Delta_{\mathrm{TR}}^{\{k\}})^2 \leqslant 0, \tag{3.24b}$$

$$\lambda_{\mathrm{TR}}^{\star} \geqslant 0, \tag{3.24c}$$

where $\lambda_{\mathrm{TR}}^{\star}$ is the optimal Lagrange multiplier for (3.23), which will be equal to zero if the model minimizer is located inside the trust region. Since the exact solution of this constrained sub-problem and especially the identification of the correct Lagrange multiplier value can quickly become very expensive, there exist also a number of truncated algorithms for large systems of equations which avoid the computation of the exact solution. See for example the truncated conjugate gradient method, the dogleg or double-dogleg methods, or the Lanczos approach for more information. All algorithms can be found in the mentioned book chapter. Combinations of a trust region method and a line search method are also available as proposed by Nocedal and Yuan [205].

3.1.2.3. Levenberg–Marquardt Method

The Levenberg–Marquardt method named after Levenberg [174], Marquardt [187], or also known as the *damped least-squares method* can be seen as a direct ancestor of the trust region algorithm. The theoretical changes are only minor and, as the second name already suggests, the typical derivation is based on a least-squares problem, i.e., (3.2) and the Gauss-Newton model (3.12). Thus, the following system of equations is usually considered

$$\{[\underline{\underline{J}}^{\{k\}}]^{\mathrm{T}}\, \underline{\underline{J}}^{\{k\}} + \mu_{\mathrm{LM}}^{\{k\}}\, \underline{\underline{I}}\}\ \underline{p}^{\{k\}} = -[\underline{\underline{J}}^{\{k\}}]^{\mathrm{T}}\, \underline{r}^{\{k\}}. \tag{3.25}$$

As one can see, this system is closely related to (3.24). The only difference is that the position of the optimal trust region Lagrange multiplier is now taken by the Levenberg-Marquardt parameter $\mu_{\mathrm{LM}}^{\{k\}} \geqslant 0$. In addition, instead of updating the trust-region radius (3.22), the mentioned parameter is updated. For more information, especially concerning a meaningful choice of the initial Levenberg-Marquardt parameter, the reader is referred to the literature on this topic. See for example Fan and Pan [80], Fan [81], Yamashita and Fukushima [287].

3.1.2.4. Pseudo-Transient Continuation

Lastly, the so-called Pseudo-Transient Continuation (PTC) method, also sometimes abbreviated by ΨTC, shall be mentioned. A detailed introduction can be found in Fowler and Kelley [99], Gee et al. [105], Kelley and Keyes [155]. Even though the derivation is very different from the previously discussed trust region and Levenberg-Marquardt method, the evolving system of equations shows large similarities. Consequently, it shall be introduced at this point. Again, the solution of a non-linear set of equations $\underline{r}(\underline{x}) \stackrel{!}{=} \underline{0}$ shall be determined. The idea is to treat the way from the initial point $\underline{x}^{\{0\}}$ to the solution $\underline{x}^{\star}$ as a dynamic transient problem formulated in a pseudo-time $\tau \in \mathbb{R}$ with $\tau \geqslant 0$ which is incremented by the pseudo time increment $\delta_\tau^{\{k\}}$. The final solution of the non-linear problem is the steady-state solution of the pseudo dynamic problem, i.e., it is reached for $\tau \to \infty$. Since only the steady-state solution is important, the possibly introduced time integration error becomes irrelevant. Therefore, the root finding demand can be reformulated as

$$\frac{\partial \underline{x}(\tau)}{\partial \tau} = -[\underline{\underline{V}}(\tau)]^{-1} \underline{r}(\underline{x}(\tau)), \qquad \text{with } \underline{x}(\tau = 0) = \underline{x}^{\{0\}}, \tag{3.26}$$

where $\underline{\underline{V}}$ is a newly introduced, invertible square scaling matrix. Next, the so-called *Rosenbrock time integration* scheme proposed by Rosenbrock [231] is applied to this problem which is an extension of the explicit Runge–Kutta process as described in Gear [103, Sec. 11.2, p. 223]. The simplest form of this scheme is given by

$$\underline{k}_1 = \delta_\tau^{\{k\}} \{ \underline{f}(\underline{x}^{\{k\}}) + \underline{\underline{B}}_1 \, \underline{\underline{A}}(\underline{x}^{\{k\}}) \, \underline{k}_1 \}, \tag{3.27}$$

$$\underline{x}^{\{k+1\}} = \underline{x}^{\{k\}} + \underline{\underline{C}}_1 \underline{k}_1, \tag{3.28}$$

where the vector $\underline{f}(\underline{x}^{\{k\}})$ represents the right hand side of (3.26) and the matrix $\underline{\underline{A}}$ denotes the Jacobian matrix of this right-hand side, viz. $\underline{\underline{A}}(\underline{x}(\tau)) = [\underline{\underline{V}}(\tau)]^{-\mathrm{T}} \underline{\underline{J}}(\underline{x}(\tau))$, where $\underline{\underline{J}}$ denotes the Jacobian matrix of $\underline{r}$. Furthermore, the matrices $\underline{\underline{B}}_1$ and $\underline{\underline{C}}_1$ are here identified as the identity matrix. Thus, after a quick problem reformulation the following update formula is obtained as

$$\underline{x}^{\{k+1\}} = \underline{x}^{\{k\}} - \{ [\delta_\tau^{\{k\}}]^{-1} \underline{\underline{V}}^{\{k\}} + \underline{\underline{J}}^{\{k\}} \}^{-1} \, \underline{r}^{\{k\}}. \tag{3.29}$$

Note that for a pseudo time step $\delta_\tau \to \infty$ the classical Newton scheme is obtained. Additionally, the presented method shows a similarity with (3.24) and (3.25). The main difference lies in the fact that no matrix-matrix product must be formed and that the right-hand side remains unmodified. This can have a beneficial impact if the Jacobian matrix is (nearly) singular close to the

solution such that the right hand side of (3.25) could become distorted and may indicate a non-stationary point falsely as stationary. Classically, the PTC method uses the so-called *switched evolution relaxation* (SER) method or the so-called *temporal truncation error* (TTE) to update the pseudo time-step δ_τ. Further details can be found in Kelley et al. [156]. However, since these correction schemes are applied after the new step has already been computed, they can still lead to divergence. This can happen since the evaluation at the new iterate may lead to a non-physical and highly distorted state.

An alternative interpretation which enables the access via the presented optimization algorithms is given by formulating a so-called transient residual

$$\underline{r}_\mathrm{t}(\underline{x}) = [\delta_\tau^{\{k\}}]^{-1} \underline{\underline{V}}^{\{k\}} \, (\underline{x} - \underline{x}^{\{k\}}) + \underline{r}(\underline{x}) \stackrel{!}{=} \underline{0}. \tag{3.30}$$

In this way the PTC method can be accessed by globalization methods such as line search. The interested reader is referred to Ceze and Fidkowski [44], Modisette [198].

3.2. Constrained Optimization

In this section it will be demonstrated that many of the presented unconstrained optimization approaches can be generalized such that they become applicable to constrained problems. Since the focus of this thesis is on contact problems which can be interpreted as a special type of an inequality constrained optimization problem, the following general problem shall be considered

$$\underset{\underline{x} \in \mathbb{R}^n}{\text{minimize}} \quad f(\underline{x}) \tag{3.31a}$$

$$\text{subject to} \quad g^i(\underline{x}) \geqslant 0 \qquad \forall i \in \{1, \ldots, m\}, \tag{3.31b}$$

where m is used to denote the number of considered inequality constraints and $m < n$ shall hold throughout the entire thesis. In the remainder of this work $(\underline{x}^\star, \underline{\lambda}^\star) \in \mathbb{R}^n \times \mathbb{R}^m$ indicates a so-called *Karush–Kuhn–Tucker* (KKT) pair of problem (3.31), i.e., a pair of primal and dual variables which satisfies the following conditions

$$\nabla_{\underline{x}} \mathscr{L}(\underline{x}^\star, \underline{\lambda}^\star) = \underline{0}, \qquad g^i(\underline{x}^\star) \geqslant 0, \qquad [\lambda^\star]^i \geqslant 0, \qquad \langle \underline{\lambda}^\star, \, \underline{g}(\underline{x}^\star) \rangle = 0, \tag{3.32}$$

for all $i \in \{1, \ldots, m\}$. Thus, at the solution all constraints must be feasible and their associated Lagrange multipliers λ^i with $i \in \{1, 2, \ldots, m\}$ must be non-negative. Note that the last condition in (3.32), the so-called *complementarity condition*, ensures on the one hand that all Lagrange multipliers associated to inactive constraints, i.e., constraints with a positive value, must be zero. On the other hand, it enforces that the constraint values of all strictly active constraints, i.e., constraints with a positive Lagrange multiplier value, must be zero. Thus, if $\mathcal{S}$ is the index set of all constraints, viz., $\mathcal{S} = \{1, \ldots, m\}$, then the index set of all active constraints at the solution $\underline{x}^\star$ follows as

$$\mathcal{A}_0 = \{i \in \mathcal{S} : \; g^i(\underline{x}^\star) = 0\}, \tag{3.33}$$

and the index set of all strongly active constraints can be defined as a sub-set of $\mathcal{A}_0$ by

$$\mathcal{A}_+ = \{i \in \mathcal{A}_0 : \ [\lambda^\star]^i > 0\}. \tag{3.34}$$

If the sets coincide at the solution, i.e., $\mathcal{A}_0 \equiv \mathcal{A}_+$, then *strict complementarity* holds. Finally, the *Lagrangian function*, denoted as $\mathscr{L} : \mathbb{R}^{n\times m} \to \mathbb{R}$, is obtained by

$$\mathscr{L}(\underline{x}, \underline{\lambda}) = f(\underline{x}) - \lambda^i \ \min\{g_i(\underline{x}), \ \frac{1}{c}\lambda_i\} \overset{(\dagger)}{=} f(\underline{x}) - \langle \underline{\lambda}^{\mathcal{A}}, \ \underline{g}^{\mathcal{A}} \rangle - \frac{1}{c}\|\underline{\lambda}^{\mathcal{I}}\|^2, \tag{3.35}$$

where a new scalar constant $c > 0$, the so-called *regularization parameter*, has been introduced. Furthermore, the current index set $\mathcal{A}$ is used in the last step $(\dagger)$ as a superscript which is defined as

$$\mathcal{A}(\underline{x}, \underline{\lambda}) = \{i \in \mathcal{S} : \ \lambda^i - cg^i \geqslant 0\}. \tag{3.36}$$

This set can be interpreted as an educated guess for the final active set. This guess is necessary as long as the final KKT pair has not yet been determined and, hence, the KKT conditions do not yet hold. As a direct consequence $\mathcal{A}_+ \subseteq \mathcal{A} \subseteq \mathcal{A}_0$ can be expected to hold at some point during the convergence to the optimal KKT pair. Another set is the index set of all inactive constraints which is directly obtained by $\mathcal{I} = \mathcal{S}\backslash\mathcal{A}$.

A detailed motivation for the regularization parameter or the inequalities in (3.35) and (3.36) will be presented in a moment. First the attention is on (3.32). The Karush–Kuhn–Tucker conditions represent the *necessary first order optimality conditions for constrained problems*, where the first condition in (3.32) yields

$$\underline{0} = \nabla_{\underline{x}}\mathscr{L}(\underline{x}^\star, \underline{\lambda}^\star) = \nabla_{\underline{x}} f(\underline{x}^\star) - \nabla_{\underline{x}}\underline{g}^{\mathcal{A}}(\underline{x}^\star) \ \underline{\lambda}^{\star\mathcal{A}} \quad \Leftrightarrow \quad \nabla_{\underline{x}} f(\underline{x}^\star) = \nabla_{\underline{x}}\underline{g}^{\mathcal{A}}(\underline{x}^\star) \ \underline{\lambda}^{\star\mathcal{A}}, \tag{3.37}$$

since $\lambda_i^\star = 0, \ \forall \ i \in \mathcal{I}$. The illustrative meaning of this formula is that the gradient of the objective function and the gradients of the active constraints must be collinear at the KKT point. Now, the task of the associated optimal Lagrange multipliers is to scale the magnitudes of the constraint gradients in such a way that the objective function gradient and the linear combination of the constraint gradients vanishes at the solution. It is obvious that for this linear combination a variety of possible very different values of Lagrange multipliers might be suitable. However, if the so-called *linear independence constraint qualification* (LICQ) holds, then the optimal Lagrange multiplier vector $\underline{\lambda}^\star$ is unique. The definition of this important qualification is stated as follows:

Definition 3.1. Assume a feasible point $\underline{x}$ as well as its associated active set $\mathcal{A}$ of feasible active constraints are given. Then the *linear independence constraint qualification* (LICQ) will hold if the set of active constraint gradients $\{\nabla_{\underline{x}} g^i(\underline{x}), i \in \mathcal{A}\}$ is linearly independent.

This important definition is taken from Nocedal and Wright [204, Sec. 12.2]. Furthermore, the interested reader is kindly referred to Nocedal and Wright [204, Sec. 12.4] for more information

on the proof that the KKT-conditions are indeed the first order optimality conditions. Here, the attention is next drawn to the *necessary second order optimality conditions for constrained optimization*. Firstly, a cone with first order feasible directions at a feasible point $\underline{x}$ must be defined. This cone of feasible directions as well as the related cone of critical directions at the KKT-pair is defined as

$$\mathcal{C}(\underline{x}) = \{\underline{v} : \langle \underline{v},\ \nabla_{\underline{x}} g^i(\underline{x})\rangle \geqslant 0 \quad \forall i \in \mathcal{A}\}, \tag{3.38a}$$

$$\mathcal{C}^\star(\underline{x}^\star, \underline{\lambda}^\star) = \{\underline{v} \in \mathcal{C}(\underline{x}^\star) : \langle \underline{v},\ \nabla_{\underline{x}} g^i(\underline{x}^\star)\rangle = 0,\ \forall i \in \mathcal{A}^\star \text{ with } \lambda_i^\star > 0\}. \tag{3.38b}$$

It must be noted that (3.38) directly implies that $\langle \underline{v},\ \nabla_{\underline{x}} g^i(\underline{x}^\star)\rangle \geqslant 0$ for $\lambda^\star = 0$ and $i \in \mathcal{A}^\star$. However, these sets need some further explanation: The set of feasible directions given in (3.38a) is considered first: At the solution it is known from the first order optimality condition that $\nabla_{\underline{x}} f^\star$ is equal to $\langle \nabla_{\underline{x}} \underline{g}^\star,\ \lambda^\star\rangle$ and, hence, each direction $\underline{v} \in \mathcal{C}(\underline{x}^\star)$ under consideration of a linear model for the objective function leads to either an increase or to no change due to $\langle \nabla_{\underline{x}} f^\star, \underline{v}\rangle \geqslant 0$. Therefore, in the latter case this first order information is not enough to identify $\underline{x}^\star$ as an optimal point. Actually, all feasible directions for which this decision is difficult are summarized in the critical cone (3.38b) due to the fact that from (3.37) it directly follows

$$\underline{v} \in \mathcal{C}^\star \Rightarrow \qquad \langle \underline{v}, \nabla_{\underline{x}} f^\star\rangle = \langle \underline{v}, \nabla_{\underline{x}} \underline{g} \underline{\lambda}^\star\rangle = 0. \tag{3.39}$$

This at hand it becomes possible to state the following theorem:

Theorem 3.4. *Second-Order Necessary Conditions for Constrained Optimization.* Under the assumption that the objective function f as well as the constraints g^i are twice continuously differentiable in a neighborhood around the KKT-pair, and by assuming that the LICQ is fulfilled, then the point $\underline{x}^\star$ is a local solution and $\underline{\lambda}^\star$ is its associated Lagrange multiplier vector such that the KKT conditions are satisfied, whenever

$$\langle \underline{v}, \nabla^2_{\underline{x}\,\underline{x}} \mathscr{L}(\underline{x}^\star, \underline{\lambda}^\star)\ \underline{v}\rangle \geqslant 0, \qquad \forall \underline{v} \in \mathcal{C}^\star(\underline{x}^\star, \underline{\lambda}^\star), \underline{v} \neq \underline{0} \tag{3.40}$$

holds.

Finally, similar to the unconstrained case, the second order sufficient conditions are stated. Note that these conditions do not require that the LICQ hold. They can be formulated as

Theorem 3.5. *Second-Order Sufficient Conditions for Constrained Optimization.* It shall be assumed that for some feasible $\underline{x}^\star$ there exists a Lagrange multiplier vector $\underline{\lambda}^\star$ such that the KKT conditions are fulfilled. Furthermore,

$$\langle \underline{v}, \nabla^2_{\underline{x}\,\underline{x}} \mathscr{L}(\underline{x}^\star, \underline{\lambda}^\star)\ \underline{v}\rangle > 0, \qquad \forall \underline{v} \in \mathcal{C}^\star(\underline{x}^\star, \underline{\lambda}^\star),\ \underline{v} \neq \underline{0} \tag{3.41}$$

shall hold. Then $\underline{x}^\star$ is a strict local minimizer for the constrained problem (3.31) and $\underline{\lambda}^\star$ is its associated optimal Lagrange multiplier.

For a much deeper insight, for example concerning the Mangasarian–Fromovitz constraint qualification (MFCQ) which represents a generalization of the LICQ, as well as for a proof of all these theorems, the reader is kindly referred to Nocedal and Wright [204, Ch. 12].

Up to here, mainly the KKT point has been considered where $\lambda_i^\star = 0,\ \forall i \in \mathcal{I}$ holds such that the inactive part in (3.35) vanishes naturally. However, during an iterative non-linear solution scheme this inactive part as well as the shown active-set decision in (3.36) is a crucial ingredient. Therefore, next a short derivation for the used inactive extension is given. This derivation is achieved by a reformulation of the problem (3.31) as an equality constrained problem. In a first step, a vector $\underline{z} \in \mathbb{R}^m$ with additional auxiliary variables shall be formally introduced leading to

$$\underset{\underline{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\underline{x}) \tag{3.42a}$$

$$\text{subject to} \quad g^i(\underline{x}) - (z^i)^2 = 0 \qquad \forall i \in \{1, \dots, m\}. \tag{3.42b}$$

Under consideration of this equality constrained problem the so-called *augmented Lagrangian* can be stated

$$\mathscr{L}_\mathrm{c}(\underline{x}, \hat{\underline{s}}, \underline{\lambda}) = f(\underline{x}) - \langle \underline{\lambda},\ \underline{g}(\underline{x}) - \hat{\underline{s}} \rangle + \frac{c}{2} \|\underline{g}(\underline{x}) - \hat{\underline{s}}\|^2, \tag{3.43}$$

where the vector $\hat{\underline{s}}$ with the positive components $\hat{s}^i = (z^i)^2$ is inserted. These variables are called *slack variables*, or just *slacks*. Now, the augmented Lagrangian (3.43) is minimized with respect to these non-negative slack variables yielding the new problem

$$\underset{\hat{\underline{s}}\geqslant 0}{\text{minimize}} \quad \langle \underline{\lambda},\ \underline{g}(\underline{x}) - \hat{\underline{s}} \rangle - \frac{c}{2} \|\underline{g}(\underline{x}) - \hat{\underline{s}}\|^2, \tag{3.44a}$$

$$\Rightarrow \quad -\underline{\lambda} + c(\underline{g}(\underline{x}) - \tilde{\underline{s}}) = \underline{0}, \tag{3.44b}$$

$$\Rightarrow \quad \tilde{\underline{s}} = \underline{g}(\underline{x}) - \frac{1}{c}\underline{\lambda}. \tag{3.44c}$$

Under the prerequisite that $\hat{\underline{s}} \geqslant 0$ must hold, the final result follows as $(\hat{s}^i)^\star = \max\{0,\ \tilde{s}^i\}$. This result inserted into (3.43) leads to

$$\mathscr{L}_\mathrm{c}(\underline{x}, \underline{\lambda}) = \mathscr{L}_\mathrm{c}(\underline{x}, \hat{\underline{s}}^\star(\underline{x}, \underline{\lambda}), \underline{\lambda}) = f(\underline{x}) + \frac{1}{2c} \sum_{i=1}^{m} \{[\max\{0, \lambda^i - c g^i\}]^2 - (\lambda^i)^2\} \tag{3.45a}$$

$$= f(\underline{x}) - \langle \underline{\lambda}^{\mathcal{A}},\ \underline{g}^{\mathcal{A}} \rangle + \frac{c}{2} \|\underline{g}^{\mathcal{A}}\|^2 - \frac{1}{2c} \|\underline{\lambda}^{\mathcal{I}}\|^2. \tag{3.45b}$$

From this derivation and under consideration of the optimal slack variable $\hat{\underline{s}}^\star$, the Lagrangian stated in (3.35) can be directly deduced. The derivation above can be also found in a slightly different form in Bertsekas [23, Sec. 3.1]. The augmented Lagrangian for inequality constraints is well-known in the literature and has also been successfully used by Bertsekas [22], Buys [41], Gill et al. [107], Glad and Polak [111], Rockafellar [230], for instance.

3.2.1. Penalty Approach

There are different ways how to tackle the constraints in problem (3.31). Actually, even the presented way incorporating the constraints by building up a Lagrangian or an augmented Lagrangian function is by far not the only possibility. For example, one truncated variant of the augmented Lagrangian is the so-called *penalty approach*. In contrast to (3.43) the slacks depending on the Lagrange multiplier values can no longer be used. Instead the penalty function is defined as

$$\mathscr{P}_{\mathrm{c}}(\underline{x}) = f(x) + \frac{c}{2}\sum_{i=1}^{m}[\max\{0, -g^i(\underline{x})\}]^2, \tag{3.46}$$

where $c > 0$ is now called a *penalty parameter* rather than regularization parameter. This quadratic penalty function has been introduced for the first time by Courant [55]. The obvious advantage is that this function only depends on the primal solution variable $\underline{x}$. The big drawback is that the solution quality depends strongly on the chosen penalty parameter. To reach the exact solution of (3.31), the penalty parameter must tend to infinity, since the enforcement of the constraints becomes stricter by increasing the penalty parameter.

For the application considered here, it shall be assumed that a first order optimal point for (3.46) can be found for each global iteration, i.e., for each global iteration k it is possible to find a solution such that $\|\nabla_{\underline{x}}\mathscr{P}_{\mathrm{c}}\| < \mathrm{TOL}_{\mathrm{p}}^{\{k\}}$ with $\nabla_{\underline{x}}\mathscr{P}_{\mathrm{c}}(\underline{x}) = \nabla_{\underline{x}}f(\underline{x}) + c^{\{k\}}\nabla_{\underline{x}}\underline{g}^{\mathcal{A}_{\mathrm{p}}}(\underline{x})\ \underline{g}^{\mathcal{A}_{\mathrm{p}}}(\underline{x})$, where $\mathcal{A}_{\mathrm{p}} = \{i \in \mathcal{S} : \ g^i \leqslant 0\}$ is defined in accordance with (3.36). Furthermore, for the considered sequences $\{c^{\{k\}}\}$ and $\{\mathrm{TOL}_{\mathrm{p}}^{\{k\}}\}$ it shall hold that $\mathrm{TOL}_{\mathrm{p}}^{\{k\}} \to 0$ and $c^{\{k\}} \to \infty$. If these assumptions are fulfilled then each limit point $\underline{x}^\star$ of the generated sequence $\{\underline{x}^{\{k\}}\}$ is either a non-optimal stationary point of the penalty term $\|\underline{g}_i^{\mathcal{A}}\|^2$, or a feasible point. If the point is feasible and the constraint gradients are linearily independent, then the point $\underline{x}^\star$ is the desired KKT point. For KKT points and under consideration of an infinite sub-sequence of iterations denoted by $\mathcal{K}_{\mathrm{p}}$ it follows

$$\lim_{k\in\mathcal{K}_{\mathrm{p}}} \underline{x}^{\{k\}} = \underline{x}^\star \qquad \text{and} \qquad \lim_{k\in\mathcal{K}_{\mathrm{p}}} c^{\{k\}}\max\{0, -g_i(\underline{x}^{\{k\}}) = \lambda_i^\star, \tag{3.47}$$

where $\lambda_i^\star$ is the associated optimal Lagrange multiplier value (see Nocedal and Wright [204, Theorem 17.2]). This identification will become handy when the condition number of the Hessian shall be investigated in more detail. Therefore, (3.47) will be reconsidered in Section 3.2.3.2. For a more detailed discussion of the penalty approach the reader is again referred to Nocedal and Wright [204, Ch. 17.1].

3.2.2. Lagrange Multiplier Function

Within this thesis the dual Lagrange multiplier variables will be mainly treated as primary variables, however, this is not strictly necessary. It is also possible to construct a so-called Lagrange multiplier function $\underline{\lambda}(\underline{x})$ which is twice continuously differentiable and is expected to converge to the correct Lagrange multiplier vector $\underline{\lambda}^\star$ for $\underline{x}^{\{k\}} \to \underline{x}^\star$ as long as the MFCQ hold. This idea was proposed for equality constraints by Fletcher [92], Tapia [259] and was later extended to

inequality constrained problems by Glad and Polak [111], Lucidi [182]. At this point the extension used by Facchinei and Lucidi [79], Lucidi [182] shall be briefly presented which leads to the following definition of the Lagrange multiplier function

$$\underline{\lambda}(\underline{x}) = [\underline{\underline{N}}(\underline{x})]^{-1}[\nabla_{\underline{x}}\underline{g}(\underline{x})]^{\mathrm{T}}\nabla_{\underline{x}}f(\underline{x}), \tag{3.48}$$

where the matrix $\underline{\underline{N}}(\underline{x}) \in \mathbb{R}^{m\times m}$ is defined as

$$\underline{\underline{N}}(\underline{x}) = [\nabla_{\underline{x}}\underline{g}(\underline{x})]^{\mathrm{T}}\nabla_{\underline{x}}\underline{g}(\underline{x}) + \gamma_1^{\lambda}\,\mathrm{diag}[\{g^i(\underline{x})\}^2] + \gamma_2^{\lambda}\sum_{i=1}^{m}\max\{0,-g^i(\underline{x})\}^3\underline{\underline{I}}. \tag{3.49}$$

The presented definition of matrix $\underline{\underline{N}}(\underline{x})$ might look quite complex at a first glance, but luckily the different terms can easily be explained. Therefore, the following least-squares minimization problem is considered

$$\breve{\underline{\lambda}}(\underline{x}) = \underset{\underline{\lambda}\in\mathbb{R}^m}{\arg\min}\,\frac{1}{2}\|\nabla_{\underline{x}}f(\underline{x}) - \nabla_{\underline{x}}\underline{g}(\underline{x})\;\underline{\lambda}\|^2 \tag{3.50}$$

which directly leads to

$$\breve{\underline{\lambda}}(\underline{x}) = \{[\nabla_{\underline{x}}\underline{g}(\underline{x})]^{\mathrm{T}}\nabla_{\underline{x}}\underline{g}(\underline{x})\}^{-1}[\nabla_{\underline{x}}\underline{g}(\underline{x})]^{\mathrm{T}}\nabla_{\underline{x}}f(\underline{x}). \tag{3.51}$$

This result coincides with (3.48) for $\gamma_1^{\lambda} = \gamma_2^{\lambda} = 0$. Next, γ_1^{λ} is set to a value larger than zero while γ_2^{λ} shall be still equal to zero. Then, the multiplier function introduced in Glad and Polak [111] is revealed which is defined at any point where the gradients of the active constraints are linearly independent. Finally, the last regularization term is added to ensure a well-defined Lagrange multiplier function also when at least one constraint is violated. This last term is especially important if globalization techniques are addressed [182]. The presented Lagrange multiplier function asks for the solution of a $m \times m$ system of equations. Fortunately, this is easily achievable in the case of contact problems since the number of constraints is almost always much smaller than the number of primal degrees of freedom and, hence, the system stays small and can be solved efficiently.

Remark 3.2. The reason why it is not straight forward to apply this interesting approach to contact problems is that this method asks for the evaluation of the constraint gradients not only at infeasible but also at feasible points, i.e., at local positions where the gap is positive. That can be a problem since the necessary projections between master and slave surface might not be defined, since the slave body slides over an edge or the projection is not evaluated due to the chosen search radius of the contact pairing algorithm. However, in a student project by Schulze [239] this idea has been successfully applied under consideration of the mortar-like contact formulation presented in Chapter 4. During that study it has become apparent that the regularization terms are of major importance to avoid non-physical oscillations of the Lagrange multiplier values.

3.2.3. Local Iterative Solution Methods

In this section a basic iterative solution approach for constrained optimization problems shall be presented. Therefore, the discussion is mainly restricted to Newton's method. However, some possible modification will be mentioned as well.

3.2.3.1. Sequential Quadratic Programming

The system of equations which naturally evolves when Newton's method is applied to (3.31) can also be interpreted in a different way. The reader is reminded of the quadratic model in (3.9) and the discussion about the fact that the Newton step represents a step to the minimizer of the underlying quadratic model function. This idea leads to the so-called *Recursive Quadratic Programming* (RQP), *Sequential Quadratic Programming* or even sometimes called *Successive Quadratic Programming* (SQP) approaches. The SQP approach considers the following quadratic problem in each step

$$\underset{\underline{p}_x \in \mathbb{R}^n}{\text{minimize}} \qquad f^{\{k\}} + \langle \nabla_{\underline{x}} f^{\{k\}}, \underline{p}_x \rangle + \frac{1}{2} \langle \underline{p}_x, \nabla^2_{\underline{x}\,\underline{x}} \mathscr{L}^{\{k\}} \underline{p}_x \rangle \tag{3.52a}$$

$$\text{s. t.} \qquad g^i(\underline{x}^{\{k\}}) + \langle \nabla_{\underline{x}} g^i(\underline{x}^{\{k\}}), \underline{p}_x \rangle \geqslant 0, \qquad i \in \mathcal{S}, \tag{3.52b}$$

where $\underline{p}_x$ denotes the step of the primal variables. Based on this quadratic constrained minimization problem, the corresponding Lagrangian function reads

$$\begin{aligned} \mathscr{L}(\underline{x}^{\{k\}} + \underline{p}_x, \underline{\lambda}_+) = & f^{\{k\}} + \langle \nabla_{\underline{x}} f^{\{k\}}, \underline{p}_x \rangle + \frac{1}{2} \langle \underline{p}_x, \nabla^2_{\underline{x}\,\underline{x}} \mathscr{L}^{\{k\}} \underline{p}_x \rangle \\ & - \langle \underline{g}^{\mathcal{A}\{k\}}, \underline{\lambda}^{\mathcal{A}}_+ \rangle - \langle \nabla_{\underline{x}} \underline{g}^{\mathcal{A}\{k\}}\, \underline{\lambda}^{\mathcal{A}}_+, \underline{p}_x \rangle \\ & - \frac{1}{c} \langle \underline{\lambda}^{\mathcal{I}}_+, \underline{\lambda}^{\mathcal{I}}_+ \rangle \end{aligned} \tag{3.53}$$

where $\underline{\lambda}_+$ denotes the (trial) Lagrange multiplier value for the quadratic optimization problem which is supposed to coincide with the Lagrange multiplier vector at the new iterate $\underline{\lambda}^{\{k+1\}}$ if a simple local scheme is applied, i.e., without considering any globalization technique. Since this Lagrange multiplier enters the equation only linearly, it is also possible to split the trial Lagrange multiplier into an old, previously accepted part and an incremental update by replacing $\underline{\lambda}_+$ with $\underline{\lambda}^{\{k\}} + \underline{p}_\lambda$. This split can become handy for globalization methods. Under consideration of the Lagrangian model (3.53) derived from (3.52), the related first order optimality condition can be generally written as

$$\underline{\underline{K}}(\underline{x}^{\{k\}}, \underline{\lambda}^{\{k\}})\, \underline{p} = -\underline{r}^{\mathrm{SQP}}(\underline{x}^{\{k\}}, \underline{\lambda}^{\{k\}}), \tag{3.54}$$

where the matrix $\underline{\underline{K}}$ is identified by

$$\underline{\underline{K}}^{\{k\}} = \begin{pmatrix} \nabla^2_{\underline{x}\,\underline{x}} \mathscr{L}^{\{k\}} & -\nabla_{\underline{x}} \underline{g}^{\mathcal{A}\{k\}} & \underline{\underline{0}} \\ -[\nabla_{\underline{x}} \underline{g}^{\mathcal{A}\{k\}}]^{\mathrm{T}} & \underline{\underline{0}} & \underline{\underline{0}} \\ \underline{\underline{0}} & \underline{\underline{0}} & \underline{\underline{I}} \end{pmatrix} \tag{3.55}$$

and the right hand side or residual vector is either given by

$$\underline{r}^{\mathrm{SQP}} = \begin{pmatrix} -\nabla_{\underline{x}} f^{\{k\}} \\ \underline{g}^{\mathcal{A}\{k\}} \\ \underline{0} \end{pmatrix} \tag{3.56a}$$

or by

$$\underline{r}^{\mathrm{SQP}} = \begin{pmatrix} -\nabla_{\underline{x}} f^{\{k\}} + \nabla_{\underline{x}} \underline{g}^{\mathcal{A}\{k\}} \underline{\lambda}^{\{k\}} \\ \underline{g}^{\mathcal{A}\{k\}} \\ -\underline{\lambda}^{\mathcal{I}\{k\}} \end{pmatrix} \tag{3.56b}$$

depending on the definition of the solution vector $\underline{p} = (\underline{p}^{\mathcal{A}},\ \underline{p}^{\mathcal{I}})^{\mathrm{T}}$ in (3.54). If the active solution vector is defined as $\underline{p}^{\mathcal{A}} = (\underline{p}_x,\ \underline{\lambda}_+^{\mathcal{A}})^{\mathrm{T}}$, then (3.56a) holds. Otherwise, if $\underline{p}^{\mathcal{A}} = (\underline{p}_x,\ \underline{p}_\lambda^{\mathcal{A}})^{\mathrm{T}}$ holds, then (3.56b) must be considered. The inactive contributions can be easily added due to the fact that the corresponding equations are decoupled and dependent solely on the inactive Lagrange multipliers, thus,

$$\underline{\lambda}_+^{\mathcal{I}} = \underline{0} \qquad \text{or} \qquad \underline{p}_\lambda^{\mathcal{I}} = -\lambda^{\mathcal{I}\{k\}} \tag{3.57}$$

follows. As long as the KKT-matrix (3.54) is non-singular, the iteration scheme is well-defined. In contrast to an equality constrained problem, an inequality constrained formulation asks for a meaningful guess of the final active set. As long as this current active set $\mathcal{A}^{\{k\}}$ following (3.36) is not yet equal to the final active set, the system of equations shows still a quite heavy change each time a constraint joins or leaves the current active set estimate. Only as soon as the active set has converged, the remaining system (3.54) acts like a typical Newton scheme and quadratic convergence can be expected under certain preliminaries. This is summarized in a theorem originally proposed by Robinson [228] and restated in Nocedal and Wright [204, Theorem 18.1]:

Theorem 3.6. Suppose that $\underline{x}^\star$ is a local solution of (3.31) at which the KKT conditions are fulfilled for some $\underline{\lambda}^\star$. Furthermore, suppose that the LICQ, the strict complementarity condition and the second-order sufficient conditions hold at the KKT pair $(\underline{x}^\star, \underline{\lambda}^\star)$. Then, if the iterate $(\underline{x}^{\{k\}}, \underline{\lambda}^{\{k\}})$ is sufficiently close to the solution, there exists a local solution of the subproblem (3.52) whose current active set $\mathcal{A}^{\{k\}}$ coincides with the active set $\mathcal{A}(\underline{x}^\star) = \mathcal{A}_+$ of the non-linear program (3.31) at $\underline{x}^\star$.

Remark 3.3. The crucial part of Theorem 3.6 for contact problems is the prerequisite that strict complementarity must hold. If this assumption is not satisfied and the two contacting bodies are only in touch, i.e., the gap between them is zero, but no (meaningful) force is transferred between them and thus the final Lagrange multiplier values $\lambda_i^{\mathcal{A}_0}$ are zero or close to zero, then the presented iterative method often shows a very poor performance and converges only very slowly to the final solution or the active set starts to cycle.

By the way, the quasi-Newton methods presented in Section 3.1.1 are also applicable to the KKT system and are typically used to generate an estimate for the $\nabla^2_{\underline{x}\,\underline{x}}\mathscr{L}$ matrix. However,

the direct application might be more complicated. For example, as long as the matrix $\nabla^2_{\underline{x}\,\underline{x}}\mathscr{L}$ is positive definite, an estimate via the BFGS method will be mostly a good idea. As soon as the real curvature matrix exhibits negative eigenvalues, the classical BFGS method might perform poorly. Therefore, different remedies have been proposed. For instance, the BFGS update is sometimes skipped if the secant condition implies it or a so-called damped BFGS updating is considered (see Nocedal and Wright [204, Sec. 18.3]). However, even though the last option seems to work pretty well in many cases, it is still not able to properly represent the case that the underlying Hessian is no longer positive definite. Therefore, it might be a better idea to switch to so-called symmetric rank-one (SR1) update routines which do not rely on the positive definiteness assumption.

Furthermore, the linear system (3.54) together with (3.56a) can also be applied if a Lagrange multiplier function following (3.48) shall be used. In this case, the second order derivative matrix is evaluated dependent on this multiplier function, i.e. $\nabla^2_{\underline{x}\,\underline{x}}\mathscr{L}[\underline{x}^{\{k\}}, \underline{\lambda}(\underline{x}^{\{k\}})]$, and the solution vector parts $\underline{\lambda}_+$ or $\underline{p}_\lambda$ are just used as auxiliary variables which are discarded after the solution of the linear system. This approach is quite common, see for example Facchinei and Lucidi [79], Ulbrich [264]. Alternatively, it is also possible to formulate the entire system completely in dependence on the primal variable by inserting the Lagrange multiplier function into the Lagrangian definition. The latter idea has been followed in the original work by Fletcher [92] subject to equality constrained problems. However, in contrast to (3.54) the latter attempt can become more involved since terms such as $\nabla_{\underline{x}}\underline{\lambda}$ must be computed, which is possible but can be computationally quite expensive.

3.2.3.2. Solution of the Penalty Approach

In case of the penalty method (3.46) a formulation based on the primal variables $\underline{x}$ is given. Therefore, the related Hessian follows as

$$\nabla^2_{\underline{x}\,\underline{x}}\mathscr{P}_\mathrm{c}(\underline{x}) = \nabla^2_{\underline{x}\,\underline{x}}f(\underline{x}) + c\sum_{i=1}^{|\mathcal{A}_\mathrm{P}|} g_i(\underline{x})\ \nabla^2_{\underline{x}\,\underline{x}}g_i(\underline{x}) + c\nabla_{\underline{x}}\underline{g}^{\mathcal{A}_\mathrm{P}}(\underline{x})\ [\nabla_{\underline{x}}\underline{g}^{\mathcal{A}_\mathrm{P}}(\underline{x})]^\mathrm{T}. \tag{3.58}$$

Typically, the associated Newton system has the drawback that the conditioning of the Hessian is badly influenced by a rising penalty parameter. Luckily, the impact on the condition number can actually easily be avoided: Under consideration of (3.47), the first two terms in (3.58) can be immediately identified as an estimate for $\nabla^2_{\underline{x}\,\underline{x}}\mathscr{L}$. Thus, it can be concluded that only the eigenvalues of the rightmost matrix in (3.58) are potentially badly influenced by the penalty parameter close to the solution. If, as suggested, a penalty parameter sequence $\{c^{\{k\}}\}$ with $c^{\{k\}} \to \infty$ is followed, the computation of the Newton direction $\underline{p}$ will become more and more inaccurate due to this term. This is especially critical if iterative solution methods for the linear system are applied. Fortunately, an easy reformulation is possible by introducing an auxiliary vector $\underline{z} = c\nabla_{\underline{x}}\underline{g}^{\mathcal{A}_\mathrm{P}}\underline{p}$ yielding the new system of equations

$$\begin{pmatrix} \nabla^2_{\underline{x}\,\underline{x}}f(\underline{x}) + c\sum\limits_{i=1}^{|\mathcal{A}_\mathrm{P}|} g_i(\underline{x})\ \nabla^2_{\underline{x}\,\underline{x}}g_i(\underline{x}) & \nabla_{\underline{x}}\underline{g}^{\mathcal{A}_\mathrm{P}}(\underline{x}) \\ [\nabla_{\underline{x}}\underline{g}^{\mathcal{A}_\mathrm{P}}(\underline{x})]^\mathrm{T} & -\frac{1}{c}\underline{\underline{I}} \end{pmatrix} \begin{pmatrix} \underline{p} \\ \underline{z}_\mathrm{P} \end{pmatrix} = \begin{pmatrix} -\nabla_{\underline{x}}\mathscr{P}_\mathrm{c}(\underline{x}) \\ \underline{0} \end{pmatrix}. \tag{3.59}$$

Hence, it can be concluded that (3.59) represents a well conditioned reformulation. However, the computed Newton direction can still be of bad quality if the identification (3.47) does not (yet) hold and, thus, the upper left block in the system matrix of (3.59) is only a rough estimate for the Hessian of the Lagrangian (3.35). Another drawback is that the better conditioning asks for the solution of a saddle point system of equations which might demand for special solution techniques. The interested reader is referred to Gould [113] for a more comprehensive discussion.

Remark 3.4. The auxiliary vector $\underline{z}_{\mathrm{p}}$ can be easily identified as an estimate for $-\underline{\lambda}$ close to the solution. This idea will be taken into account again in Chapter 5.

3.2.3.3. Interior Point Methods

Within this thesis only ideas from the SQP and the penalty approach will be used. Both of these approaches rely on an active set strategy, which is either built up on (3.36), or an inactive-active decision solely based on the constraint value (3.46). However, there is also another class of methods which circumvents these combinatorial difficulties and transforms these active-set methods to a continuation method [2] called *interior-point* or *barrier method*. Even though these methods are not used in this thesis, their underlying idea shall be briefly explained and why they are not used for the contact problems considered here. First of all, there is a number of different interpretations and ideas which all lead to the final interior-point system of equations, but probably the best way to start the discussion is by considering an already introduced idea. Therefore, (3.42) shall be revisited. The reformulation with slacks can be also stated as

$$\underset{\underline{x}\in\mathbb{R}^n,\underline{s}\in\mathbb{R}^m}{\text{minimize}} \qquad f(\underline{x}) \tag{3.60a}$$

$$\text{subject to} \qquad g^i(\underline{x}) - s^i = 0, \qquad s^i \geqslant 0 \qquad \forall i \in \{1,\dots,m\}. \tag{3.60b}$$

In this way the inequality constrained problem has been shifted from the constraints $\underline{g}$ to the slack vector $\underline{s}$ and, besides that, not much has changed. However, the KKT conditions for this reformulated system can be written as

$$\nabla_{\underline{x}} f(\underline{x}) - \nabla_{\underline{x}} \underline{g}(\underline{x})\, \underline{\lambda} = \underline{0}, \qquad \operatorname{diag}[\underline{s}]\, \underline{\lambda} - \mu_{\mathrm{IP}}\, \underline{e} = \underline{0}, \qquad \underline{g}(\underline{x}) - \underline{s} = \underline{0}, \tag{3.61}$$

where the vector $\underline{e} \in \mathbb{R}^m$ shall be defined by $e^i = 1$ for all $i \in \mathcal{S}$. Due to the introduction of the positive scalar $\mu_{\mathrm{IP}} \geqslant 0$, the inequality constraints $\lambda \geqslant \underline{0}$ and $\underline{s} \geqslant \underline{0}$ must only be explicitly added if $\mu_{\mathrm{IP}} = 0$. Now, this parameter μ_{IP} is used as a continuation parameter and thus the homotopy approach follows under consideration of a sequence $\{\mu_{\mathrm{IP}}^{\{k\}}\}$ which converges to zero, while the positivity of $\underline{\lambda}^{\{k\}}$ and $\underline{s}^{\{k\}}$ is always maintained. A second alternative problem formulation for (3.60) is given by

$$\underset{\underline{x}\in\mathbb{R}^n,\underline{s}\in\mathbb{R}^m}{\text{minimize}} \qquad f(\underline{x}) - \mu_{\mathrm{IP}} \sum_{i=1}^{m} \log(s^i) \tag{3.62}$$

$$\text{subject to} \qquad g^i(\underline{x}) - s^i = 0, \qquad \forall i \in \{1,\dots,m\}. \tag{3.63}$$

where the usage of the log-function naturally enforces the positivity of the slack variables and the combinatoral difficulties of active set strategies are avoided. This formulation leads to slightly different KKT conditions, namely

$$\nabla_{\underline{x}} f(\underline{x}) - \nabla_{\underline{x}} \underline{g}(\underline{x})\, \underline{\lambda} = \underline{0}, \qquad \underline{\lambda} - \mu_{\mathrm{IP}}\, \underline{\underline{S}}^{-1}\, \underline{e} = \underline{0}, \qquad \underline{g}(\underline{x}) - \underline{s} = \underline{0}, \tag{3.64}$$

where the abbreviation $\underline{\underline{S}} = \mathrm{diag}[\underline{s}]$ has been introduced.

Remark 3.5. So why is it called *interior point method*? This question can easily be answered: In the early publications, e.g. by Fiacco and McCormick [89], no slacks have been used and instead the inequality constraints have been incorporated by the so-called *barrier function*, viz.

$$f(\underline{x}) - \mu_{\mathrm{IP}} \sum_{i=1}^{m} \log(g^i). \tag{3.65}$$

That is an additional restriction in contrast to the formulation with slacks, since the natural logarithm completely prevents the iterates to leave the feasible region. In other words, the objective function forces the iterates to become interior points of the feasible region. However, the introduction of slacks relaxes the situation and allows the start also from an infeasible point.

Finally, the evolving system of equations shall be presented, which can be obtained by simply applying Newton's method to (3.61) or (3.64), followed by a reformulation such that the presented symmetric form yields

$$\begin{pmatrix} \nabla^2_{\underline{x}\,\underline{x}} \mathscr{L} & \underline{\underline{0}} & \nabla_{\underline{x}} \underline{g} \\ \underline{\underline{0}} & \underline{\underline{\Sigma}} & -\underline{\underline{I}} \\ \nabla_{\underline{x}} \underline{g}^{\mathrm{T}} & -\underline{\underline{I}} & \underline{\underline{0}} \end{pmatrix} \begin{pmatrix} \underline{p}_x \\ \underline{p}_s \\ -\underline{p}_\lambda \end{pmatrix} = - \begin{pmatrix} \nabla_{\underline{x}} f - \nabla_{\underline{x}} \underline{g}\, \underline{\lambda} \\ \underline{\lambda} - \mu_{\mathrm{IP}} \underline{\underline{S}}^{-1} \underline{e} \\ \underline{g} - \underline{s} \end{pmatrix}. \tag{3.66}$$

For the so-called *primal-dual* system derived from (3.61) the matrix $\underline{\underline{\Sigma}}$ is identified as $\underline{\underline{S}}^{-1}\underline{\underline{\Lambda}}$ and for the so-called *primal* form derived from (3.64) the matrix $\underline{\underline{\Sigma}}$ can be identified as $\mu_{\mathrm{IP}}\underline{\underline{S}}^{-2}$. Therefore, in the latter case, the Lagrange multiplier matrix $\underline{\underline{\Lambda}} = \mathrm{diag}[\underline{\lambda}]$ is implicitly given by $\mu_{\mathrm{IP}}\underline{\underline{S}}^{-1}$, such that the Lagrange multiplier is no longer a true independent variable. Note that it is possible to eliminate $\underline{p}_s$ and $\underline{p}_\lambda$ in (3.66) and obtain a system which must be only solved for the unknown $\underline{p}_x$ vector. For more information the reader is referred to Nocedal and Wright [204, Ch. 19]. Finally, a last note shall be added before the discussion about the interior point method is concluded: In general, the presented system is solved for the search directions and, subsequently, a line search method must be applied which ensures that the variables $\underline{s}$ and $\underline{z}$ do not exceed their lower bound and stay positive. This rule is called *fraction to the boundary* and can be found in the related mathematical literature (cf. Gould et al. [115], Nocedal and Wright [204], Ulbrich et al. [263], Wächter and Biegler [272]).

Next, the question shall be answered why the interior point method is not considered in this thesis. Therefore, a short summary of the related publications shall be briefly given. For more information, the reader is also referred to the references in the mentioned publications. The interior-point method has already been applied to contact problems: One of the first relevant

references can be found in Oden and Kim [207]. Therein, a barrier function formulation has been considered for the solution of the 2-D Signorini problem, i.e., contact between a linearily elastic and a rigid body. The performance has been compared with a classical penalty approach. Some years later, the interior point method including slack variables has been considered by Christensen et al. [47] and has been compared to a semi-smooth Newton approach for frictional contact problems. Those results suggest a superior performance of the semismooth Newton approach, where especially the treatment of frictional 3-D contact problems has been mentioned as difficult to achieve with the interior point method. While Christensen et al. [47] only considered small displacements and linear elasticity, Kloosterman et al. [159] made the step to large deformations. Therein a frictionless 2-D contact problem is discussed. The solution approach uses a modified barrier function and a special updating scheme for the newly introduced parameters. In Tanoh et al. [258] one supposed advantage of interior point methods compared to active-set strategies is addressed, which is the better performance for large scale constrained problems, i.e., for non-linear problems considering many constraints simultaneously. Therefore, again the Signorini problem has been used with the mesh successively refined. In this reference the primal-dual as well as the primal approach including slacks are considered for 2-D and 3-D problems. Then, Miyamura et al. [197] combined the ideas of interior point methods and active-set strategies in a new algorithm. They proposed a method which uses the interior point method at the beginning of a new load step to achieve an educated guess for the active set at the solution and afterwards switch to a semi-smooth Newton approach. Especially interesting are the presented examples which provide a class of numerical examples with a tough to identify active-set distribution. In Kučera et al. [166] an approach is presented for frictional 3-dimensional contact under small deformations. Therein, a primal-dual interior point method is applied and two different condensation strategies are discussed. Furthermore, also the mathematical community has considered to some extent contact problems, e.g. Herskovits and Mazorche [128], Stadler [254]. Finally, Temizer et al. [261] made the step to large deformations and mortar-type frictionless contact formulations as they are also considered in this work. The mentioned publication put much more emphasis on the isogeometric part and the contact formulation uses a closest-point rather than a ray-tracing approach. However, therein the interior point method in its primal-dual and primal variant is compared to an augmented Lagrangian formulation. The presented results suggest that the primal variant is superior. This contradicts the results of many mathematical publications (see e.g. Nocedal and Wright [204], Wächter and Biegler [272]) and might be due to the used formulation. Furthermore, another important point is mentioned in Temizer et al. [261] even though not a lot of emphasis is put on this fact: The interior point method asks for an integration over the entire potential contact surface in each evaluation step in contrast to an active set strategy which requires only the evaluation of the active contributions on the slave *and* master sides. This and the fact that no significant advantages are proposed but instead the additional treatment of the interior point parameter becomes necessary, the discussion in this thesis will be restricted to the SQP and semi-smooth Newton methods. However, an extension to interior point methods in the future is possible and especially a combination of both approaches as suggested by Miyamura et al. [197] seems worth trying.

3.2.4. Globalization Techniques

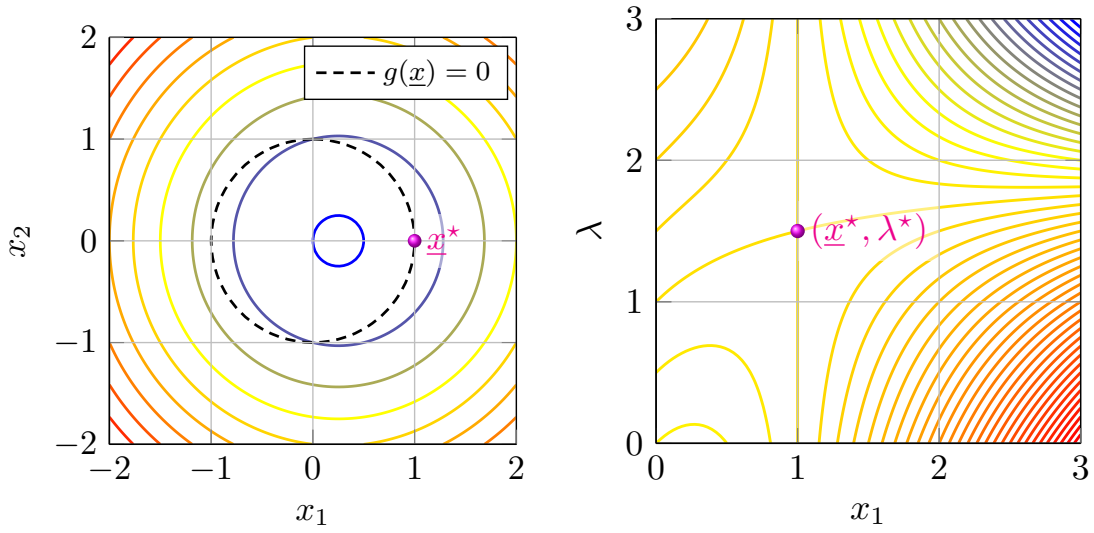
Similar to the case of unconstrained optimization there exists a variety of methods for constrained optimization. Thus, direct methods, as well as gradient based methods are applicable to constrained problems and these ideas are also often combined with each other. In the field of gradient-based method two main classes of algorithms can be found once more: the trust region and the line search methods. Both of them have advantages and disadvantages such that none is clearly superior to the other. For instance, one drawback of trust region methods is that the inherent additional constraint, namely the restriction of the step length, might lead to an infeasible sub-problem. Thus, trust-region methods must implement special strategies to prevent such a scenario. On the other hand, they have the clear advantage that they can easily handle indefinite Hessian matrices, while a line search method might be in trouble in such cases and would ask for a modified Hessian (see Section 5.6.8 for an example).

Another general topic which is much more involved in the constrained case is the decision about acceptance or rejection of a step. While for unconstrained problems this question is usually directly answered by the used merit function which often coincides with the objective function, the constrained optimization has no clear representative measure. Instead, it can be quite difficult to decide if a step is meaningful or should be rejected. Classically, this decision has been based on a special merit function which incorporates the constraints via a penalty/regularization term. Two popular variants of such merit functions are the non-smooth ℓ_1 and ℓ_2 penalty functions which can be directly deduced from (3.46), namely by removing the square, multiplication of the penalty term by two and inserting the respective norm. The drawback of such penalty functions is that the inherent penalty parameter must be chosen sufficiently large such that the prediction of this functions becomes reliable or more precisely *exact* (see Nocedal and Wright [204, Definition 15.1]). For the ℓ_1 merit function this is the case if

$$c_{\ell_1}^\star = \max\{|\lambda_i^\star|,\ i \in \mathcal{S}\}. \tag{3.67}$$

The problem of (3.67) is obvious: In general the value of the optimal Lagrange multipliers is not known in advance. Thus, it can be a difficult task to provide an educated guess or to implement an adaptive strategy to increase the penalty parameter if necessary. Under the assumption that such an update strategy is accessible, the Armijo rule (3.16) can be directly applied where the directional derivative of the underlying non-smooth merit function in search direction must be inserted. Similar ideas can be also used for the acceptance by the trust region methods. Here, the role of the objective function is simply taken by the exact merit function and the quadratic model is some model for the Lagrangian, for instance.

Since the choice of the penalty parameter is such a crucial ingredient, a different strategy shall be followed in this thesis. This alternative strategy is the so-called filter method. The filter method uses ideas from multi-objective optimization. The two goals of minimizing the objective function and maintaining feasibility of the solution are considered separately, instead of relying on a scalar merit function which asks for a correctly chosen weighting of the constraints compared to the objective function. The literature on this topic is huge. A brief summary can be found in Section 1.2.2. In this thesis a line search filter method shall be considered and the extension to this type of algorithms has been given in Wächter and Biegler [270, 271, 272]. The topic will be comprehensively revisited in Chapter 6.



(a) (b)

Figure 3.2.: In Figure 3.2a the contour lines of (3.68a) are visualized as well as the zero-isoline of the equality constraint (3.68b). The solution is marked by ●. Furthermore, a slice through the corresponding Lagrangian $\mathscr{L}(x_1, 0, \lambda)$ inclusive the KKT pair $(\underline{x}^\star, \lambda^\star)$ is presented in Figure 3.2b.

The Maratos effect

There is another important observation which must be addressed in conjunction with globalization techniques for constrained optimization problems. As already mentioned, the two aimed for goals of minimizing the objective function value and simultaneously maintaining or improving the feasibility of an iterate can be a conflicting task. In this context, a phenomenon called the *Maratos effect* named after its discoverer Maratos [186] exists. It describes the scenario that a obviously good step, which may even imply a quadratic rate of convergence, is refused due to the fact that both the objective function value as well as the infeasibility measure are increased at the new trial point position. Such a case can not be handled by a classical merit function or a filter approach. Since this scenario is something which is not quite intuitive at first glance, a short example introduced by Powell [219] shall be presented. Therein, the following simple equality constrained optimization example is considered:

$$\underset{\underline{x}\in\mathbb{R}^2}{\text{minimize}} \qquad f(\underline{x}) = 2(x_1^2 + x_2^2 - 1) - x_1, \tag{3.68a}$$

$$\text{s. t.} \qquad g(\underline{x}) = x_1^2 + x_2^2 - 1 = 0. \tag{3.68b}$$

By a look at Figure 3.2a it can be easily seen that the optimal solution point is equal to $\underline{x}^\star = (1, 0)^{\mathrm{T}}$ and the associated Lagrange multiplier is obtained by $\lambda^\star = 3/2$. This fact can be also checked visually in Figure 3.2b where the associated saddle-point becomes clearly visible. Furthermore, a short calculation reveals that the Hessian $\nabla^2_{\underline{x}\,\underline{x}}\mathscr{L}^\star$ at that solution is equal to the identity matrix $\underline{\underline{I}}$ based on the Lagrangian $\mathscr{L}(\underline{x}, \lambda) = f(\underline{x}) - \lambda g(\underline{x})$, thus the second order sufficient conditions from Theorem 3.5 hold. Powell [219] suggests now to consider any feasible (non-optimal) point on the circle defined by (3.68b), viz. $\underline{x}^{\{k\}} = (\cos(\varphi), \sin(\varphi))^{\mathrm{T}}$ as

starting point for the calculation. Furthermore, the Lagrange multiplier shall be set to its optimal value $\lambda^{\{k\}} = \lambda^\star$. If now the Newton direction is computed, the new trial point $\underline{x}^+ = (\cos(\varphi) - \sin^2(\varphi), \sin(\varphi) - \sin(\varphi)\cos(\varphi))^{\mathrm{T}}$ is obtained. Under consideration of (3.7) it can be directly shown that the Newton direction for a step length equal to one implies quadratic convergence since

$$\frac{\|\underline{x}^+ - \underline{x}^\star\|}{\|\underline{x}^{\{k\}} - \underline{x}^\star\|^2} = \frac{2\sin^2(\frac{\varphi}{2})}{[2\,|\sin(\frac{\varphi}{2})|]^2} = \frac{1}{2}. \tag{3.69}$$

However, the objective function value as well as the constraint function value rises at the trial point $\underline{x}^+$, yielding

$$f(\underline{x}^+) = \sin^2(\varphi) - \cos(\varphi) > -\cos(\varphi) = f^{\{k\}}, \tag{3.70}$$

$$g(\underline{x}^+) = \sin^2(\varphi) > 0 = g^{\{k\}}. \tag{3.71}$$

Thus, both of the proposed globalization methods would fail and reject the iterate. In the literature different remedies have been suggested to circumvent this problem, such as:

1. Usage of a merit function which is better suited and does not suffer under the Maratos effect. One possibility is Fletcher's augmented Lagrangian function which is a combination of (3.43) and the Lagrange multiplier function (3.48) for equality constrained problems, i.e. without the necessity for slacks. See Nocedal and Wright [204] and [53] for more information. Alternatively, it is also possible to extend the filter method by choosing suitable filter entries such that the original trust region SQP steps can be applied without the need for any second order correction. This method is also based on the Lagrange multiplier function (3.48) and is described in Ulbrich [264].

2. Implementation of a *second-order correction* step which augments the current search direction and aims for a reduction of the constraint violation. For more information the reader is referred to the literature on this topic such as Fletcher [93], Nocedal and Wright [204], Wächter and Biegler [270] as well as to Section 6.4 where the idea of a second order correction step will be reconsidered in the context of a filter method suitable for contact problems.

3. Consideration of a non-monotone strategy and allowing a certain increase in the merit function for a certain number of iterations (see also the discussion in Section 3.1.2). See for example Chamberlain et al. [45] for a classical introduction.

As already mentioned, the idea of a *second-order correction* step has been considered in this thesis and is discussed in Section 6.4. Furthermore, it should be noted that the presented example might seem some what artificial due to the explicit choice of the Lagrange multiplier equal to its optimal value at iteration k. In the original work [219] a quasi-Newton method is considered and the actual Hessian is estimated by the optimal Hessian at the solution, i.e., by the identity matrix. In this way the point of view changes slightly. However, the observation stays valid also for a classical Newton approach and similar problems could also be observed during numerical experiments in the field of contact mechanics (see Section 6.10).

4. Mortar-Based Contact Methods for Finite Deformation Solid Mechanics

In this chapter two mortar-based segment-to-segment contact formulations will be developed for the frictionless finite deformation case: While the first one is derived by consistent variation of all active contributions of a scalar-valued potential subject to inequality constraints, thus resulting in a truly variationally consistent and symmetric formulation, the second approach is designed in such a way that it is less computationally expensive, but still conserves important quantities such as linear and angular momentum. Since both formulations are derived side by side, the introduced simplifications can be specifically analyzed and quantified. Based on a Lagrange multiplier approach the corresponding inequality constraint terms are introduced in two popular ways: Firstly, in form of a standard Lagrangian formulation and, secondly, via an augmented Lagrangian formulation. Both variational forms as well as both solution procedures will be consistently linearized and discussed. Finally, the obtained results will be compared to each other as well as to a well-established, yet slightly inconsistent, mortar-based contact formulation. It must be noted that the content of this chapter has already been published in Hiermeier et al. [131] with exception of the discussion of the inherent instability of variationally inconsistent contact formulations presented in Section 4.7.4.

4.1. Motivation

Contact mechanics is a field of continuing wide interest. This research interest often stems from problems arising during the numerical simulation of complex real-world applications, where one source of complexity is the efficient computational treatment of the non-linear contact conditions. Therefore, it is not surprising that many researchers have an engineering background and several interesting research developments are primarily built up on heuristic observations and ideas. One outstanding and very successful representative of such a heuristic idea is e.g. [294], where a start-up procedure for contact problems with large load steps is presented. Even though interesting results are often given, the underlying theory may lack rigorous mathematical proofs. At the same time, the mathematical community of constrained optimization also contributes very interesting improvements, but often demonstrates the effectiveness only with a set of standardized examples (see e.g. Gould and Toint [116], Wächter and Biegler [272]). This is of course beneficial with regard to a better comparability of different approaches, but the direct transferability of the results to other disciplines may be limited. Even though it is to say that there are recently some very promising first attempts which aim in this direction such as Temizer et al. [261], Youett et al. [291], the observation still raises the question why the two research fields seem to develop almost independently from each other without much deeper interaction? One reason for the lack of interdisciplinary exchange might be the fact that it is not always possible to

quantify the introduced errors in the different discretized contact formulations. This is especially true when large deformations are considered. Almost all available proofs and theoretical theorems concern the small deformation case (see e.g. Kikuchi and Oden [157], Wohlmuth [279]), which is on the one hand obvious, since it is far more difficult to find general mathematical results when there are strong non-linearities involved. On the other hand, however, it might also be due to certain underlying assumptions of some formulations, which are hidden at first sight.

For example, it is not immediately obvious why most mortar-type contact formulations for frictionless contact lead to a non-symmetric Jacobian matrix in their consistently linearized form (e.g., Popp et al. [215, 216]), even though the formulation is usually based on a scalar valued potential formulation. Another question might be why some formulations lack exact angular momentum conservation even in the quasi-static case (e.g. Popp et al. [215, 216], Puso and Laursen [222]). Only a deeper insight reveals the neglected terms and certain assumptions in the published formulations. Therefore, a rigorous and consistent mortar-based contact formulation will be developed in this work, which leads to a symmetric system of equations with respect to all active set contributions. The constraints will be enforced by Lagrange multipliers, thus leading to a standard Lagrange saddle-point system as well as an augmented Lagrangian formulation [1, 65, 212]. Specifically, all variations of the active contributions will be considered and the related second order derivative terms are given as well. Since the arising terms can become computationally quite expensive, a second, incomplete formulation is developed simultaneously, which leads to a non-symmetric Jacobian, but still fulfills conservation of linear and angular momentum. Since both formulations are derived side by side, the introduced errors and simplifications can be quantified, and the final formulations will be compared in terms of robustness and accuracy. It will also be shown that a completely consistent variational approach can actually lead to unwanted side effects when the numerical evaluation of interface integrals is addressed. Finally, a comparison to one exemplary implementation of a well-established and widely spread mortar-type contact formulation is given by considering the work of Popp et al. [215, 216, 218]. In summary, the intention of this work is by no means to show that certain published formulations lead to inaccurate results or lack consistency, but instead the primary objective is to provide a deeper understanding of the different formulations and their respective limitations. It is the authors' firm belief that this will lead to new interdisciplinary progress in the mentioned fields.

The remainder of this chapter is organized as follows: Section 4.2 gives a summary of the employed frictionless contact formulation and the general discretized problem statement. Furthermore, the differences between the two constraint enforcement strategies, namely the standard and the augmented Lagrangian formulation, are discussed in detail in Section 4.3. Section 4.4 puts the focus on the complete and incomplete variational approaches, emphasizes their differences and presents the consistent linearization of the derived terms. Section 4.5 is entirely devoted to mechanical conservation laws: In particular, the conservation of angular momentum of both discretized variational formulations is proven theoretically. In Section 4.6 the attention is on the numerical time integration for contact problems and the discussion started in Section 2.4 about the Generalized-α time integration scheme is continued. Next, three representative numerical examples with varying objectives are discussed in Section 4.7. Therein, the previously derived contact formulations and constraint enforcement strategies are compared to each other as well as to the already mentioned well-established mortar-based formulation from the literature Popp et al. [215, 216]. Finally, some conclusions are drawn in Section 4.8.

4.2. Contact Formulation

This section is meant to give a brief summary of the numerical background of the employed frictionless contact formulation. For convenience, the presented formulation only considers the special case of unilateral contact between two elastic bodies undergoing large deformations. Nevertheless, the presented algorithms are also capable of more complex contact situations, such as self-contact or contact between multiple bodies. A detailed description of the continuous frictionless contact formulation can be found in 2.2.2. Therefore, the attention is next directly drawn to the discretized formulation.

4.2.1. Discretized Contact Kinematics

Before the full discrete contact problem is stated, a finite element discretization of all important contact quantities shall be introduced. This way is taken to give a better insight into the actual implementation. Moreover, the number of necessary redefinitions of continuous quantities is reduced to a minimum. Starting with a very brief summary of the basic ideas of the finite element method and, right after, it will be applied to the contact problem. For more information, the interested reader is referred to the corresponding literature (see e.g. Bathe [12], Zienkiewicz and Taylor [297]) and the introduction in Section 2.3. Throughout this chapter mainly a Lagrange interpolation for the shape functions will be considered. Consequently, the problem description is restricted to Lagrange interpolations only, even though a Non-Uniform Rational Basis Spline (NURBS) discretization would be possible as well (see e.g. De Lorenzis et al. [65], Farah et al. [82], Hughes et al. [141], Seitz et al. [241]). One reason is that a pure NURBS discretization does not exhibit some of the typical error sources, since some neglected terms would vanish naturally, which is not the case for Lagrange shape functions. However, one numerical example with a NURBS discretization is also given and will be discussed in Section 4.7.3.

The basic idea of the finite element method is that the considered problem is not solved on the potentially geometrically very complex domain Ω_0, but instead a finite number of degrees of freedom are defined at $\mathrm{N_n}$ discrete points. These points are called nodes and are connected by $\mathrm{N_e}$ elements, where each element occupies a closed sub-domain of the entire problem domain $\Omega_0^{(e)} \subseteq \Omega_{0,\mathrm{h}}$. The disjoint union of all element domains represents the domain estimate for each body. The approximation error of the solution is supposed to decrease with an increasing number of elements $\mathrm{N_e}$ (known as h-convergence). Since the formulation proposed here is restricted to Lagrange basis polynomials as shape functions, the value of an arbitrary nodal quantity inside an element e can be computed by interpolating its nodal values. For instance, the current position $\underline{x}(\underline{X}^{[b]}, t)$ at $\underline{X}^{[b]} \in \Gamma_{\mathrm{c}}^{[b]}$ on the contact boundaries is, just in accordance with (2.72), obtained by

$$\underline{x}(\xi^{[b](e)k}) = N^{[b](e)}{}_{j}\, \underline{x}^{[b](e)j} = N^{[b]}{}_{j}\, (\underline{X}^{[b]j} + \underline{d}^{[b]j}), \qquad \forall j \in \{1, \dots, \mathrm{N}_{\mathrm{n}}^{[b](e)}\}, \tag{4.1}$$

where (4.1) takes advantage of the Einstein summation convention. In the final step, the element dependency has been omitted for the sake of brevity. The index e identifies one of the surface elements on $\Gamma_{\mathrm{c}}^{[b]}$. The set of all surface nodes on the contact interfaces shall be split into one set $\mathcal{S} = \{l \in \{1, \dots, \mathrm{N}_{\mathrm{n}}^{[1]}\} \mid \underline{X}^{l} \in \Gamma_{\mathrm{c}}^{[1]}\}$ containing all slave nodes, and a second set $\mathcal{M}$ containing all master nodes following an equivalent set definition for body $b = 2$. The matrices, e.g. $\underline{\underline{x}}^{[b](e)}$,

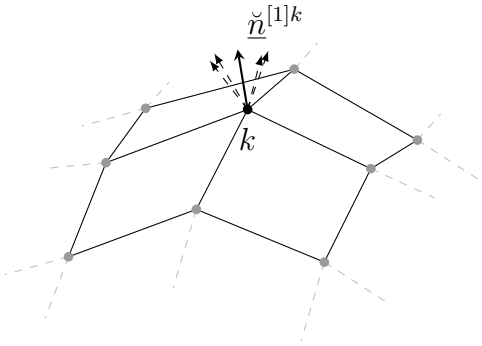


Figure 4.1.: Visualization of the smooth nodal unit normal given in (4.3) at the node k.

contain in each column the current coordinates of the respective quantity at one node of the element e. The coordinates $\{\xi^{[b](e)k}\}_{k\in\{1,2\}}$ denote the surface parametrization of the element e on body $\Omega^{[b]}$ and $N^j(\xi^{[b](e)k}) : \mathbb{R}^2 \to \mathbb{R}$ is one of the nodal Lagrange basis polynomials at node j of the surface element e. Equivalent to (4.1), the discrete covariant base vectors can be defined as

$$\underline{\tau}^{[b](e)}{}_j = \frac{\partial \underline{x}^{[b](e)}}{\partial \xi^j} = N^{[b](e)}{}_{k,\xi^j}\, \underline{x}^{[b](e)k} = N^{[b]}{}_{k,\xi^j}\, \underline{x}^{[b]k}. \tag{4.2}$$

The contravariant counterparts as well as the according metrics can be computed by following the continuous definitions.

The discretization of the contact quantities is based on the mortar method. In contrast to Popp et al. [218], where first the weak form is derived from the strong form and, in a second step, Lagrange interpolation of each term is inserted, here, the discretized nodal contact quantities are defined first and, subsequently, they are directly inserted into the discrete potential. However, the details concerning the continuous approach can be found again in Section 2.2.2. The attentive reader will notice that in the case of a Lagrangian functional the difference will be restricted to the interpretation of the different terms. However, if an augmented Lagrangian formulation De Lorenzis et al. [65] or a penalty formulation Yang et al. [290] is considered, this approach will make the consistent introduction of regularization terms much easier. Another advantage is that all introduced simplifications can be easily identified, especially the possible loss of symmetry of the evolving system of equations. Lastly, the transition to the mathematical optimization literature seems more fluent with this choice.

Let us start with the definition of the considered smooth normal field Popp et al. [215], Yang et al. [290]. At each slave node $k \in \mathcal{S}$ an averaged nodal unit normal $\breve{\underline{n}}^{[1]k}$ is obtained by

$$\breve{\underline{n}}^{[1]k} = \frac{\tilde{\underline{n}}^{[1]k}}{\|\tilde{\underline{n}}^{[1]k}\|} \text{ with } \tilde{\underline{n}}^{[1]k} = \sum_{e=1}^{\mathrm{N}_k^{\mathrm{adj}}} \underline{n}^{[1](e)k}, \tag{4.3}$$

where $\underline{n}^{[1](e)k}$ is the outward-pointing unit normal vector following the definitions (2.53) and (4.2) of the adjacent slave element e evaluated at the slave node k (see Figure 4.1 for a visualization). It is also possible to use weights for the different element normals. Still, the simplest,

so-called mean weighted equally, variant is used here Gouraud [117]. A comprehensive study of the different weightings can be found, e.g. in Jin et al. [149], Neto et al. [202]. Through Lagrange interpolation of the averaged nodal unit normals (4.3), a C^0-continuous normal field is achieved, viz.

$$\breve{\underline{n}}^{[1]} = \breve{\underline{n}}(\underline{x}^{[1]}) = N^{[1]}{}_i \; \breve{\underline{n}}^{[1]i}, \qquad i \in \mathcal{S}. \tag{4.4}$$

In general, the interpolation scheme does not maintain the unit length property of the nodal normal vectors inside the element, and thus an additional normalization will become necessary if the unit length property of the interpolated normal field needs to be maintained. This interpolated unit length smooth normal field is denoted by

$$\hat{\underline{n}}^{[1]} = \frac{\breve{\underline{n}}^{[1]}}{\|\breve{\underline{n}}^{[1]}\|} = \frac{N^{[1]}{}_i \; \breve{\underline{n}}^{[1]i}}{\|N^{[1]}{}_i \; \breve{\underline{n}}^{[1]i}\|}, \qquad i \in \mathcal{S}. \tag{4.5}$$

By inserting (4.1) and (4.5) into (2.55), the discrete normal gap definition is obtained by

$$g_{\mathrm{N}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}) = \left\langle \hat{\underline{n}}^{[1]}, \; N^j(\bar{\xi}^{[2](\bar{e})l}) \; \underline{x}^{[2](\bar{e})}{}_j - N^k(\xi^{[1](e)n}) \; \underline{x}^{[1](e)}{}_k \right\rangle, \tag{4.6}$$

where $k \in \mathcal{S}$ and $j \in \mathcal{M}$. The element superscript e denotes the slave element index, while the superscript $\bar{e}$ represents the index of the master element which has been found by the projection and search algorithm as described in Yang and Laursen [288]. The function $g_{\mathrm{N}} : \mathbb{R}^2 \to \mathbb{R}$ represents the discrete form of the geometrical gap evaluated at the slave parameter space coordinates $\{\xi^{[1]i}\}_{i\in\{1,2\}}$ and their corresponding projected parametric coordinates $\{\bar{\xi}^{[2]i}\}_{i\in\{1,2\}}$ on the master side. In the spirit of the mortar method, the nodal *averaged weighted gap* $\hat{g}^i_{\mathrm{N}}$ of a slave node $i \in \mathcal{S}$ is defined as

$$\hat{g}^i_{\mathrm{N}} = \frac{\tilde{g}^i_{\mathrm{N}}}{A^i}, \text{ where } \tilde{g}^i_{\mathrm{N}} = \iint\limits_{\gamma^{[1]}_{\mathrm{c,h}}} N^{[1]i} \; g_{\mathrm{N}} \; \mathrm{d}\underline{a} \qquad \text{and} \qquad A^i = \iint\limits_{\gamma^{[1]}_{\mathrm{c,h}}} N^{[1]i} \; \mathrm{d}\underline{a}, \tag{4.7}$$

where $\mathrm{d}\underline{a}$ is an infinitesimal interface area segment on the slave side. The quantity $\tilde{g}^i_{\mathrm{N}}$ is called the *weighted gap* of the slave node and A^i is its *tributary area*. For more details, especially related to the tributary area, the interested reader is kindly referred to De Lorenzis et al. [65]. A more mathematical explanation for the necessity of this additional scaling can be found in Hüeber [138, Remark 2.5]: The additional scaling is especially important when the standard Lagrangian shape functions are considered for the discrete test function space of the Lagrange multipliers, instead of the bi-orthogonal shape functions as e.g. applied by [87, 138, 218, 279].

4.2.2. Problem Statement

In this section, the problem statement shall be reviewed from a mathematical point of view as an optimization problem subject to inequality constraints. The already discretized problem can be stated in a similar way as (3.31) by

$$\min_{\underline{x}\in\mathbb{R}^n} \quad \mathscr{U}(\underline{x}) = \mathscr{U}^{[1]}(\underline{x}^{[1]}) + \mathscr{U}^{[2]}(\underline{x}^{[2]}), \tag{4.8a}$$

$$\text{subject to} \quad \hat{\underline{g}}_{\mathrm{N}}(\underline{x}) \geqslant \underline{0}, \tag{4.8b}$$

where the objective function $\mathscr{U}(\underline{x}) : \mathbb{R}^n \to \mathbb{R}$ contains the sum of the elastic potentials of the considered elastic bodies. For the sake of brevity, the abbreviation $\underline{x} = ((\underline{x}^{[1]})^T, (\underline{x}^{[2]})^T)^T$, introduced in (4.8a), is used also in connection with all contact related terms, where it becomes $\underline{x} = ((\underline{x}^{[1]})^T, (\underline{\bar{x}}^{[2]})^T)^T$ such that $\hat{\underline{g}}_{\mathrm{N}}(\underline{x}) \in \mathbb{R}^m$ with $m < n$ represents the global vector of averaged weighted normal gap constraints. Note that n denotes the number of displacement degrees of freedom and is equal to $3 \cdot \mathrm{N_n}$, while m is equal to the cardinality of the slave set $|\mathcal{S}|$. Moreover, any additional labeling of global quantities is intentionally omitted, since the author is convinced that the meaning can easily be derived from the context. Furthermore, it is assumed for now that the objective function as well as the constraint equations are sufficiently smooth.

In the sequel, $(\underline{x}^\star, \underline{\lambda}_{\mathrm{N}}^\star) \in \mathbb{R}^n \times \mathbb{R}^m$ will indicate a KKT pair of problem (4.8), i.e. a pair of primal and dual variables, which satisfies the following conditions

$$\nabla_{\underline{d}} \mathscr{L}(\underline{x}^\star, \underline{\lambda}_{\mathrm{N}}^\star) = \underline{0}, \qquad \hat{\underline{g}}_{\mathrm{N}}(\underline{x}^\star) \geqslant \underline{0}, \qquad \underline{\lambda}_{\mathrm{N}}^\star \geqslant \underline{0}, \qquad \langle \underline{\lambda}_{\mathrm{N}}^\star,\ \hat{\underline{g}}_{\mathrm{N}}(\underline{x}^\star) \rangle_{\underline{\underline{A}}} = \underline{0}, \tag{4.9}$$

where $\underline{\lambda}_{\mathrm{N}} \in \mathbb{R}^m$ is the vector of nodal Lagrange multiplier values and the notation $\langle \underline{u}, \underline{v} \rangle_{\underline{\underline{A}}} = \underline{u}^{\mathrm{T}} \underline{\underline{A}}\, \underline{v}$ denotes an inner product of two vectors scaled by the quadratic diagonal matrix $\underline{\underline{A}}$ containing the tributary area values. The Lagrangian for inequality constraints $\mathscr{L}(\underline{x}, \underline{\lambda}_{\mathrm{N}}) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is defined as

$$\begin{aligned} \mathscr{L}(\underline{x}, \underline{\lambda}_{\mathrm{N}}) &= \mathscr{U}(\underline{x}) - \mathscr{C}(\underline{x}, \underline{\lambda}_{\mathrm{N}}) \\ &= \mathscr{U}(\underline{x}) - \langle \underline{\lambda}_{\mathrm{N}},\ \hat{\underline{g}}_{\mathrm{N}}(\underline{x}) - \underline{s}(\underline{x}, \underline{\lambda}_{\mathrm{N}}) \rangle_{\underline{\underline{A}}} \qquad \text{with } s^i(\underline{x}, \underline{\lambda}_{\mathrm{N}}) \geqslant 0, \end{aligned} \tag{4.10}$$

where the components of the vector $\underline{s}(\underline{x}, \underline{\lambda}_{\mathrm{N}}) \in \mathbb{R}^m_+$ follow as

$$s^i(\underline{x}, \underline{\lambda}_{\mathrm{N}}) = \begin{cases} \max\{0,\ \hat{g}^i_{\mathrm{N}}(\underline{x})\} & \text{if } c_{\mathrm{N}} = 0, \\ \max\left\{0,\ \hat{g}^i_{\mathrm{N}}(\underline{x}) - \frac{\lambda^i_{\mathrm{N}}}{c_{\mathrm{N}}}\right\} & \text{otherwise,} \end{cases} \tag{4.11}$$

and $c_{\mathrm{N}} \in \mathbb{R}$ with $c_{\mathrm{N}} \geqslant 0$ is a regularization parameter. This notation follows Gill et al. [107], Rockafellar [229] and is closely related to the comprehensive introduction in 3.2. In conjunction with the inherent active-set decision, a steady transition between the active and the inactive branch of the Lagrangian function becomes possible (see Figures 2.3 and 2.4). Again, $\mathcal{A}$ denotes the set of active constraints, i.e., the index set of contact constraints with positive augmented Lagrangian multipliers, viz.

$$\mathcal{A} = \{i \in \{1, \ldots, m\} \mid \lambda^i_{\mathrm{N}} - c_{\mathrm{N}} \hat{g}^i_{\mathrm{N}} \geqslant 0\}. \tag{4.12}$$

and further, $\mathcal{I} = \{1, \ldots, m\} \backslash \mathcal{A}$ with $\mathcal{I} \cap \mathcal{A} = \varnothing$ indicates the set of the remaining inactive constraints. This is the discrete contact variant of (3.36). Alternatively, it is also possible to formulate an augmented Lagrangian rather than a standard Lagrangian. Therefore, the discrete form of (2.61) together with (2.59) results in

$$\mathscr{L}_{c_\mathrm{N}}(\underline{x}, \underline{\lambda}_\mathrm{N}, \underline{\hat{s}}) = \mathscr{U}(\underline{x}) - \langle \underline{\lambda}_\mathrm{N}, \underline{\hat{g}}_\mathrm{N}(\underline{x}) - \underline{\hat{s}} \rangle_{\underline{\underline{A}}} + \frac{c_\mathrm{N}}{2} \| \underline{\hat{g}}_\mathrm{N}(\underline{x}) - \underline{\hat{s}} \|^2_{\underline{\underline{A}}}, \tag{4.13}$$

where $\underline{\hat{s}}$ takes on the role of a formally still independent variable. This becomes more important in Chapter 6.

Either way, if the Mangasarian-Fromowitz constraint qualifications are fulfilled, the KKT conditions (4.9) are the first order optimality conditions for the problem (4.8), see also Nocedal and Wright [204] and Section 3.2 for more details. The mentioned qualifications are typical assumptions for many optimization algorithms. Important to note is that these qualifications only need to be fulfilled at the solution $(\underline{x}^\star, \underline{\lambda}^\star_\mathrm{N})$. If the gradients of the constraints become (nearly) linear dependent during the solution procedure, a fall-back solving strategy (see e.g. Wächter and Biegler [272]) can take care. For more information on the mathematical background the reader is referred to Section 3.2 and the references therein.

4.3. Non-Linear and Linearized Systems of Equations

In this section the optimality conditions given in (4.9) will be discussed in more detail with the focus on the Lagrangian formulation. Since the augmented Lagrangian formulation, first introduced in the context of contact problems by Alart and Curnier [1], is also quite often in use, the necessary augmentation steps will be briefly given afterwards in Section 4.3.2. Possible advantages and drawbacks of the two formulations during the non-linear solution procedure will be discussed in Section 4.7.

4.3.1. Lagrangian Formulation

For the computation of the first order optimality conditions (4.9), the directional derivative of the Lagrangian (4.10) become necessary, which follows as

$$\mathrm{D}\mathscr{L}((\underline{x}, \underline{\lambda}_\mathrm{N}); (\delta\underline{d}, \delta\underline{\lambda}_\mathrm{N})) = \mathrm{D}\mathscr{U}(\underline{x}; \delta\underline{d}) - \mathrm{D}\mathscr{C}((\underline{x}, \underline{\lambda}_\mathrm{N}); (\delta\underline{d}, \delta\underline{\lambda}_\mathrm{N})), \tag{4.14a}$$

where

$$\mathrm{D}\mathscr{U}(\underline{x}; \delta\underline{d}) = \sum_{b=1}^{2} \langle \nabla_{\underline{d}^{[b]}} \mathscr{U}^{[b]}(\underline{x}^{[b]}), \delta\underline{d}^{[b]} \rangle, \tag{4.14b}$$

$$\begin{aligned} \mathrm{D}\mathscr{C}((\underline{x}^{[1]}, \underline{x}^{[2]}, \underline{\lambda}_\mathrm{N}); (\delta\underline{d}^{[1]}, \delta\underline{d}^{[2]}, \delta\underline{\lambda}_\mathrm{N})) = \; & \langle \nabla_{\underline{d}^{[1]}} \mathscr{C}(\underline{x}^{[1]}, \underline{x}^{[2]}, \underline{\lambda}_\mathrm{N}), \delta\underline{d}^{[1]} \rangle \\ & + \langle \nabla_{\underline{d}^{[2]}} \mathscr{C}(\underline{x}^{[1]}, \underline{x}^{[2]}, \underline{\lambda}_\mathrm{N}), \delta\underline{d}^{[2]} \rangle \\ & + \langle \nabla_{\underline{\lambda}_\mathrm{N}} \mathscr{C}(\underline{x}^{[1]}, \underline{x}^{[2]}, \underline{\lambda}_\mathrm{N}), \delta\underline{\lambda}_\mathrm{N} \rangle. \end{aligned} \tag{4.14c}$$

Equation (4.14b) describes the virtual work contributions of the two bodies $\Omega^{[b]}$, $b \in \{1,2\}$ in their already discretized form. These are assumed to be known at this point. However, a detailed derivation can be found e.g. in Bathe [12], Holzapfel [136] as well as in Section 2.1.4. Here, the focus is solely put on the contact contributions summarized in the directional derivative of the contact potential (4.14c), therefore representing the contact virtual work generated by the contact forces acting on the contact interfaces between the slave and master bodies. Under renewed consideration of the abbreviation $\underline{x} = ((\underline{x}^{[1]})^T, (\underline{\bar{x}}^{[2]})^T)^T$, it can be stated

$$\nabla_{\underline{d}} \mathscr{C}(\underline{x}, \underline{\lambda}_{\mathrm{N}}) = \begin{cases} \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}) \; \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} = [\tilde{g}_{\mathrm{N}}]^i{}_{,\tilde{d}^j} \; [\lambda_{\mathrm{N}}]_i \; \underline{e}^j & \forall i \in \mathcal{A}, \\ \frac{1}{c_{\mathrm{N}}} \nabla_{\underline{d}} \langle \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}}, \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} \rangle_{\underline{\underline{A}}^{\mathcal{I}}} = \frac{1}{c_{\mathrm{N}}} [\lambda_{\mathrm{N}}]^i \, A_{ii,d^j} \, [\lambda_{\mathrm{N}}]^i \, \underline{e}^j \triangleq \underline{0} & \forall i \in \mathcal{I}, \end{cases} \tag{4.15}$$

where the respective set affiliation of the vector components is indicated by the corresponding superscript. Note that the neglected gradient calculation of the tributary diagonal matrix $\underline{\underline{A}}^{\mathcal{I}}(\underline{x}^{[1]})$ is an usual assumption, since otherwise the inactive contact forces would influence the balance of linear momentum during the non-linear solution scheme, which is clearly non-physical. However, at the solution point, $[\lambda_{\mathrm{N}}^{\star}]^i = 0 \; \forall i \in \mathcal{I}$ holds true and the contributions would vanish anyway.

Remark 4.1. The operator $\tilde{\nabla}(\cdot)$ is introduced to distinguish between two different variational approaches. While the first one is based on the rigorous variation of the underlying potential, thus leading to a fully symmetric linearized system of equations with respect to all active quantities, the second approach will introduce some simplifications. These simplifications cause on the one hand slightly different results, but also a non-symmetric linearized system. The introduced operator is supposed to emphasize the influence of these two options. Anyway, in the first case (i.e. consistent variation), it can be read as the well-known gradient operator. The details will be discussed in Section 4.4.1.

The gradient with respect to the Lagrange multipliers follows as

$$\nabla_{\underline{\lambda}_{\mathrm{N}}} \mathscr{C}(\underline{x}, \underline{\lambda}_{\mathrm{N}}) = \begin{cases} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}) = \delta_{ij} \; [\tilde{g}_{\mathrm{N}}]^i \, \underline{e}^j = [\tilde{g}_{\mathrm{N}}]^i \, \underline{e}_i & \forall i \in \mathcal{A}, \\ \frac{2}{c_{\mathrm{N}}} \underline{\underline{A}}^{\mathcal{I}} \; \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} = \frac{2}{c_{\mathrm{N}}} \delta_{ij} \; A^{ii} \, [\lambda_{\mathrm{N}}]_i \, \underline{e}^j = \frac{2}{c_{\mathrm{N}}} A^{ii} \, [\lambda_{\mathrm{N}}]_i \, \underline{e}_i & \forall i \in \mathcal{I}. \end{cases} \tag{4.16}$$

Under consideration of the fact that the virtual quantities $\delta\underline{d}$ and $\delta\underline{\lambda}_{\mathrm{N}}$ are arbitrary, the first order optimality conditions can be split into displacement related equations, representing the contact virtual work principle, and equations for the Lagrange multipliers, which incorporate the constraint equations. The set of non-linear equations is given by

$$\begin{aligned} &\mathrm{D}\mathscr{L}((\underline{x}^{[1]}, \underline{x}^{[2]}, \underline{\lambda}_{\mathrm{N}}); (\delta\underline{d}^{[1]}, \delta\underline{d}^{[1]}, \delta\underline{\lambda}_{\mathrm{N}})) = \\ &\begin{pmatrix} \delta\underline{d}^{[1]} \\ \delta\underline{d}^{[2]} \\ \delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \\ \delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \nabla_{\underline{d}^{[1]}} \mathscr{U}^{[1]}(\underline{x}^{[1]}) - \tilde{\nabla}_{\underline{d}^{[1]}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}^{[1]}, \underline{\bar{x}}^{[2]}) \; \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \\ \nabla_{\underline{d}^{[2]}} \mathscr{U}^{[2]}(\underline{x}^{[2]}) - \tilde{\nabla}_{\underline{d}^{[2]}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}^{[1]}, \underline{\bar{x}}^{[2]}) \; \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \\ -\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}^{[1]}, \underline{\bar{x}}^{[2]}) \\ -\frac{2}{c_{\mathrm{N}}} \underline{\underline{A}}^{\mathcal{I}} \; \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} \end{pmatrix} \overset{!}{=} 0. \end{aligned} \tag{4.17}$$

The unknowns in the set of non-linear equations (4.17) are all nodal values $(\underline{d}^{[1]}, \underline{d}^{[2]}, \underline{\lambda}_{\mathrm{N}})$. Usually, these equations are solved by applying a variant of Newton's method, thus incorporating some kind of information about the second order derivatives. For this purpose, a consistent linearization of the involved terms at the current iterate $(\underline{d}^{[1]}, \underline{d}^{[2]}, \underline{\lambda}_{\mathrm{N}})|_{\{k\}}$ has to be computed, where $k = 1, 2, \ldots$ denotes the current Newton iteration index and is placed into curly braces to distinguish it more easily from the other indices. The linearized form of (4.17) is formally obtained by

$$0 = \mathrm{D}\mathscr{L}((\underline{x}, \underline{\lambda}_{\mathrm{N}}); (\delta\underline{d}, \delta\underline{\lambda}_{\mathrm{N}}))|_{\{k\}} + \Delta(\mathrm{D}\mathscr{L}((\underline{x}, \underline{\lambda}_{\mathrm{N}}); (\delta\underline{d}, \delta\underline{\lambda}_{\mathrm{N}})); (\Delta\underline{d}, \Delta\underline{\lambda}_{\mathrm{N}}))|_{\{k\}}\,, \tag{4.18a}$$

where the linearized contact term follows as

$$\begin{aligned}
&\Delta(\mathrm{D}\mathscr{C}((\underline{x}^{[1]}, \bar{\underline{x}}^{[1]}, \underline{\lambda}_{\mathrm{N}}); (\delta\underline{d}^{[2]}, \delta\underline{d}^{[2]}, \delta\underline{\lambda}_{\mathrm{N}})); (\Delta\underline{d}^{[1]}, \Delta\underline{d}^{[2]}, \Delta\underline{\lambda}_{\mathrm{N}}))\Big|_{\{k\}} \\
&= \begin{pmatrix} \delta\underline{d}^{[1]} \\ \delta\underline{d}^{[2]} \\ \delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \\ \delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} \end{pmatrix}^{T} \underline{\underline{K}}_{\mathscr{C}}(\underline{x}^{[1]}, \underline{x}^{[2]}, \underline{\lambda}_{\mathrm{N}})\Big|_{\{k\}} \begin{pmatrix} \Delta\underline{d}^{[1]} \\ \Delta\underline{d}^{[2]} \\ \Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \\ \Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} \end{pmatrix}_{\{k\}}, \\
\underline{\underline{K}}_{\mathscr{C}}\Big|_{\{k\}} &= \begin{pmatrix} (\nabla_{\underline{d}^{[1]}}(\tilde{\nabla}_{\underline{d}^{[1]}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\ \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}))^{T} & (\nabla_{\underline{d}^{[2]}}(\tilde{\nabla}_{\underline{d}^{[1]}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\ \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}))^{T} & \tilde{\nabla}_{\underline{d}^{[1]}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}} & \underline{\underline{0}} \\ (\nabla_{\underline{d}^{[1]}}(\tilde{\nabla}_{\underline{d}^{[2]}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\ \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}))^{T} & (\nabla_{\underline{d}^{[2]}}(\tilde{\nabla}_{\underline{d}^{[2]}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\ \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}))^{T} & \tilde{\nabla}_{\underline{d}^{[2]}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}} & \underline{\underline{0}} \\ (\nabla_{\underline{d}^{[1]}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}})^{T} & (\nabla_{\underline{d}^{[2]}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}})^{T} & \underline{\underline{0}} & \underline{\underline{0}} \\ \frac{2}{c_{\mathrm{N}}}(\nabla_{\underline{d}^{[1]}}(\underline{\underline{A}}^{\mathcal{I}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}}))^{T} & \underline{\underline{0}} & \underline{\underline{0}} & \frac{2}{c_{\mathrm{N}}}\underline{\underline{A}}^{\mathcal{I}} \end{pmatrix}_{\{k\}}.
\end{aligned} \tag{4.18b}$$

Equivalent to (4.15), the variation of the diagonal tributary area matrix $\underline{\underline{A}}^{\mathcal{I}}$ with respect to the displacement degrees of freedom is omitted during the derivation of the set of non-linear equations. A direct consequence of this simplification is that the inactive solution part can be condensed from the evolving system of equations (4.18) by a simple post-processing step, i.e.

$$\Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}}\Big|_{k} = -\left\{\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} + (\underline{\underline{A}}^{\mathcal{I}}(\underline{x}^{[1]}))^{-1}(\nabla_{\underline{d}^{[1]}}(\underline{\underline{A}}^{\mathcal{I}}(\underline{x}^{[1]})\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}}))^{T}\Delta\underline{d}^{[1]}\right\}\Big|_{k}. \tag{4.19}$$

Since the matrix $\underline{\underline{A}}^{\mathcal{I}}$ is diagonal, the inverse can easily be formed. It is also quite common in the literature [65] to omit the linearization of the tributary area matrix in the linear system of equations (4.18). In this case, the inactive nodal Lagrange multipliers would be set to zero within one full Newton step. However, it has been noticed that the second term on the right hand side of (4.19) can improve the robustness of the applied non-linear solution method, since it delays the decrease of the absolute value of the inactive nodal Lagrange multipliers. In numerical examples it could be observed that this delay can avoid unnecessary active set changes and therefore improves the overall performance of the algorithm. Furthermore, the gradient computation is cheap, since it depends only on variables completely defined on the slave side. Due to these beneficial features, this term is kept in the system of equations. Finally, it is to mention that the introduced

non-symmetry has no effect whatsoever on iterative linear solution strategies, which expect a symmetric linear system of equations, since the non-symmetry can be easily removed from the system under consideration of (4.19).

4.3.2. Augmented Formulation

As mentioned by (4.13), another possibility is to derive the system of equations from the augmented Lagrangian potential (4.13) including the regularization term resulting in equations very similar to the derivations in Alart and Curnier [1], De Lorenzis et al. [65], Pietrzak and Curnier [212]. In the following, the corresponding equations will be derived. First of all, the virtual work expression in (4.14a) stays formally unchanged except for the fact that the Lagrangian potential $\mathscr{L}$ is replaced by its augmented counterpart $\mathscr{L}_{c_\mathrm{N}}$ and the contact potential $\mathscr{C}$ by the augmented contact potential $\mathscr{C}_{c_\mathrm{N}}$. Next, the gradient with respect to the displacements of the augmented contact potential is considered, i.e.

$$\nabla_{\underline{d}} \mathscr{C}_{c_\mathrm{N}} = \begin{cases} \left\{ [\tilde{g}_\mathrm{N}]^i_{,\tilde{d}^j} \, [\lambda_\mathrm{N}]_i - c_\mathrm{N} [\tilde{g}_\mathrm{N}]^i_{,\tilde{d}^j} [\hat{g}_\mathrm{N}]_i + \frac{c_\mathrm{N}}{2} A^{ii}_{\ ,\tilde{d}^j} [\hat{g}_\mathrm{N}]_i \, [\hat{g}_\mathrm{N}]_i \right\} \underline{e}^j & \forall i \in \mathcal{A}, \\ \frac{1}{2c_\mathrm{N}} A^{ii}_{\ ,d^j} \, [\lambda_\mathrm{N}]_i \, [\lambda_\mathrm{N}]_i \, \underline{e}^j \triangleq \underline{0} & \forall i \in \mathcal{I}, \end{cases} \tag{4.20}$$

where the contributions are again split into their active and inactive part. As before, the dependency on the displacements of the tributary area in the inactive part is ignored and the tilde accent over the displacement derivatives again represents the $\tilde{\nabla}(\cdot)$ operator (see Remark 4.1).

The gradient with respect to the Lagrange multipliers stays almost untouched and only the inactive part exhibits a small variation in terms of the appearing scaling factor. But since the inactive displacement dependency is ignored in (4.20), the influence of the changed factor on the linear system of equations will vanish. Now, the augmented set of non-linear equations can be stated as

$$\begin{aligned} \mathrm{D}\mathscr{L}_{c_\mathrm{N}}((\underline{x}^{[1]}, \underline{x}^{[2]}, \underline{\lambda}_\mathrm{N}); (\delta\underline{d}^{[1]}, \delta\underline{d}^{[1]}, \delta\underline{\lambda}_\mathrm{N})) &= \mathrm{D}\mathscr{L}((\underline{x}^{[1]}, \underline{x}^{[2]}, \underline{\lambda}_\mathrm{N}); (\delta\underline{d}^{[1]}, \delta\underline{d}^{[1]}, \delta\underline{\lambda}_\mathrm{N})) \\ &+ \begin{pmatrix} \delta\underline{d}^{[1]} \\ \delta\underline{d}^{[2]} \\ \delta\underline{\lambda}_\mathrm{N}^{\mathcal{A}} \\ \delta\underline{\lambda}_\mathrm{N}^{\mathcal{I}} \end{pmatrix}^\mathrm{T} \begin{pmatrix} \frac{c_\mathrm{N}}{2} \tilde{\nabla}_{\underline{d}^{[1]}} \langle \hat{\underline{g}}_\mathrm{N}^{\mathcal{A}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}), \ \hat{\underline{g}}_\mathrm{N}^{\mathcal{A}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}) \rangle_{\underline{\underline{A}}^{\mathcal{A}}} \\ \frac{c_\mathrm{N}}{2} \tilde{\nabla}_{\underline{d}^{[2]}} \langle \hat{\underline{g}}_\mathrm{N}^{\mathcal{A}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}), \ \hat{\underline{g}}_\mathrm{N}^{\mathcal{A}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}) \rangle_{\underline{\underline{A}}^{\mathcal{A}}} \\ \underline{0} \\ \frac{1}{c_\mathrm{N}} \underline{\underline{A}}^{\mathcal{I}} \, \underline{\lambda}_\mathrm{N}^{\mathcal{I}} \end{pmatrix} \overset{!}{=} 0, \end{aligned} \tag{4.21}$$

where the directional derivative of the Lagrangian potential defined in (4.17) is included as well. It can easily be seen that the influence of the regularization term vanishes at the solution. The inactive Lagrange multipliers are forced to zero, while the active weighted gap values and, consequently, the active averaged weighted gap values are equal to zero at the KKT pair fulfilling (4.9). Hence, the effect on the non-linear solution strategy is restricted to the pre-asymptotic solution path and does not change the final result.

The unknowns in the regularized set of non-linear equations are the same as in the standard Lagrangian case. Again, the intention is to use a consistently linearized Newton scheme to solve

(4.21). The augmented tangential stiffness matrix of the contact potential follows from a superposition of the Lagrangian contributions and the augmented contributions, i.e.

$$\underline{\underline{K}}_{\mathscr{C}_{c_\mathrm{N}}}\Big|_k = \underline{\underline{K}}^{\{k\}}_{\mathscr{C}} - \tilde{\underline{\underline{K}}}^{\{k\}}_{\mathscr{C}_{c_\mathrm{N}}}, \tag{4.22}$$

with the augmentation matrix

$$\tilde{\underline{\underline{K}}}_{\mathscr{C}_{c_\mathrm{N}}} = \begin{pmatrix} \frac{c_\mathrm{N}}{2}(\nabla_{\underline{d}^{[1]}}(\tilde{\nabla}_{\underline{d}^{[1]}}\langle \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N},\, \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N}\rangle_{\underline{\underline{A}}^{\mathcal{A}}}))^\mathrm{T} & \frac{c_\mathrm{N}}{2}(\nabla_{\underline{d}^{[2]}}(\tilde{\nabla}_{\underline{d}^{[1]}}\langle \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N},\, \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N}\rangle_{\underline{\underline{A}}^{\mathcal{A}}}))^\mathrm{T} & \underline{\underline{0}} & \underline{\underline{0}} \\ \frac{c_\mathrm{N}}{2}(\nabla_{\underline{d}^{[1]}}(\tilde{\nabla}_{\underline{d}^{[2]}}\langle \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N},\, \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N}\rangle_{\underline{\underline{A}}^{\mathcal{A}}}))^\mathrm{T} & \frac{c_\mathrm{N}}{2}(\nabla_{\underline{d}^{[2]}}(\tilde{\nabla}_{\underline{d}^{[2]}}\langle \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N},\, \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N}\rangle_{\underline{\underline{A}}^{\mathcal{A}}}))^\mathrm{T} & \underline{\underline{0}} & \underline{\underline{0}} \\ \underline{\underline{0}} & \underline{\underline{0}} & \underline{\underline{0}} & \underline{\underline{0}} \\ \frac{1}{c_\mathrm{N}}(\nabla_{\underline{d}^{[1]}}(\underline{\underline{A}}^{\mathcal{I}}\underline{\lambda}^{\mathcal{I}}_\mathrm{N}))^T & \underline{\underline{0}} & \underline{\underline{0}} & \frac{1}{c_\mathrm{N}}\underline{\underline{A}}^{\mathcal{I}} \end{pmatrix}.$$

First, it is to note that the introduced condensation of the inactive part (4.19) stays unchanged. Secondly, the involved second order derivatives of the scaled inner product of two averaged weighted gap vectors stand out and can be expressed by quantities also occurring in the standard approach. Therefore, the already implicitly used first order derivative of the averaged weighted gap is reconsidered, viz.

$$\nabla_{\underline{d}}\hat{\underline{g}}_\mathrm{N} = [\hat{g}_\mathrm{N}]^i{}_{,d^j}\, \underline{e}_i \otimes \underline{e}^j = [A^{-1}]^i{}_i \, \Big\{ [\tilde{g}_\mathrm{N}]^i{}_{,d^j} - A^i{}_{i,d^j}[\hat{g}_\mathrm{N}]^i \Big\}\, \underline{e}_i \otimes \underline{e}^j. \tag{4.23}$$

Keeping this in mind, the second order derivative of the scaled inner product results in

$$\begin{aligned} \frac{c_\mathrm{N}}{2}\nabla_{\underline{d}}[\tilde{\nabla}_{\underline{d}}\langle \hat{\underline{g}}_\mathrm{N},\, \hat{\underline{g}}_\mathrm{N}\rangle_{\underline{\underline{A}}}] = c_\mathrm{N}\Big\{ & [\tilde{g}_\mathrm{N}]^i{}_{,\tilde{d}^j d^k}\, [\hat{g}_\mathrm{N}]_i + [\tilde{g}_\mathrm{N}]^i{}_{,\tilde{d}^j}\, [\hat{g}_\mathrm{N}]_{i,d^k} \\ & - \frac{1}{2}A^{ii}{}_{,\tilde{d}^j d^k}\, [\hat{g}_\mathrm{N}]_i\, [\hat{g}_\mathrm{N}]_i - A^{ii}{}_{,\tilde{d}^j}[\hat{g}_\mathrm{N}]_{i,d^k}\, [\hat{g}_\mathrm{N}]_i \Big\}\, \underline{e}^j \otimes \underline{e}^k. \end{aligned} \tag{4.24}$$

One final note concerning the regularization parameter: In contrast to the standard Lagrangian scheme, c_N has a direct influence on the augmented Lagrangian solution procedure. In the standard Lagrangian case, c_N only has a minor effect, since the only row of the linear system (4.18) that contains the regularization parameter is the last row. This row, however, corresponds to the inactive Lagrange multipliers, where the parameter will cancel out as (4.19) shows. Thus, the only remaining impact is during the active set decision. In contrast, the augmented system of equations is directly influenced in form of the scaled regularization term. In summary, it can be concluded that a change of c_N should have a more remarkable impact on the non-linear solver performance.

Remark 4.2. Finally, the opportunity is taken to motivate the introduced, but maybe unfamiliar, averaged weighted gap notation in combination with the tributary area once again. First of all, the active part of the discretized Lagrangian given in (4.10) coincides completely with the result of a traditional successive variation and discretization approach. The element-wise interpolated Lagrange multiplier $\lambda_\mathrm{N}^{[1](e)}(\underline{x}) = N^{[1](e)i}(\underline{x})[\lambda_\mathrm{N}^{[1](e)}]_i$ evolves naturally from the formulation even though a pure node-based one in combination with an averaged weighted gap and a tributary

area has been inserted. The main reasons for the used definition can be found in the active set decision, which develops more fluently from the mathematical formulation (see Hiermeier et al. [131]). Furthermore, the extension to an augmented Lagrangian formulation is also much easier to achieve, since the formal lumping via the tributary area acts beneficially here [138]. Also, with regard to more sophisticated non-linear solution strategies, the approach presented here is much more accessible as e.g. shown by Temizer et al. [261]. Since mainly linearly interpolated finite elements will be considered, the positivity property remains. If a higher order interpolation is necessary, the Lagrange interpolation can be replaced by a NURBS interpolation as proposed by De Lorenzis et al. [65], Temizer et al. [261], or, alternatively, a bi-orthogonal lumping approach in combination with the necessary adaptions for higher order interpolation schemes has to be taken into account Popp et al. [218].

Concerning the convergence properties, the formulation shown here is supposed to behave exactly as the formulations summarized in Wohlmuth [279]. The only difference is the convergence to a different solution in the case of strong geometrical non-linearities due to the more consistent formulation.

4.4. Variation and Linearization of Discretized Contact Kinematics

This section is meant to provide a brief summary of all involved contact variations and linearizations. The focus is on the differences between the complete and incomplete variational approaches as already mentioned in Remark 4.1. It is important to understand what the underlying assumptions and simplifications are and what the direct consequences will be.

Remark 4.3. The numerical integration is performed in the parameter space of each slave element using the well-known Gaussian quadrature. In general, this will introduce an integration error. For a comparison between element-based and segment-based integration in the context of mortar methods in contact mechanics the interested reader is referred to Farah et al. [83]. Since the influence of the introduced integration error on the full variational approach is remarkable, this topic will be discussed in Section 4.7.2.

The following notation is introduced for the first and second order total directional derivatives, for example in $\underline{v}$ and $\underline{w}$ direction:

$$\mathrm{D}_{\underline{v}}(\cdot) = \langle \nabla_{\underline{d}}(\cdot), \underline{v} \rangle, \qquad \mathrm{D}_{\underline{w}}(\mathrm{D}_{\underline{v}}(\cdot)) = \langle \underline{v}, [\nabla^2_{\underline{d}\,\underline{d}}(\cdot)]^T \underline{w} \rangle. \tag{4.25}$$

4.4.1. Variation

In (4.17), the variation of the normal gap vector occurs, which can be expressed as the consistent directional derivative of the weighted gap vector

$$\mathrm{D}_{\delta \underline{d}}(\tilde{g}_{\mathrm{N}}) = \mathop{\mathsf{A}}_{e=1}^{|\mathcal{E}^{[1]}|} \Big\{ \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1](e)i}\, \mathrm{D}_{\delta \underline{u}}(g_{\mathrm{N}}(\underline{x}^{[1](e)}, \underline{x}^{[2](\bar{e})}))\, j^{[1](e)}\, \mathrm{d}\xi^{[1](e)j} \tag{4.26a}$$

$$+ \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1](e)i}\, g_{\mathrm{N}}(\underline{x}^{[1](e)}, \underline{x}^{[2](\bar{e})})\, \mathrm{D}_{\delta \underline{u}}(j^{[1](e)})\, \mathrm{d}\xi^{[1](e)j} \Big\}. \tag{4.26b}$$

Note that the integration is executed in the parameter space of each slave interface element in its current configuration, which is supposed to be indicated by the given integration bounds. Afterwards, the element contributions are assembled over all slave elements $\mathcal{E}^{[1]}$ to form the global directional derivative. The first summand (4.26a) contains the directional derivative of the discrete gap (4.6). To improve the readability, the element superscript will be omitted in the following and the derivative yields

$$\mathrm{D}_{\delta \underline{u}}(g_{\mathrm{N}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]})) = \langle \mathrm{D}_{\delta \underline{u}}(\hat{\underline{n}}^{[1]}), \bar{\underline{x}}^{[2]} - \underline{x}^{[1]} \rangle + \langle \hat{\underline{n}}^{[1]}, \mathrm{D}_{\delta \underline{u}}(\bar{\underline{x}}^{[2]}) - \mathrm{D}_{\delta \underline{u}}(\underline{x}^{[1]}) \rangle. \tag{4.27}$$

Here, the directional derivative of the interpolated smooth normal field is obtained by

$$\mathrm{D}_{\delta \underline{u}}(\hat{\underline{n}}^{[1]}) = \mathrm{D}_{\delta \underline{u}} \left(\frac{\breve{\underline{n}}^{[1]}}{\|\breve{\underline{n}}^{[1]}\|} \right) = \frac{1}{\|\breve{\underline{n}}^{[1]}\|} \left(\underline{\underline{I}} - \hat{\underline{n}}^{[1]} \otimes \hat{\underline{n}}^{[1]} \right) \cdot N^{[1]}{}_{i} \mathrm{D}_{\delta \underline{u}}(\breve{\underline{n}}^{[1]i}), \tag{4.28}$$

where (4.3) and (4.4) have been inserted. For more details, the reader is referred to Popp et al. [215], Yang et al. [290]. Since the distance vector between a slave point and its projected counterpart on the master side is parallel to the smooth normal vector, the first summand in (4.27) vanishes. At this point it must be emphasized, however, that this will only be strictly true if the (more costly) smooth unit normal (4.5) is considered, while for the projection algorithm the smooth non-unit normal (4.4) would be sufficient. If the second order derivative of (4.5) has to be evaluated as well, this can lead to a significant performance loss based on the actual implementation. The second summand of (4.27) contains the variation of the slave point $\underline{x}^{[1]}$ and its projected counterpart $\bar{\underline{x}}^{[2]}$. Yielding

$$\mathrm{D}_{\delta \underline{u}}(\underline{x}^{[b]}) = \frac{\partial x^{[b]i}}{\partial u^r} \delta u^r + \underline{\tau}^{[b]}{}_{i}\, \mathrm{D}_{\delta \underline{u}}(\xi^{[b]i}) = \delta \underline{u}^{[b]} + \underline{\tau}^{[b]}{}_{i}\, \mathrm{D}_{\delta \underline{u}}(\xi^{[b]i}), \tag{4.29}$$

where the last term is equal to zero for the slave side contribution. However, for the master side, the parametric coordinates are deformation dependent and the projection (2.54) has to be taken into account. This leads to a local system of equations, which has to be solved to obtain the directional derivatives. For more details the interested reader is referred to Appendix A.1.

The second term (4.26b) depends on the directional derivative of the determinant of the slave element Jacobian $j^{[1](e)} := \det(\underline{\underline{J}}(\xi^{[1](e)i})) : \mathbb{R}^2 \to \mathbb{R}$. This and all remaining variations can be also found in the already mentioned Appendix A.1. Thus, all necessary terms have been computed and the consistent variation of the weighted normal gap is fully defined.

4.4.2. Incomplete Variational Approach

If the consistent first order directional derivative of the weighted gap vector is considered in the set of non-linear equations (4.17), the necessary second order derivatives in (4.18) can become computationally expensive (see Section 4.4.3). Due to this fact, a second, so-called incomplete, but computationally much cheaper variational approach is introduced with respect to the weighted gap in (4.26). By skipping the variation of the element Jacobian determinant and the variation of the master parameter space coordinate, the following expression is obtained:

$$\tilde{\mathrm{D}}_{\delta \underline{d}}(\tilde{\underline{g}}_{\mathrm{N}}) = \overset{|\mathcal{E}^{[1]}|}{\underset{e=1}{\mathsf{A}}} \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^i(\xi^{[1](e)j})\ \tilde{\mathrm{D}}_{\delta \underline{u}}(g_{\mathrm{N}}^{(e)})\ j^{[1](e)}\ \mathrm{d}\xi^{[1](e)j}, \tag{4.30}$$

where $\tilde{\mathrm{D}}_{\delta \underline{u}}(g_{\mathrm{N}}^{(e)}) = \langle \hat{\underline{n}}^{[1]}, \delta \bar{\underline{u}}^{[2]} - \delta \underline{u}^{[1]} \rangle$ holds. In doing so, the derived linear system of equations (4.18) will become non-symmetric (see also Remark 4.1). For a short explanation and interpretation of the used assumptions leading to (4.30), two major simplifications of the variation have to be considered. The first one is the variation of the Jacobian determinant. By consideration of (4.26b), it can be argued that the influence of this term will vanish at the solution point, since it is demanded that the weighted gap $\tilde{\underline{g}}_{\mathrm{N}}(\underline{x}^\star)$ is equal to zero and therefore $g_{\mathrm{N}}(\underline{x}^\star)$ is equal to zero as well, at least in a weak sense. The second simplification concerns the variation of the projected parametric master coordinates. With respect to the correct variational form defined in (4.27) and by insertion of (4.29), it can be seen that the cumbersome directional derivatives are pointing in the tangential directions of the active master contact interface denoted by the corresponding convective base vectors. Now, the assumption is used that in the converged state and under the prerequisite that the slave and master sides are discretized sufficiently fine, $\langle \hat{\underline{n}}^{[1]}(\underline{x}^\star), \underline{\tau}^{[2]}{}_i(\bar{\underline{x}}^\star) \rangle \approx 0$, $\forall i \in \{1, 2\}$ holds true. But anyhow, this is only strictly true if a non-smooth closest point projection is used (see e.g. Laursen [170]) or if the special case of infinitesimal sliding occurs Puso and Laursen [222]. In any other case it remains an approximation, which, however, becomes better and better with an increasing number of elements, such that the spatial solutions of both cases are supposed to asymptotically coincide. Since the influence of the neglected terms on the discrete solution cannot be fully predicted, both approaches will be followed. To the best of the author's knowledge, this is together with [131] the first comprehensive comparison of numerical results for both approaches. The comparison can be found within Section 4.7. Furthermore, an additional example has been added in Section 4.7.4 revealing a possible instability which is clearly based on the neglected variation of the projected parametric master coordinate.

4.4.3. Linearization

In the following, the directional derivatives of the set of non-linear equations (4.17) are going to be derived. Starting with the linearization of the product of the tributary area matrix and the nodal inactive Lagrange multiplier vector, which is defined as

$$\begin{aligned}
&\mathrm{D}_{\Delta\underline{d}}(\underline{\underline{A}}^{\mathcal{I}}(\underline{x}^{[1]})\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}}) = \langle \nabla_{\underline{d}^{[1]}}\underline{\underline{A}}^{\mathcal{I}}(\underline{x}^{[1]})\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}}, \Delta\underline{d}\rangle \\
&= \mathop{\mathsf{A}}_{e=1}^{|\mathcal{E}^{[1]}|} [\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}}]^{(e)i} \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N_i(\xi^{[1](e)j})\, \mathrm{D}_{\Delta\underline{u}}(j^{[1](e)})\, \mathrm{d}\xi^{[1](e)j}, \qquad \forall i \in \mathcal{S}, \qquad (4.31)
\end{aligned}$$

where again the directional derivative of the already known slave element Jacobian determinant comes into play. Next, the consistent mixed second order directional derivative of the weighted gap vector is obtained by repeated application of the chain rule

$$\begin{aligned}
\mathrm{D}_{\Delta\underline{d}}(\mathrm{D}_{\delta\underline{d}}(\tilde{\underline{g}}_{\mathrm{N}})) = \mathop{\mathsf{A}}_{e=1}^{|\mathcal{E}^{[1]}|} \Big\{ &\iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1](e)i}\, \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(g_{\mathrm{N}}^{(e)}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]})))\, j^{[1](e)}\, \mathrm{d}\xi^{[1](e)j} && (4.32a)\\
&+ \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1](e)i}\, \mathrm{D}_{\delta\underline{u}}(g_{\mathrm{N}}^{(e)}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}))\, \mathrm{D}_{\Delta\underline{u}}(j^{[1](e)})\, \mathrm{d}\xi^{[1](e)j} && (4.32b)\\
&+ \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1](e)i}\, \mathrm{D}_{\Delta\underline{u}}(g_{\mathrm{N}}^{(e)}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}))\, \mathrm{D}_{\delta\underline{u}}(j^{[1](e)})\, \mathrm{d}\xi^{[1](e)j} && (4.32c)\\
&+ \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1](e)i}\, g_{\mathrm{N}}^{(e)}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]})\, \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(j^{[1](e)}))\, \mathrm{d}\xi^{[1](e)j} \Big\}. && (4.32d)
\end{aligned}$$

The element affiliation is omitted and the consistent second order directional derivative of the variation of the discrete gap is derived by

$$\begin{aligned}
\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(g_{\mathrm{N}}(\underline{x}^{[1]}, \bar{\underline{x}}^{[2]}))) = \; &\langle \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]})), \bar{\underline{x}}^{[2]} - \underline{x}^{[1]}\rangle \\
&+ \langle \mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}), \mathrm{D}_{\Delta\underline{u}}(\bar{\underline{x}}^{[2]}) - \mathrm{D}_{\Delta\underline{u}}(\underline{x}^{[1]})\rangle \\
&+ \langle \mathrm{D}_{\Delta\underline{u}}(\hat{\underline{n}}^{[1]}), \mathrm{D}_{\delta\underline{u}}(\bar{\underline{x}}^{[2]}) - \mathrm{D}_{\delta\underline{u}}(\underline{x}^{[1]})\rangle \\
&+ \langle \hat{\underline{n}}^{[1]}, \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\bar{\underline{x}}^{[2]})) - \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\underline{x}^{[1]}))\rangle, \qquad (4.33)
\end{aligned}$$

where the first order directional derivative of the interpolated smooth normal is given in (4.28), while the mixed second order directional derivative is given by

$$\begin{aligned}
\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]})) = \frac{1}{\|\breve{\underline{n}}^{[1]}\|} \Big\{ &- \langle \hat{\underline{n}}^{[1]}, \mathrm{D}_{\Delta\underline{u}}(\breve{\underline{n}}^{[1]})\rangle\, \mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}) \\
&- \Big[\mathrm{D}_{\Delta\underline{u}}(\hat{\underline{n}}^{[1]}) \otimes \hat{\underline{n}}^{[1]} + \hat{\underline{n}}^{[1]} \otimes \mathrm{D}_{\Delta\underline{u}}(\hat{\underline{n}}^{[1]})\Big] \cdot \mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]}) \\
&+ \Big(\underline{\underline{I}} - \hat{\underline{n}}^{[1]} \otimes \hat{\underline{n}}^{[1]}\Big) \cdot \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]})) \Big\}. \qquad (4.34)
\end{aligned}$$

The mixed second order directional derivative of the variation of an interface point, defined in (4.29), follows as

$$\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\underline{x}^{[b]})) = \mathrm{D}_{\Delta\underline{u}}(\delta\underline{u}^{[b]}) + \mathrm{D}_{\Delta\underline{u}}(\underline{\tau}^{[b]}{}_{i})\mathrm{D}_{\delta\underline{u}}(\xi^{[b]i}) + \underline{\tau}^{[b]}{}_{i}\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\xi^{[b]i})), \tag{4.35}$$

$$\mathrm{D}_{\Delta\underline{u}}(\delta\underline{u}^{[b]}) = \frac{\partial(\delta\underline{u}^{[b]})}{\partial\xi^{[b]i}}\mathrm{D}_{\Delta\underline{u}}(\xi^{[b]i}). \tag{4.36}$$

The entire derivative (4.35) is equal to zero for the slave side. However, for a point on the master body the terms do not vanish. The last term can be derived from the linearized version of the variation of the projection, which involves the second order derivative of the smooth normal field (4.4). Under consideration of all zero-terms on the slave side, (4.33) can be simplified and

$$\begin{aligned}\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(g_{\mathrm{N}})) = \; & \langle \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]})), \bar{\underline{x}}^{[2]} - \underline{x}^{[1]} \rangle + \langle \mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}), \mathrm{D}_{\Delta\underline{u}}(\bar{\underline{x}}^{[2]}) - \Delta\underline{u}^{[1]} \rangle \\ & + \langle \mathrm{D}_{\Delta\underline{u}}(\hat{\underline{n}}^{[1]}), \mathrm{D}_{\delta\underline{u}}(\bar{\underline{x}}^{[2]}) - \delta\underline{u}^{[1]} \rangle + \langle \hat{\underline{n}}^{[1]}, \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\bar{\underline{x}}^{[2]})) \rangle \end{aligned} \tag{4.37}$$

is obtained. The remaining linearized variation of the element Jacobian in (4.32d), as well as all other relations, such as the detailed derivation of the linearized variation of the projection, including the smooth normal field, can be found in Appendix A.2.

4.4.4. Linearization of the Incomplete Variational Approach

Finally, the consistently linearized version of the simplified variational approach yields

$$\begin{aligned}\mathrm{D}_{\Delta\underline{d}}(\tilde{\mathrm{D}}_{\delta\underline{d}}(\tilde{\underline{g}}_{\mathrm{N}})) = \mathop{\mathsf{A}}_{e=1}^{|\mathcal{E}^{[1]}|} \Big\{ & \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1](e)i}\; \mathrm{D}_{\Delta\underline{u}}(\tilde{\mathrm{D}}_{\delta\underline{u}}(g_{\mathrm{N}}^{(e)}))\; j^{[1](e)}\; \mathrm{d}\xi^{[1](e)j} \\ & + \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1](e)i}\; \tilde{\mathrm{D}}_{\delta\underline{u}}(g_{\mathrm{N}}^{(e)})\; \mathrm{D}_{\Delta\underline{u}}(j^{[1](e)})\; \mathrm{d}\xi^{[1](e)j} \Big\}, \end{aligned} \tag{4.38}$$

where the directional derivative of the incomplete variation of the normal gap follows as

$$\begin{aligned}\mathrm{D}_{\Delta\underline{u}}(\tilde{\mathrm{D}}_{\delta\underline{u}}(g_{\mathrm{N}}^{(e)})) = \; & \langle \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]})), \bar{\underline{x}}^{[2]} - \underline{x}^{[1]} \rangle + \langle \mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}), \mathrm{D}_{\Delta\underline{u}}(\bar{\underline{x}}^{[2]}) - \Delta\underline{u}^{[1]} \rangle \\ & + \langle \mathrm{D}_{\Delta\underline{u}}(\hat{\underline{n}}^{[1]}), \delta\bar{\underline{u}}^{[2]} - \delta\underline{u}^{[1]} \rangle + \langle \hat{\underline{n}}^{[1]}, \mathrm{D}_{\Delta\underline{u}}(\delta\bar{\underline{u}}^{[2]}) \rangle. \end{aligned} \tag{4.39}$$

The linearized smooth normal field is defined in (4.28), while the derivative of the variation of the master point is given in (4.36).

4.5. Conservation Laws

In this section, it will be shown that the complete as well as the incomplete approach both fulfill the conservation of angular momentum with respect to the semi-discrete system, i.e. only

the spatially discretized system will be considered and any side-effects of an underlying time integration scheme are excluded for this discussion. The conservation of linear and angular momentum of the semi-discrete system provides a basis for the consistency proof of the employed finite element discretization scheme. In detail, all contact contributions to the force balance of the non-linear set of equations given in (4.17) and (4.21) will be investigated: On the one hand, there is the Lagrange multiplier term and on the other hand the regularization term.

In the following, it will be firstly demonstrated that it is easily possible to prove conservation of linear momentum for both formulations and all terms, subsequently, the focus is turned to the conservation of angular momentum. The reader is also kindly referred to the literature on this topic (e.g. Laursen [170], Puso and Laursen [222]).

4.5.1. Linear Momentum

As proposed by Puso and Laursen [222], conservation of linear momentum can be shown by inserting a constant discrete vector $\underline{w} \neq \underline{0}$ into (4.17) and (4.21) for all nodal displacement weighting coefficients, i.e. $\delta \underline{d}^{[b]i}$, $\forall i \in \{1, \dots, n\}$, $b \in \{1, 2\}$ is replaced by the vector $\underline{w}$, such that

$$\langle \underline{w}, \nabla_{\underline{x}} \mathscr{C}_{c_\mathrm{N}} \rangle = \langle \underline{w},\ [\tilde{g}_\mathrm{N}]^i{}_{,\tilde{d}^j}\ \{[\lambda_\mathrm{N}]_i - c_\mathrm{N}[\hat{g}_\mathrm{N}]_i\}\, \underline{e}^j \rangle + \frac{c_\mathrm{N}}{2} \langle \underline{w},\ A^{ii}{}_{,\tilde{d}^j} [\hat{g}_\mathrm{N}]_i\ [\hat{g}_\mathrm{N}]_i \underline{e}^j \rangle = 0, \tag{4.40}$$

where the definition of (4.20) is inserted. The proof of the linear momentum conservation begins under consideration of the second inner product in (4.40). For this term, it is sufficient to consider only one arbitrary slave element. Thus, the demand

$$\left([\hat{g}_\mathrm{N}^{\mathcal{A}}]^i\right)^2 \iint\limits_{\gamma^{[1](e)}_{\mathrm{c},[-1,1]}} N^{[1]}{}_i(\xi^{[1]j})\, \langle \underline{n}^{[1]}(\xi^{[1]j}), \mathrm{D}_{\underline{w}}(\hat{\underline{n}}^{[1]}(\xi^{[1]j})) \rangle \, \mathrm{d}\xi^{[1]j} = 0 \tag{4.41}$$

can be formulated. Now, solely the inner product of the directional derivative of the non-unit element normal and the unit element normal are taken into account and (A.6) is inserted. This yields

$$\langle \underline{n}^{[1]}, \mathrm{D}_{\underline{w}}(\hat{\underline{n}}^{[1]}) \rangle = \langle \frac{\underline{\tau}^{[1]}{}_1 \times \underline{\tau}^{[1]}{}_2}{\| \underline{\tau}^{[1]}{}_1 \times \underline{\tau}^{[1]}{}_2 \|}, \mathrm{D}_{\underline{w}}(\underline{\tau}^{[1]}{}_1) \times \underline{\tau}^{[1]}{}_2 + \underline{\tau}^{[1]}{}_1 \times \mathrm{D}_{\underline{w}}(\underline{\tau}^{[1]}{}_2) \rangle = 0 \tag{4.42}$$

since the following holds

$$\mathrm{D}_{\underline{w}}(\underline{\tau}^{[1]}{}_k) = N^{[1]l}{}_{,\xi^k} \underline{w}_l = \sum_l N^{[1]l}{}_{,\xi^k} \underline{w} = \underline{0}, \qquad \forall \underline{w} \in \mathbb{R}^3, \tag{4.43}$$

where the partition of unity is used such that $\sum_l N^{[1]l}{}_{,\xi^k} = 0$ is true for each element.

Next, the first term in (4.40) is considered leading to

$$[\tilde{\lambda}_{\mathrm{N}}]^i \Big\{ \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1]}{}_i \langle \hat{\underline{n}}^{[1]}, \mathrm{D}_{\underline{w}}(\bar{\underline{x}}^{[2]} - \underline{x}^{[1]}) \rangle \, j^{[1]} \, \mathrm{d}\xi^{[1]j} + \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1]}{}_i \langle \hat{\underline{n}}^{[1]}, \bar{\underline{x}}^{[2]} - \underline{x}^{[1]} \rangle \, \mathrm{D}_{\underline{w}}(j^{[1]}) \, \mathrm{d}\xi^{[1]j} \Big\} = 0, \tag{4.44}$$

where the abbreviation $[\tilde{\lambda}_{\mathrm{N}}]^i = ([\lambda_{\mathrm{N}}]^i - c_{\mathrm{N}}[\hat{g}_{\mathrm{N}}]^i)$ is introduced and, further, the second term in (4.44) vanishes following the same idea as above. Now, the attention is drawn to the first summand in (4.44). Inserting (4.27) and (4.29), the term yields

$$[\tilde{\lambda}_{\mathrm{N}}]^i \Big\{ \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1]}{}_i \langle \hat{\underline{n}}^{[1]}, \bar{N}^{[2]}{}_k \underline{w}^k - N^{[1]}{}_l \underline{w}^l \rangle \, j^{[1]} \, \mathrm{d}\xi^{[1]j} + \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1]}{}_i \langle \hat{\underline{n}}^{[1]}, \bar{\underline{\tau}}^{[2]}{}_r \mathrm{D}_{\underline{w}}(\bar{\xi}^{[2]r}) \rangle \, j^{[1]} \, \mathrm{d}\xi^{[1]j} \Big\} = 0, \tag{4.45}$$

where the first term is equal to zero since the partition of unity holds for each element resulting in

$$\bar{N}^{[2]}{}_k \underline{w}^k - N^{[1]}{}_l \underline{w}^l = \left(\sum_k \bar{N}^{[2]}{}_k - \sum_l N^{[1]}{}_l \right) \underline{w} = (1-1)\underline{w} = \underline{0}, \qquad \forall \underline{w} \in \mathbb{R}^3. \tag{4.46}$$

Concerning the second term in (4.45), the reformulation

$$\begin{aligned} \bar{\underline{\tau}}^{[2]}{}_i \mathrm{D}_{\underline{w}}(\bar{\xi}^{[2]i}) =& \bar{\underline{\tau}}^{[2]}{}_i [L_\chi^{-1}]^{ij} \left(N^{[1]}{}_l w^{lj} - \bar{N}^{[2]}{}_k w^{kj} + \bar{\alpha}_\chi \mathrm{D}_{w^j}(\breve{\underline{n}}^{[1]}) \right) \\ =& \bar{\underline{\tau}}^{[2]}{}_i [L_\chi^{-1}]^{ij} \left((1-1)\, w^j + \bar{\alpha}_\chi \mathrm{D}_{w^j}(\breve{\underline{n}}^{[1]}) \right) = \bar{\alpha}_\chi \bar{\underline{\tau}}^{[2]}{}_i [L_\chi^{-1}]^{ij} \mathrm{D}_{w^j}(\breve{\underline{n}}^{[1]}), \end{aligned} \tag{4.47}$$

is possible, where again the partition of unity for the first part comes into play. Finally, it is shown that the directional derivative of the non-unit smooth normal field will become zero, if a constant direction $\underline{w}$ is considered for all nodes. Indeed, this is the case since

$$\begin{aligned} \mathrm{D}_{\underline{w}}(\breve{\underline{n}}^{[1]}) =& N^{[1]}{}_k \frac{1}{\|\tilde{\underline{n}}^{[1]k}\|} (\underline{\underline{I}} - \breve{\underline{n}}^{[1]k} \otimes \breve{\underline{n}}^{[1]k}) \\ & \left\{ \sum_{e=1}^{\mathrm{N}_{\mathrm{adj}}^k} \frac{1}{\|\hat{\underline{n}}^{[1](e)}\|} (\underline{\underline{I}} - \underline{n}^{[1](e)} \otimes \underline{n}^{[1](e)}) \mathrm{D}_{\underline{w}}(\hat{\underline{n}}^{[1](e)}) \right\} \end{aligned} \tag{4.48}$$

and

$$\begin{aligned}\mathrm{D}_{\underline{w}}(\underline{\hat{n}}^{[1]}) =& \mathrm{D}_{\underline{w}}(\underline{\tau}^{[1]}{}_1) \times \underline{\tau}^{[1]}{}_2 + \underline{\tau}^{[1]}{}_1 \times \mathrm{D}_{\underline{w}}(\underline{\tau}^{[1]}{}_2) \\ =& \sum_i N^{[1]}{}_{i,\xi^{[1]1}} \underline{w} \times \underline{\tau}^{[1]}{}_2 + \underline{\tau}^{[1]}{}_1 \times \sum_j N^{[1]}{}_{j,\xi^{[1]2}} \underline{w} = \underline{0} \times \underline{\tau}^{[1]}{}_2 + \underline{\tau}^{[1]}{}_1 \times \underline{0} = \underline{0} \end{aligned} \tag{4.49}$$

hold true. Hence, the conservation of the linear momentum is satisfied.

4.5.2. Angular Momentum

To prove the conservation of angular momentum, $\underline{w} \times \underline{x}^{[b]i}$ is inserted into $\delta \underline{d}^{[b]i}$, $\forall i \in \{1, \dots, \mathrm{N_n}\}$, $b \in \{1, 2\}$. In summary, the former demand (4.40) slightly changes to

$$\langle \underline{w} \times \underline{x}^j,\ [\tilde{g}_\mathrm{N}]^i{}_{,\tilde{d}^j}\ \{[\lambda_\mathrm{N}]_i - c_\mathrm{N}[\hat{g}_\mathrm{N}]_i\}\, \underline{e}^j \rangle + \frac{c_\mathrm{N}}{2} \langle \underline{w} \times \underline{x}^j,\ A^{ii}{}_{,\tilde{d}^j}[\hat{g}_\mathrm{N}]_i\ [\hat{g}_\mathrm{N}]_i \underline{e}^j \rangle = 0. \tag{4.50}$$

The discussion shall begin with the gradient of the tributary area as part of the regularization term. Under consideration of the Jacobian identity

$$\begin{aligned}\underline{0} =& \underline{\tau}^{[1]}{}_1 \times (\underline{\tau}^{[1]}{}_2 \times \underline{w}) + \underline{\tau}^{[1]}{}_2 \times (\underline{w} \times \underline{\tau}^{[1]}{}_1) + \underline{w} \times (\underline{\tau}^{[1]}{}_1 \times \underline{\tau}^{[1]}{}_2) \\ =& \underline{w} \times \underline{\hat{n}}^{[1]} - \underline{\tau}^{[1]}{}_1 \times (\underline{w} \times \underline{\tau}^{[1]}{}_2) - (\underline{w} \times \underline{\tau}^{[1]}{}_1) \times \underline{\tau}^{[1]}{}_2, \end{aligned} \tag{4.51}$$

the directional derivative of the non-unit element normal yields

$$\mathrm{D}_{\underline{w} \times \underline{x}}(\underline{\hat{n}}^{[1]}) = (\underline{w} \times \underline{\tau}^{[1]}{}_1) \times \underline{\tau}^{[1]}{}_2 + \underline{\tau}^{[1]}{}_1 \times (\underline{w} \times \underline{\tau}^{[1]}{}_2) = \underline{w} \times \underline{\hat{n}}^{[1]}, \tag{4.52}$$

where $\mathrm{D}_{\underline{w} \times \underline{x}^k}(\underline{\tau}^{[1]}{}_i) = N^{[1]}{}_{k,\xi^{[1]i}}(\underline{w} \times \underline{x}^k) = \underline{w} \times \underline{\tau}^{[1]}{}_i$ has been used. Inserting (4.52) into the directional derivative of the Jacobian results in

$$\langle \underline{n}^{[1]},\ \mathrm{D}_{\underline{w} \times \underline{x}^k}(\underline{\hat{n}}^{[1]}) \rangle = \langle \underline{n}^{[1]}, \underline{w} \times \underline{\hat{n}}^{[1]} \rangle = 0, \tag{4.53}$$

since the unit slave element normal and the non-unit slave element normal point into the same direction and therefore the enclosed volume of the triple product is zero. Looking at the first term in (4.50) reveals that the angular momentum check for (4.26b), as the part of the weighted gap variation concerning the variation of the slave Jacobian determinant, vanishes for the same reason. Only the directional derivative of the discrete gap remains and can be expressed as

$$\begin{aligned}[\tilde{\lambda}_\mathrm{N}]^i \Big\{ & \iint\limits_{\gamma^{[1](e)}_{\mathrm{c},[-1,1]}} N^{[1]}{}_i \langle \underline{\hat{n}}^{[1]}, \bar{N}^{[2]}{}_k (\underline{w} \times \underline{x}^k) - N^{[1]}{}_l (\underline{w} \times \underline{x}^l) \rangle\, j^{[1]}\, \mathrm{d}\xi^{[1]j} \\ & + \iint\limits_{\gamma^{[1](e)}_{\mathrm{c},[-1,1]}} N^{[1]}{}_i \langle \underline{\hat{n}}^{[1]}, \bar{\underline{\tau}}^{[2]}{}_r \mathrm{D}_{\underline{w} \times \underline{x}}(\bar{\xi}^{[2]r}) \rangle\, j^{[1]}\, \mathrm{d}\xi^{[1]j} \Big\} = 0, \end{aligned} \tag{4.54}$$

where again the abbreviation $[\tilde{\lambda}_{\mathrm{N}}]^i = ([\lambda_{\mathrm{N}}]^i - c_{\mathrm{N}}[\hat{g}_{\mathrm{N}}]^i)$ is used. Due to the fact that this must hold for all constant $\underline{w} \neq \underline{0}$, the demand $\langle \underline{a}, \underline{w} \times \underline{b} \rangle = 0$ for two given vectors $\underline{a}, \underline{b} \in \mathbb{R}^3$ becomes equivalent to the alternative demand $\underline{a} \times \underline{b} = \underline{0}$, i.e. the two vectors must be parallel. Following this idea,

$$[\tilde{\lambda}_{\mathrm{N}}]^i \iint\limits_{\gamma_{\mathrm{c},[-1,1]}^{[1](e)}} N^{[1]}{}_i \left[\underline{\hat{n}}^{[1]} \times (\bar{N}^{[2]}{}_k \underline{x}^k - N^{[1]}{}_l \underline{x}^l) \right] j^{[1]} \, \mathrm{d}\xi^{[1]j} = \underline{0}, \tag{4.55}$$

is obtained from the first summand, where (4.55) vanishes since the vectors $\underline{\hat{n}}^{[1]}$ and $\bar{\underline{x}}^{[2]} - \underline{x}^{[1]}$ are collinear at each evaluation point as a result of the employed projection rule (2.54). Further, the reformulated inner product of the second summand in (4.54) leads to

$$\begin{aligned} \langle \underline{\hat{n}}^{[1]}, \bar{\underline{\tau}}^{[2]}{}_i \mathrm{D}_{\underline{w} \times \underline{x}}(\bar{\xi}^{[2]i}) \rangle = \frac{1}{\det(\bar{\underline{\underline{L}}}_\chi)} \Big\{ & \langle \underline{\hat{n}}^{[1]}, \bar{\underline{\tau}}^{[2]}{}_1 \rangle \langle \bar{\underline{\tau}}^{[2]}{}_2 \times \breve{\underline{n}}^{[1]}, \, \underline{r}^{\chi^\star}_{\underline{w} \times \underline{x}} \rangle \\ & - \langle \underline{\hat{n}}^{[1]}, \bar{\underline{\tau}}^{[2]}{}_2 \rangle \langle \bar{\underline{\tau}}^{[2]}{}_1 \times \breve{\underline{n}}^{[1]}, \, \underline{r}^{\chi^\star}_{\underline{w} \times \underline{x}} \rangle \Big\} = 0, \end{aligned} \tag{4.56}$$

where the identities of the first and second row of the $\bar{\underline{\underline{L}}}_\chi^{-1} \in \mathbb{R}^{3 \times 3}$ matrix have been inserted (see (A.10)). For the proof, a closer look at the introduced residual expression of the fulfilled projection becomes necessary following as

$$\underline{r}^{\chi^\star}_{\underline{w} \times \underline{x}} = \underline{w} \times (\underline{x}^{[1]} - \bar{\underline{x}}^{[2]}) + \bar{\alpha}_\chi \mathrm{D}_{\underline{w} \times \underline{x}}(\breve{\underline{n}}^{[1]}), \tag{4.57}$$

where the directional derivative of the interpolated smooth non-unit normal field is given by

$$\begin{aligned} \mathrm{D}_{\underline{w} \times \underline{x}}(\breve{\underline{n}}^{[1]}) = & N^{[1]}{}_k \frac{1}{\|\tilde{\underline{n}}^{[1]k}\|} (\underline{\underline{I}} - \breve{\underline{n}}^{[1]k} \otimes \breve{\underline{n}}^{[1]k}) \\ & \left\{ \sum_{e=1}^{\mathrm{N}^k_{\mathrm{adj}}} \frac{1}{\|\hat{\underline{n}}^{[1](e)}\|} (\underline{\underline{I}} - \underline{n}^{[1](e)} \otimes \underline{n}^{[1](e)}) \mathrm{D}_{\underline{w} \times \underline{x}}(\hat{\underline{n}}^{[1](e)}) \right\} \end{aligned} \tag{4.58}$$

under consideration of (4.52). In the next considered derivation, the abbreviation $\underline{v} = \bar{\underline{\tau}}^{[2]}{}_i \times \breve{\underline{n}}^{[1]}$, for $i = \{1, 2\}$ is used. Again, the fact comes into play that (4.56) has to hold for all vectors $\underline{w} \neq \underline{0}$, such that it can be written as

$$\underline{v} \times (\underline{x}^{[1]} - \bar{\underline{x}}^{[2]}) + \bar{\alpha}_\chi v^i N^{[1]}{}_k \frac{1}{\|\tilde{\underline{n}}^{[1]k}\|} \breve{T}^k{}_{il} \sum_{e=1}^{\mathrm{N}^k_{\mathrm{adj}}} \left\{ \frac{1}{\|\hat{\underline{n}}^{[1](e)}\|} T^{(e)lm} \, \varepsilon^r{}_{mq} \hat{n}^{(e)q} \right\} \underline{e}_r = \underline{0}, \tag{4.59}$$

where $\varepsilon^r{}_{mq}$ denotes the Levi-Civita symbol. The matrix $\breve{\underline{\underline{T}}}^k$ is the node-wise orthographic projection matrix onto the smooth tangential plain

$$\breve{\underline{\underline{T}}}^{k} = (\delta^{ij} - \breve{n}^{[1]ki}\breve{n}^{[1]kj})\, \underline{e}_i \otimes \underline{e}_j, \tag{4.60}$$

while $\underline{\underline{T}}^{(e)}$ is the element-wise orthographic projection matrix onto the tangential plain of each adjacent element $e \in \{1, \dots, \mathrm{N}^k_{\mathrm{adj}}\}$ evaluated at the current node k, i.e.

$$\underline{\underline{T}}^{(e)} = (\delta^{ij} - n^{[1](e)i}\, n^{[1](e)j})\, \underline{e}_i \otimes \underline{e}_j. \tag{4.61}$$

To prove (4.59), first, the following identity has to be considered

$$\begin{aligned}\tilde{\underline{\underline{W}}}^{[1]k}_{\times} &= \sum_{e=1}^{\mathrm{N}^k_{\mathrm{adj}}} \left\{ \frac{1}{\|\hat{\underline{n}}^{[1](e)}\|} T^{(e)lm}\, \varepsilon^{r}{}_{mq} \hat{n}^{(e)q} \right\} \underline{e}_l \otimes \underline{e}_r = \sum_{e=1}^{\mathrm{N}^k_{\mathrm{adj}}} \left\{ T^{(e)lm}\, \varepsilon^{r}{}_{mq} n^{(e)q} \right\} \underline{e}_l \otimes \underline{e}_r \\ &= \sum_{e=1}^{\mathrm{N}^k_{\mathrm{adj}}} \sum_{i=1}^{3} (\underline{n}^{[1](e)} \times \underline{e}_i) \otimes \underline{e}_i = \begin{pmatrix} 0 & -\tilde{n}^{[1]k3} & \tilde{n}^{[1]k2} \\ \tilde{n}^{[1]k3} & 0 & -\tilde{n}^{[1]k1} \\ -\tilde{n}^{[1]k2} & \tilde{n}^{[1]k1} & 0 \end{pmatrix}, \end{aligned} \tag{4.62}$$

where $\tilde{\underline{\underline{W}}}^{[1]k}_{\times}$ denotes the so-called skew-symmetric cross-product matrix $[\tilde{\underline{n}}^{[1]k}]_{\times}$ of each element node k. Inserting (4.62) into the second part of (4.59) yields

$$\bar{\alpha}_\chi v^i \sum_k N^{[1]}{}_k \frac{1}{\|\tilde{\underline{n}}^{[1]k}\|} \breve{T}^k_{il} [\tilde{W}^{[1]}_{\times}]^{klr} \underline{e}_r = \bar{\alpha}_\chi v_i [\breve{W}^{[1]}_{\times}]^{ir} \underline{e}_r = \bar{\alpha}_\chi (\underline{v} \times \breve{\underline{n}}^{[1]}), \tag{4.63}$$

where $\breve{\underline{\underline{W}}}^{[1]}_{\times}$ denotes the cross-product matrix $[\breve{\underline{n}}^{[1]}]_{\times}$. Finally, under consideration of the distance relation $(\underline{x}^{[1]} - \bar{\underline{x}}^{[2]}) = -\bar{\alpha}_\chi \breve{\underline{n}}^{[1]}$ that follows from the satisfied projection rule (2.54), the left hand side of (4.59) vanishes for all vectors $\underline{v} \in \mathbb{R}^3$. Hence, the conservation of angular momentum is fulfilled for all involved terms.

4.5.3. Final Remarks

Towards the end of this section, it is highly emphasized that the conservation of both, linear and angular momentum, is not automatically true for all contact formulations. Especially for the case of mortar-based methods, the fulfillment is strongly coupled to the way how the normal gap definition is integrated into the force balance equations. But at the same time, it is also not necessarily true that all variations have to be considered to achieve full conservation properties for the semi-discrete system. The incomplete approach provides full conservation, while it neglects parts of the full variation. Thus, even the approach denoted here as incomplete can actually be considered as being more consistent for large deformations than many well-established mortar-based approaches in the literature (e.g. Popp et al. [215, 216], Puso and Laursen [222], Yang and Laursen [288], Yang et al. [290]), which use similar simplifications, but additionally lack the conservation of angular momentum as discussed in Puso and Laursen [222]. In all cases, the conservation of angular momentum can become troublesome, if the fully discrete system, i.e., discretized in space and time, is considered. For more details the reader is kindly referred to the literature on energy-momentum conserving time integrators, e.g. Hesch and Betsch [129, 130], as well as to the discussion about numerical time integration following in the next Section 4.6.

4.6. Numerical Time Integration

Next, the attention is on the necessary extensions for the treatment of dynamic structural contact problems under consideration of the Generalized-α time integration scheme. Other schemes are not considered. The discussion starts where Section 2.4 has ended. By adding the contact contributions, the dynamic balance equations (2.85) become

$$\underline{r}_{\mathrm{cg}\alpha} = \underline{\underline{M}}\,\underline{a}^{\{n+1-\alpha_\mathrm{m}\}} + \underline{\underline{C}}\,\underline{v}^{\{n+1-\alpha_\mathrm{f}\}} + \nabla_{\underline{d}}\mathscr{U}(\underline{d}^{\{n+1-\alpha_\mathrm{f}\}}) \tag{4.64a}$$

$$+ \nabla_{\underline{d}}\mathscr{V}_{\mathrm{ext}}(\underline{d}^{\{n+1-\alpha_\mathrm{f}\}}) - \nabla_{\underline{d}}\tilde{\underline{g}}_\mathrm{N}(\underline{d}^{\{n+1-\alpha_\mathrm{f}\}})(\underline{\lambda}_\mathrm{N}^{\{n+1-\alpha_\mathrm{f}\}} - c_\mathrm{N}\hat{\underline{g}}_\mathrm{N}^{\{n+1-\alpha_\mathrm{f}\}}). \tag{4.64b}$$

This is the most general version which also considers the regularization term of the augmented Lagrangian formulation. As already mentioned in Section 2.4, there exist at least two possibilities to evaluate the different terms at the respective mid-time points $t_{n+1-\alpha_\mathrm{f}}$ and $t_{n+1-\alpha_\mathrm{m}}$. The first one is the so-called *implicit mid-point rule*, which asks for the evaluation of the different terms at these time points. The second one uses a *trapezoidal rule* to approximate the corresponding quantities. Throughout this thesis only the second variant is followed, thus, (4.64a) can be rewritten as

$$\underline{r}_{\mathrm{cg}\alpha} \approx \underline{\underline{M}}\,[(1-\alpha_\mathrm{m})\underline{a}^{\{n+1\}} + \alpha_\mathrm{m}\underline{a}^{\{n\}}] \tag{4.65a}$$

$$+ (1-\alpha_\mathrm{f})\,\nabla_{\underline{d}}\mathscr{U}|_{\underline{d}^{\{n+1\}}} + \alpha_\mathrm{f}\,\nabla_{\underline{d}}\mathscr{U}|_{\underline{d}^{\{n\}}} \tag{4.65b}$$

$$+ (1-\alpha_\mathrm{f})\,\nabla_{\underline{d}}\mathscr{V}_{\mathrm{ext}}|_{\underline{d}^{\{n+1\}}} + \alpha_\mathrm{f}\,\nabla_{\underline{d}}\mathscr{V}_{\mathrm{ext}}|_{\underline{d}^{\{n\}}} \tag{4.65c}$$

$$- (1-\alpha_\mathrm{f})\,\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_\mathrm{N}\Big|_{\underline{d}^{\{n+1\}}}(\underline{\lambda}_\mathrm{N}^{\{n+1\}} - c_\mathrm{N}\hat{\underline{g}}_\mathrm{N}^{\{n+1\}}) - \alpha_\mathrm{f}\,\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_\mathrm{N}\Big|_{\underline{d}^{\{n\}}}(\underline{\lambda}_\mathrm{N}^{\{n\}} - c_\mathrm{N}\hat{\underline{g}}_\mathrm{N}^{\{n\}}). \tag{4.65d}$$

This is the implemented residual which is also consistently linearized. Consequently, the solution is reached as soon as $\underline{r}_{\mathrm{cg}\alpha} = \underline{0}$ holds and the remaining KKT-conditions (4.9) are fulfilled as well. Now, while the choice of the contact contributions in (4.65) seems completely reasonable, the question may rise how to formulate the constraint conditions. In a continuous, time dependent set-up the *Hertz-Signorini-Moreau* conditions proposed in (2.56) must be extended by

$$p_\mathrm{N}(\underline{x}(\underline{X},t)\,\dot{g}_\mathrm{N}(\underline{x}(\underline{X},t)) = 0, \tag{4.66}$$

where this newly added condition is the so-called *persistency condition*, which demands for non-zero contact pressure only during persistent contact. This equation plays a crucial role when it comes to the construction of energy conserving discrete time integration schemes incorporating contact. For a much deeper insight the reader is referred to Laursen and Chawla [171], Laursen and Love [172]. Therein, the so-called *discrete energy-momentum* method, firstly introduced by Simo and Tarnow [250], has been applied to contact problems. In their first publication [170] the energy conservation has been achieved with the drawback of a remaining finite interpenetration of the contacting bodies. Although this penetration does not affect the energy and momentum transfer, it is still something which should be avoided. Therefore, in their second work [172] the penetration issue could be resolved but this time the used velocity update scheme destroyed

the second order accuracy of the underlying time integration scheme. With regard to the energy momentum method a variety of extensions can be found in the literature. The interested reader is for instance referred to [5, 6, 167] and the references therein. A method which combines energy momentum conservation and second order accuracy in the context of mortar-type contact formulations is presented in Hesch and Betsch [129, 130].

There are even more issues, however, which ask for a special algorithmic treatment when it comes to unilateral dynamic contact and all of them are somehow related to the force balance and the constraint equations. For example, significant artificial oscillations might be produced at the contact interface. This problem can be even more severe when the time step size and/or the spatial discretization are refined as described in Carpenter [43]. One possible counteraction can be found in Deuflhard et al. [70] based on Kane et al. [152]. In Deuflhard et al. [70] the force balance is modified by an additional predictor displacement state. This modification asks for a necessary L_2-projection. In order to avoid this additional computational effort, the authors of [70] suggest to use a lumped mass matrix. A completely different idea is proposed by Hager et al. [122], Hager and Wohlmuth [123]. Herein, a modified integration scheme is used to remove the mass from the contact interface.

This short interlude had just the purpose to highlight some of the difficulties which come along with dynamic contact simulations. However, the topic of this thesis is the robust treatment of contact simulations with the focus on the non-linear solver part, therefore, all these mentioned possible modifications shall be excluded for now. Nevertheless, they are not forgotten and should be addressed in the future. In this thesis the satisfaction of the constraints is enforced at the new point in time t_{n+1}, yielding

$$[\hat{g}_{\mathrm{N}}(\underline{d}^{\star})]^{i}_{n+1} \geqslant 0, \quad [\lambda^{\star}_{\mathrm{N}}]^{i}_{n+1} \geqslant \underline{0}, \quad [\lambda^{\star}_{\mathrm{N}}]^{i}_{n+1}\,[A(\underline{d}^{\star})]^{n+1}_{ii}\,[\hat{g}_{\mathrm{N}}(\underline{d}^{\star})]^{i}_{n+1} = 0, \quad \forall i \in \mathcal{S}. \tag{4.67}$$

So, at first glance this stands in contrast to the contact force balance presented in (4.65d) and, indeed, the unsatisfied persistency condition (4.66) is one of the major drawbacks of this formulation. In Section 6.8.1, however, a possible extension of the complementarity condition will be proposed which makes it possible to find a capable objective function which allows a consistent derivation of (4.65) and (4.67). Thus, in combination with the velocity update method [172] and the energy momentum method [167, 250], an extension to a globally convergent, energy conserving scheme should be possible. The remaining drawback would be still only first order accuracy in time. An advantage of (4.67) is that $\hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{n\}} = \underline{0}$ holds, thus, (4.65d) can be simplified.

To conclude this section the resulting system of equations for dynamic contact problems shall be briefly presented. Therefore, the system is horizontally split into one part corresponding to the variation with respect to displacements and the second one corresponding to the variation with respect to the Lagrange multipliers. The first part becomes

$$[\underline{\underline{K}}_{\mathrm{g}\alpha} - (1-\alpha_{\mathrm{f}})\{[\nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}})]^{\mathrm{T}} \tag{4.68a}$$

$$-\frac{c_{\mathrm{N}}}{2}[\nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\langle\hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}},\hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\rangle_{\underline{\underline{A}}^{\mathcal{A}}})]^{\mathrm{T}}\}]\Big|_{(\underline{d}^{\{n+1\}}_{\{k\}},\underline{\lambda}^{\{n+1\}}_{\mathrm{N}\{k\}})}\Delta\underline{d}^{\{n+1\}}_{\{k\}} \tag{4.68b}$$

$$-(1-\alpha_{\mathrm{f}})\,[\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]\Big|_{(\underline{d}^{\{n+1\}}_{\{k\}},\underline{\lambda}^{\{n+1\}}_{\mathrm{N}\{k\}})}\Delta\underline{\lambda}^{\mathcal{A}\{n+1\}}_{\mathrm{N}\{k\}} = -\,\underline{r}_{\mathrm{cg}\alpha}\Big|_{(\underline{d}^{\{n+1\}}_{\{k\}},\underline{\lambda}^{\{n+1\}}_{\mathrm{N}\{k\}})}, \tag{4.68c}$$

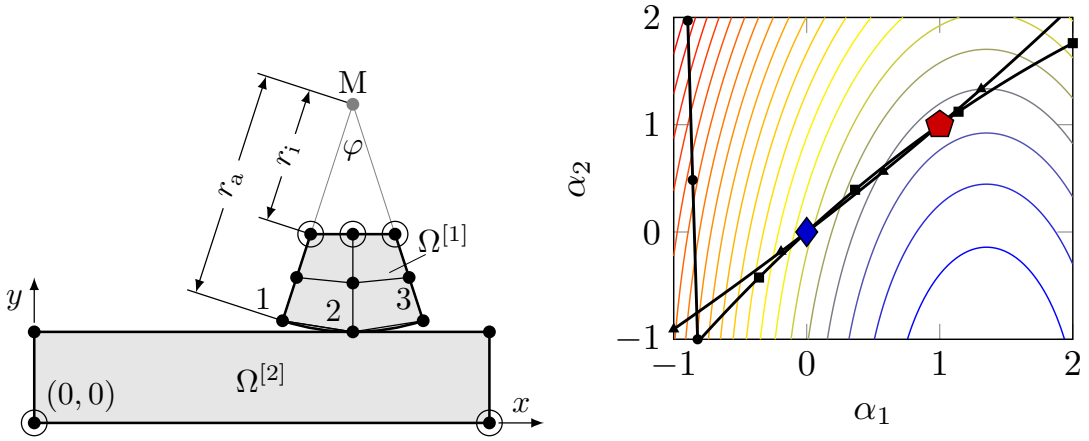


Figure 4.2.: Left: reference configuration of the circular segment pressed onto a rectangle; right: Contour lines of the parametrized objective function $\mathscr{U}(\alpha^{[1]}, \alpha^{[2]})$. The points ♦ and ⬟ denote the solutions of the incomplete and complete variational approaches, respectively. The zero isolines —▲—, —■—, —●— correspond to the nodal weighted gap constraints of the slave nodes 1, 2 and 3.

where $\underline{\underline{K}}_{g\alpha}$ has been defined in (2.96). The second part stays unchanged in comparison with (4.67), i.e., the constraint rows coincide with the (quasi-)static formulation. By neglecting all terms in (4.65) and (4.68) which are explicitly multiplied with c_N, the system of equations for the standard Lagrangian case is obtained. Further investigations will follow in Section 6.8.1 and within the examples in the Sections 6.10.6 and 6.10.7.

4.7. Numerical Examples

In the following, four different examples with varying objectives are presented. In the first example, the incomplete and complete approach are compared to each other and a detailed error analysis is presented. In the second example, the influence of the numerical integration error is investigated and possible remedies are provided. The third example is supposed to demonstrate the numerical robustness of the different methods and constraint enforcement strategies. Furthermore, a comparison between a well-established mortar method Popp et al. [215, 216] and the incomplete variational approach is drawn. Finally, a possible instability of variationally inconsistent formulation will be revealed in the last example. All presented results have been computed with the in-house C++ research code of the Institute for Computational Mechanics Wall and Kronbichler [274]. Finally, it is mentioned that more information about the often applied tangential predictor at iteration zero can be found in Appendix B.

4.7.1. Circular Segment and Rectangle

First, the influence of the different variational approaches onto the final numerical results shall be investigated. Therefore, a very simple two dimensional example is considered, which is shown in Figure 4.2. The master body is a rectangle of height 2 and length 10. The slave body is a segment of a circle, where the xy-coordinates of the center point M are chosen as $(7, 7)$, while the angle φ is equal to $\pi/5$. The radius r_a is 5 and r_i is 3. The circled nodes ⊙ denote Dirichlet

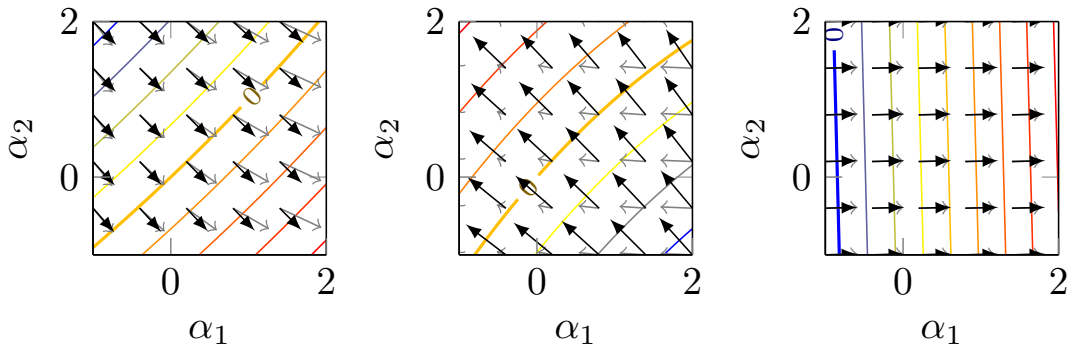


Figure 4.3.: Left to right: Contour lines of the weighted gap constraints incl. their corresponding gradients for the slave nodes 1, 2 and 3. The consistently computed gradients are denoted by black arrows, while the incomplete gradients are represented by gray arrows. Note that the gradient magnitudes of slave node 2 are scaled differently with a ratio of incomplete to complete of $1:8$.

nodes. The bottom degrees of freedom of the master body are completely fixed and the degrees of freedom at the top of the slave body are fixed in x-direction, while in y-direction a prescribed displacement is applied. First the displacement is ramped from 0 to -1 in one load step. A simple compressible Neo-Hookean material law (2.27) under a plane strain assumption is used, where $\nu^{[b]} = 0.27$ and $E^{[1]} = E^{[2]} > 0$ hold.

For the following discussion, the coarse mesh from Figure 4.2 will be considered, where the final displacement vectors $\underline{d}_{\mathrm{c}}^{\star}$ of the complete variational approach and $\underline{d}_{\mathrm{i}}^{\star}$ of the incomplete approach differ in the ℓ_2-norm to the amount of $8.226\,\mathrm{E}{-2}$. For illustration reasons the two-dimensional parametrization

$$\underline{d}_{\mathrm{i}\to\mathrm{c}}^{\star}(\alpha^{[1]}, \alpha^{[2]}) = \underline{d}_{\mathrm{i}}^{\star} + \sum_{b=1}^{2} \alpha^{[b]} (\underline{d}_{\mathrm{c}}^{[b]\star} - \underline{d}_{\mathrm{i}}^{[b]\star}) \tag{4.69}$$

shall be introduced. If both parameters are zero, the incomplete solution is obtained, while both parameters equal to one recover the complete solution. The parameterization makes it possible to visualize a two-dimensional slice through the underlying multi-dimensional objective function $\mathscr{U} : \mathbb{R}^{16} \to \mathbb{R}$.

As visualized in the right part of Figure 4.2, the solution of the complete approach reaches the minimal feasible energy level, while the incomplete approach converges to a non-optimal point. Nevertheless, both solutions are feasible with respect to the active constraints. Note that the constraint of slave node 3 is inactive at the solution. A closer look at the directional derivatives in Figure 4.3 reveals the reason for the different solutions. The varying directions of the gradients cause a different displacement solution, while any perturbation in magnitude and direction becomes manifest in different values of the corresponding Lagrange multipliers. In this case the normalized Lagrange multiplier values $\bar{\lambda}_{\mathrm{N}}^{i}$ of the active slave nodes $i \in \{1, 2\}$ for the complete approach result in $\bar{\lambda}_{\mathrm{N}}^{1} = 1.642\mathrm{e}{-1}$, $\bar{\lambda}_{\mathrm{N}}^{2} = 1.0$, while $\bar{\lambda}_{\mathrm{N}}^{1} = 1.271\mathrm{e}{-1}$, $\bar{\lambda}_{\mathrm{N}}^{2} = 1.027$ are obtained for the incomplete approach.

In a next step, this very simple example is taken into account for quantitative error estimation. Therefore, the slave body is successively refined while the master body stays discretized with only one quadrilateral linear finite element. As mentioned in Section 4.4.2 two major influences

can be distinguished: The first one stems from the neglected variation of the Jacobian determinant. Measures for the introduced error per active slave node as well as for the total mean square error are given by

$$e_{\mathrm{jac}} = \sqrt{\frac{\sum_i^{|\mathcal{A}|} (e^i_{\mathrm{jac}})^2}{|\mathcal{A}|}}, \qquad e^i_{\mathrm{jac}} = \| \mathop{\mathsf{A}}_{e=1}^{|\mathcal{E}^{[1]}|} \iint\limits_{\gamma^{[1](e)}_{\mathrm{c},[-1,1]}} N^{[1](e)i}\, g_{\mathrm{N}}\, j^{[1](e)}{}_{,\underline{d}^k}\, \mathrm{d}\xi^{[1](e)j} \|. \tag{4.70}$$

The second error source is the neglected variation of the convective master parameter coordinate, which is measured by

$$e_{\mathrm{ma}} = \sqrt{\frac{\sum_i^{|\mathcal{A}|} (e^i_{\mathrm{ma}})^2}{|\mathcal{A}|}}, \tag{4.71a}$$

$$e^i_{\mathrm{ma}} = \| \mathop{\mathsf{A}}_{e=1}^{|\mathcal{E}^{[1]}|} \iint\limits_{\gamma^{[1](e)}_{\mathrm{c},[-1,1]}} N^{[1](e)i}\, \langle \hat{\underline{n}}^{[1]}, \bar{\underline{\tau}}^{[2]}{}_r\, \bar{\xi}^{[2]r}{}_{,\underline{d}^k} \rangle j^{[1](e)}\, \mathrm{d}\xi^{[1](e)j} \|. \tag{4.71b}$$

Both quantities are evaluated in Figure 4.4. In Figures 4.4a and 4.4b the evolution of the two nodal errors is plotted over a total of eleven Newton iterations and over the reference position angles of the active nodes expressed in terms of the angle $\varphi \in (0, \frac{\pi}{5})$ defined in Figure 4.2. Additionally, the predictor step, i.e. $\#0$, is highlighted in red. Here, a 100×100 element mesh for the slave body has been considered as well as a reduced prescribed displacement of -0.4 in y-direction that is applied in one load step.

The error due to the missing Jacobian determinant in Figure 4.4a flattens over the iterations. The exponentially decreasing error of the Jacobian determinant towards the center of the active contact zone seems to be originating from a slight oscillation of the discrete gap (4.6), which decays with a sufficient distance to the contact zone boundary. Note that these oscillations of the discrete gap can occur even though the norm of the accumulated weighted gap vector already more than fulfills the predefined tolerance with a value below $1.0\mathrm{E}{-17}$ in the final converged configuration.

The variation of the parametric master coordinate does also include a direct linear coupling to the current penetration in terms of the auxiliary distance factor $\bar{\alpha}_\chi$ (see e.g. (4.56) and (4.57)). Nevertheless, the effect is almost completely canceled out during the Newton iterations. The explanation becomes obvious by looking at the corresponding deformed configurations. Only the very first predictor step shows a characteristic penetration shape (see Figure 4.4d), which is also clearly represented in the related error plot (see the red highlighted curve in Figure 4.4b). In all following Newton iterations the active slave surface is approximately parallel to the master surface (see Figures 4.4e and 4.4f). Hence, the only remaining error contributions are almost completely based on constant geometrical quantities.

A convergence study of the two mean square error measures reveals that both terms decay slowly in linear fashion. This is shown in Figure 4.4c. Further investigations show that a refinement of the master side only has a very slight influence on the error magnitude. The slope

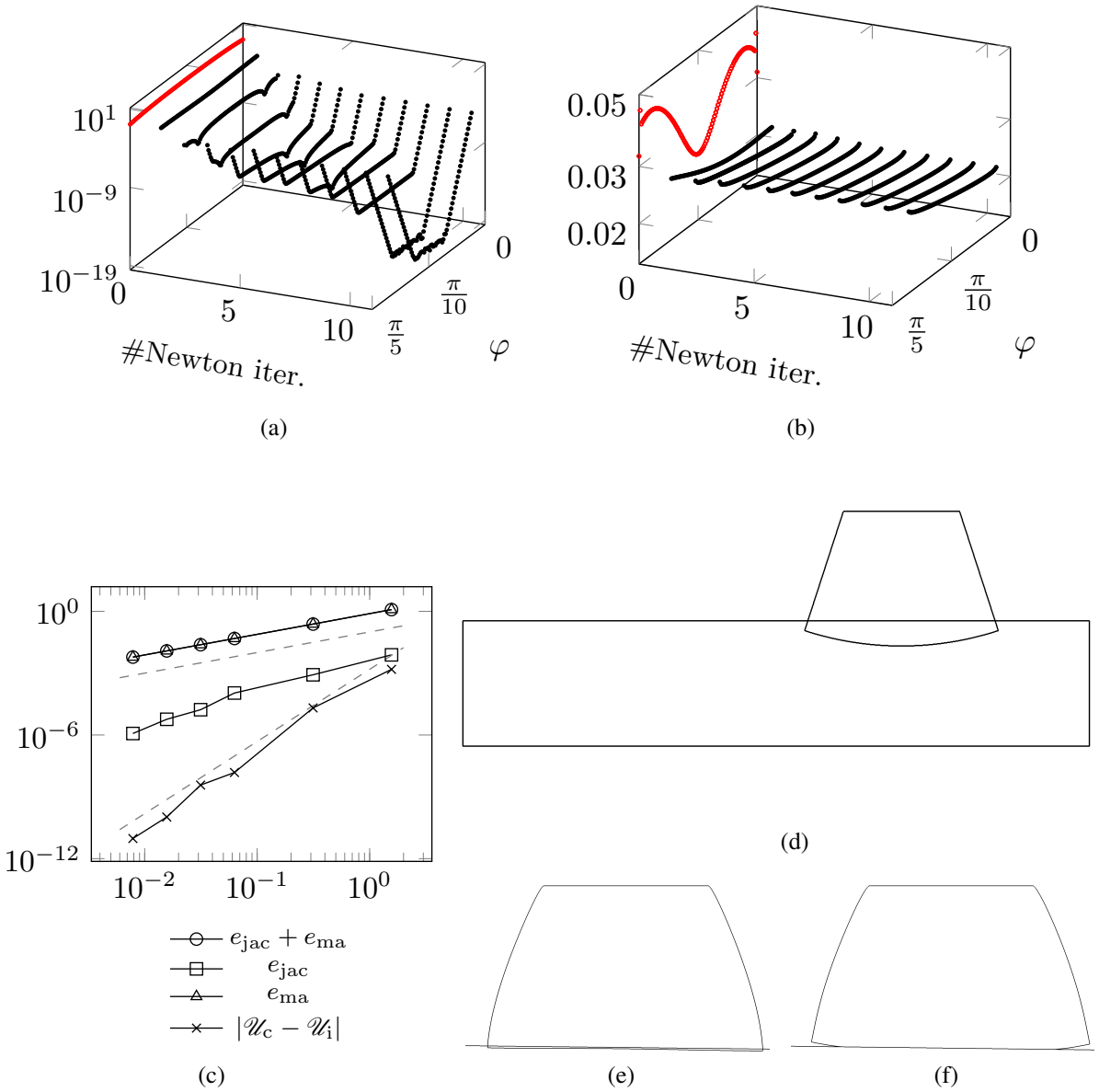


Figure 4.4.: Figure 4.4a shows the nodal errors due to the neglected variation of the Jacobian determinant and Figure 4.4b due to the neglected projected parametric master coordinates. The predictor iteration (#0) is highlighted in red. Figure 4.4c shows the total errors over the characteristic slave element length h. All three plots in the first row use a logarithmic scale on the z-axis. In Figures 4.4d to 4.4f the corresponding deformed configurations of the predictor (#0), the first Newton iteration (#1) and the final converged iteration (#11) are presented, respectively, where the two last figures show a close-up view around the slave body.

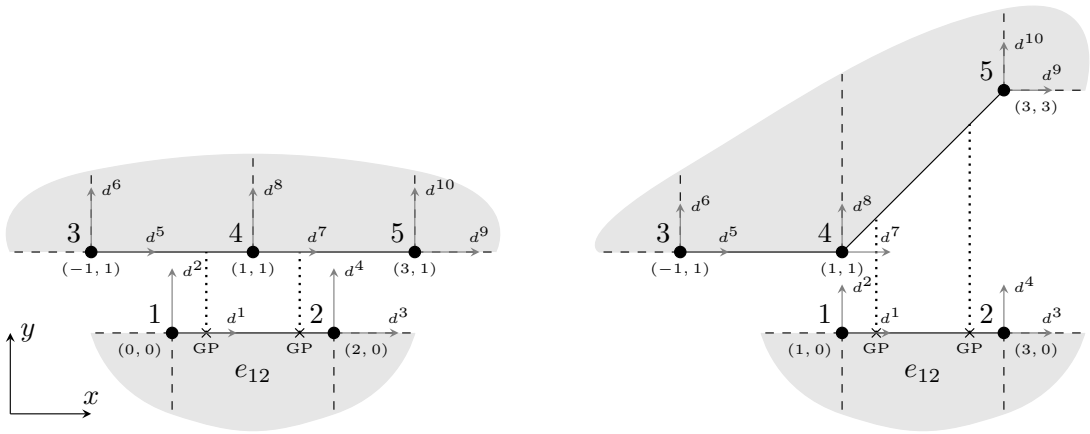


Figure 4.5.: Figure 4.5a sketches the geometrical initial configuration, while Figure 4.5b shows the final deformed configuration for β equal to 1.0.

remains at one. Finally, the difference in the total strain energy of the two approaches is observed. Interestingly, the strain energy difference decays very rapidly with a slope of 3.5 in the double logarithmic convergence plot.

In summary it can be concluded that the final solution is almost not changed for a sufficiently fine mesh and the example considered here. Furthermore, the necessary level of refinement is not extraordinarily high to achieve reliable results with the incomplete approach. However, during the pre-asymptotic phase the impact of these errors can be remarkably higher. Therefore, even so the solution might stay almost the same, the way to the solution can severely change. The possible impact will be presented in Section 4.7.4.

4.7.2. Influence of the Integration Error

Before the discussion moves on to more complex examples, the influence of the integration error mentioned in Remark 4.3 must be addressed. Therefore, the very simple example in Figure 4.5 is going to be investigated. The displacement field $d^1 = d^3 = \beta$ and $d^{10} = 2\beta$ for $\beta \in (0, 1)$ shall be considered. The remaining displacements shall be equal to zero. The lower body is supposed to be the slave body and the upper is the master body, respectively. The analytically calculated contribution of the slave element e_{12} between node 1 and 2 to the parametrized weighted gap value of node 1 yields

$$\tilde{g}_{\mathrm{N}}^{(e_{12})1}(\beta) = \int_{-1}^{1} N^{[1]1}(\xi^{[1]})\, g_N(\xi^{[1]}, \beta)\, \mathrm{d}\xi^{[1]} = \frac{1}{12}(\beta^4 + 3\beta^3 + 3\beta^2 + \beta + 12), \tag{4.72}$$

with respect to the piecewise defined gap function

$$g_{\mathrm{N}}(\xi^{[1]}, \beta) = 1 + \max\{0, \beta(1 + \bar{\xi}^{[2]}(\xi^{[1]}))\} = 1 + \max\{0, \beta(\beta + \xi^{[1]})\} \tag{4.73}$$

for all $\xi^{[1]} \in [-1,1]$. Since the analytically calculated weighted gap expression is readily available, the exact derivative with respect to the deformation parameter β is easily obtained as

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\tilde{g}_{\mathrm{N}}^{(e_{12})1} = \frac{1}{12}(4\beta^3 + 9\beta^2 + 6\beta + 1). \tag{4.74}$$

To achieve the same exact result in a finite element context, a lot more effort has to be taken. The Jacobian determinant of the slave element is deformation independent in this example. Nevertheless, the projection onto two master elements makes it necessary to introduce two new parameter coordinates for the evolving segments, namely $\eta, \zeta \in [-1,1]$: The first one for the segment between nodes 1 and 4, and the second one for the segment between nodes 4 and 2. The contribution of the first segment to the total variation of the weighted gap of node 1 follows as

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\tilde{g}_{\mathrm{N}}^{(e_{12})1,\mathrm{seg1}} = \int_{-1}^{1} \frac{\partial N^{[1]}}{\partial \xi^{[1]}} \frac{\partial \xi^{[1]}}{\partial \beta} \frac{\partial \xi^{[1]}}{\partial \eta} + N^{[1]} \frac{\partial^2 \xi^{[1]}}{\partial \eta \partial \beta} \mathrm{d}\eta = -\frac{1}{2}(1+\beta), \tag{4.75}$$

where the gap is equal to one over the whole segment. The second segment contributes with

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\beta}\tilde{g}_{\mathrm{N}}^{(e_{12})1,\mathrm{seg2}} =& \int_{-1}^{1} \frac{\partial N^{[1]}}{\partial \xi^{[1]}} \frac{\partial \xi^{[1]}}{\partial \beta} g_{\mathrm{N}}^{\mathrm{seg2}} \frac{\partial \xi^{[1]}}{\partial \zeta} + N^{[1]} \frac{\mathrm{d} g_{\mathrm{N}}^{\mathrm{seg2}}}{\mathrm{d}\beta} \frac{\partial \xi^{[1]}}{\partial \zeta} + N^{[1]} g_{\mathrm{N}}^{\mathrm{seg2}} \frac{\partial^2 \xi^{[1]}}{\partial \zeta \partial \beta} \mathrm{d}\zeta \\ =& \frac{1}{3}\beta^3 + \frac{3}{4}\beta^2 + \beta + \frac{7}{12}, \end{aligned} \tag{4.76}$$

where the total derivative of the gap function is given by

$$\frac{\mathrm{d} g_{\mathrm{N}}^{\mathrm{seg2}}}{\mathrm{d}\beta} = \left[\frac{\partial g_{\mathrm{N}}^{\mathrm{seg2}}}{\partial \beta} + \frac{\partial g_{\mathrm{N}}^{\mathrm{seg2}}}{\partial \bar{\xi}^{[2]}} \frac{\partial \bar{\xi}^{[2]}}{\partial \beta} + \frac{\partial g_{\mathrm{N}}^{\mathrm{seg2}}}{\partial \bar{\xi}^{[2]}} \frac{\partial \bar{\xi}^{[2]}}{\partial \xi^{[1]}} \frac{\partial \xi^{[1]}}{\partial \beta} \right]. \tag{4.77}$$

By summation of both contributions, the result in (4.74) is recovered. A closer look at the part coming from the first segment (4.75) shows, that in the limit case $\beta = 1$ the integral degenerates to a point evaluation, which has to be considered since the second Jacobian determinant is now deformation dependent and contributes to the final derivative value. To compute a comparative solution for the incomplete approach, the total derivative given in (4.77) is reduced to the first term and all derivatives with respect to the slave parameter space coordinate are neglected as well. This yields

$$\frac{\tilde{\mathrm{d}}}{\tilde{\mathrm{d}}\beta}\tilde{g}_{\mathrm{N}}^{(e_{12})1} = \int_{-1}^{1} N^{[1]}(\xi^{[1]})\ \max\{0, \beta + \xi^{[1]}\}\ \mathrm{d}\xi^{[1]} = \frac{1}{12}(\beta^3 + 3\beta^2 + 3\beta + 1). \tag{4.78}$$

Next, the attention is drawn to the element-wise Gaussian quadrature used here and the following expression

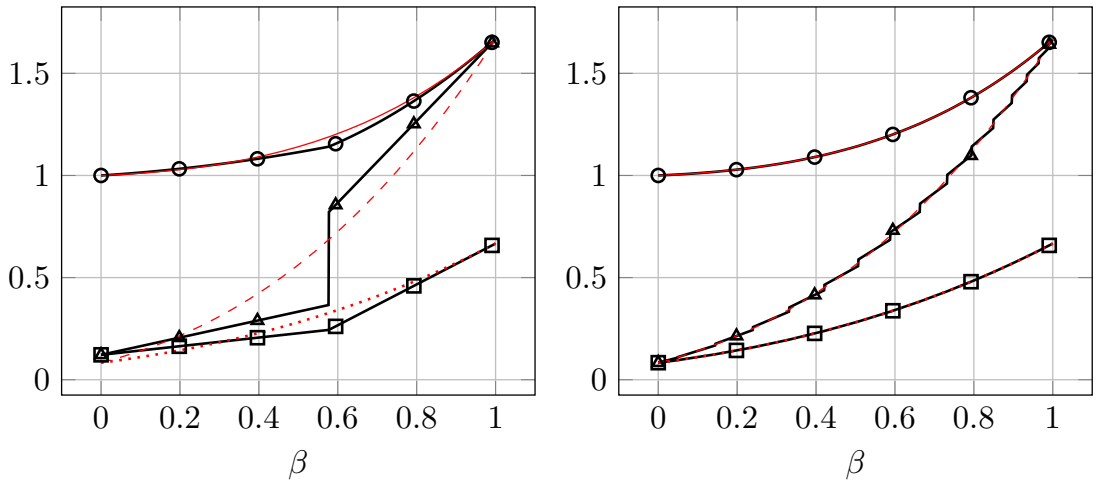


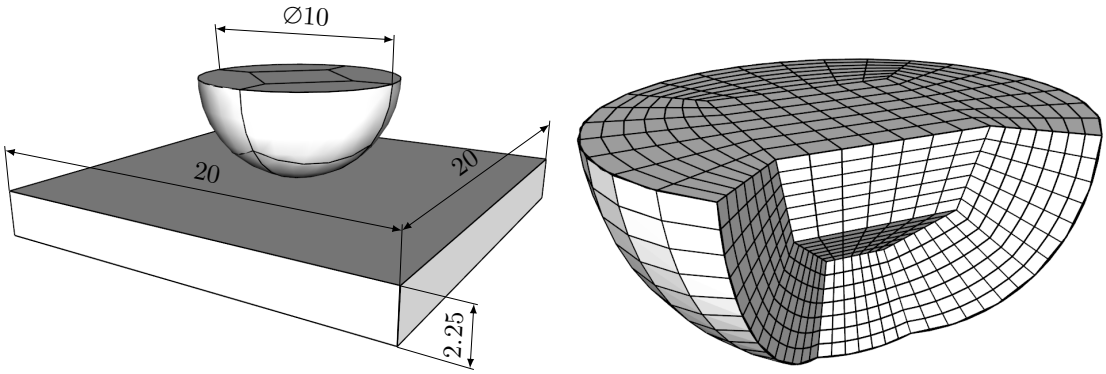
Figure 4.6.: Visualization of the weighted gap and the related gradients with respect to the deformation parameter β. (a): Result of Gaussian quadrature with 2 Gauss-points for the weighted gap $-\circ-$, the complete weighted gap gradient $-\triangle-$ and the incomplete weighted gap gradient $-\square-$. (b): The results for a Gaussian quadrature with 32 Gauss-points. The corresponding analytical solutions from (4.72), (4.74) and (4.78) are plotted with solid, dashed and dotted red lines, respectively.

$$\frac{\tilde{\mathrm{d}}}{\tilde{\mathrm{d}}\beta}\tilde{g}_{\mathrm{N}}^{(e_{12})1,\mathrm{GQ}} = \sum_{g=1}^{\mathrm{N_g}} w_g N^{[1]}(\xi_g^{[1]}) \begin{cases} 0, & \text{for } \xi_g^{[1]} \leqslant -\beta \\ a\beta + \xi_g^{[1]}, & \text{otherwise} \end{cases} \tag{4.79}$$

is obtained for this simple example, where $a = 1$ represents the integrand for the incomplete and $a = 2$ for the complete variational approach. $\mathrm{N_g}$ is the number of used Gauss points. A short examination reveals that the integrand for the first case shows a weak discontinuity, and for the second case a strong discontinuity at $\xi^{[1]}$ equal to $-\beta$.

The results are illustrated in Figure 4.6. While a segment-based integration leads to a smooth deformation dependent gradient, the element-wise integration causes discontinuities in the approximated gradient fields. The kink or jump discontinuity appears each time a slave side Gauss-point switches its master target element during its motion from left to right (see Figure 4.5). While the amplitude of the jump decreases with an increasing number of Gauss-points, the number of discontinuities increases. For example, the jump/kink for two Gauss-points appears exactly for $\beta = 1/\sqrt{3}$. In the case with 32 Gauss points, there are already 16 jumps for this simple example. The complete approach converges to the exact gradient solution, while the incomplete gradient approximation converges to (4.78). In summary, the appearing discontinuities are part of both variational approaches, but only for the complete variational approach the strong discontinuity becomes a part of the right-hand side and deteriorates convergence. In some cases the scheme even fails entirely, because the insufficient continuity can cause a never ending cycling of the non-linear solution method.

At first glance, the easiest solution for this problem is to use a smooth representation of the master side, e.g. with Hermite splines in 2-D [262], or via Gregory patches [221] in 3-D. It is also possible to use discretization methods based on NURBS firstly introduced by Hughes et al.



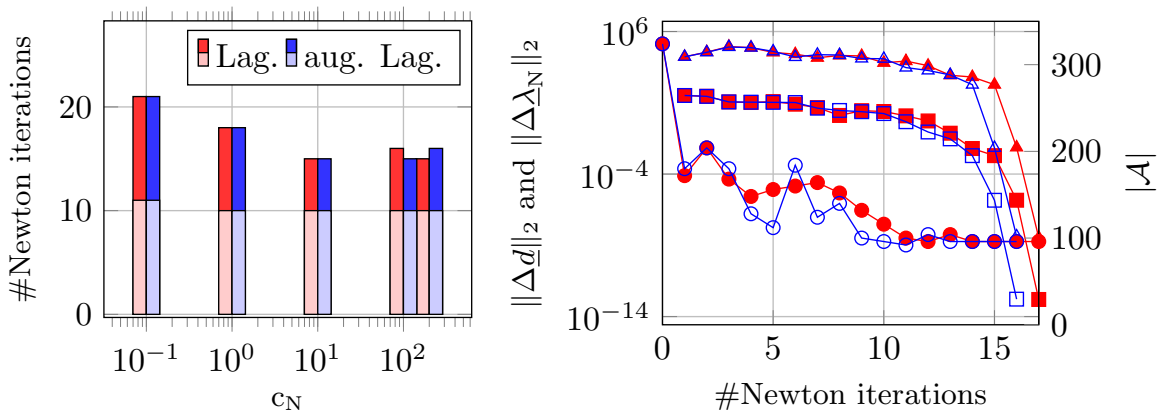
(a) (b)

Figure 4.7.: (a): initial configuration, (b): structured mesh of the indentor.

[141]. A possible approach for arbitrary 3-D surface meshes has been given by Neto et al. [202]. But unfortunately, all of these approaches bring along some kind of burden. The application of Gregory patches is bound to quadrilateral meshes only. The Hermite spline interpolation is hardly extendable to 3-D and in the case of NURBS the construction of suitable patches for complex geometries is quite intricate. Finally, the approach in Neto et al. [202] has its restriction if inflection points appear. The necessary modification destroys locally the C^1-continuity [203]. Furthermore, a pure smoothing approach would encounter its limits as soon as a strong discontinuity in the real geometry appears leading to the possible circumstance that a slave facet slides off the edge of the opposing master surface. Therefore, on second sight, the only real solution to counter the problem is to use a fully segment-based integration as proposed by Puso and Laursen [222] and [290]. Nevertheless, all given contact references with segment-based integration schemes only deal with a simplified variational approach and cannot be applied directly to the formulations considered here, without the loss of some properties. For instance, the 3-D segment-based integration typically introduces a so-called auxiliary plane [222]. These planes are defined by the normals at the center of each considered slave element. Now, this auxiliary normal rather than the smooth normal field is used for the projection and, subsequently, a tessellation algorithm is applied to obtain non-warped, usually triangle-, or recently also quadrilateral-shaped integration domains [278]. The introduction of such an auxiliary plane can imply the loss of angular momentum conservation if the contact variation is not adapted accordingly. In general, the adaption becomes necessary since the projection is no longer collinear to the smooth normal field. Consequently, (4.55) will no longer hold. To regain conservation the first term in (4.27) must be reintroduced into the system. Furthermore, higher order derivatives would become necessary if the problematic full variational approach is to be addressed.

4.7.3. Sliding Hemisphere

Next, a 3-D example is investigated: A hemisphere of radius 5 is pressed onto a plate of dimension $20 \times 20 \times 2.25$, which is oriented along the Cartesian axes. In the reference position, the hemisphere lies exactly in the center of the plate surface and is in forceless contact, i.e. gap and contact pressure are both zero (see Figure 4.7). All degrees of freedom of the plate at the four sides are fixed, while the remaining degrees of freedom are still free to move. In the first load



(a) (b)

Figure 4.8.: (a): Number of Newton iterations for varying c_N values. The light colors represent the sliding step, the dark colors the penetration step. (b): Evolution of the incremental norms w.r.t. the Lagrange multipliers —△— and the displacements —□—. The cardinality of the active set is shown by —○—. Filled, red stands for the Lagrange, unfilled, blue for the augmented incomplete formulation.

step, the nodes on the circular top surface of the hemisphere are moved downwards by an amount of 2.5. In a subsequent load step, they are moved in the in-plane x- and y-directions by an amount of 1.5, respectively. The material parameters are chosen as $E^{[1]} = 25{,}000$ and $\nu^{[1]} = 0.25$ for the stiffer spherical indentor, while the plate consists of softer material ($E^{[2]} = 2{,}500$, $\nu^{[2]} = 0.25$). As material model, the coupled form of the compressible neo-Hookean material model (2.26) is considered. The top surface of the plate shall be the master and the opposing spherical surface the slave side. All chosen parameters are, as far as possible, equivalent to the 2-D semicircular indentor example in Zavarise et al. [294]. Only the thickness as well as the boundary conditions of the plate were changed to reduce the computational costs. Nevertheless, the initial contact situation stays comparable, since a tangential predictor is used, which causes a rigid body motion of the initially unconstrained indentor (see Figure 4.9). Also, the mesh of the hemisphere was adapted to fit the reference 2-D test case (see Figure 4.7). The plate is discretized by 32 elements in x- and y-direction and with 3 elements in thickness direction. The test case is supposed to demonstrate the superior robustness of the mortar-type contact formulations. To detect convergence, the relevant criterion is that the relative ℓ_2-norm of the Lagrange multiplier increment becomes lower than $1.0\mathrm{E}{-10}$.

First, the influence of the regularization parameter c_N on the performance of the incomplete formulation will be investigated. The complete formulation is skipped for the moment due to the problems mentioned in the previous subsection. The regularization parameter is varied within a range from 0.1 to 200. For a c_N value equal to 300 and higher, neither the Lagrangian nor the augmented Lagrangian formulation did converge. The comparison in Figure 4.8 shows no noteworthy difference in the dependence on the regularization parameter with respect to the two formulations. While the augmented Lagrangian formulation needs one iteration less in the first load step for a $c_N = 100$, it is the other way around for $c_N = 200$. The second load step seems almost completely independent from the chosen regularization parameter. In another study, c_N is chosen as 10 and the behavior for more moderate load states is investigated. Therefore, the load is reduced by a factor of 2 multiple times and the number of necessary iterations are summarized

load fac.	1	1/2	1/4	1/8	1/16	1/32
Lag.	(16, 10)	(17, 9)	(13, 9)	(12, 4)	(7, 4)	(7, 3)
aug. Lag.	(15, 10)	(17, 9)	(13, 9)	(12, 4)	(7, 4)	(7, 3)

Table 4.1.: Newton iteration study for the incomplete variational approach by steady reduction of the load. The first braced value corresponds to the penetration load step, the second one to the sliding step.

in Table 4.1. While the first reduction even leads to an increasing number of iterations for the first load step, a monotonically decreasing behavior can be observed for all subsequent reductions. Again, the Lagrangian and the augmented Lagrangian formulations show an almost identical behavior in terms of Newton iterations.

Next, the maximal possible initial penetration is tested. In Zavarise et al. [294], a 2-dimensional penalty node-to-segment formulation was investigated. For the consistent linearization a maximum initial penetration of 0.8 has been reported, while the there newly developed large penetration approach reaches convergence up to a penetration of 4.0. The variationally incomplete mortar-based approach presented here shows convergence up to an initial penetration of 2.9, whether applied as standard Lagrangian or augmented Lagrangian formulation. Both formulations are consistently linearized and the Lagrangian takes 17, while the augmented scheme takes 16 iterations. The tangential predictor step and the first three Newton steps as well as the converged state for the Lagrangian formulation are shown in Figure 4.9. The arrows represent the scalar nodal Lagrange multiplier values mapped onto the surface normal direction of the spherical indentor. The color-map corresponds to the projected forces acting on the master surface. Furthermore, the detailed step lengths with respect to the displacements and Lagrange multiplier degrees of freedom as well as the cardinality of the active set are shown in the right part of Figure 4.8. Again, no severe differences besides a slightly stronger fluctuating active set in the augmented case can be noticed.

Figure 4.9 reveals that the magnitude and the distribution of the Lagrange multiplier values in the non-converged states can be far off from the final values. In turn, the corresponding forces lead to severe mesh distortions and the simulation is close to crash. The reason is the considered second order update scheme for the Lagrange multipliers, which is known for fast convergence near the solution, but inferior performance during the pre-asymptotic phase compared to a first order update scheme (see Bertsekas [23] and Chapter 5 for more details). For a constant penetration of 2.9, the sliding step could be increased up to a movement of 1.9 in the x- and y-directions, respectively, until divergence would hit. Both formulations took 10 Newton iterations for the second load step.

In summary, the segment-to-segment approach shows a superior robustness compared to the consistently linearized node-to-segment approach in Zavarise et al. [294]. The obvious drawback is its higher computational cost.

Second Order NURBS Discretization

To be able to investigate also the variationally complete approach, the soft block is now discretized with second order NURBS, while the linear Lagrange discretization of the indentor stays unchanged. The sufficiently smooth master surface now makes it possible to apply the

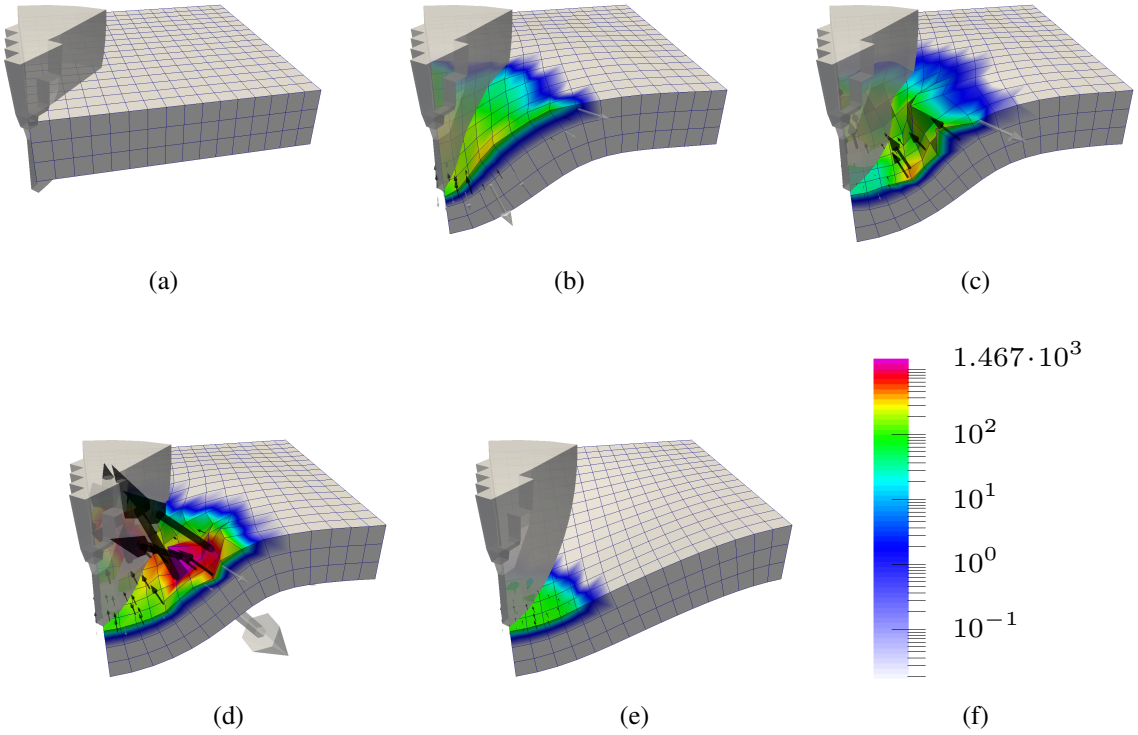
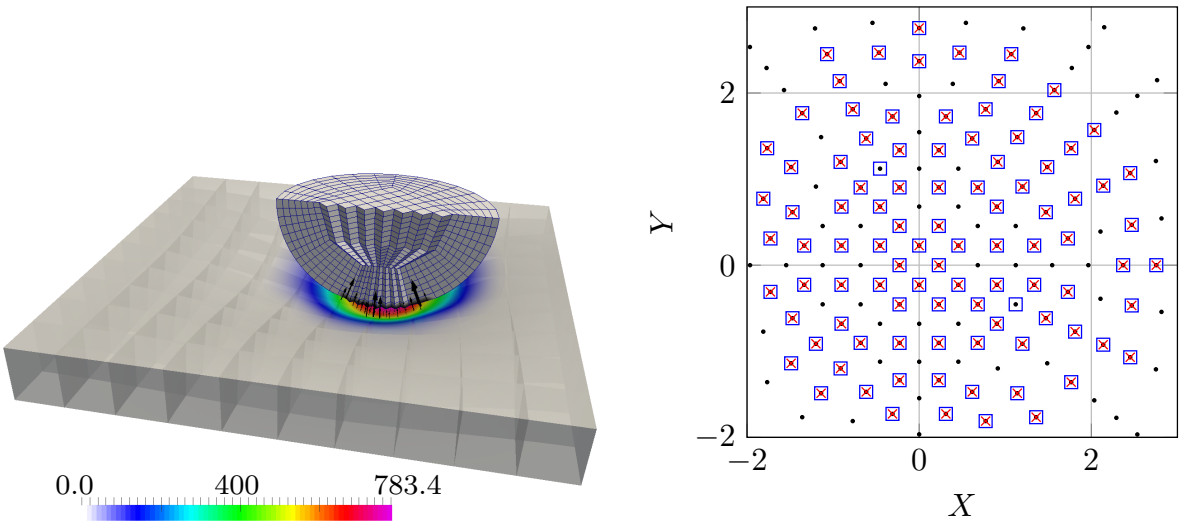


Figure 4.9.: Visualization of $1/8$ of the indentor and $1/4$ of the soft plate. The initial configuration after the tangential predictor step is shown in Figure 4.9a and the first three Newton iterations in Figures 4.9b to 4.9d, as well as the converged configuration in Figure 4.9e. The color-map represents the logarithmic scaled projected Lagrange multiplier forces acting on the master side, while the arrows represent the Lagrange multiplier values on the slave side mapped onto the smooth normal direction. Black arrows correspond to active, white to inactive nodes.

	penetration						sliding	
load step	1	2	3	4	5	6	7	8
NURBS c.	13 (4)	5 (12)	4 (16)	8 (52)	7 (80)	5 (100)	13 (112)	
NURBS i.	17 (4)	5 (12)	4 (16)	7 (52)	8 (80)	5 (100)	12 (112)	
LIN i.	13 (12)	9 (28)	6 (28)	7 (52)	7 (76)	7 (96)	8 (96)	10 (98)
LIN Popp	12 (12)	6 (28)	6 (28)	7 (52)	7 (76)	7 (96)	8 (96)	9 (96)

Table 4.2.: Comparison of the plain Newton performance for different formulations. "c." refers to the complete, "i." to the incomplete augmented formulation and "Popp" to the reference formulation described in Popp [213], Popp et al. [215, 216]. The values in braces are the final numbers of active nodes at the end of each load step.

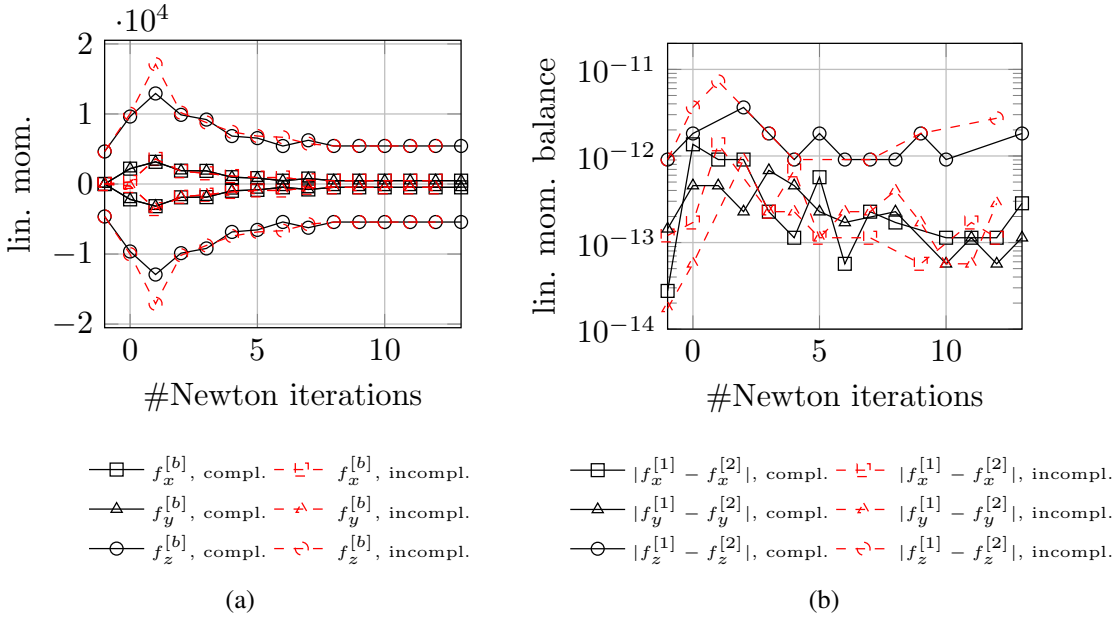


(a) (b)

Figure 4.10.: In Figure 4.10a the converged configuration for the half NURBS, half Lagrange discretization is visualized. The indentor was moved 3.0 in z-direction and 2.0 in x- and y-direction. In Figure 4.10b a comparison of the final active set distribution for the pure linear Lagrange discretization mapped into the XY-reference plain is performed. The × markers refer to the reference formulation by Popp [213], Popp et al. [215, 216], while the □ markers represent the active set for the incomplete variational approach.

complete variational approach. Furthermore, c_{N} is set to 10 and the augmented formulation is used for all following investigations. A short study of the incomplete approach reveals that the maximal step size of the first step must be reduced to 2.2. For higher loads divergence has been detected. In contrast, the complete approach only allows an initial step size up to 0.5. For slightly higher initial loads, a cycling of the active set could be observed, thus leading to a stagnation of the plain Newton approach. Interestingly, in a follow-up step the load can be increased much further, such that the supposition is obvious that the cycling is caused by the initial geometrical configuration rather than due to any numerical integration issues. However, the incomplete approach does not show this undesirable behavior. To compare the performance of both variational approaches the total penetration is set to 3.0 and will be applied in 6 load steps. In a seventh step, the indentor is then moved by 2.0 in x- and y-direction. The final configuration is shown in Figure 4.10a. The required numbers of iterations per load step are given in Table 4.2. For these moderate load levels, both algorithms perform quite similar for the mixed discretizations. Only in the very first step, the complete approach manages to take 4 iterations less.

In Section 4.5, the conservation of linear and angular momentum was proven theoretically. As mentioned, the conservation of angular momentum is strongly bound to the way how the gradient of the active weighted gap is incorporated into the force balance. Here, the opportunity is taken to demonstrate conservation of linear and angular momentum for the sliding load step in Figures 4.11a, 4.11b, 4.12a and 4.12b. Since the test for conservation involves subtraction operations of the nodal forces, loss of significant digits plays a role and the values are bound somewhere below $1.1\mathrm{E}{-11}$ for this example. Another source of error for the angular momentum conservation is the local Newton scheme for the projection algorithm where the tolerance is set to $1.0\mathrm{E}{-12}$. Nonetheless, the results confirm the conservation of linear and angular momentum



(a) (b)

Figure 4.11.: Conservation of linear momentum under consideration of the half NURBS, half Lagrange example.

conservation practically to machine precision. Furthermore, a closer look will reveal that some points are missing in the balance plots. In these cases, the numerical result was equal to 0.0, i.e. all significant digits vanished.

Comparison with an Established Mortar Contact Formulation

In the following, the results are compared to the well-established mortar contact formulation of Popp [213], Popp et al. [215, 216]. For the comparison, standard first order Lagrange shape functions are considered. Again, only the incomplete variational approach is taken into account. In terms of Newton iterations, both approaches are very similar with some slight advantages for the well-established formulation for this specific example, see Table 4.2. The sliding step had to be halved in step size to obtain convergence. Interestingly, in the very last step even the number of active nodes differs. The related distribution can be seen in Figure 4.10b. The scattered pattern originates from the linear discretization. In the case of the combined NURBS and Lagrange discretization, the pattern changes to a more continuous distribution. The conservation of angular momentum, however, is not achieved by the reference formulation (see Figures 4.12c and 4.12d). The relative difference in the converged resultant forces and moments is around 1%. A look at the final internally stored elastic energy of the two bodies reveals a value of 6247.40 for the reference solution and a value of 6247.36 for the incomplete variant. Again, the latter one is slightly more consistent, but the absolute difference in the results is negligible. In summary, the introduced errors seem to behave as in the 2-D example (see Figure 4.4).

4.7.4. Instability of the Variationally Inconsistent Formulation

In the last example of this chapter, a rare but intrinsic instability of the variationally inconsistent formulation is presented and briefly discussed. Therefore, a stiff wedge indentor is taken into

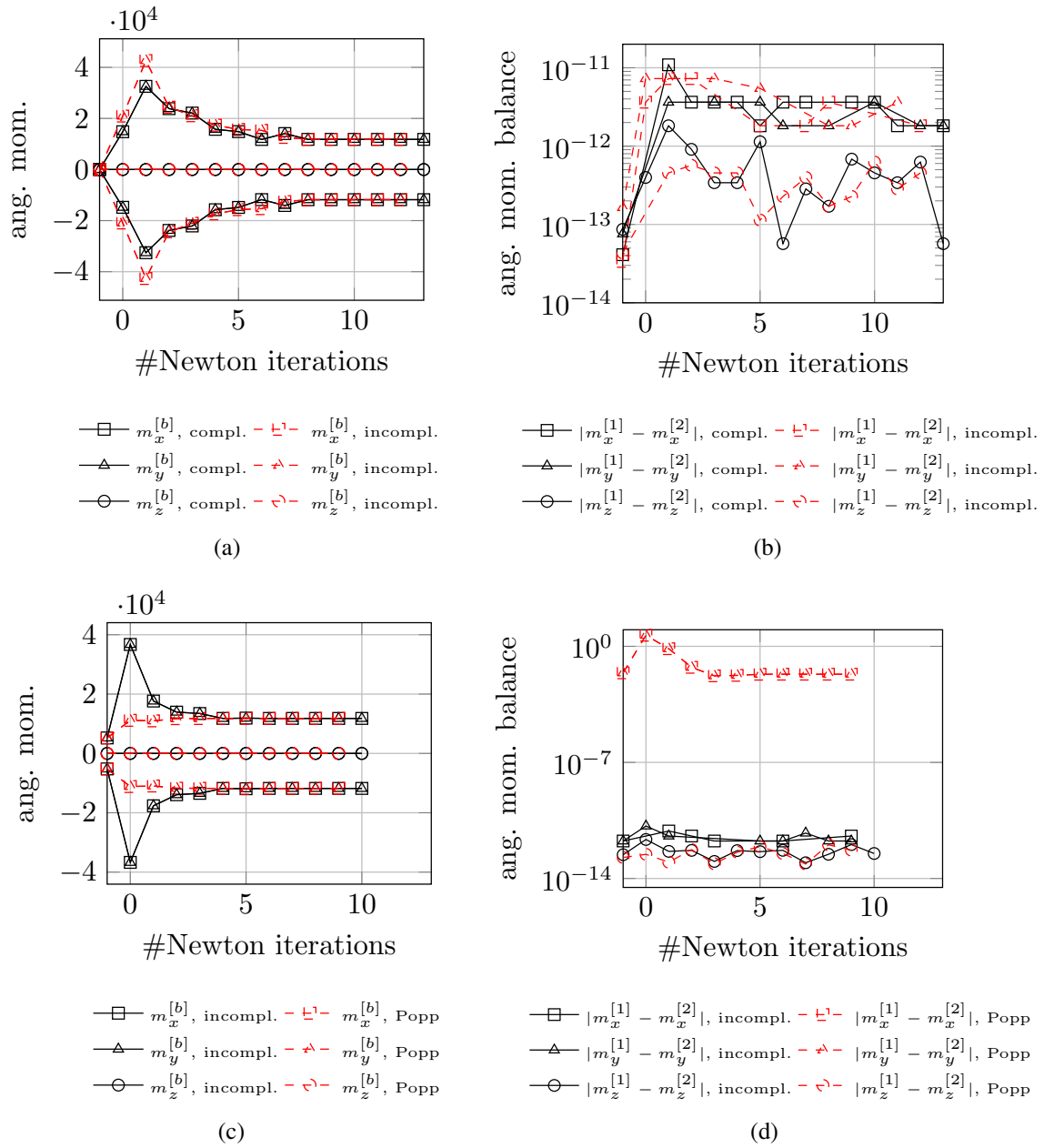


Figure 4.12.: Conservation of angular momentum. In Figures 4.12a and 4.12b the half NURBS, half Lagrange example is considered, while Figures 4.12c and 4.12d address linear Lagrange elements.

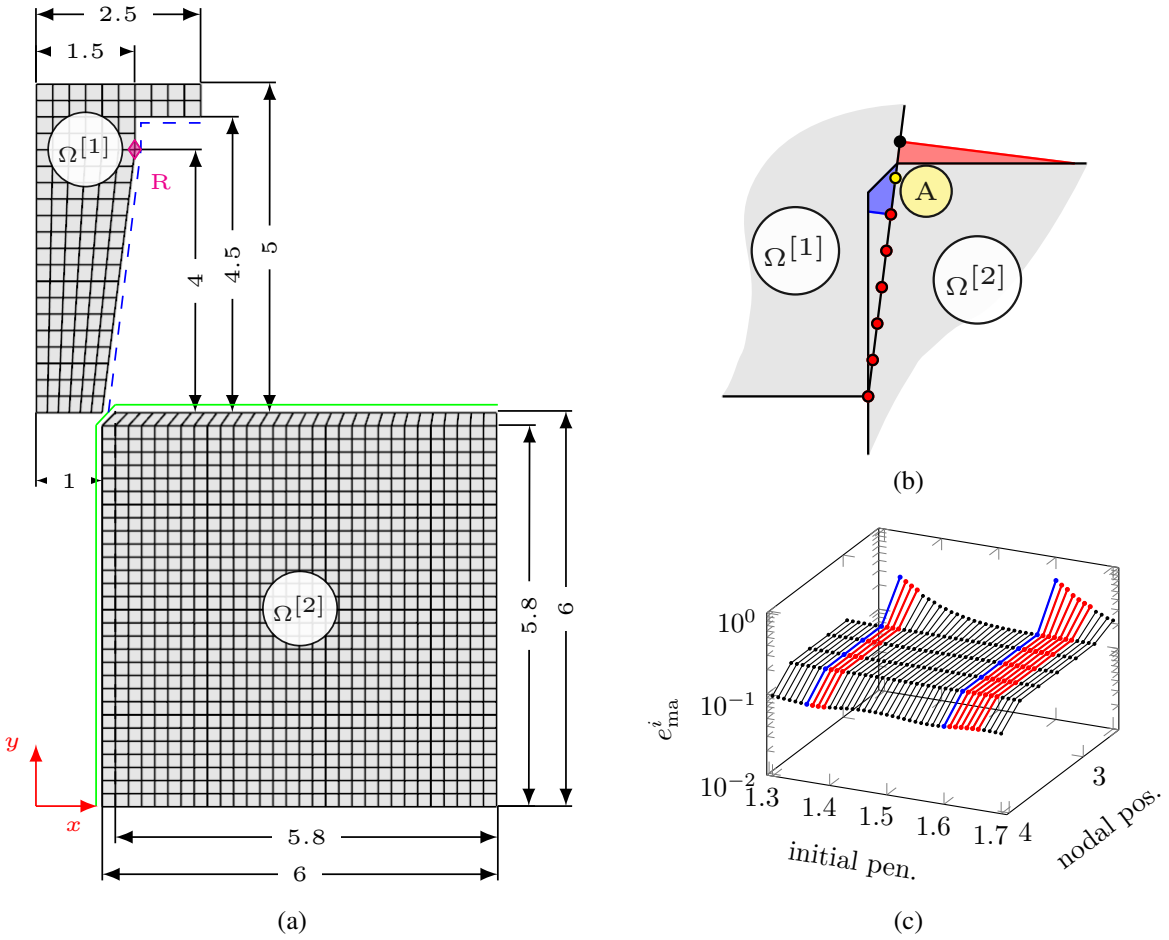


Figure 4.13.: In Figure 4.13a the geometries of a wedge indentor and a chamfered plate are presented. The potential slave contact surface is marked by a blue dashed line - - - , while its master counterpart is represented by a green solid line ——. The considered projection is sketched in Figure 4.13b. Highlighted in blue is the uncritical and in red the critical projection zone of the two adjacent elements contributing to the newly active node A. Finally, Figure 4.13c shows the magnitude of the missing parameteric master coordinate variation for all active nodes. The stated nodal position is measured from the reference point R in Figure 4.13a.

account which is moved into a ten times softer plate. The plate has a chamfer angle of 45° at the upper left corner. The dimensioned geometry as well as the used mesh are presented in Figure 4.13a. The boundary conditions are chosen such that the plate is compressed in x-direction through the applied deformations. Therefore, the right edge of the plate must be fixed in x- and y-direction and the wedge indentor is restrained on its left edge in x-direction, while the top edge is moved by a prescribed motion in negative y-direction. The material parameters are chosen as $E^{[1]} = 2{,}500$, $E^{[2]} = 250$ and $\nu^{[1]} = \nu^{[2]} = 0.25$ based on a compressible Neo-Hookean material law as introduced in (2.27). Furthermore, quadraliteral two-dimensional elements under a plane strain assumption are used.

In a first step the variationally inconsistent formulation is applied to the problem. A variation of the initial penetration reveals that the consistently linearized incomplete approach shows convergence only up to a penetration of 1.36. For a constant regularization parameter c_{N} of 10 around 8 to 9 Newton iterations are necessary to solve this problem dependent on the actual penetration. Similar to the observations in Section 4.7.3 no noteworthy difference between a standard Lagrangian and an augmented Lagrangian formulation could be detected. An incremental increase of only 0.01 to an initial penetration of 1.37 leads to divergence. Surprisingly, the standard Lagrangian as well as the augmented formulation converge again for a higher penetration between 1.40 to 1.60. Afterwards no convergence can be achieved with both formulations for another six incremental load increases between 1.61 and 1.65. For even higher load increases, however, convergence is again observed.

This pattern repeats for further increases and can be also noticed for smaller penetrations. The only difference lies in the fact that the non-linear solver does not fail entirely for smaller overlaps. In order to understand the origin of this phenomenon, the number of Gauss points (GP) has been set to 50 per slave element for the consistently linearized approaches to exclude any artifacts caused by the numerical integration scheme. However, the higher GP number provokes no change in behavior. Furthermore, the well-established mortar method introduced in [213, 215] shows exactly the same undesired convergence pattern for a segment-based integration. Deeper investigations reveal that the variationally consistent mortar contact formulation does not show this initially devastating distortion, see Figure 4.14b for a comparison. It is therefore reasonable to assume that the origin can be found in the neglected terms of the incomplete variational approach. Thus, the associated nodal error due to the neglected variation of the convective master parameter coordinate (4.71) has been calculated and is presented in Figure 4.13c. Therein, the suddenly rising influence of this missing variation can be clearly recognized each time a new slave node becomes active during the motion from the initial wedge position in negative y-direction. A look at Figure 4.13b helps to understand what is happening: The sketch shows the geometrical configuration right after the tangential predictor step for an initial motion of 1.6 in negative y-direction. Node A newly joins the active set between a prescribed displacement of 1.59 and 1.60, since the corresponding averaged weighted gap value becomes negative.

It is now important to underline that the used contact interface considers the blue as well as the red projections for the necessary integrations. The critical contribution is the red one: A slightly changing slave displacement field drastically changes the position of the projected Gauss point position on the master side due to the flat angle between slave normal and top master surface. Exactly this effect is considered by the neglected variation of the projected parametric master coordinate. Since this variation is missing in both, i.e., the incomplete approach introduced in Chapter 4 as well as in the mentioned well-established variant of [213, 215], the heavy mesh

distortions appear right after the tangential predictor step. Although this effect could have been avoided by a different interface pairing, it seems important to address this issue since it is an impressive demonstration for the possible influence of the neglected variations. Furthermore, it must be mentioned that a simple node-to-segment approach would not suffer from this problem, since only the slave node position and no longer the surrounding Gauss point positions on the adjacent elements would play a role. In the same turn it is further to highlight that this is one of the rare issues which gets worse with a better integration method or a higher number of Gauss points, since a high Gauss point number uncovers the missing variational parts even to a greater extent by adding more and more contributions as part of the red projection zone in Figure 4.13b.

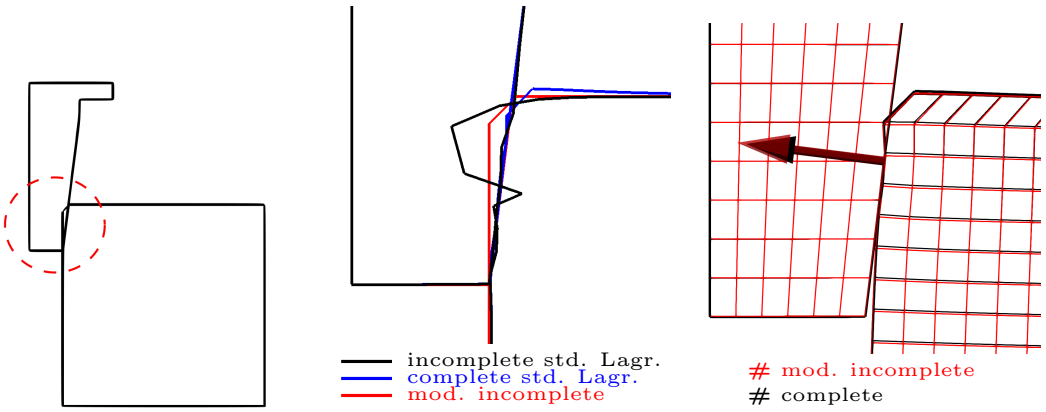
In Figures 4.14a to 4.14c, a few snap shots of the simulations for an initial penetration of 1.37 can be found, where Figure 4.14b shows a comparison of the very first Newton step right after the predictor step. The variationally consistent approach works, as expected, flawlessly and the incomplete approach shows a devastating mesh distortion near the newly active node next to the upper boundary of the active contact zone.

Fortunately, it is possible to resolve the issue at least partly by either changing the coupling and interface conditions or by considering a special modification of the system of equations, which will be discussed in Chapter 5. This modified approach will allow to solve this specific example even with the variationally inconsistent formulation by making the entire non-linear solution approach significantly more robust. A comparison of the obtained results can be found in Figure 4.14c, where next to the displacement fields also the associated Lagrange multiplier values are shown. Thus, if the first critical point can be passed, the final results differ again only slightly. However, the method presented in Chapter 5 does not resolve the origin of the instability but rather helps to pass the cumbersome regime during the way to the solution. That means that the instability will not be completely gone by adding the modification. Another possible remedy might be given by a switch to a closest point projection instead of the ray-tracing method considered within this thesis. To put it in a nutshell: The presented issue asks definitely for further investigations in the future.

4.8. Conclusion

Two new discrete contact formulations based on mortar finite element methods have been developed and analyzed in detail: the so-called variationally complete and incomplete approaches. While the version with a fully consistent variation leads to a symmetric system of equations for all active contributions, it could be shown that the often neglected variations indeed do not significantly distort the converged solution if sufficiently fine meshes are used. In the analyzed examples, this level of refinement has not been inconveniently high for most of the examples. The introduced errors due to the missing variation of the Jacobian determinant and the missing variation of the projected master coordinate as well as the influence on the internally stored strain energy have been quantified. Furthermore, a proof for the conservation of angular momentum for both formulations has been given.

The variationally consistent formulation exhibits severe problems for C^0-continuous Lagrange discretizations if an element-based rather than a segment-based numerical integration scheme is considered. The origin could be traced back to the integration error, which leads to strong discontinuities in the gradient field of the weighted gap. Different remedies have been proposed.



(a) predictor (b) #1 (c) converged

Figure 4.14.: Wedge indentor for an initial penetration of 1.37. Figure 4.14a shows the initial configuration directly after the tangential predictor step; in Figure 4.14b a comparison of the first non-linear iteration among the three listed methods is presented. Note that the solution method from [213] coincides completely with the variationally inconsistent solution in this very first step. The final converged states are shown in Figure 4.14c, where the solution of the incomplete variant could only be reached for the modified approach of Chapter 5.

To put it in a nutshell, the influence of the two variational formulations on the non-linear solution approach (Newton-Raphson method) can be very different. For example, a load level that is slightly too high could lead to a cycling of the active set in case of the complete variational approach, while no such behavior became apparent for the incomplete variational approach. A comparison of the incomplete approach with an established mortar method Popp et al. [215, 216] even revealed a slightly different active set distribution in the converged state, although the global results were very similar.

Concerning the two variational approaches, it can be concluded that for many examples the differences in the results are negligible. Yet, there are rare cases where the neglected variational terms can lead to truly troublesome situations. This is impressively shown in Section 4.7.4. Furthermore, since the influence on the non-linear solver can be remarkable, the behavior of an incomplete approach might become less predictive when optimization methods based on scalar-valued merit functions are transferred to contact problems. Within this context most model equations assume that at least the gradient of the merit function is consistent, but exactly this is not the case if the shown simplifications are applied. This will be also a sub-topic of Chapter 6.

Concerning the two constraint enforcement strategies, namely the standard Lagrangian and the augmented Lagrangian formulations, only a very slight difference could be observed, without any clear tendency for one or the other. Even the influence of a varying regularization parameter was rather weak for the investigated problems. However, this observation might change in the future as soon as the variationally consistent formulation can be applied more easily to a larger number of problems. Currently, mainly the variationally incomplete formulation has been investigated in depth.

5. A Variant of Newton's Method for Constrained Problems

In this chapter a novel modification of the classical implicit non-linear Newton-Raphson solution scheme will be developed. The modification is easily applicable to a wide range of existing algorithms which use Lagrange multipliers to enforce active constraints. Furthermore, it does not require any specific discretization type. Instead, the variant of the classical Newton's method shown here has a number of appealing properties: Firstly, it enables a formulation completely restricted to the primal degrees of freedom without the need for any special shape functions. The underlying Lagrange multiplier update follows in a post-processing step. Alternatively, a modified saddle-point formulation, which simultaneously contains displacements and Lagrange multipliers as solution variables, is described. The differences concerning the solvability of the two linear system types are studied and discussed. Secondly, the new variant significantly improves robustness and reliability of the non-linear solution method compared to a plain Newton-Raphson approach. Thirdly, the new solution method is based on a strong mathematical foundation leading to reliable local convergence results. Finally, a novel switching strategy will be proposed that can be used to switch from the modified to the classical Newton's method close to the solution. In this way the typical second order convergence rate can be elegantly combined with a higher robustness and predictability of the non-linear solution behavior in the pre-asymptotic regime. For a meaningful switching a number of necessary conditions will be given.

5.1. Motivation

The work presented in this chapter is inspired by Zavarise et al. [294] and Bertsekas [23]. In [294] a non-consistent start-up procedure for a penalty contact formulation was proposed and discussed. However, in contrast to [294], a formulation based on Lagrange multipliers is considered here. In particular, the focus lies on the solution of the variationally incomplete variant comprehensively discussed in Chapter 4, even though the method is not restricted to this formulation. Another difference is that the approach proposed here is entirely based on a standard Lagrangian contact formulation and, instead of removing terms from the consistently linearized system during the pre-asymptotic phase, a single term is added. In the following it will be shown that this term alone is sufficient to significantly improve the robustness of the underlying contact formulation. Furthermore, the introduced term allows a smooth transition between the modified system and the original one. The necessary adaption as well as appropriate switching strategies will be comprehensively discussed. Furthermore, the conditioning and solvability of the emerging linear systems of equations will be addressed in detail. The mathematical foundation of the presented method is mainly based on the work of Bertsekas [23]. Thus, it is to emphasize that the presented modification is not restricted to contact problems but can be applied to any con-

strained optimization problem which uses Lagrange multipliers. In fact, it should be straight forward modifying the node-to-segment approach presented in [1, 212] or a mortar-type contact formulation considering a closest-point projection, e.g., used by [65, 261]. Furthermore, it can be also applied to the complete variational approach introduced in [131] and Chapter 4. This would be very appealing as soon as the numerical integration issue is resolved by considering a suitable integration scheme, see the discussion in Section 4.7.2 for more details. However, first attempts of applying the method discussed here to a simple example, which uses the consistent formulation, led to some new unexpected, but interesting observations which will be described and discussed in Section 5.6.8.

The remainder of this chapter is organized as follows: In Section 5.2, the modified system of equations will be presented by starting with a derivation and a subsequent discussion. Since the modification and its performance are strongly coupled to the regularization parameter c_{N} and its reliable correction during the non-linear solution procedure, two novel correction schemes will be presented and discussed in Section 5.3. Subsequently, the attention is on the convergence analysis of the modified formulation in Section 5.4. Local convergence as well as the boundedness of the regularization parameter c_{N} will be proven in dependence on one of the proposed correction schemes. In Section 5.5 a method for switching from the modified system to the consistently linearized system of equations will be presented. Afterwards, the Numerical Examples Section 5.6 will demonstrate the great applicability and performance of the proposed method. Therein, a comprehensive numerical study of the newly proposed method will be given. In addition, a detailed analysis concerning the conditioning of the modified system matrix will follow in Section 5.6.7 based on a challenging 3-D example. A short conclusion can be found at the end of this chapter in Section 5.7.

5.2. Modification of Newton's Method

The consistently linearized second order systems of equations presented in Section 4.3 will only show a second order rate of convergence, if the current state is close enough to the solution point. In general, however, convergence towards the solution is only guaranteed if the initial state, consisting of displacements and Lagrange multiplier values, is part of a bounded region around the solution. This behavior is well-known and thus it is not surprising that a similar observation was also documented for a node-to-segment penalty algorithm in Zavarise et al. [294]. The mentioned paper suggested a non-consistent start-up procedure to overcome possible convergence problems during the first Newton iterations in the case of large initial penetrations. As soon as a state sufficiently close to the actual solution had been reached, the algorithm was supposed to switch to a consistently linearized scheme. A major part of [294] was about the proper switching point. The proposed criterion was based on an user-specified estimate of the maximally expected contact pressure. This value was then used for the switch as well as a cut-off for the maximal admissible contact pressure during the non-linear solution scheme. Since a good estimate was not always readily available, an adaptive update routine was additionally proposed.

Even though the problem origin is the same, a mathematically more profound approach is developed here in order to also address the additional complexity arising from the Lagrange multipliers. The goal is to achieve a superior performance during the pre-asymptotic phase compared to the standard consistently linearized Newton approach from Chapter 4 without the need

for any problem dependent tuning parameters, or user-specified estimates. Meanwhile the correctness of the solution is supposed to be maintained under all circumstances.

Furthermore, the algorithm presented here also shows convergence for the modified system of equations, however, with a reduced convergence rate, i.e., maximal (super-)linear. Thus, it is to underline that it is not necessary to switch to a consistent linearization to reach the (numerical) exact solution. However, the switch is very beneficial since it helps to circumvent an ill-conditioned system matrix and it is very beneficial with respect to the local convergence rate. A meaningful set of switching criteria will be presented in Section 5.5.

5.2.1. Modified Linear System of Equations

The discussion starts with the Lagrangian potential defined in (4.10). Its consistently linearized form with respect to all active contributions follows as

$$\left\{\nabla^2_{\underline{x}\,\underline{x}}\mathscr{U} - [\nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}\,\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}})]^{\mathrm{T}}\right\}\Delta\underline{d} \quad - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}\Delta\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}} = -\nabla_{\underline{d}}\mathscr{U}(\underline{x}) + \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}}, \tag{5.1a}$$

$$-[\nabla_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}]^{\mathrm{T}}\Delta\underline{d} = \tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}. \tag{5.1b}$$

The idea is to adopt a modification described in [23] for the linearization of the active constraints (5.1b) by adding the relaxation term $-\frac{1}{c_{\mathrm{N}}}\underline{\underline{A}}^{\mathcal{A}}\Delta\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}}$, where $\underline{\underline{A}}^{\mathcal{A}}$ denotes again the tributary area matrix introduced for the first time in (4.7). In this way, (5.1b) becomes

$$-[\nabla_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}]^{\mathrm{T}}\Delta\underline{d} - \frac{1}{c_{\mathrm{N}}}\underline{\underline{A}}^{\mathcal{A}}\Delta\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}} = \tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}. \tag{5.2}$$

There are now two options how to progress: The first one is that (5.1a) and (5.2) are combined and the emerging modified saddle-point system is used throughout the non-linear solution procedure. Alternatively, a condensation approach is applied to develop a system completely formulated in displacement degrees of freedom. For the latter approach it is necessary to multiply (5.2) by $-c_{\mathrm{N}}\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}[\underline{\underline{A}}^{\mathcal{A}}]^{-1}$ and add the result to (5.1a). This yields

$$\begin{aligned}&\left\{\nabla^2_{\underline{d}\,\underline{d}}\mathscr{U} - [\nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}\,\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}})]^{\mathrm{T}} + c_{\mathrm{N}}\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}[\underline{\underline{A}}^{\mathcal{A}}]^{-1}[\nabla_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}]^{\mathrm{T}}\right\}\Delta\underline{d}\\ &= -\nabla_{\underline{d}}\mathscr{U}(\underline{x}) + \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}(\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}} - c_{\mathrm{N}}\hat{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}).\end{aligned} \tag{5.3}$$

In this way the saddle-point structure is removed. In a post-processing step the corresponding Lagrange multiplier increments can be calculated via

$$\Delta\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}} = -c_{\mathrm{N}}[\underline{\underline{A}}^{\mathcal{A}}]^{-1}\left(\tilde{\underline{g}}_{\mathrm{N}} + [\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}]^{\mathrm{T}}\Delta\underline{d}\right). \tag{5.4}$$

Both approaches, i.e., the modified saddle-point formulation and the condensed formulation, have advantages, however, the modified saddle-point system can be seen as beneficial with respect to the solvability of the linear system of equations in case of larger problems. For a detailed discussion the reader is kindly referred to Section 3.2.3.2 where a similar observation is made

during the discussion of the penalty method. Especially, (3.47) in connection with (3.58) and (3.59) reveal a close relationship between the method proposed here and the penalty approach. A more comprehensive computational investigation of the conditioning of the system matrix will follow in Section 5.6.7.

Remark 5.1. Note that the *modified* saddle point system will still simply be denoted as saddle point system in the following. In fact, this nomenclature agrees with the classification provided in Benzi et al. [19]. Furthermore, the modified system changes back into the original one defined in (5.1) or (4.18) when $c_{\mathrm{N}} \to \infty$. Lastly, it should be noted that in the case of a non-symmetric saddle-point system, which occurs here in case of the incomplete variational approach, the expression *generalized* saddle-point system is often used. This additional term, however, is avoided for simplicity's sake.

5.2.2. Properties of the Modified System

The post-processing step (5.4) is closely related to Uzawa's well-known staggered augmented Lagrangian method [170, 267], but instead of solving the displacement problem for a fixed set of Lagrange multipliers, the update is performed *on the fly* during the non-linear iteration. This type of update is called a first order iteration scheme, which comes in theory with a superior robustness in comparison to a second order iteration scheme [23]. Furthermore, the entire modification has a number of appealing properties as originally stated by Bertsekas [23, pp. 240–241]:

1. As the regularization parameter c_{N} tends to infinity, the system becomes in the limit the one corresponding to the consistent Newton's method for the Lagrangian (4.10). Furthermore, if $c_{\mathrm{N}} \to \infty$, a super-linear convergence rate can be expected.
2. If c_{N} is sufficiently large, then the sequences $\{\underline{d}^{\{k\}}\}$ and $\{\underline{\lambda}_{\mathrm{N}}^{\{k\}}\}$ converge locally to a KKT-pair fulfilling (4.9). Note that this does not necessarily mean that c_{N} must be infinite at the solution. However, if c_{N} does not tend to infinity, the convergence rate will decrease. Further information will follow in Section 5.4.1.
3. The system of equations has a unique solution if either

$$[\tilde{\nabla}^2_{\underline{d}\,\underline{d}} \mathscr{L}(\underline{x}^\star, \underline{\lambda}_{\mathrm{N}}^\star)]^{-1} = \left\{ \nabla^2_{\underline{x}\,\underline{x}} \mathscr{U}(\underline{x}^\star) - [\nabla_{\underline{d}} (\tilde{\nabla}_{\underline{d}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}^\star)\, \underline{\lambda}_{\mathrm{N}}^\star)]^{\mathrm{T}} \right\}^{-1} \tag{5.5}$$

   or

$$[\tilde{\nabla}^2_{\underline{d}\,\underline{d}} \mathscr{L}_{c_{\mathrm{N}}}(\underline{x}^\star, \underline{\lambda}_{\mathrm{N}}^\star)]^{-1} = \left\{ \tilde{\nabla}^2_{\underline{d}\,\underline{d}} \mathscr{L}^\star + c_{\mathrm{N}} \tilde{\nabla}_{\underline{d}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}^\star) [\underline{\underline{A}}^{\mathcal{A}}(\underline{x}^\star)]^{-1} [\nabla_{\underline{d}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}^\star)]^{\mathrm{T}} \right\}^{-1} \tag{5.6}$$

   exists at the solution. Note that (5.6) is truly the incomplete second order derivative of the augmented Lagrangian at the solution with respect to the displacement degrees of freedom, since $\underline{\tilde{g}}_{\mathrm{N}}(\underline{x}^\star) = \underline{0}$ holds.
4. The gradient of the active constraints is allowed to become partially linear dependent and a unique solution is still maintained. This weakens the Mangasarian-Fromowitz constraint qualifications [204].

5. If the matrix defined in (5.6) is positive definite, then the calculated search direction for the displacements, i.e. $\Delta\underline{d}$, is a descent direction for the augmented Lagranigan functional (4.13).
6. The condensed system matrix can be easily identified as an approximated Hessian of the augmented Lagrangian functional.

However, to understand the later shown highly beneficial effect on the non-linear solver behavior of this simple modification, the active part of the modified saddle-point system of equations proposed in (5.1a) and (5.2) shall be reformulated once more. Therefore, (5.1a) is reordered such that it becomes possible to express $\Delta\underline{d}$ in terms of $\Delta\underline{\lambda}_{\mathrm{N}}$ by

$$\Delta\underline{d} = -\tilde{\nabla}^2_{\underline{d}\underline{d}}\mathscr{L}^{-1}\{\nabla_{\underline{d}}\mathscr{U} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}(\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}} + \Delta\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}})\}, \tag{5.7}$$

where it is assumed that $\tilde{\nabla}^2_{\underline{d}\underline{d}}\mathscr{L}$ is non-singular. Now, by inserting (5.7) into (5.2) a new system of equations is obtained which can be solved for $\Delta\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}}$, viz.

$$\begin{aligned}&\{[\nabla_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}]^{\mathrm{T}}\tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}^{-1}\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}} + \frac{1}{c_{\mathrm{N}}}\underline{\underline{A}}^{\mathcal{A}}\}\ \Delta\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}}\\ &= -\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}} + [\nabla_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}]^{\mathrm{T}}\tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}^{-1}\{\nabla_{\underline{d}}\mathscr{U} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}\underline{\lambda}^{\mathcal{A}}_{\mathrm{N}}\}.\end{aligned} \tag{5.8}$$

The idea for this reformulation can also be found in Gould [113], for instance. These equations, i.e., (5.8) and (5.7), are known as the *range-space equations*. A more detailed discussion of the range-space and the closely related *null-space* method can be found in Conn and Gould [52] and Gill et al. [108, p. 183]. In a nutshell: In the application discussed here, the range-space method is theoretically more appealing than the null-space method, since the number of constraints m is almost always much smaller than the number of displacement degrees of freedom n. Therefore, if the inverse of the Hessian matrix is readily at hand, the system (5.8) will become small. A possible variant, where this approach can become very attractive for the practical use, is given if the exact Hessian is replaced by an approximation, e.g., obtained by a suitable BFGS or Broyden method, see Section 3.1.1. A discussion of range-space methods which use a suitable positive definite approximation of $\nabla^2_{\underline{d}\underline{d}}\mathscr{L}$ can be found in Biggs [24, 25, 26]. However, even though the range-space method itself does not play a practical role in this thesis, the reformulation (5.8) reveals the true nature of the modification: The scaled tributary area matrix $\frac{1}{c_{\mathrm{N}}}\underline{\underline{A}}^{\mathcal{A}}$ acts as an additive regularization which helps to make the entire matrix on the left hand side of (5.8) more positive definite. The influence of the regularization becomes smaller and smaller with a rising regularization parameter c_{N}. Furthermore, (5.8) reveals that the step length of the Lagrange multiplier increment can be expected to become smaller compared to the non-modified system as long as the Lagrangian Hessian matrix mapped into the range space of the constraint gradients is positive definite. In fact, the structure of (5.8) reminds one of the trust region or Levenberg–Marquardt method presented in Sections 3.1.2.2 and 3.1.2.3, respectively. A major difference is only that it is formulated with respect to the Lagrange multipliers instead of the primal variables.

Another effect can be shown by looking at the range-space part of the displacement solution vector. Therefore, the solution vector shall be formally split into a range-space and a null-space part

$$\Delta \underline{d} = \underline{\underline{R}} \Delta \underline{d}_{\mathrm{R}} + \underline{\underline{Z}} \Delta \underline{d}_{\mathrm{Z}}, \tag{5.9}$$

where $\underline{\underline{R}} \in \mathbb{R}^{n \times m}$ is a matrix whose columns span the range space of $\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}$, while $\underline{\underline{Z}} \in \mathbb{R}^{n \times (n-m)}$ is a matrix whose columns span the null space of $[\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}$, i.e. $\underline{\underline{R}}^{\mathrm{T}} \underline{\underline{Z}} = \underline{\underline{0}}$. A suitable choice for $\underline{\underline{R}}$ is $\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}$, for instance. The null space matrix $\underline{\underline{Z}}$ can not so easily be obtained and asks for a singular value decomposition or the reduced row echelon form of the matrix $[\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}$. For more information the reader is kindly referred to Gill et al. [108] and Nocedal and Wright [204, p. 457 and p. 538]. However, the split makes it possible to write the displacement solution as two distinct parts. The first part belongs to the range-space of the constraint gradients and is obtained by inserting (5.9) into (5.2) yielding

$$[\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}} \underline{\underline{R}} \, \Delta \underline{d}_{\mathrm{R}} = -\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}} - \frac{1}{c_{\mathrm{N}}} \underline{\underline{A}}^{\mathcal{A}} \Delta \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}. \tag{5.10}$$

The second part belongs to the null-space of the constraint gradients and follows after inserting (5.9) into (5.1a) and multiplying the obtained equation by $\underline{\underline{Z}}^{\mathrm{T}}$ from left resulting in

$$\underline{\underline{Z}}^{\mathrm{T}} \, \nabla^2_{\underline{d}\,\underline{d}} \mathscr{L} \, \underline{\underline{Z}} \, \Delta \underline{d}_{\mathrm{Z}} = -\underline{\underline{Z}}^{\mathrm{T}} \nabla_{\underline{d}} \mathscr{U} - \underline{\underline{Z}}^{\mathrm{T}} \, \nabla^2_{\underline{d}\,\underline{d}} \mathscr{L} \, \underline{\underline{R}} \, \Delta \underline{d}_{\mathrm{R}}. \tag{5.11}$$

Especially, (5.10) reveals a second regulative effect of the modification. Under certain assumptions, such as the positive definiteness of the modified matrix on the left side of (5.8), it can be expected that the Lagrange multiplier increments point in the opposite direction of the active constraint violations and thus reduce the right-hand side in (5.10). This reduces also the solution increment $\Delta \underline{d}_{\mathrm{R}}$. In summary, a damped but, therefore, more reliable system answer to (huge) constraint violations, i.e. large initial penetrations, can be expected which is very beneficial during the *pre-asymptotic phase* of displacement controlled problems. For the overall performance of the method, the updating strategy for the c_{N} parameter will be of major importance. This is going to be discussed in Section 5.3.

Remark 5.2. Notice that the split of the displacement solution vector into null-space and range-space as presented in (5.9), (5.10) and (5.11) is technically only possible as long as the complete variational approach is used. However, the made observations hold also for the incomplete approach. After all, the incomplete approach is a truncated variant of its variationally consistent counterpart and, interestingly, the modified approach works even better for the variationally inconsistent than for the consistent approach. This will be further investigated in Section 5.6.8.

5.3. Dynamic Correction of the Regularization Parameter

A crucial point for the applicability of the modified system is a reliable correction strategy for the involved regularization parameter $c_{\mathrm{N}} > 0$. This correction routine must avoid updates during

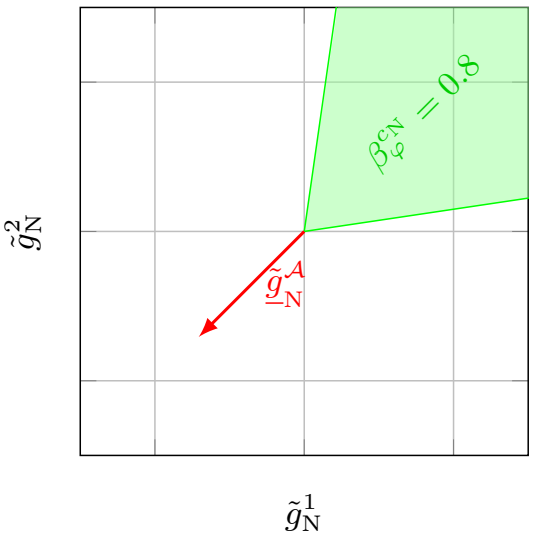


Figure 5.1.: Fundamental idea of the sufficient angle demand for $\beta_\varphi^{c_N} = 0.8$. The c_N parameter would stay unchanged as long as the directional derivative of the weighted gap $\langle \tilde{\nabla}_{\underline{d}} \underline{\tilde{g}}_N^{\mathcal{A}\{k\}}, \Delta \underline{d} \rangle$ is part of the green cone.

the non-linear solution procedure in already cumbersome situations, but has to be activated as soon as the modification of the system of equations leads to a remarkable performance loss. Most of the proposed update routines in the literature are based on heuristic considerations and include some initially unknown tuning parameters. See for example Bertsekas [23, Sec. 2.2.5] or Glad and Polak [111]. These heuristics make it often difficult to transfer the routine to another application. In the following, two novel approaches are presented which make it possible to carry out a correction without such heuristics.

Both ideas rely on the fact that the computed displacement increment as solution of (5.3) must be at least a descent direction for the active constraint equations, such that

$$\langle \underline{\tilde{g}}_N^{\mathcal{A}}, [\nabla_{\underline{d}} \underline{\tilde{g}}_N^{\mathcal{A}}]^T \Delta \underline{d} \rangle < 0 \tag{5.12}$$

holds true. This very basic demand is always fulfilled for the consistently linearized system (5.1) since

$$[\nabla_{\underline{d}} \underline{\tilde{g}}_N^{\mathcal{A}}]^T \Delta \underline{d} = -\underline{\tilde{g}}_N^{\mathcal{A}} \tag{5.13}$$

is enforced by (5.1b). However, in the modified case it can be used to derive a reliable correction routine for the regularization parameter.

5.3.1. Sufficient Enclosed Angle

The first correction strategy can be derived by reordering the terms in the solved modified linear system of equations (5.3). Therefore, the system is multiplied by the displacement solution increment vector from the left side, bringing up the scalar-valued demand

$$c_{\mathrm{N}}\langle \hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, [\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}\rangle = -\langle \Delta\underline{d}, \nabla_{\underline{d}}\mathscr{U} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}\rangle - \langle \Delta\underline{d}, \nabla^2_{\underline{d}\,\underline{d}}\mathscr{U}\,\Delta\underline{d}\rangle + \langle \Delta\underline{d}, [\nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}})]^{\mathrm{T}}\Delta\underline{d}\rangle - c_{\mathrm{N}}\|[\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}\|^2_{\underline{\underline{A}}^{-1}} \overset{!}{<} 0, \tag{5.14}$$

where the scaled quadratic norm $\|\underline{v}\|^2_{\underline{\underline{A}}} = \langle \underline{v}, \underline{\underline{A}}\,\underline{v}\rangle$ is used. Notice that the reformulation (5.14) proposed here contains a slight simplification compared to (5.3), since in general

$$\|[\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}\|^2_{\underline{\underline{A}}^{-1}} \neq \Delta\underline{d}^{\mathrm{T}}\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\,\underline{\underline{A}}^{-1}\,[\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d} \tag{5.15}$$

for the incomplete formulation, see Remark 4.1 for more information. The simplified inequality (5.14) asks for a descent direction of the active constraints, but in a quite weak sense: Any descent direction will be admissible, even if it is almost orthogonal to the steepest descent direction. In order to become more applicable, the original inequality must be strengthened by introducing a control parameter $\beta_{\varphi}^{c_{\mathrm{N}}} \in (0,1)$, such that the following new demand is obtained

$$\cos(\measuredangle(\hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, [\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d})) \overset{!}{\leqslant} -\beta_{\varphi}^{c_{\mathrm{N}}} \;\Leftrightarrow\; \langle \hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, [\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}\rangle \overset{!}{\leqslant} -\beta_{\varphi}^{c_{\mathrm{N}}}\|\hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\|\,\|[\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}\|. \tag{5.16}$$

See also Figure 5.1 for an illustrative sketch of the condition. By inserting (5.16) into (5.14), the new request

$$-c_{\mathrm{N}}^{\{k\}}\beta_{\varphi}^{c_{\mathrm{N}}}\|\hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\|\,\|\langle\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, \Delta\underline{d}\rangle\| \overset{!}{\geqslant} \langle \Delta\underline{d}, [\nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}})]^{\mathrm{T}}\Delta\underline{d}\rangle - \langle \Delta\underline{d}, \nabla^2_{\underline{d}\,\underline{d}}\mathscr{U}\,\Delta\underline{d}\rangle - \langle \Delta\underline{d}, \nabla_{\underline{d}}\mathscr{U} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}\rangle - c_{\mathrm{N}}\|\langle\Delta\underline{d}, \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\rangle\|^2_{\underline{\underline{A}}^{-1}}. \tag{5.17}$$

is created. This relation can now be used to define a correction equation for the regularization parameter, viz.

$$c_{\mathrm{N}} \overset{!}{\geqslant} \frac{1}{\|\langle\Delta\underline{d}, \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\rangle\|^2_{\underline{\underline{A}}^{-1}}}\Big\{c_{\mathrm{N}}^{\{k\}}\beta_{\varphi}^{c_{\mathrm{N}}}\|\hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\|\,\|[\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}\| + \langle \Delta\underline{d}, [\nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}})]^{\mathrm{T}}\Delta\underline{d}\rangle - \langle \Delta\underline{d}, \nabla_{\underline{d}}\mathscr{U} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}\rangle - \langle \Delta\underline{d}, \nabla^2_{\underline{d}\,\underline{d}}\mathscr{U}\,\Delta\underline{d}\rangle\Big\}. \tag{5.18}$$

Under the typical assumption that the Hessians of the elastic bodies are symmetric and positive definite [297], i.e., $\langle \Delta\underline{d}, \nabla^2_{\underline{d}\,\underline{d}}\mathscr{U}\,\Delta\underline{d}\rangle > 0$ for $\Delta\underline{d} \neq \underline{0}$ holds true, the inequality can be further simplified. In fact, the expensive matrix vector multiplication can be removed from the lower bound estimate of the regularization parameter c_{N} and the final, computationally less expensive estimate follows as

$$c_{\mathrm{N}}^{\mathrm{low}} = \frac{c_{\mathrm{N}}^{\{k\}}\beta_{\varphi}^{c_{\mathrm{N}}}\|\hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\|\,\|[\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}\| + \langle \Delta\underline{d}, [\nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}})]\Delta\underline{d}\rangle - \langle \Delta\underline{d}, \nabla_{\underline{d}}\mathscr{U} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}\rangle}{\|[\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}\|^2_{\underline{\underline{A}}^{-1}}}. \tag{5.19}$$

The evaluation of each term can be performed with limited effort. The most expensive remaining operation is the vector matrix product, where the dimensions of the matrix are bounded by the displacement degrees of freedom on the slave and master interfaces. While the denominator is always positive, the numerator can become both, positive and negative. For the latter case no action is necessary, while for positive values an adaption may be initiated. Furthermore, it is to note that the numerator tends to zero close to the solution. Unfortunately, the denominator follows the same trend, since the displacement increment tends to zero as well and some safe guarding strategy should be invoked to avoid numerical issues. Finally, the correction routine is given by $c_{\mathrm{N}}^{\{k+1\}} = \max\{c_{\mathrm{N}}^{\{k\}}, c_{\mathrm{N}}^{\mathrm{low}}\}$.

Remark 5.3. The two simplifications introduced during the derivation, namely once (5.15) and secondly the disregard of $\langle \Delta \underline{d}, \nabla^2_{\underline{d}\,\underline{d}} \mathscr{U}\, \Delta \underline{d} \rangle$ makes the behavior of this control mechanism less predictive. Actually, (5.15) causes an inconsistency throughout the entire derivation since even the demands (5.16) and (5.12) do not coincide for the incomplete formulation. The later considered examples in Section 5.6.2 will reveal that these slight inconsistencies can, at least under certain circumstances, lead to a bad performance of the correction scheme proposed here. Therefore, an alternative will be presented in Section 5.3.2 which does not suffer from these fundamental problems. However, the *sufficient enclosed angle* (SEA) scheme will be useful for comparison reasons.

5.3.2. Sufficient Infeasibility Reduction

The second idea is based on a model for the infeasibility measure, for example

$$\Theta^{\{k\}} = \|\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x}^{\{k\}})\| = \sqrt{\langle \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}, \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}} \rangle}. \tag{5.20}$$

The reduction of such a model can be enforced by introducing a regularization parameter and choosing an appropriate value for this parameter. For instance, the predicted model reduction for a step length of one is claimed to be at least

$$\Theta^{\{k\}} + m_\Theta(1) \leqslant (1 - \beta_\Theta^{c_{\mathrm{N}}})\,\Theta^{\{k\}}, \tag{5.21}$$

where $\beta_\Theta^{c_{\mathrm{N}}} \in (0, 1)$ is a given constant. This basic idea is inspired by the literature on filter methods, see e.g. [96, 98, 270, 271] and Chapter 6. In the following, a linear model $m_\Theta(\alpha) : \mathbb{R} \to \mathbb{R}$ is used, such that

$$m_\Theta(\alpha) = \frac{\alpha}{\Theta^{\{k\}}} \left(\langle \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}, [\nabla_{\underline{d}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}} \Delta \underline{d} \rangle \right) \tag{5.22}$$

follows. First of all, it should be noted that $m_\Theta(1) = -\Theta^{\{k\}}$ holds for the fully consistent Newton approach due to (5.13) and, therefore, the inequality (5.21) is always fulfilled in this case. However, this is no longer true for the modified approach. Instead, the inequality (5.21) yields

$$
\begin{aligned}
(1-\beta_{\Theta}^{c_{\mathrm{N}}})\,\Theta^{\{k\}} &\geqslant \Theta^{\{k\}} + \frac{1}{\Theta^{\{k\}}}\langle \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, -\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}} - \frac{1}{c_{\mathrm{N}}}\underline{\underline{A}}^{\{k\}}\Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}\{k\}}\rangle \\
&= -\frac{1}{c_{\mathrm{N}}\,\Theta^{\{k\}}}\langle \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \underline{\underline{A}}^{\mathcal{A}\{k\}}\Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}\{k\}}\rangle,
\end{aligned}
\tag{5.23}
$$

where (5.2) is used to replace the term $[\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}]^{\mathrm{T}}\Delta\underline{d}$ in (5.22). The idea is now to choose the regularization parameter $c_{\mathrm{N}} = c_{\mathrm{N}}^{\{k+1\}}$ for the next step in such a way that it would have been at least sufficient for the current step, i.e.

$$
c_{\mathrm{N}}^{\{k+1\}} \geqslant \frac{\langle \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \underline{\underline{A}}^{\mathcal{A}\{k\}}\Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}\{k\}}\rangle}{(\beta_{\Theta}^{c_{\mathrm{N}}}-1)[\Theta^{\{k\}}]^2}. \tag{5.24}
$$

Finally, by reinserting (5.4) into (5.24), it is possible to obtain an explicit relationship between the previous regularization parameter and the new one

$$
c_{\mathrm{N}}^{\{k+1\}} = c_{\mathrm{N}}^{\{k\}}\max\left\{1, \frac{1}{1-\beta_{\Theta}^{c_{\mathrm{N}}}}\left[1 + \frac{\langle \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \langle \nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \Delta\underline{d}\rangle\rangle}{[\Theta^{\{k\}}]^2}\right]\right\}. \tag{5.25}
$$

An alternative interpretation of this result is that the regularization parameter will be increased as soon as the current search direction is or is close to be no longer a descent direction for the currently active constraints. This is true, since the cosine of the enclosed angle between the active constraints and their respective directional derivatives is claimed to be bounded in the negative half plane:

$$
\cos(\measuredangle(\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, \langle \nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, \Delta\underline{d}\rangle)) \leqslant -\beta_{\Theta}^{c_{\mathrm{N}}}\frac{\Theta^{\{k\}}}{\|\langle \nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, \Delta\underline{d}\rangle\|} \leqslant 0. \tag{5.26}
$$

As soon as this condition is violated, the regularization parameter is going to be increased. Furthermore, it is possible to deduce a lower bound for the norm of the Lagrange multiplier increment by assuming that (5.21) does not hold in a Newton iteration k. By multiplication with $\|\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}\|$, (5.21) yields

$$
\langle \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}} + [\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}]^{\mathrm{T}}\Delta\underline{d}^{\{k\}}\rangle > (1-\beta_{\Theta}^{c_{\mathrm{N}}})\langle \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}\rangle. \tag{5.27}
$$

Under the stated assumption the following simple estimate can be derived

$$
\begin{aligned}
0 < (1-\beta_{\Theta}^{c_{\mathrm{N}}})\|\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}\|^2 &< -\frac{1}{c_{\mathrm{N}}^{\{k\}}}\langle \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \underline{\underline{A}}^{\mathcal{A}\{k\}}\Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}\{k\}}\rangle \\
&\leqslant \frac{1}{c_{\mathrm{N}}^{\{k\}}}\|\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}\|\|\underline{\underline{A}}^{\mathcal{A}\{k\}}\Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}\{k\}}\| \\
&\leqslant \frac{1}{c_{\mathrm{N}}^{\{k\}}}\|\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}\|\|\underline{\underline{A}}^{\mathcal{A}\{k\}}\|\|\Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}\{k\}}\|
\end{aligned}
\tag{5.28}
$$

which instantly leads to the lower bound for the Lagrange multiplier increment norm

$$\|\Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}\{k\}}\| > (1-\beta_{\Theta}^{c_{\mathrm{N}}})c_{\mathrm{N}}^{\{k\}}\frac{\|\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}\|}{\|\underline{\underline{A}}^{\mathcal{A}\{k\}}\|}. \tag{5.29}$$

Thus, a minimal slope of $(1-\beta_{\Theta}^{c_{\mathrm{N}}})$ can be expected, if the regularization parameter is continuously corrected during the non-linear solution approach. To reach this minimal slope, however, the generated sequence of regularization parameters $\{c_{\mathrm{N}}^{\{k\}}\}$ must stay (nearly) constant. This, at first glance, opposing result will become clearer under consideration of Section 5.4.2. Furthermore, it is demonstrated in Section 5.6.5.

Visualization

To get an idea of the SIR method a short visualization shall be presented and discussed. The underlying contact problem will follow in Section 5.6.1.2. For the moment, the details are of minor importance. Let us assume that the simulation reached Newton iteration $\#4$ and the correction parameter $\beta_{\Theta}^{c_{\mathrm{N}}}$ is set to 0.9 for the entire solution sequence. Now, two different directions are considered: Once the modified direction $(\Delta\underline{d}_{\mathrm{mod}}, \Delta\underline{\lambda}_{\mathrm{N}}^{\mathrm{mod}})$ obtained from (5.3) and (5.4) and, secondly, the default/unmodified direction $(\Delta\underline{d}_{\mathrm{def}}, \Delta\underline{\lambda}_{\mathrm{N}}^{\mathrm{def}})$ obtained from (5.1). These two directions span the space for Figure 5.2a by

$$\Theta(\alpha_1,\alpha_2) = \Theta(\underline{x}^{\{4\}} + \alpha_1\Delta\underline{d}_{\mathrm{mod}} + \alpha_2\Delta\underline{d}_{\mathrm{def}}). \tag{5.30}$$

The one-dimensional slices along these directions, i.e. $\Theta(\alpha_1,0)$ and $\Theta(0,\alpha_2)$, are shown in (5.2b) and (5.2c), respectively. Notice that the Lagrange multiplier direction influences the function indirectly since the Lagrange multipliers are part of the active set decision (4.12). First, the attention is on the default direction: As expected, for a linear model this direction will reduce the infeasibility measure to zero (see the red line in Figure 5.2c). However, the true behavior of the function can be severely different. In this case the linear model fits well for $\alpha_2 < 0.8$, but afterwards the true function values follow a highly non-linear path, thus, the gap between both curves rises rapidly. In contrast, if the modified direction is followed the difference between the linear model and the actual infeasibility value is only very small, i.e., the model quality is high in this direction and for the considered step-length. At the end of iteration $\#4$ the inequality (5.21) is tested and fails, since

$$\Theta^{\{4\}} + m_{\Theta}^{\{4\}}(1) = 2.617 - 1.966 = 0.651 \nleqslant 0.2617 = (1-\beta_{\Theta}^{c_{\mathrm{N}}})\Theta^{\{4\}}. \tag{5.31}$$

Thus, c_{N} must be increased by a factor $c_{\mathrm{N}}^{\{5\}}/c_{\mathrm{N}}^{\{4\}} = 2.062$ following (5.25). This brief example demonstrates two things: Firstly, the SIR update is easy to execute. Secondly, the modified direction might lead to less change in the linear model but can still be superior compared to the default direction. This is very obvious in Figure 5.2a, where the modified direction is clearly advantageous compared to the default direction by coming much closer to the local sub-space minimum.

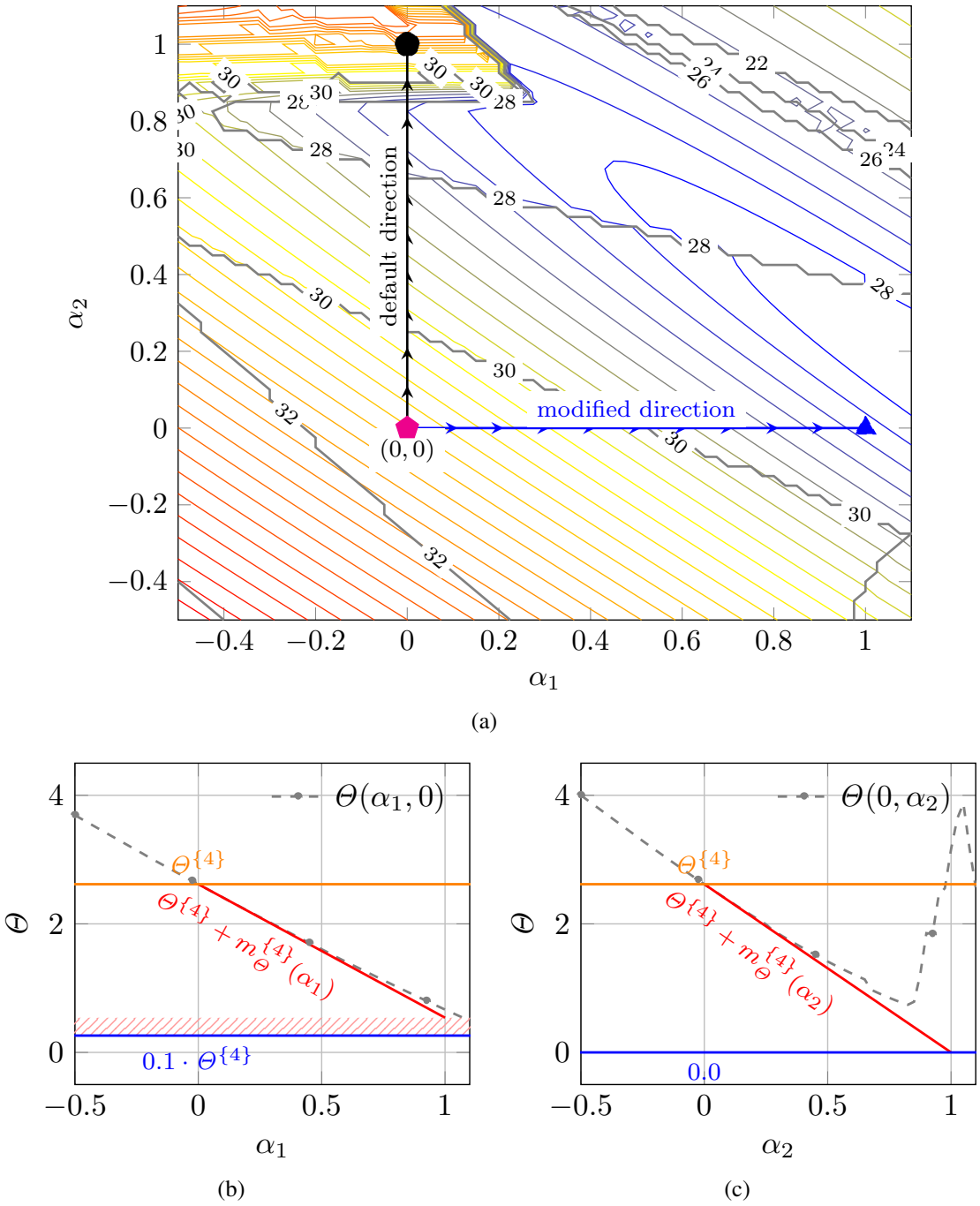


Figure 5.2.: Figure 5.2a shows the colorful sub-space contour lines of (5.30) as well as the two parametrized search directions. The gray solid lines surround areas with a constant active set. In addition, one-dimensional slices through the two-dimensional function (5.30) are presented in the Figures 5.2b and 5.2c. In these figures not only the function graph is shown as - - - , but also the linear model is given by —— as well as the initial function value —— and the at least expected reduction level ——.

Remark 5.4. It must be noticed that the discontinuities of (5.20) visualized in Figure 5.2 are partially expected whenever the active set changes. The reason is that only the active weighted gap values are considered in (5.20) and as long as the corresponding scaled nodal Lagrange multipliers, i.e. $\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}/c_{\mathrm{N}}$, are unequal to zero, the active set decision (4.12) will cause a jump in (5.20) during the transition from active to inactive and vice versa. However, not all of these jumps are caused solely by this effect. Others are triggered by a distorted mesh and missing projections. However, the important point is that the SIR correction stays in all cases unaffected since it is completely based on the local linear model.

5.4. Local Convergence Analysis and Boundedness of the SIR Correction Scheme

The main objective of this section is to prove the boundedness of the sequence $\{c_{\mathrm{N}}^{\{k\}}\}$ for $k \to \infty$. However, the proof relies on the convergence properties of the entire modified sequence of generated iterates $\{(\underline{d}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}})\}$. In a first step, the attention is on the local convergence analysis. To begin with, required key properties of the constrained optimization introduction in Section 3.2 are briefly recapitulated. For example, the active set definition given in (4.12) can be seen as an educated guess of the true active set at the solution pair $(\underline{d}^{\star}, \underline{\lambda}_{\mathrm{N}}^{\star})$, defined by

$$\mathcal{A}_0 = \{i \in \mathcal{S} \,|\, [\hat{g}_{\mathrm{N}}(\underline{d}^{\star})]_i = 0\} \quad \text{and} \quad \mathcal{A}_+ = \{i \in \mathcal{A}_0 \,|\, [\lambda_{\mathrm{N}}^{\star}]_i > 0\}. \tag{5.32}$$

This is just in accordance with (3.33) and (3.34). The optimization algorithm often performs poorly near the solution, if any of the active constraints have a Lagrange multiplier near zero, i.e., if they are not strongly active, see also Theorem 3.6 and the associated Remark 3.3. Therefore, it shall be assumed during the following derivations that $\underline{x}^{\star}$ and its associated Lagrange multipliers $\underline{\lambda}_{\mathrm{N}}^{\star}$ represent a strict local minimizer of the considered optimization problem presented in Section 4.2.2. Furthermore, $\underline{x}^{\star}$ shall be a regular point, i.e., the active constraints are linearly independent at the solution. In addition, the functions $\mathscr{U}$ and $\underline{\hat{g}}_{\mathrm{N}}$ shall be twice continuously differentiable in some open neighborhood $\mathcal{V}^{\star}$ centered at the solution point and defined by

$$\mathcal{V}^{\star} = \mathcal{V}^{\star}(\delta_1, \delta_2, \bar{c}_{\mathrm{N}}) := \{(\underline{d},\, \underline{\lambda}_{\mathrm{N}},\, c_{\mathrm{N}}) :\ \|\underline{d} - \underline{d}^{\star}\| < \delta_1,\ \|\underline{\lambda}_{\mathrm{N}} - \underline{\lambda}_{\mathrm{N}}^{\star}\| < \delta_2 c_{\mathrm{N}},\ c_{\mathrm{N}} > \bar{c}_{\mathrm{N}}\}, \tag{5.33}$$

with the positive scalars δ_1, δ_2 and $\bar{c}_{\mathrm{N}}$. The neighborhood $\mathcal{V}^{\star}$ is chosen such that the corresponding active set, viz.

$$\mathcal{A}_{c_{\mathrm{N}}}(\underline{d}, \underline{\lambda}_{\mathrm{N}}) = \{i :\ \underline{\lambda}_{\mathrm{N}} - c_{\mathrm{N}} \underline{\hat{g}}_{\mathrm{N}}(\underline{d}) > 0\}, \qquad \forall (\underline{d}, \underline{\lambda}_{\mathrm{N}}, c_{\mathrm{N}}) \in \mathcal{V}^{\star} \tag{5.34}$$

is a subset of the active set at the solution, i.e., $\mathcal{A}_+ \subseteq \mathcal{A}_{c_{\mathrm{N}}}(\underline{d}, \underline{\lambda}_{\mathrm{N}}) \subseteq \mathcal{A}_0$. Furthermore, Theorem 3.5 shall hold, which is captured in the following assumption:

AS 5.1. The Karush-Kuhn-Tucker (KKT) pair $(\underline{d}^{\star}, \underline{\lambda}_{\mathrm{N}}^{\star})$ satisfies

$$\underline{v}^{\mathrm{T}}\, \tilde{\nabla}^2_{\underline{d}\underline{d}} \mathscr{L}(\underline{d}^{\star}, \underline{\lambda}_{\mathrm{N}}^{\star})\, \underline{v} > 0, \qquad \forall \underline{v} \neq \underline{0} \text{ with } \tilde{\nabla}_{\underline{d}} \tilde{g}_{\mathrm{N}}^{\mathcal{A}_+}(\underline{d}^{\star})\Big|^{\mathrm{T}} \underline{v} = \underline{0}. \tag{5.35}$$

For more details the reader is referred to Section 3.2 and references therein.

Remark 5.5. Note that a slightly different function declaration will be used within this section. The current displacement field $\underline{d}$ is used as an input variable rather then the current position vector $\underline{x}$. However, this is only a small notation detail to denote the displacement dependencies since $\underline{x} = \underline{X} + \underline{d}$ holds and $\underline{X}$ is constant. This slight change has been applied to reduce the number of additional auxiliary variables during the following proof and does in no way affect the meaning of the expressions, i.e. $\mathscr{L}(\underline{d}, \underline{\lambda}_\mathrm{N}) \equiv \mathscr{L}(\underline{x}, \underline{\lambda}_\mathrm{N})$ etc.

5.4.1. Local Convergence Rate

In the following, important local convergence results for the modified Newton approach are stated and proven. These results as well as their asscoiated proofs are largely inspired by *Proposition 2.4* and the related proof in [23]. However, a different update scheme for the Lagrange multipliers is considered here which implies some modifications. Furthermore, the proof shall be formally extended to the considered inequality constrained problem.

Theorem 5.1. Under consideration of Assumption 5.1, let $\bar{c}_\mathrm{N}$ be a positive regularization parameter such that

$$\tilde{\nabla}^2_{\underline{d}\underline{d}} \mathscr{L}_{c_\mathrm{N}}(\underline{d}^\star, \underline{\lambda}^\star_\mathrm{N}) > 0. \tag{5.36}$$

In addition, it is assumed that the tributary area matrix $\underline{\underline{A}}(\underline{d})$ contains only positive values as well as that it has full rank $|\mathcal{S}|$.

Then, positive scalars δ_1, δ_2, ε and $\breve{M}$ exist such that:

a) The considered problem

$$\min_{\underline{d}} \max_{\underline{\lambda}_\mathrm{N}} \mathscr{L}_{c_\mathrm{N}}(\underline{d}, \underline{\lambda}_\mathrm{N}), \qquad \forall (\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N}) \in \mathcal{V}^\star \tag{5.37}$$

has a unique solution $\breve{\underline{d}}(\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N})$ and $\breve{\underline{\lambda}}_\mathrm{N}(\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N})$. These functions $\breve{\underline{d}}(\cdot,\cdot,\cdot)$ and $\breve{\underline{\lambda}}_\mathrm{N}(\cdot,\cdot,\cdot)$ are continuously differentiable in the interior of $\mathcal{V}^\star$ defined in (5.33).

b) For all $(\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N}) \in \mathcal{V}^\star$, the inequalities

$$\|\breve{\underline{d}}(\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N}) - \underline{d}^\star\| \leqslant \breve{M} \frac{\|\underline{\lambda}_\mathrm{N} - \underline{\lambda}^\star_\mathrm{N}\|}{c_\mathrm{N}}, \quad \|\breve{\underline{\lambda}}_\mathrm{N}(\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N}) - \underline{\lambda}^\star_\mathrm{N}\| \leqslant \breve{M} \frac{\|\underline{\lambda}_\mathrm{N} - \underline{\lambda}^\star_\mathrm{N}\|}{c_\mathrm{N}}, \tag{5.38}$$

hold, where $\breve{\underline{\lambda}}_\mathrm{N}$ is defined as in (5.4):

$$\breve{\underline{\lambda}}_\mathrm{N}(\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N}) = \underline{\lambda}_\mathrm{N} - c_\mathrm{N}[\underline{\underline{A}}(\underline{d})]^{-1}\{\tilde{\underline{g}}_\mathrm{N}(\underline{d}) + [\nabla_{\underline{d}} \tilde{\underline{g}}_\mathrm{N}(\underline{d})]^\mathrm{T}[\breve{\underline{d}}(\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N}) - \underline{d}]\}. \tag{5.39}$$

c) For all $(\underline{d}, \underline{\lambda}_\mathrm{N}, c_\mathrm{N})$ in $\mathcal{V}^\star$, the matrix $\tilde{\nabla}^2_{\underline{d}\,\underline{d}} \mathscr{L}_{c_\mathrm{N}}[\breve{\underline{d}}, \breve{\underline{\lambda}}_\mathrm{N}]$ is positive definite and the gradient of the active constraints $\nabla_{\underline{d}} \tilde{\underline{g}}^{\mathcal{A}}_\mathrm{N}(\breve{\underline{d}})$ as well as the incomplete gradient $\tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}^{\mathcal{A}}_\mathrm{N}(\breve{\underline{d}})$ have full rank $|\mathcal{A}|$.

Proof. For the following proof the modified system of equations defined by (5.1a) and (5.2) is replaced by its continuous counterpart formulated in $(\underline{d}, \breve{\underline{d}}, \underline{\lambda}_{\mathrm{N}}, \breve{\underline{\lambda}}_{\mathrm{N}}, c_{\mathrm{N}})$. Hence, the ansatz

$$\begin{aligned}
\underline{0} =& \nabla_{\underline{d}} \mathscr{U}(\breve{\underline{d}}) - \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\breve{\underline{d}})\, \breve{\underline{\lambda}}_{\mathrm{N}}, && (5.40a)\\
\underline{0} =& \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}) + (\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}))^{\mathrm{T}}(\breve{\underline{d}} - \underline{d}^{\star}) - (\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}))^{\mathrm{T}}(\underline{d} - \underline{d}^{\star})\\
&+ \underline{\underline{A}}(\underline{d}) \frac{\breve{\underline{\lambda}}_{\mathrm{N}} - \underline{\lambda}_{\mathrm{N}}^{\star}}{c_{\mathrm{N}}} - \underline{\underline{A}}(\underline{d}) \frac{\underline{\lambda}_{\mathrm{N}} - \underline{\lambda}_{\mathrm{N}}^{\star}}{c_{\mathrm{N}}}, && (5.40b)
\end{aligned}$$

follows, where (5.40b) represents the continuous counterpart of (5.2), since it can be easily rewritten as

$$\underline{0} = \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}) + (\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}))^{\mathrm{T}}(\breve{\underline{d}} - \underline{d}) + \underline{\underline{A}}(\underline{d}) \frac{\breve{\underline{\lambda}}_{\mathrm{N}} - \underline{\lambda}_{\mathrm{N}}}{c_{\mathrm{N}}}. \tag{5.41}$$

Next, a reparametrization of (5.40) is introduced under consideration of

$$\underline{\Delta} = \underline{d} - \underline{d}^{\star}, \quad \underline{l} = \frac{\underline{\lambda}_{\mathrm{N}} - \underline{\lambda}_{\mathrm{N}}^{\star}}{c_{\mathrm{N}}}, \quad \omega = \frac{1}{c_{\mathrm{N}}}. \tag{5.42}$$

This yields

$$\begin{aligned}
\underline{0} =& \nabla_{\underline{d}} \mathscr{U}(\breve{\underline{d}}) - \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\breve{\underline{d}})\, \breve{\underline{\lambda}}_{\mathrm{N}}, && (5.43a)\\
\underline{0} =& \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})] + (\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})])^{\mathrm{T}}(\breve{\underline{d}} - \underline{d}^{\star}) - (\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})])^{\mathrm{T}} \underline{\Delta}\\
&+ \omega \underline{\underline{A}}[\underline{d}(\underline{\Delta})]\{\breve{\underline{\lambda}}_{\mathrm{N}} - \underline{\lambda}_{\mathrm{N}}^{\star}\} - \underline{\underline{A}}[\underline{d}(\underline{\Delta})]\underline{l}, && (5.43b)
\end{aligned}$$

where $\underline{d}(\underline{\Delta}) = \underline{\Delta} + \underline{d}^{\star}$ and thus $\underline{d}(\underline{0}) = \underline{d}^{\star}$ holds. At the solution point, i.e., for $\underline{\Delta} = \underline{0}$, $\underline{l} = \underline{0}$ and $\omega \in [0, 1/\bar{c}_{\mathrm{N}}]$, (5.43) leads to

$$\begin{aligned}
\nabla_{\underline{d}} \mathscr{U}(\breve{\underline{d}}) - \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\breve{\underline{d}})\, \breve{\underline{\lambda}}_{\mathrm{N}} &= \underline{0}, && (5.44a)\\
\tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^{\star}) + (\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^{\star}))^{\mathrm{T}}(\breve{\underline{d}} - \underline{d}^{\star}) + \omega \underline{\underline{A}}(\underline{d}^{\star})\{\breve{\underline{\lambda}}_{\mathrm{N}} - \underline{\lambda}_{\mathrm{N}}^{\star}\} &= \underline{0}. && (5.44b)
\end{aligned}$$

This system has the solution $\breve{\underline{d}} = \underline{d}^{\star}$ and $\breve{\underline{\lambda}}_{\mathrm{N}} = \underline{\lambda}_{\mathrm{N}}^{\star}$, while the Jacobian with respect to $\breve{\underline{d}}$ and $\breve{\underline{\lambda}}_{\mathrm{N}}$ at such an optimal point follows as

$$\begin{pmatrix} \nabla^2_{\underline{d}\,\underline{d}} \mathscr{U}(\underline{d}^{\star}) - \nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^{\star}) \underline{\lambda}_{\mathrm{N}}^{\star})^{\mathrm{T}} & -\tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^{\star}) \\ -\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^{\star})^{\mathrm{T}} & -\omega \underline{\underline{A}}(\underline{d}^{\star}), \end{pmatrix} \tag{5.45}$$

where $\underline{\underline{A}}(\underline{d}^{\star})$ represents the diagonal tributary area matrix evaluated at the displacement solution. For convenience, the second row has been multiplied by -1. Now, (5.45) is multiplied from the right side by a vector $(\underline{v}, \underline{w})^{\mathrm{T}}$ with $\underline{v} \in \mathbb{R}^n$ and $\underline{w} \in \mathbb{R}^m$ and the emerging system of equations is set equal to zero, yielding

$$\begin{pmatrix} \nabla^2_{\underline{d}\,\underline{d}}\mathscr{U}(\underline{d}^\star) - \nabla_{\underline{d}}(\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^\star)\underline{\lambda}^\star_{\mathrm{N}})^{\mathrm{T}} & -\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^\star) \\ -\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^\star)^{\mathrm{T}} & -\omega\underline{\underline{A}}(\underline{d}^\star), \end{pmatrix} \begin{pmatrix} \underline{v} \\ \underline{w} \end{pmatrix} = \begin{pmatrix} \underline{0} \\ \underline{0} \end{pmatrix}. \tag{5.46}$$

Next, the condensation step which leads to (5.3) can be directly applied to (5.46), leading to

$$\{\tilde{\nabla}_{\underline{d}\,\underline{d}}\mathscr{L}(\underline{d}^\star, \underline{\lambda}^\star_{\mathrm{N}}) + \omega^{-1}\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^\star)\,\underline{\underline{A}}^{-1}(\underline{d}^\star)\,(\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^\star))^{\mathrm{T}}\}\,\underline{v} = \underline{0}. \tag{5.47}$$

At the solution, i.e. for $\tilde{\underline{g}}_{\mathrm{N}}(\underline{d}^\star) = \hat{\underline{g}}_{\mathrm{N}}(\underline{d}^\star) = \underline{0}$, this matrix completely coincides with the Jacobian with respect to the displacements of the incomplete augmented Lagrangian approach presented in Section 4.3.2. Consequently, since $\tilde{\nabla}_{\underline{d}\,\underline{d}}\mathscr{L}_{c_{\mathrm{N}}}(\underline{d}^\star, \underline{\lambda}^\star_{\mathrm{N}}) > 0$ for $c_{\mathrm{N}} > \bar{c}_{\mathrm{N}}$ (see (5.36)), it follows $\underline{v} = \underline{0}$ and thus system (5.46) directly implies $\underline{w} = \underline{0}$. Thus, (5.45) is invertible for all $\omega \in [0, 1/\bar{c}_{\mathrm{N}}]$.

This, together with the second implicit function theorem defined in Bertsekas [23, p. 12], enables the existence of a neighborhood around the solution $(\underline{d}^\star, \underline{\lambda}^\star_{\mathrm{N}})$. Thus, it exists a number of scalars $\varepsilon > 0$, $\delta_1 > 0$, $\delta_2 > 0$ as well as continuously differentiable functions $\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)$ and $\breve{\underline{\lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega)$ defined on

$$\mathcal{V}^\star(\delta_1, \delta_2, \bar{c}_{\mathrm{N}}) = \{(\underline{\Delta},\ \underline{l},\ \omega) :\ \|\underline{\Delta}\| < \delta_1,\ \|\underline{l}\| < \delta_2,\ \omega \in [0, 1/\bar{c}_{\mathrm{N}}]\}, \tag{5.48}$$

such that $(\|\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega) - \underline{d}^\star\|^2 + \|\breve{\underline{\lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega) - \underline{\lambda}^\star_{\mathrm{N}}\|^2)^{1/2} < \varepsilon$ for all $(\underline{\Delta},\ \underline{l},\ \omega) \in \mathcal{V}^\star(\delta_1, \delta_2, \bar{c}_{\mathrm{N}})$. Further, these continuously differentiable functions fulfill (5.40) resulting in

$$\underline{0} = \nabla_{\underline{d}}\mathscr{U}[\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)] - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}[\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)]\ \breve{\underline{\lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega), \tag{5.49a}$$

$$\begin{aligned} \underline{0} =& \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})] + (\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})])^{\mathrm{T}}(\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega) - \underline{d}^\star) - (\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})])^{\mathrm{T}}\underline{\Delta} \\ &+ \omega\underline{\underline{A}}[\underline{d}(\underline{\Delta})]\{\breve{\underline{\lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega) - \underline{\lambda}^\star_{\mathrm{N}}\} - \underline{\underline{A}}[\underline{d}(\underline{\Delta})]\underline{l}, \end{aligned} \tag{5.49b}$$

where again the relationship $\underline{d}(\underline{\Delta}) = \underline{\Delta} + \underline{d}^\star$ is used. Without loss of generality, the introduced scalars δ_1, δ_2 and ε can be set sufficiently small such that matrix (5.45) remains invertible and $\nabla_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}$ as well as the incomplete gradient $\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}$ have full rank $|\mathcal{A}|$ in the entire neighborhood $(\underline{\Delta},\ \underline{l},\ \omega) \in \mathcal{V}^\star(\delta_1, \delta_2, \bar{c}_{\mathrm{N}})$. Consequently, the considered Jacobian matrix (5.45) stays positive definite.

Thus, only the proof for the convergence rate result (5.38) is still missing. To obtain this, (5.49) is differentiated with respect to $\underline{\Delta}$, $\underline{l}$ and ω. First, (5.49a) is considered yielding

$$\underline{0} = \tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}[\breve{\underline{D}}, \breve{\underline{\lambda}}_{\mathrm{N}}]\ \nabla_{\underline{\Delta}}\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}[\breve{\underline{D}}]\ \nabla_{\underline{\Delta}}\breve{\underline{\lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}}, \tag{5.50a}$$

$$\underline{0} = \tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}[\breve{\underline{D}}, \breve{\underline{\lambda}}_{\mathrm{N}}]\ \nabla_{\underline{l}}\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}[\breve{\underline{D}}]\ \nabla_{\underline{l}}\breve{\underline{\lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}}, \tag{5.50b}$$

$$\underline{0} = \tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}[\breve{\underline{D}}, \breve{\underline{\lambda}}_{\mathrm{N}}]\ \nabla_{\omega}\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}[\breve{\underline{D}}]\ \nabla_{\omega}\breve{\underline{\lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}}. \tag{5.50c}$$

Secondly, (5.49b) leads to

$$\begin{aligned}
\underline{0} =& \nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})]^{\mathrm{T}} + (\breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega) - \underline{d}^{\star})^{\mathrm{T}} \nabla^2_{\underline{d}\,\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})] + \nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})]^{\mathrm{T}} \nabla_{\underline{\Delta}} \breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} \\
& - \underline{\Delta}^{\mathrm{T}} \nabla^2_{\underline{d}\,\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})] - \nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})]^{\mathrm{T}} + \omega(\breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega) - \underline{\lambda}^{\star}_{\mathrm{N}})^{\mathrm{T}} \nabla_{\underline{d}} \underline{\underline{A}}[\underline{d}(\underline{\Delta})]^{\mathrm{T}} \\
& + \omega \underline{\underline{A}}[\underline{d}(\underline{\Delta})] \nabla_{\underline{\Delta}} \breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} - \underline{l}^{\mathrm{T}} \nabla_{\underline{d}} \underline{\underline{A}}[\underline{d}(\underline{\Delta})]^{\mathrm{T}}, \qquad (5.51a) \\
\underline{0} =& \nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})]^{\mathrm{T}} \nabla_{\underline{l}} \breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} + \omega \underline{\underline{A}}[\underline{d}(\underline{\Delta})] \nabla_{\underline{l}} \breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} - \underline{\underline{A}}[\underline{d}(\underline{\Delta})], \qquad (5.51b) \\
\underline{0} =& \nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}[\underline{d}(\underline{\Delta})]^{\mathrm{T}} \nabla_{\omega} \breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} + \underline{\underline{A}}[\underline{d}(\underline{\Delta})] (\breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}} - \underline{\lambda}^{\star}_{\mathrm{N}}) \\
& + \omega \underline{\underline{A}}[\underline{d}(\underline{\Delta})] \nabla_{\omega} \breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega)^{\mathrm{T}}. \qquad (5.51c)
\end{aligned}$$

These equations can be reformulated such that a closed form for the desired derivatives is obtained, viz.

$$\begin{pmatrix} \nabla_{\underline{\Delta}} \breve{\underline{D}}^{\mathrm{T}} & \nabla_{\underline{l}} \breve{\underline{D}}^{\mathrm{T}} & \nabla_{\omega} \breve{\underline{D}}^{\mathrm{T}} \\ \nabla_{\underline{\Delta}} \breve{\underline{\Lambda}}_{\mathrm{N}}^{\mathrm{T}} & \nabla_{\underline{l}} \breve{\underline{\Lambda}}_{\mathrm{N}}^{\mathrm{T}} & \nabla_{\omega} \breve{\underline{\Lambda}}_{\mathrm{N}}^{\mathrm{T}} \end{pmatrix} = \underline{\underline{S}}^{-1}(\underline{\Delta}, \underline{l}, \omega)\, \underline{\underline{\breve{R}}}(\underline{\Delta}, \underline{l}, \omega), \qquad (5.52a)$$

where

$$\underline{\underline{S}}^{-1}(\underline{\Delta}, \underline{l}, \omega) = \begin{pmatrix} \tilde{\nabla}^2_{\underline{d}\,\underline{d}} \mathscr{L}(\underline{\Delta}, \underline{l}, \omega) & -\tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{\Delta}) \\ -\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{\Delta})^{\mathrm{T}} & -\omega \underline{\underline{A}}(\underline{\Delta}), \end{pmatrix}^{-1}, \qquad (5.52b)$$

$$\underline{\underline{\breve{R}}}(\underline{\Delta}, \underline{l}, \omega) = \begin{pmatrix} \underline{0} & \underline{0} & \underline{0} \\ \{\breve{\underline{D}} - \underline{d}^{\star} - \underline{\Delta}\}^{\mathrm{T}} \nabla^2_{\underline{d}\,\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}(\underline{\Delta}) & & \\ +\omega(\breve{\underline{\Lambda}}_{\mathrm{N}} - \underline{\lambda}^{\star}_{\mathrm{N}})^{\mathrm{T}} \nabla_{\underline{d}} \underline{\underline{A}}(\underline{\Delta})^{\mathrm{T}} & -\underline{\underline{A}}(\underline{\Delta}) & \underline{\underline{A}}(\underline{\Delta})\{\breve{\underline{\Lambda}}_{\mathrm{N}} - \underline{\lambda}^{\star}_{\mathrm{N}}\} \\ -\underline{l}^{\mathrm{T}} \nabla_{\underline{d}} \underline{\underline{A}}(\underline{\Delta})^{\mathrm{T}} & & \end{pmatrix}. \qquad (5.52c)$$

By applying equation (5.52a), it becomes possible to express the distance between the continuously differentiable function values and the solution as

$$\begin{aligned}
\begin{pmatrix} \breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega) - \underline{d}^{\star} \\ \breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega) - \underline{\lambda}^{\star}_{\mathrm{N}} \end{pmatrix} &= \begin{pmatrix} \breve{\underline{D}}(\underline{\Delta}, \underline{l}, \omega) - \breve{\underline{D}}(\underline{0}, \underline{0}, 0) \\ \breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta}, \underline{l}, \omega) - \breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{0}, \underline{0}, 0) \end{pmatrix} \\
&= \int_0^1 \underline{\underline{S}}^{-1}(\zeta\underline{\Delta}, \zeta\underline{l}, \zeta\omega)\, \underline{\underline{\breve{R}}}(\zeta\underline{\Delta}, \zeta\underline{l}, \zeta\omega) \begin{pmatrix} \underline{\Delta} \\ \underline{l} \\ \omega \end{pmatrix} \mathrm{d}\zeta. \qquad (5.53)
\end{aligned}$$

Since (5.45) is invertible for all $\omega \in [0, 1/\bar{c}_{\mathrm{N}}]$, δ_1 and δ_2 can be sufficiently small, such that $\underline{\underline{S}}^{-1}(\underline{\Delta}, \underline{l}, \omega)$ is uniformly bounded in $\mathcal{V}^{\star}$. Defining a $\breve{\mu}$ thus $\|\underline{\underline{S}}^{-1}(\underline{\Delta}, \underline{l}, \omega)\| < \breve{\mu}$ holds, leads to the estimate

$$
\begin{aligned}
&(\|\breve{\underline{D}}(\underline{\Delta},\underline{l},\omega)-\underline{d}^{\star}\|^2+\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta},\underline{l},\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|^2)^{1/2}\\
\leqslant&\breve{\mu}\{\max_{\zeta\in[0,1]}\|\nabla^2_{\underline{d}\,\underline{d}}\tilde{g}_{\mathrm{N}}(\zeta\underline{\Delta})\|[\max_{\zeta\in[0,1]}\|\breve{\underline{D}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{d}^{\star}\|+\max_{\zeta\in[0,1]}(\zeta\|\underline{\Delta}\|)]\,\|\underline{\Delta}\|\\
&+[\max_{\zeta\in[0,1]}(\zeta\omega)\max_{\zeta\in[0,1]}\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|+\max_{\zeta\in[0,1]}(\zeta\|\underline{l}\|)]\max_{\zeta\in[0,1]}\|\nabla_{\underline{d}}\underline{\underline{A}}(\zeta\underline{\Delta})^{\mathrm{T}}\underline{\Delta}\|\\
&+\max_{\zeta\in[0,1]}\|\underline{\underline{A}}(\zeta\underline{\Delta})\|\|\underline{l}\|+\omega\max_{\zeta\in[0,1]}\|\underline{\underline{A}}(\zeta\underline{\Delta})\|\max_{\zeta\in[0,1]}\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|\}.
\end{aligned}
\tag{5.54}
$$

By the positivity property of the tributary area matrix, there must exist a $\delta_1>0$ such that

$$
\max_{\zeta\in[0,1]}\{\|\nabla_{\underline{d}}\underline{\underline{A}}(\zeta\underline{\Delta})^{\mathrm{T}}\underline{\Delta}\|,\ \|\underline{\underline{A}}(\zeta\underline{\Delta})\|\}\leqslant\breve{a}
\tag{5.55}
$$

holds for all $\|\underline{\Delta}\|<\delta_1$. Furthermore, a upper bound for $\max_{\zeta\in[0,1]}\|\nabla^2_{\underline{d}\,\underline{d}}\tilde{g}_{\mathrm{N}}(\zeta\underline{\Delta})\|\leqslant\breve{G}$ is introduced and all terms linearly scaled by ζ are approximated by their maximal absolute value. This is achieved by setting ζ equal to one in all linear terms. This yields

$$
\begin{aligned}
&(\|\breve{\underline{D}}(\underline{\Delta},\underline{l},\omega)-\underline{d}^{\star}\|^2+\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta},\underline{l},\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|^2)^{1/2}\\
\leqslant&2\breve{\mu}\,\breve{a}\,\|\underline{l}\|+\breve{\mu}\,\breve{G}\,\|\underline{\Delta}\|\,[\max_{\zeta\in[0,1]}\|\breve{\underline{D}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{d}^{\star}\|+\|\underline{\Delta}\|]\\
&+2\breve{\mu}\,\breve{a}\,\omega\ \max_{\zeta\in[0,1]}\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|.
\end{aligned}
\tag{5.56}
$$

Without loss of generality, $\delta_2>0$ is sufficiently small such that at least $2\breve{\mu}\breve{a}\delta_2<1$ holds. Next, the demand

$$
\begin{aligned}
\|\breve{\underline{D}}(\underline{\Delta},\underline{l},\omega)-\underline{d}^{\star}\|\leqslant\ &2\breve{\mu}\,\breve{a}\,\|\underline{l}\|+\breve{\mu}\,\breve{G}\,\|\underline{\Delta}\|\max_{\zeta\in[0,1]}\|\breve{\underline{D}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{d}^{\star}\|+\breve{\mu}\,\breve{G}\,\|\underline{\Delta}\|^2\\
&+2\breve{\mu}\,\breve{a}\,\omega\ \max_{\zeta\in[0,1]}\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|
\end{aligned}
\tag{5.57}
$$

is formulated and the left hand side is evaluated at $(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)$, thus, the estimate

$$
\begin{aligned}
\max_{\zeta\in[0,1]}\|\breve{\underline{D}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{d}^{\star}\|\leqslant\ &\frac{2\breve{\mu}\,\breve{a}}{1-\breve{\mu}\breve{G}\,\|\underline{\Delta}\|}\left\{\|\underline{l}\|+\omega\max_{\zeta\in[0,1]}\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|\right\}\\
&+\frac{\breve{\mu}\breve{G}\|\underline{\Delta}\|^2}{1-\breve{\mu}\breve{G}\,\|\underline{\Delta}\|}
\end{aligned}
\tag{5.58}
$$

is obtained from which $\delta_1<\min\{(\breve{\mu}\breve{G})^{-1},(\breve{\mu}\breve{G})^{-1/2}\}$ can be directly deduced. Next, this result is inserted into (5.56) and the procedure is repeated to get an estimate for $\|\breve{\underline{\Lambda}}_{\mathrm{N}}-\underline{\lambda}_{\mathrm{N}}^{\star}\|$. Inserting into (5.56) and replacing the left hand side yields

$$\max_{\zeta\in[0,1]} \|\breve{\underline{\Lambda}}_{\mathrm{N}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\| \leqslant 2\breve{\mu}\breve{a}\frac{\|\underline{l}\|+\omega\max\limits_{\zeta\in[0,1]}\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|}{1-\breve{\mu}\breve{G}\,\|\underline{\Delta}\|} + \frac{\breve{\mu}\breve{G}\|\underline{\Delta}\|^2}{1-\breve{\mu}\breve{G}\,\|\underline{\Delta}\|} \tag{5.59}$$

and, subsequently,

$$\max_{\zeta\in[0,1]} \|\breve{\underline{\Lambda}}_{\mathrm{N}}(\zeta\underline{\Delta},\zeta\underline{l},\zeta\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\| \leqslant \frac{2\breve{\mu}\breve{a}\|\underline{l}\|+\breve{\mu}\breve{G}\|\underline{\Delta}\|^2}{1-\breve{\mu}(\breve{G}\|\underline{\Delta}\|+2\breve{a}\omega)}. \tag{5.60}$$

Finally, the results (5.58) and (5.60) can be combined by inserting them together into (5.56) to obtain

$$(\|\breve{\underline{D}}(\underline{\Delta},\underline{l},\omega)-\underline{d}^{\star}\|^2+\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta},\underline{l},\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|^2)^{1/2} \leqslant \frac{2\breve{\mu}\breve{a}\|\underline{l}\|+\breve{\mu}\breve{G}\|\underline{\Delta}\|^2}{1-\breve{\mu}(\breve{G}\|\underline{\Delta}\|+2\breve{a}\omega)}, \tag{5.61}$$

where $\omega\in[0,1/\bar{c}_{\mathrm{N}}]$ and additionally the demand $0<\omega<\delta_1$ must hold. By taking δ_1 sufficiently small if necessary, such that $\breve{\mu}(\breve{G}\|\underline{\Delta}\|+2\breve{a}\omega)<1/2$ and $\|\underline{\Delta}\|^2<\|\underline{l}\|$ hold, the following reformulation becomes valid:

$$\begin{aligned}(\|\breve{\underline{D}}(\underline{\Delta},\underline{l},\omega)-\underline{d}^{\star}\|^2+\|\breve{\underline{\Lambda}}_{\mathrm{N}}(\underline{\Delta},\underline{l},\omega)-\underline{\lambda}_{\mathrm{N}}^{\star}\|^2)^{1/2} &\leqslant 4\breve{\mu}\breve{a}\|\underline{l}\|+2\breve{\mu}\breve{G}\|\underline{l}\| \\ &= 2\breve{\mu}(2\breve{a}+\breve{G})\|\underline{l}\|.\end{aligned} \tag{5.62}$$

It can be concluded that for

$$\delta_1<\min\{[2\breve{\mu}(2\breve{a}+\breve{G})]^{-1},\sqrt{\delta_2}\}\quad \delta_2<[2\breve{\mu}(2\breve{a}+\breve{G})]^{-1},\quad c_{\mathrm{N}}>\max\{\bar{c}_{\mathrm{N}},1/\delta_1\} \tag{5.63}$$

with $\|\lambda-\lambda^{\star}\|<\delta_2 c_{\mathrm{N}}$, the following holds

$$\|\breve{\underline{d}}(\underline{d},\underline{\lambda}_{\mathrm{N}},c_{\mathrm{N}})-\underline{d}^{\star}\| \leqslant 2\breve{\mu}(2\breve{a}+\breve{G})\frac{\|\underline{\lambda}_{\mathrm{N}}-\underline{\lambda}_{\mathrm{N}}^{\star}\|}{c_{\mathrm{N}}}, \tag{5.64a}$$

$$\|\breve{\underline{\lambda}}_{\mathrm{N}}(\underline{d},\underline{\lambda}_{\mathrm{N}},c_{\mathrm{N}})-\underline{\lambda}_{\mathrm{N}}^{\star}\| \leqslant 2\breve{\mu}(2\breve{a}+\breve{G})\frac{\|\underline{\lambda}_{\mathrm{N}}-\underline{\lambda}_{\mathrm{N}}^{\star}\|}{c_{\mathrm{N}}}. \tag{5.64b}$$

This directly implies that (5.38) holds for $\breve{M}=2\breve{\mu}(2\breve{a}+\breve{G})$ and $c_{\mathrm{N}}>\max\{\bar{c}_{\mathrm{N}},\ 1/\delta_1\}$. Furthermore, it is also possible to find a $\breve{M}$ such that (5.38) holds for $\bar{c}_{\mathrm{N}}\leqslant c_{\mathrm{N}}\leqslant\max\{\bar{c}_{\mathrm{N}},\ 1/\delta_1\}$, since $\breve{\underline{d}}$ and $\breve{\underline{\lambda}}_{\mathrm{N}}$ are continuous functions. Hence, the proof is complete. □

Note that the assumptions concerning matrix $\underline{\underline{A}}(\underline{d})$ are crucial for the proof: The tributary area matrix must always have full rank, i.e. $\mathrm{rank}(\underline{\underline{A}})=m=|\mathcal{S}|$. However, this must hold for any finite element discretized contact problem. Otherwise the underlying mesh would be already so heavily distorted that it is impossible to find a solution. Therefore, the requirements concerning $\underline{\underline{A}}$ do not restrict the applicability of Theorem 5.1 and are rather mandatory prerequisites for any finite element simulation.

5.4.2. Bounded Regularization Parameter

In the following the SIR update of Section 5.3.2 is considered. Under the prerequisite that the sequence $\{\underline{\lambda}_{\mathrm{N}}^{\{k\}}\}$ remains bounded, it is possible to show that the sequence $\{c_{\mathrm{N}}^{\{k\}}\}$ remains bounded as well. This can be summarized in the following theorem:

Theorem 5.2. Let $\{\underline{\lambda}_{\mathrm{N}}^{\{k\}}\}$ be a bounded sequence and assume that the inequalities

$$c_{\mathrm{g}}(1-\beta_{c_{\mathrm{N}}}^{\theta})\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\| < \|\tilde{\underline{g}}_{\mathrm{N}}^{\{k+1\}}\|, \tag{5.65}$$

$$\|\underline{\underline{A}}^{\{k+1\}}[\underline{\underline{A}}^{\{k\}}]^{-1}\| \leqslant c_{\mathrm{a}} \tag{5.66}$$

hold for some $0 < \beta_{c_{\mathrm{N}}}^{\theta} < 1$, where $c_{\mathrm{g}}, c_{\mathrm{a}} \in (0,\infty)$ are positive scalars, which are bounded away from zero and infinity. Furthermore, let $\|\tilde{\underline{g}}_{\mathrm{N}}\| > 0$ hold and let (5.21) be violated for all considered iterations $k \in \mathcal{K}_{c_{\mathrm{N}}}$, where $\mathcal{K}_{c_{\mathrm{N}}}$ is a finite sub-sequence of all iterations, such that c_{N} must be increased. Lastly, assume that Theorem 5.1 holds.

Then, for all $(\underline{d}, \underline{\lambda}_{\mathrm{N}}, c_{\mathrm{N}}) \in \mathcal{V}^{\star}(\delta_1, \delta_2, \bar{c}_{\mathrm{N}})$, where $\delta_1, \delta_2, \bar{c}_{\mathrm{N}} > 0$ are defined in Theorem 5.1, the regularization parameter is bounded by

$$c_{\mathrm{N}}^{\{k+1\}} \leqslant \breve{M}\left\{1 + \frac{2c_a}{c_{\mathrm{g}}(1-\beta_{c_{\mathrm{N}}}^{\theta})^2}\right\}. \tag{5.67}$$

Proof. To prove Theorem 5.2, assume that the sequence of the regularization parameter $\{c_{\mathrm{N}}^{\{k\}}\}$ becomes unbounded. Then, at some finite $k \geqslant \bar{k}$ the corresponding $c_{\mathrm{N}}^{\{k\}}$ will be large enough, such that $c_{\mathrm{N}}^{\{k\}} > \bar{c}_{\mathrm{N}}$ and, additionally, $c_{\mathrm{N}}^{\{k\}} > \breve{M}$ hold. For all $k > \bar{k}$ and due to the fact that c_{N} is increased in each considered iteration $k \in \mathcal{K}_{c_{\mathrm{N}}}$, (5.21) is multiplied by $\Theta^{\{k\}}$ resulting in

$$\begin{aligned}(1-\beta_{c_{\mathrm{N}}}^{\theta})\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\|^2 &\leqslant \langle \tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}, \tilde{\underline{g}}_{\mathrm{N}}^{\{k\}} + \nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}\Big|_{\{k\}}^{\mathrm{T}} (\underline{d}^{\{k+1\}} - \underline{d}^{\{k\}})\rangle \\ &\leqslant \|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\| \, \|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}} + \nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}\Big|_{\{k\}}^{\mathrm{T}} (\underline{d}^{\{k+1\}} - \underline{d}^{\{k\}})\|,\end{aligned} \tag{5.68}$$

where the Cauchy-Schwarz inequality has been applied. Next, (5.68) is divided by $\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\|$ on both sides and (5.4) is used to reformulate the right side, such that

$$\begin{aligned}(1-\beta_{c_{\mathrm{N}}}^{\theta})\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\| &\leqslant \frac{\|\underline{\underline{A}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k+1\}} - \underline{\lambda}_{\mathrm{N}}^{\{k\}})\|}{c_{\mathrm{N}}^{\{k\}}} \\ &\leqslant \frac{\|\underline{\underline{A}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k+1\}} - \underline{\lambda}_{\mathrm{N}}^{\star})\| + \|\underline{\underline{A}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k\}} - \underline{\lambda}_{\mathrm{N}}^{\star})\|}{c_{\mathrm{N}}^{\{k\}}} \\ &\leqslant \frac{2\|\underline{\underline{A}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k\}} - \underline{\lambda}_{\mathrm{N}}^{\star})\|}{c_{\mathrm{N}}^{\{k\}}}\end{aligned} \tag{5.69}$$

follows under consideration of the triangle inequality. The last inequality follows since $c_{\mathrm{N}}^{\{k\}} > \breve{M}$ holds, which implies $\|\underline{\lambda}_{\mathrm{N}}^{\{k+1\}} - \underline{\lambda}_{\mathrm{N}}^{\star}\| < \|\underline{\lambda}_{\mathrm{N}}^{\{k\}} - \underline{\lambda}_{\mathrm{N}}^{\star}\|$ according to (5.38) of Theorem 5.1. Now, under consideration of $(\underline{d}, \underline{\lambda}_{\mathrm{N}}, c_{\mathrm{N}}) \in \mathcal{V}^{\star}(\delta_1, \delta_2, \bar{c}_{\mathrm{N}})$ together with (5.34) and by applying the reverse triangle inequality, a second estimate is obtained

$$\begin{aligned}\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\| &\geqslant \|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}} + \nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}\Big|_{\{k\}}^{\mathrm{T}} (\underline{d}^{\{k+1\}} - \underline{d}^{\{k\}})\| = \frac{\|\underline{\underline{A}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k+1\}} - \underline{\lambda}_{\mathrm{N}}^{\{k\}})\|}{c_{\mathrm{N}}^{\{k\}}} \\ &\geqslant \frac{\|\underline{\underline{A}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k\}} - \underline{\lambda}_{\mathrm{N}}^{\star})\|}{c_{\mathrm{N}}^{\{k\}}} - \frac{\|\underline{\underline{A}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k+1\}} - \underline{\lambda}_{\mathrm{N}}^{\star})\|}{c_{\mathrm{N}}^{\{k\}}} \\ &\geqslant \left(\frac{1}{\breve{M}} - \frac{1}{c_{\mathrm{N}}^{\{k\}}}\right) \|\underline{\underline{A}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k+1\}} - \underline{\lambda}_{\mathrm{N}}^{\star})\|. \end{aligned} \tag{5.70}$$

In a next step, the iteration counter k in (5.69) is increased by one and the result is combined with (5.70), yielding

$$\frac{1}{2}(1 - \beta_{c_{\mathrm{N}}}^{\theta})\|\underline{\underline{A}}^{\{k+1\}}[\underline{\underline{A}}^{\{k\}}]^{-1}\|^{-1}\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k+1\}}\| \leqslant \left(\frac{c_{\mathrm{N}}^{\{k+1\}}}{\breve{M}} - \frac{c_{\mathrm{N}}^{\{k+1\}}}{c_{\mathrm{N}}^{\{k\}}}\right)^{-1} \|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\|. \tag{5.71}$$

To conclude the proof, $c_{\mathrm{N}}^{\{k\}}$ is set equal to $c_{\mathrm{N}}^{\{k+1\}}$ in (5.71) resulting in

$$c_{\mathrm{N}}^{\{k+1\}} \leqslant \breve{M}\left\{1 + \frac{2\|\underline{\underline{A}}^{\{k+1\}}[\underline{\underline{A}}^{\{k\}}]^{-1}\|}{(1 - \beta_{c_{\mathrm{N}}}^{\theta})} \frac{\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\|}{\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k+1\}}\|}\right\} \leqslant \breve{M}\left\{1 + \frac{2c_a}{c_{\mathrm{g}}(1 - \beta_{c_{\mathrm{N}}}^{\theta})^2}\right\}, \tag{5.72}$$

where $\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k+1\}}\| > c_{\mathrm{g}}(1 - \beta_{c_{\mathrm{N}}}^{\theta})\|\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\|$ and $\|\underline{\underline{A}}^{\{k+1\}}[\underline{\underline{A}}^{\{k\}}]^{-1}\| \leqslant c_a$ have been inserted to obtain the last inequality. Thus, from (5.72) follows that the assumption, that $\{c_{\mathrm{N}}^{\{k\}}\}$ is unbounded, together with $c_{\mathrm{N}}^{\{k+1\}} = c_{\mathrm{N}}^{\{k\}}$, i.e., no increase of c_{N}, leads to a finite upper bound for the regularization parameter as long as $\beta_{c_{\mathrm{N}}}^{\theta} < 1$ holds. This contradicts the assumption. Hence, it is proven for the SIR correction scheme that the sequence of regularization parameters $\{c_{\mathrm{N}}^{\{k\}}\}$ is indeed bounded as long as $\beta_{c_{\mathrm{N}}}^{\theta} < 1$. In addition, it is obvious that this upper bound rises with $\beta_{c_{\mathrm{N}}}^{\theta} \to 1$. The proof is thus complete. □

Now, the results and assumptions stated in Theorem 5.2 shall be examined in more detail. First, the positive scalar c_{g} shall be considered. This constant accounts for possible influences of neglected higher order terms, since (5.21) considers only a linear extension. However, this influence will decay with at least second order if the sequence $\{\underline{d}^{\{k\}}\}$ converges towards the solution $\underline{d}^{\star}$. Therefore, it can be assumed that c_{g} tends to 1 close enough to the solution and, consequently, it is reasonable to assume that it is bounded away from zero as long as the sequence $\{\underline{d}^{\{k\}}\}$ converges and $\|\tilde{\underline{g}}_{\mathrm{N}}\| > 0$ holds. Note that if $\|\tilde{\underline{g}}_{\mathrm{N}}\| = 0$, the solution is reached or there are no active constraints. In both cases, c_{N} would not be increased.

Next, the attention is on the positive scalar c_{a}: It is introduced with respect to the convergence of the tributary area diagonal matrix entries. All of these entries must be positive and the matrix

has always full rank. Thus, it stays invertible. In addition, the change of the tributary area matrix also decays as $\{\underline{d}^{\{k\}}\}$ converges, i.e., c_a tends to 1 close to the solution and is surely bounded away from infinity.

Finally, to the assumption that the demand (5.21) is violated for all considered iterations $k \in \mathcal{K}_{c_\mathrm{N}}$. If this prerequisite does not hold, c_N is not increased. Thus, without loss of generality, the considered sequence $\mathcal{K}_{c_\mathrm{N}}$ with $k \in \mathcal{K}_{c_\mathrm{N}}$ is a well-defined sub-sequence of all possible iterations. In fact, since c_N is bounded for $0 < \beta^{\theta}_{c_\mathrm{N}} < 1$, the sub-sequence $\mathcal{K}_{c_\mathrm{N}}$ must be finite.

Inequality (5.67) clearly shows that the upper bound rises if $\beta^{\theta}_{c_\mathrm{N}} \to 1$. This can be also seen in Figure 5.15a. Furthermore, (5.67) contains the limit case as well, since for the fulfillment of (5.21) with $\beta^{\theta}_{c_\mathrm{N}} = 1$ the modification in (5.2) must vanish and (5.1b) is reobtained. However, the modification vanishes only for $c_\mathrm{N} \to \infty$.

Remark 5.6. The entire convergence analysis presented in this section aims for Theorem 5.2. This result is of great importance, because it guarantees that the regularization parameter c_N does not rise to infinity on the way to the solution. This is crucial because in this case it is much more likely that the corresponding linear system of equations will remain solvable due to the bounded condition number of the system matrix. Without an upper bound, it could theoretically happen that the system matrix is conditioned increasingly worse with a steadily rising c_N value and the closer one comes to the solution until it is no longer possible to apply a meaningful iterative solver. This is especially true for the condensed system matrix, see also the discussion about the penalty method in Sections 3.2.1 and 3.2.3.2. Even though Theorem 5.2 guarantees an upper bound for c_N, the actual impact on the solvability remains an open question that strongly depends on the considered contact problem. Therefore, the issue of a potentially ill-conditioned system matrix will be further investigated in Section 5.6.7.

5.5. Switching Back to a Consistently Linearized System

The presented approach is advantageous as long as the discrete solution is far away from the optimal KKT-pair. If the non-linear solution procedure has already achieved a good approximation for the current load step, a switch to a consistently linearized, unmodified system of equations is still the best choice in terms of the necessary amount of Newton iterations. Despite the fact that the modification proposed here is suitable to converge comparably fast, the convergence speed is still maximally linear and strongly coupled to the applied regularization parameter correction scheme (see Sections 5.3 and 5.4). Therefore, the following switching strategy may avoid unnecessary solver steps. In addition, a well-timed switch can help to significantly reduce the impact of an increasingly worse conditioned linear system of equations, see Section 5.6.7.

First, a set of different criteria must be formulated which can be consulted to reliably identify the asymptotic regime. Therefore, three different conditions are proposed. These are tested in a consecutive manner, i.e., if one of them is not fulfilled the others can be skipped. The first one is the so-called *gap criterion*, which is quite cheap to perform. A upper bound for the absolute values of the active nodal gaps is defined dependent on the current discretization by

$$\mathcal{B}_{\text{pre}}^{\hat{g}_{\text{N}}} = \gamma_{\text{g}} \cdot \min_{e \in \mathcal{E}^{\mathcal{SM}}} \{L_{\text{edge}}^{(e)}\}, \tag{5.73}$$

where $\gamma_{\text{g}} \in (0,1)$ is a scaling factor for the minimal detected element edge length $L_{\text{edge}}^{(e)}$ of all contact surface elements on the slave and master bodies combined in the set $\mathcal{E}^{\mathcal{SM}}$. More precisely, the minimal element edge length in the reference configuration is considered. The first criterion is fulfilled as soon as the maximal value of the absolute nodal averaged weighted gap values of all active nodes is smaller than the given lower bound

$$\max_{i \in \mathcal{A}} \{|\hat{g}_{\text{N}}^{i}|\} < \mathcal{B}_{\text{pre}}^{\hat{g}_{\text{N}}}. \tag{5.74}$$

The consideration of the absolute value addresses two possible scenarios: On the one hand, this bound can be violated by a strongly overlapping region leading to large negative values. On the other hand, if the bodies are moved apart after a previous contact load step, this value can also be dominated by the still active but now positive gap values. An illustrative example will follow in Section 5.6.5.

However, since this represents a quite rough criterion, the contact contributions to the force balance are investigated in a subsequent step whenever the gap criterion is fulfilled. For this purpose, the magnitude as well as the angle of the structural and the contact gradients shall be considered. Note that only a subset of the entire structural gradient vector is considered. Namely, only entries located at degrees of freedom for which the vector $\tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\text{N}}^{\mathcal{A}} \underline{\lambda}_{\text{N}}^{\mathcal{A}}$ contains values unequal to zero. This is denoted by $(\cdot)|_{\mathcal{SM}}^{\mathcal{A}}$ in the following. The *angle criterion* will be fulfilled, iff

$$\mathcal{B}_{\text{pre}}^{\varphi} \geqslant \| \nabla_{\underline{d}} \mathscr{U}|_{\mathcal{SM}}^{\mathcal{A}} \| \| \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\text{N}}^{\mathcal{A}} \underline{\lambda}_{\text{N}}^{\mathcal{A}} \| - \langle \nabla_{\underline{d}} \mathscr{U}|_{\mathcal{SM}}^{\mathcal{A}}, \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\text{N}}^{\mathcal{A}} \underline{\lambda}_{\text{N}}^{\mathcal{A}} \rangle, \tag{5.75a}$$

$$\mathcal{B}_{\text{pre}}^{\varphi} = \text{TOL}_{\varphi} \| \nabla_{\underline{d}} \mathscr{U}|_{\mathcal{SM}}^{\mathcal{A}} \| \| \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\text{N}}^{\mathcal{A}} \underline{\lambda}_{\text{N}}^{\mathcal{A}} \| \tag{5.75b}$$

holds, where $\text{TOL}_{\varphi} > 0$ is a bound for the value of $1 - \cos(\varphi)$ and φ is the angle between the structural gradient and the product of weighted gap gradients and Lagrange multiplier values. The last criterion concerns the magnitude of the interface residual. The *magnitude criterion* will be fulfilled, iff

$$\| \nabla_{\underline{d}} \mathscr{U}|_{\mathcal{SM}}^{\mathcal{A}} - \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\text{N}}^{\mathcal{A}} \underline{\lambda}_{\text{N}}^{\mathcal{A}} \| \leqslant \mathcal{B}_{\text{pre}}^{\text{res}}, \tag{5.76a}$$

$$\mathcal{B}_{\text{pre}}^{\text{res}} = \text{TOL}_{\text{res}} \| \nabla_{\underline{d}} \mathscr{U}|_{\mathcal{SM}}^{\mathcal{A}} \| \tag{5.76b}$$

holds, where $\text{TOL}_{\text{res}} > 0$ denotes a relative tolerance for the interface contributions. Finally, the switch from the asymptotic to the pre-asymptotic phase has to be mentioned as well: Firstly, the switch back is automatically performed once at the beginning of each load step. Secondly, it is also initiated whenever the *gap criterion* is violated, or the absolute value of the averaged weighted gap value reaches a value twice as high as the smallest absolute averaged weighted gap value during the pre-asymptotic phase. The latter ones are mainly safe-guarding rules. If they are activated, the user-specified tolerances are probably too loose and should be adapted, e.g., by halving the values TOL_{res} and TOL_{φ}.

Remark 5.7. Note that in the special case $\mathcal{A}^{\{k\}} = \varnothing$ all considered contributions vanish due to the proposed selection of the considered degrees of freedom indicated by $(\cdot)|^{\mathcal{A}}_{\mathcal{SM}}$. In such a case, the algorithm switches to the asymptotic phase. However, this will become only important if the bodies come back into contact during the current load/time step.

Extensions for the Generalized-α Method

In case of dynamic contact problems the switching conditions must be slightly modified. While the gap criterion stays unchanged, the angle criterion and the magnitude criterion must be adapted. The first modification concerns the structural gradient $\nabla_{\underline{d}} \mathscr{U}|^{\mathcal{A}}_{\mathcal{SM}}$ vector which must be replaced by its dynamic counterpart, viz.

$$\underline{r}_{\mathrm{sg}\alpha}\big|^{\mathcal{A}}_{\mathcal{SM}} = \left[(1-\alpha_{\mathrm{f}})\, \nabla_{\underline{d}} \mathscr{U}\big|_{\underline{d}^{\{n+1\}}} + \alpha_{\mathrm{f}}\, \nabla_{\underline{d}} \mathscr{U}\big|_{\underline{d}^{\{n\}}} - \alpha_{\mathrm{f}}\, \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}}\Big|_{\underline{d}^{\{n\}}} \underline{\lambda}^{\{n\}}_{\mathrm{N}}\right]\Big|^{\mathcal{A}}_{\mathcal{SM}}, \tag{5.77}$$

where the DOF selection $(\cdot)|^{\mathcal{A}}_{\mathcal{SM}}$ is again based on the location of all non-zero values in the vector $[\tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}^{\mathcal{A}}_{\mathrm{N}} \underline{\lambda}^{\mathcal{A}}_{\mathrm{N}}]_{\{n+1\}}$. The second modification is related to this current contact force vector (containing all forces related to the Lagrange multiplier values) which must be simply scaled by its respective time integration parameter $(1-\alpha_{\mathrm{f}})$. Otherwise, even a possible solution point would not satisfy (5.76).

Under consideration of these two simple modifications the conditions (5.75) and (5.76) stay also valid in the dynamic case. Examples will follow in Section 6.10.6 and Section 6.10.7.

5.6. Numerical Examples

In this section the appealing properties of the novel modified approach shall be demonstrated. Therefore, the modified incomplete method will be compared to the consistently linearized method presented in Chapter 4. Firstly, a set of two-dimensional examples is presented whose purpose is to demonstrate the superior non-linear solver performance of the newly developed method. Subsequently, a detailed parameter study follows to identify an optimal parameter range for $\beta_{\theta}^{c_{\mathrm{N}}}, \beta_{\varphi}^{c_{\mathrm{N}}} \in (0,1)$. Additionally, the following points are covered: the influence of the numerical integration scheme, the major importance of the second order derivatives of the unit smooth normal field for large initial gaps as introduced in (4.34) and used in (4.39), the linear convergence rate of the plain modified Newton approach discussed in Section 5.4, as well as the impact of a possible switching strategy in the asymptotic phase which has been derived in Section 5.5. The discussion of these points is followed by a deeper look into the conditioning of the linear system of equations. Hereby, the practical difference between the condensed and the modified saddle-point formulation is going to be addressed in detail.

Finally, as already briefly mentioned in Remark 5.2, observations have been made that the modification works impressively well for the inconsistently varied approach considered here, while unexpected convergence problems occur when it is applied to the variationally consistent approach also introduced in Chapter 4. These observations will be further investigated in a final example. All presented results have been solely computed with the in-house high performance C++ code of the Institute for Computational Mechanics [274].

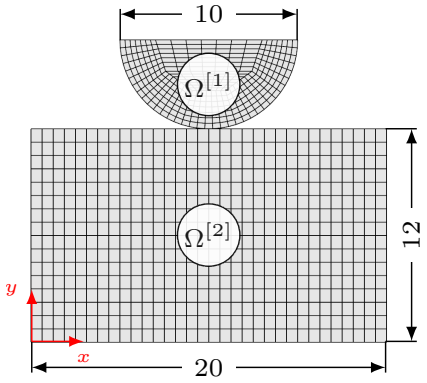


Figure 5.3.: Visualization of the geometry and mesh for the semicircular indentor onto plate example.

The stopping criteria for the non-linear solution strategy as well as the applied numerical integration scheme shall be presented next, since they will not change in the remainder of this section. The stopping criteria are given by

$$\|\nabla_{\underline{d}}\mathscr{U} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}\| < 1.0\mathrm{E}-6, \qquad \|\Delta\underline{d}^{\{k+1\}}\| < \max\{1.0, \|\underline{d}\|\} \cdot 1.0\mathrm{E}-10,$$
$$\|\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\| < 1.0\mathrm{E}-10, \qquad \|\Delta\underline{\lambda}_{\mathrm{N}}^{\{k+1\}}\| < \max\{1.0, \|\underline{\lambda}_{\mathrm{N}}^{\{k\}}\|\} \cdot 1.0\mathrm{E}-10,$$

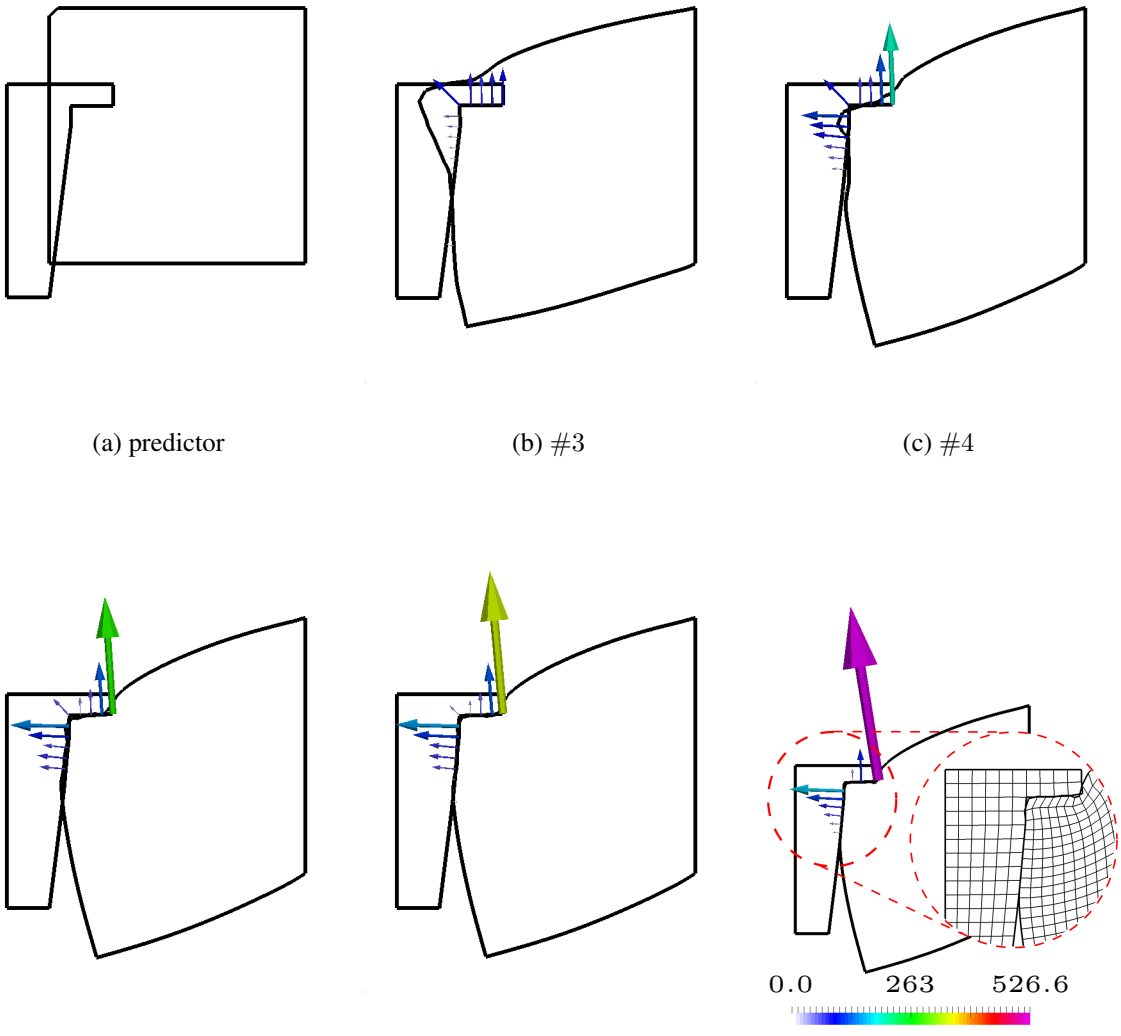
where $\|\cdot\|$ denotes always the ℓ_2-norm. The used numerical integration scheme will be the well-known element-based integration with 7 Gauss points per slave element in each dimensional direction. In other words: For two-dimensional examples 7 GPs are used and for three-dimensional examples and quadrilateral shaped surface elements 49 GPs are taken into account. A more detailed discussion about the influence of the used numerical integration scheme follows in Section 5.6.3. Finally, all of the considered example will use the so-called *tangential predictor* at iteration zero. A more comprehensive discussion of this predictor can be found in Appendix B.

5.6.1. Superior Performance for Large Initial Penetrations

Before a comprehensive parameter study of $\beta_\theta^{c_{\mathrm{N}}}$ and $\beta_\varphi^{c_{\mathrm{N}}} \in (0,1)$ is presented, the two underlying examples shall be introduced. Therefore, the regularization parameter is set to $c_{\mathrm{N}}^{\{0\}} = 1$ at the beginning of each load step. Furthermore, the attention is restricted to the *sufficient infeasibility reduction* (SIR) correction introduced in Section 5.3.2 for now and the related parameter $\beta_\theta^{c_{\mathrm{N}}}$ is set to 0.9.

5.6.1.1. Wedge Indentor

As a first example the splitting wedge shown in Figure 4.13a shall be reconsidered. Let us briefly repeat the result stated at the end of Section 4.7.4: The plain modified Newton approach converges very smoothly to the desired solution in exactly 22 iterations despite the fact that it is based on the variationally inconsistent approach. The reason is the less severe impact of the missing variations right at the beginning due to the small initial regularization parameter $c_{\mathrm{N}}^{\{0\}} = 1.0$. Then, after this critical start-up, the master body starts to bend to the right such that the red



(a) predictor (b) #3 (c) #4

(d) #5 (e) #6 (f) converged

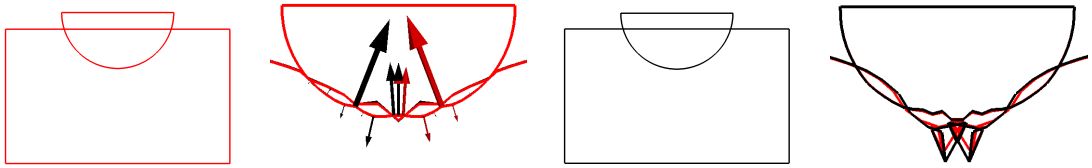
Figure 5.4.: Different snap-shots of the modified Newton approach for an initial penetration of 6.8.

projection zone slowly fades out and, finally, disappears. This circumstance helps to resolve the problem. However, since mainly the deformation helps to overcome the problem, it seems possible to construct an example where the problem remains even at the final solution, see the discussion in Section 4.7.4 for more information.

Next, the maximal possible initial penetration for the modified system of equations shall be identified. Therefore, the penetration has been increased by steps of 0.1. The results is that the simulation is still converging up to an initial penetration equal to 6.8 for the given set of boundary conditions, the stated initial parameters and the shown level of mesh refinement. At this stage, the wedge indentor has already been moved into the plate by far more than its reference height as Figure 5.4a impressively demonstrates. Obviously, such a huge penetration is challenging for the considered mesh, however, the converged solution is still successfully reached after 19 iterations. It is to highlight that this is a lower iteration number than for the only a fifth as high initial penetration which had been considered previously in Section 4.7.4. This undoubtedly indicates a superior robust performance. Furthermore, the Figures 5.4b to 5.4e show a sequence of four consecutive non-linear solver iterations which demonstrate how the approach is able to smoothly resolve an intermediate heavily distorted mesh. Actually it is possible to go even further if the correction parameter $\beta_\theta^{c_\mathrm{N}}$ is slightly reduced. Convergence up to 7.0 could be assured for an exemplary value of $\beta_\theta^{c_\mathrm{N}}$ equal to 0.8. However, an increase of the penetration gets at some level pointless, since the finally converged mesh is already quite heavily distorted for a value equal to 6.8 as Figure 5.4f demonstrates. Therefore, a further mesh refinement would become necessary, which would naturally lead to a lower barrier for the consistently linearized as well as for the modified approach.

5.6.1.2. Semicircular Indentor

As a second example the semicircular indentor from Zavarise et al. [294] shall be considered. The dimensioned geometry can be found in Figure 5.3, where the lower curved line of the semicircular body is the slave and the upper flat line of the rectangular body is the master side. The Young's moduli as well as the Poisson's ratios of the slave and master bodies are set to $E^{[1]} = 25{,}000$, $E^{[2]} = 2{,}500$ and $\nu^{[1]} = \nu^{[2]} = 0.25$. The top line of the semicircular indentor is fixed in x-direction, while a prescribed displacement in y-direction is applied. The master body is completely fixed at its bottom line. Under consideration of the consistently linearized system of equations both, the standard Lagrangian as well as the augmented Lagrangian formulation, achieve convergence up to an initial penetration of 3.5 in 12 and 11 iterations, respectively. It is to say that the reached possible initial penetration is already very high compared to the stated values in [294]. The therein used node-to-segment penalty approach converged in its consistently linearized form only up to a value of 0.8 and in its inconsistent variant up to 4.0. Anyhow, the impression can be deceiving. First of all, the chosen regularization parameter c_N plays a crucial role. In the case presented here, c_N has been set to 1.0. A value one magnitude higher, i.e. $c_\mathrm{N} = 10$, leads to divergence. Secondly, the intermediate deformation states of the contact interfaces during the non-linear iteration procedure show an already distorted displacement field. At the same time the Lagrange multiplier values scatter over a large range of non-physical values. All this can be seen in Figure 5.5. The red contour corresponds to the augmented Lagrangian solution while the black contour represents the standard Lagrangian solution. The related Lagrange multipliers are plotted over the left and right halve of the slave indentor, respectively.

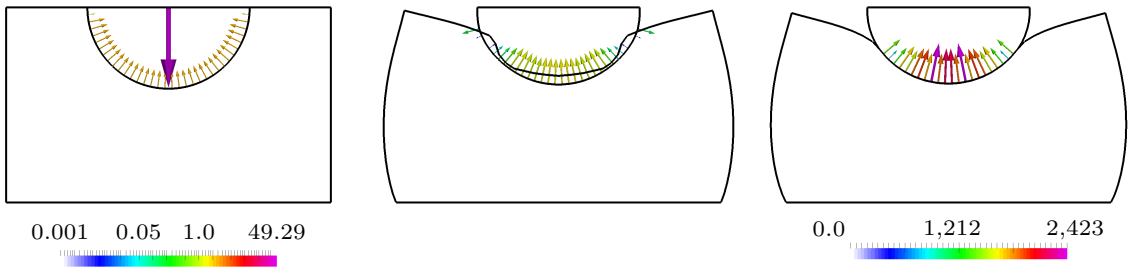


(a) pred. (pen=3.5) (b) #3 (pen=3.5) (c) pred. (pen=3.6) (d) #2 (pen=3.6)

Figure 5.5.: Figures 5.5a and 5.5b show the maximal possible initial penetration for the consistently linearized standard (black) and augmented Lagrangian (red) formulation. Figures 5.5c and 5.5d show the poor result of a further slight increase in initial penetration.

Even though the augmented Lagrangian formulation shows a slightly less severe mesh distortion on the master side, both consistently linearized formulations act very similar. This coincides with observation presented in [131]. To put it in a nutshell: The mesh distortion can become quite strong for large initial penetrations and the successful convergence can become a matter of luck. In this special case the success seems to rely largely on the simple geometry and the specifically used mesh. This becomes even more obvious if the second step for a slightly higher initial penetration is observed as presented in Figures 5.5c and 5.5d.

A selection of non-linear solution iterations of the modified Newton approach are presented in Figure 5.6. As one can easily see, no heavy mesh distortions are present. Instead, the deformation process is very smooth and predictable during the entire non-linear solution procedure. The maximal initial penetration is not restricted by the modified Newton approach, but instead by the used smooth normal field which is defined on the slave side (see also [131] and Chapter 4), i.e., on the semicircular indentor. In the limit case shown here, this leads to a concentration of the projected nodal reactive contact forces in only three master nodes, where the major part of the reaction forces is concentrated in the central node as shown in Figure 5.6a. Since the Lagrange multiplier values are initially all zero, the shown forces originate solely from the augmentation term on the modified right hand side (5.3) in form of $c_{\mathrm{N}}[\tilde{\nabla}_{\underline{d}}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}]\underline{\hat{g}}_{\mathrm{N}}^{\mathcal{A}}$. For an initial penetration larger than the circle radius the used ray-tracing along the smooth normal field would project the forces onto master elements that are not physically meaningful, such that the solution procedure is likely to fail. To make a long story short: This issue can be resolved by a smooth closest point projection or by switching the slave and master sides and has nothing to do with the modification of the linear system of equations presented here. In fact, by switching the slave and master sides an initial penetration up to -8.0 becomes possible and if at the same time the correction parameter $\beta_{\theta}^{c_{\mathrm{N}}}$ is relaxed from 0.9 to 0.8 even an initial penetration of -9.0 still converges flawlessly. Only afterwards the heavily compressed bulk elements close to the master interface start to prevent a further convergence. In order to keep things fair, the simulations are also repeated for the consistently linearized formulations. Here a switch of slave and master increases the maximal possible initial penetration up to a value of -5.4 for the standard as well as for the augmented Lagrangian formulation. A further increase leads to such a heavily distorted slave side that the iteration sequence starts to diverge. The remarkable thing is that divergence strikes very sudden in case of the consistently linearized case. While a penetration of -5.4 converges almost flawlessly, a slightly higher load completely destroys the simulation. This makes these methods hard to control, even with more sophisticated globalization approaches like line search, for example.



(a) predictor (b) #5 (c) converged

Figure 5.6.: A selection of non-linear iterations for the modified variant and an initial penetration of 5.0. The semicircular indentor acts as slave. Figure 5.6a shows the forces acting on the slave and master side due to the large initial penetration (originating from the gap). In Figures 5.6b and 5.6c the Lagrange multiplier values are visualized.

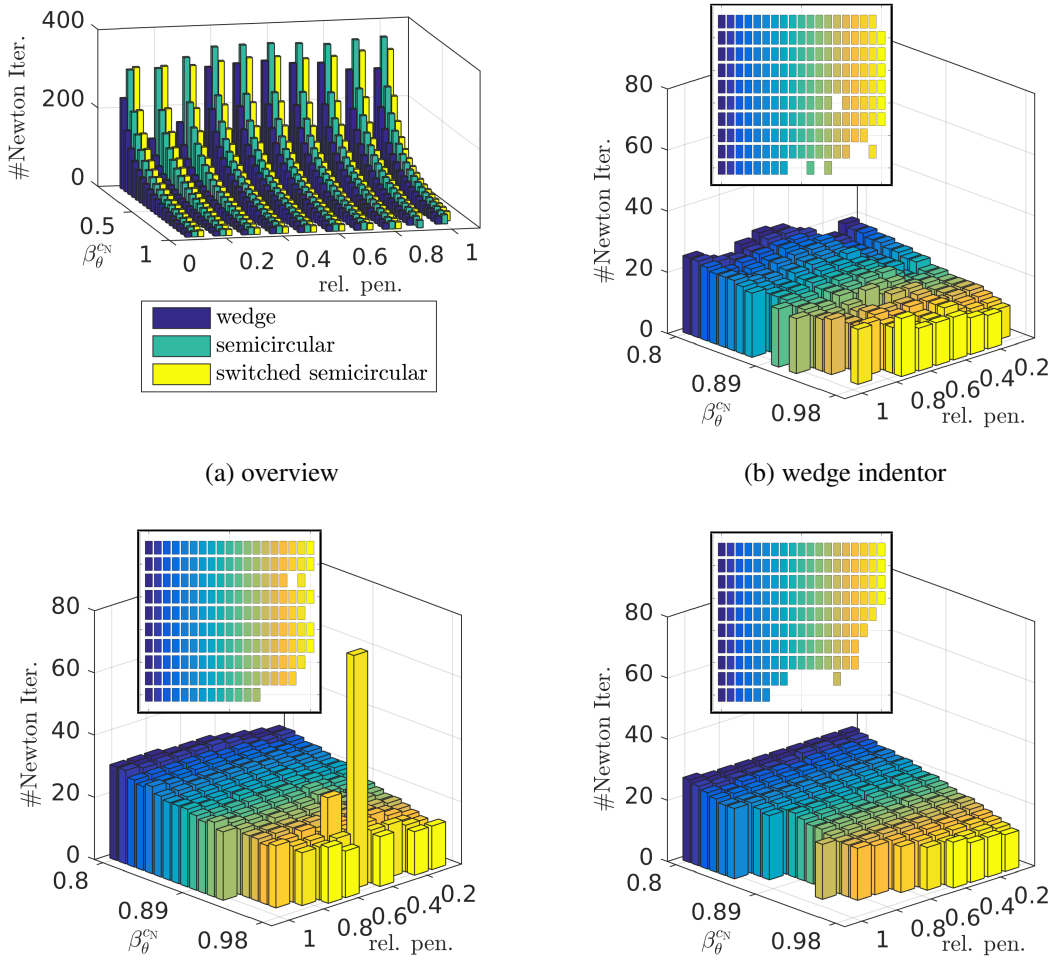
In summary it is to say that the modified Newton approach shows a superior behavior for this type of displacement controlled problems and leads to a very natural evolution of the Lagrange multiplier and displacement fields. Furthermore, the presented dynamic correction approach enormously simplifies the choice of the regularization parameter.

5.6.2. Parameter Study

The two shown examples will be now used to identify the optimal range of operation for the correction parameters $\beta_\theta^{c_\mathrm{N}}$ and $\beta_\varphi^{c_\mathrm{N}}$. Since it has been observable that a switch of slave and master in case of the semicircular indentor example can severely change the results, this variant will be included as well. In all cases the initial regularization parameter $c_\mathrm{N}^{\{0\}}$ is set to 1.0. The initial penetration is varied between 10 and 100 percent of the maximal possible penetration of the respective example.

5.6.2.1. Sufficient Infeasibility Reduction (SIR) Correction

The correction strategy called *sufficient infeasibility reduction* (SIR), which has been introduced in Section 5.3.2, comes first. Therefore, the initial penetration for the wedge indentor is in the range between 0.68 and 6.8, for the semicircular indentor as slave side, between 0.5 and 5.0 and for the switched side example, i.e., the semicircular indentor becomes the master side, between 0.9 and 9.0. The correction parameter $\beta_\theta^{c_\mathrm{N}}$ is varied between 0.1 and 0.98. Note that a value of 1.0 is not possible for the SIR correction since it would lead to a division by zero (5.25) and in theory to a transition to the consistently linearized system of equations. In total around 340 runs per example have been performed. The comprehensive results in terms of the necessary Newton iterations are summarized in Figure 5.7. Figure 5.7a gives an overview of the global development of the iterations over the entire range. At the beginning the necessary non-linear iterations are exponentially falling with a rising parameter $\beta_\theta^{c_\mathrm{N}}$. However, close to one the differences start to flatten out. In the Figures 5.7b to 5.7d a close-up view of the iterations for a $\beta_\theta^{c_\mathrm{N}}$ value in the range of 0.8 and 0.98 are shown. The wedge example in Figure 5.7b shows, in contrast to the semicircular examples, a slightly more inhomogeneous decreasing behavior to the end. For combinations of very high correction parameters, i.e. beyond 0.85, with high initial penetrations,



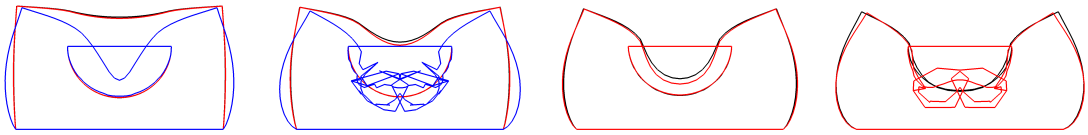
(a) overview

(b) wedge indentor

(c) semicircular indentor

(d) switched semicircular indentor

Figure 5.7.: Results of the parameter study for the SIR correction. While Figure 5.7a shows an overview over the results of the parameter study, Figures 5.7b to 5.7d present a detailed view for the higher correction parameter regime. Each figure includes a view from above such that failing simulations can be easily identified by missing squares.



(a) #3 (b) #4 (c) #5 (d) #6

Figure 5.8.: Selection of non-linear iterations for different correction parameters and an initial penetration of 9.0. The semicircular indentor represents the master side. —— corresponds to $\beta_\theta^{c_N}$ equal to 0.85, —— to 0.86 and —— to 0.99. The intermediate deformation states for the largest value, i.e. 0.99, are only shown in Figures 5.8a and 5.8b.

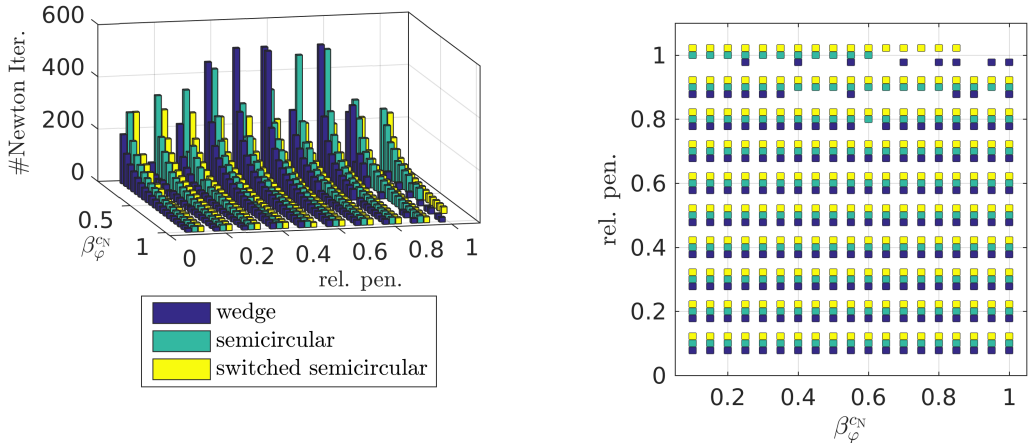
the simulations start to become unstable. In more detail, the first failing wedge simulation is detected for an initial penetration of 6.8 and a regularization parameter of 0.88. For the semicircular indentor example the first failure arises for a maximal penetration of 5.0 and a $\beta_\theta^{c_N}$ value of 0.93. Finally, the switched slave master example reaches its first limit for an initial penetration of 9.0 and a correction parameter of 0.86. In all cases the reason is a too quickly rising regularization parameter coming along with a too strong mesh distortion which leads to an undesired displacement field, see Figure 5.8 for a demonstrative example. However, the behavior stays predictable since the first failure is always detected for a combination of maximally considered penetration and high regularization parameter. Therefore, a smaller correction parameter will always lead to a more stable result.

It can be concluded that the SIR correction strategy provides a very reliable and robust update. However, the parameter $\beta_\theta^{c_N}$ must be chosen wisely dependent on the investigated example. For very high penetrations a value around 0.8 seems to be a good compromise between fast convergence and robust mesh deformation. For more usual penetrations a value around 0.9 delivers most times stable results and is slightly more efficient.

5.6.2.2. Sufficient Enclosed Angle (SEA) Correction

Now, the attention is on the *sufficient enclosed angle* (SEA) criterion which has been introduced in Section 5.3.1. The parameter study is repeated, while this time the over-all results are summarized in Figure 5.9. In Figure 5.9b a view from above is added which reveals a more unstable behavior in comparison to the SIR approach for initial penetrations above 70% of the maximal possible penetration. This is especially true in case of the wedge example. One can easily see that the method diverges, e.g., for an initial penetration of 4.08, i.e. 80%, and a correction factor of 0.6. The reason for the bad non-linear solver behavior is a too fast rising regularization parameter.

However, this unfavorable behavior is only observable for the wedge example. In case of the semicircular examples the SEA correction works competitively well compared to the SIR method. Therefore, the minimal necessary non-linear iteration numbers which still rely on a continuously stable non-linear solution path with respect to the used correction method are summarized in Table 5.1. That means if the corresponding method has been already diverged for a lower $\beta_{(\cdot)}^{c_N}$ value, a higher regularization parameter leading to a lower iteration count is ignored in the list, since the outlier on the monotone increasing $\beta_{(\cdot)}^{c_N}$ path indicates some kind of instability. First of all, Table 5.1 shows that the SIR correction leads to better or comparably good results in all considered cases. Hence, it is the better choice not only for very large penetrations but

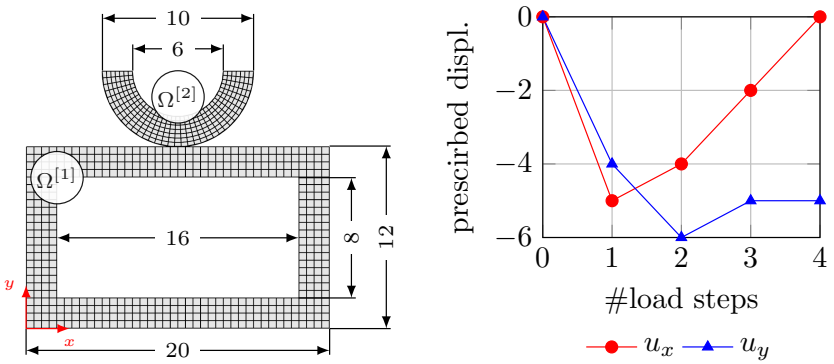


(a) (b)

Figure 5.9.: Results of the parameter study for the SEA correction scheme. Right: necessary iterations; left: view from above where failing simulations can be identified by missing squares.

rel. pen.	wedge				semicircular				s. semicircular			
	SIR		SEA		SIR		SEA		SIR		SEA	
	it.	$\beta_\theta^{c_N}$	it.	$\beta_\varphi^{c_N}$	it.	$\beta_\theta^{c_N}$	it.	$\beta_\varphi^{c_N}$	it.	$\beta_\theta^{c_N}$	it.	$\beta_\varphi^{c_N}$
0.5	14	0.98	23	0.99	18	0.96	24	0.99	14	0.97	18	1.00
0.6	18	0.91	23	0.99	15	0.98	24	1.00	15	0.96	23	0.97
0.7	15	0.96	20	0.99	17	0.97	26	0.98	16	0.95	23	0.97
0.8	15	0.93	41	0.55	17	0.97	26	0.98	17	0.93	23	0.96
0.9	17	0.92	42	0.35	19	0.95	27	0.97	21	0.87	22	0.99
1.0	21	0.87	–	–	22	0.92	43	0.5	23	0.84	22	0.89

Table 5.1.: Comparison of the SIR and SEA correction schemes. Listed are the lowest iteration numbers in a stable sequence of consecutive increases of the related correction parameter together with the associated correction parameter value. In green highlighted is the lower iteration number of the SIR or SEA approach, respectively.



(a) (b)

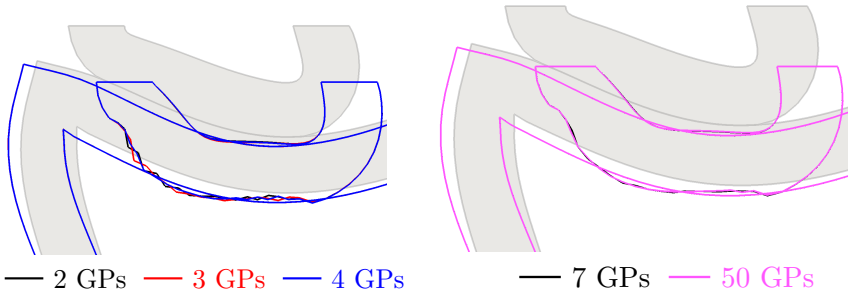
Figure 5.10.: Figure 5.10a shows the dimensioned geometry as well as the mesh in the reference configuration. Figure 5.10b presents the four considered load steps. The prescribed displacements are applied on top of the semi-circular indentor.

also for more common ones. Nevertheless, it is to say that almost all investigated cases could also be solved under consideration of the SEA correction. Only for the wedge example severe problems arose for the maximal penetration, which might be caused by the inherent instability of the incomplete variational approach as comprehensively discussed in Section 4.7.4. This suspicion comes up, since the SEA correction relies in a much greater extent on the incomplete / missing variational contributions than the SIR correction. Furthermore, for the semicircular examples, SEA performed with a quite comparable efficiency. In order to get more clarity in this point, the modified Newton approach should be transferred to the complete variational approach in the future. In addition, a formulation based on the closest-point projection should be far less affected.

To put it in a nutshell: The unambiguous result of the parameter study is that in the currently considered set-up the SIR approach is advantageous in all investigated cases and should be the method of choice. Therefore, in all up-coming examples exclusively the SIR correction scheme will be applied.

5.6.3. Successive Quasi-Static Load Steps

All previous discussions considered only the very first contact load step. During these examples an impressive performance of the modified Newton method could be observed. Nevertheless, the presented method is not restricted to the very first load step, but can also be used in all subsequent steps. However, a new ingredient must be considered which has only been mentioned very briefly: the underlying numerical integration scheme. It is to remember that the entire shown analysis is built up on the assumption that all terms are at least twice continuously differentiable. Unfortunately, this assumption does in reality not automatically hold. This assumption is exemplarily violated as soon as the quite common element-based integration is considerd for the evaluation of the active contact contributions. Even though, the underlying approach is based on the incomplete variational scheme introduced in [131], the element-based integration can still introduce artificial discontinuities in the second order derivatives of the incomplete weighted gap gradient, i.e., discontinuities which are only caused by the insufficient numerical integra-



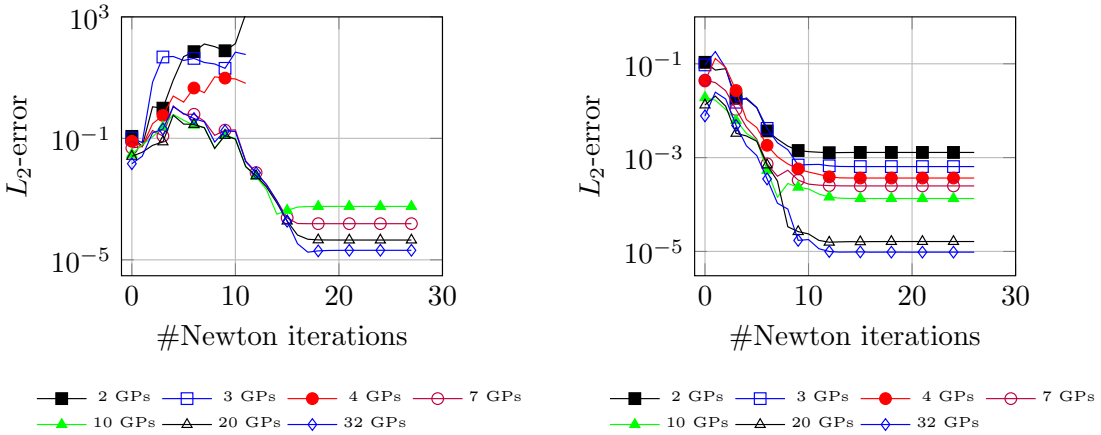
(a) (b)

Figure 5.11.: Visualization of the predictor step for different numbers of Gauss points (GP) used for the evaluation of the contact contributions. The lower GP numbers in Figure 5.11a cause heavy mesh distortion on the master side, while the higher GP numbers in Figure 5.11b show a smooth initial deformation field. The gray colored transparent silhouette in the background shows the converged configuration of the previous step, i.e. the starting point of the predictor step. Shown is the predictor step of the second load step.

tion scheme. Therefore, any discontinuities related to the active/inactive set decision is of minor importance at this point (see Figure 2.3b).

A new example shall be considered to illustrate the influence of this flaw between theory and application. The example consists of two bodies. Both use a Neo-Hookean material law. The hollow brick represents the slave body and is ten times stiffer than the semicircular body with a Young's modulus of $E^{[1]} = 2{,}500$ and $E^{[2]} = 250$, respectively. Poisson's ratio is for both bodies set equal to 0.25. While the hollow brick is fixed on the bottom line in all directions, the upper semicircular body is moved via a prescribed displacement field in x- and y-direction. The corresponding prescribed displacements are shown in Figure 5.10b. The different load steps have been chosen in such a way that they cover a wide variety of possible loading and unloading scenarios. The β_θ^{CN} value is set to 0.8. The crucial load step is the second one where the semicircular indentor is moved even further into the stiffer hollow brick while at the same time a sliding in positive x-direction is initiated. The integration is completely performed on the slave side, i.e., on the top surface of the hollow brick, and the contributions of the master side are considered by application of the well-known ray-tracing projection algorithm. Since no segment-based integration is used, the Gauss point projection from the slave onto the master surface leads to an unfavorable integration over kinks stemming from the element-wise defined Lagrangian polynomials on the master side. These kinks cannot be completely resolved by this simple integration approach and lead to a distortion of the derived derivatives and, finally, distort the (intermediate) solutions. To demonstrate this, the element-based integration is performed with respect to a changing number of Gauss points, where the range goes from 2 Gauss points per slave element to 50.

Each single load step is restarted from the solution achieved with 50 GPs, such that a coinciding initial state is given. In Figures 5.11a and 5.11b the deformed configurations after the tangential predictor load step at the very beginning of the second load step are presented (see also Figure 5.10b). The now initially active contact zone can lead to quite heavy distortions of the master surface, while the slave surface is integrated in an almost exact manner (besides the rational smooth normal expressions) and thus is not influenced in the very first tangential nonlinear solver step. Especially, in case of very low GP numbers the distortion of the master surface can become cumbersome and gets worse for a softer master body as chosen in this example. For-



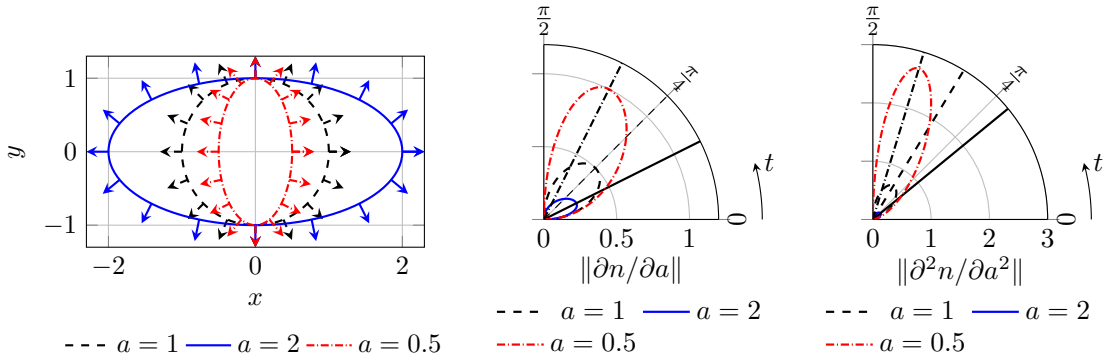
(a) #2 (b) #4

Figure 5.12.: Comparison of the L2-error among the different Gauss point numbers, the reference being the 50 GP solution. The L2-error is computed with respect to the displacement field. Figure 5.12a shows the comparison for the second load step, Figure 5.12b for the fourth load step.

tunately, the zig-zag deformation of the active master contact surface is smoothed out with an increasing number of GPs. Furthermore, also for small numbers of GPs the zig-zagging disappears quite rapidly in subsequent Newton steps. However, in the presented example the shown configuration is close to a structural instability, since the curved semicircular shaped beam tends to snap through. This circumstance can lead to a divergence of the non-linear solution method if the bodies are only slightly deflected in a wrong direction. Actually, exactly this is happening in case of the lower numbers of GPs, i.e., for 2, 3 or 4. Even though, 4 GPs are already very close to succeed. On the other hand, there is almost no detectable difference if the tangential predictor solutions between the 7 and 50 GP solutions are compared to each other in Figure 5.11b.

To put the quite rough visual comparison into perspective, the L_2-error among the displacement fields of the different GP solutions has been evaluated over the entire sequence of non-linear solution iterates. As illustrated in Figures 5.12a and 5.12b, the difficulties of the second load step are also manifested in the development of the L_2-error. In load step 2, all attempts show a rising difference in the first four non-linear iterations compared to the 50 GP solution path. This holds independently of the used number of GPs causing even a better result for 7 GPs than for 10 compared to the 50 GP result. Therefore, a non-deterministic behavior is revealed which can be traced back to the mentioned near snap through scenario. Afterwards the differences drop almost monotonically to their final values, which relates to a stable behavior. In contrast, the fourth load step presented in Figure 5.12b shows only a rise in the very first non-linear iteration, i.e., immediately after the tangential predictor, and subsequently L_2-error drops almost monotonically to its final value. Here, a high number of Gauss points correlates with a smaller error also for the pre-asymptotic phase. This is a typical deterministic behavior.

As a general rule, it can be concluded that a sufficient number of Gauss points has to be provided to achieve a meaningful behavior for successive contact loading steps. The reason why this is not crucial in the very first load step, which had been exclusively considered beforehand, is the scaling of the critical second order weighted gap derivatives by the Lagrange multipliers. In the very first step, all the Lagrange multipliers start at zero and their values are only slowly



(a) ellipses (b) first order derivatives (c) second order derivatives

Figure 5.13.: Figure 5.13a shows three different ellipses for a varying parameter a. Figures 5.13b and 5.13c show the norm of the first and second order derivatives with respect to this parameter a.

rising, such that the influence stays almost entirely undetected. The solution is already close to the asymptotic phase as soon as the Lagrange multiplier values reach a meaningful magnitude. In the asymptotic phase the inexact integration seems to play only a less significant role. As a rule of thumb: seven Gauss points for linear shape functions seem to represent a good compromise between efficiency and accuracy.

5.6.4. Second Order Derivatives of the Unit Smooth Normal Field

A detailed study of the presented formulation reveals that the introduced second order derivative term of the unit normal field in (4.39), namely,

$$\langle \mathrm{D}_{\Delta \underline{u}}(\mathrm{D}_{\delta \underline{u}}(\hat{\underline{n}}^{[1]})), \bar{\underline{x}}^{[2]} - \underline{x}^{[1]} \rangle \tag{5.78}$$

can lead to a severely high evaluation time of the considered tangential stiffness contact contributions. Thus, the question may raise how important is this term for the presented formulation? Especially, since it enters the final system of equations scaled by the current Lagrange multiplier value as shown in (5.1) or (5.3). In a first attempt, the previous examples have been executed once more without the second order derivatives of the smooth normal field and, indeed, all tests indicated that it is possible to remove this term with only minor influence on the non-linear solver behavior. But, such an empiric procedure might give a wrong idea, since the risk remains that a crucial situation is simply not triggered by the applied benchmark examples. Therefore, the following analytical investigations were initiated to obtain a better understanding of this term. An example with one degree of freedom shall be considered, given by an ellipse, viz., $\underline{e}(t, a) = [a\cos(t), \sin(t)]^{\mathrm{T}}$ for $t \in [0, 2\pi)$ and $a > 0$. In Figure 5.13a three different ellipses with three different values for the parameter a are shown. Obviously, a value equal to one leads to a circle, while a value greater than one stretches the ellipse in x-direction, whereas, a value between zero and one compresses the ellipse. Carried over to the contact problems discussed here, the different configurations can be interpreted as deformation states. One advantage of this simple example is that it is possible to directly state the unit normal field:

$$\underline{n}(t,a) = \frac{\hat{\underline{n}}(t,a)}{\|\hat{\underline{n}}(t,a)\|}, \qquad \text{where } \hat{\underline{n}}(t,a) = \big(\cos(t), a\sin(t)\big)^{\mathrm{T}}. \tag{5.79}$$

Again, the corresponding results are exemplarily shown in Figure 5.13a. Next, the first and second order derivatives with respect to the free parameter a are obtained by

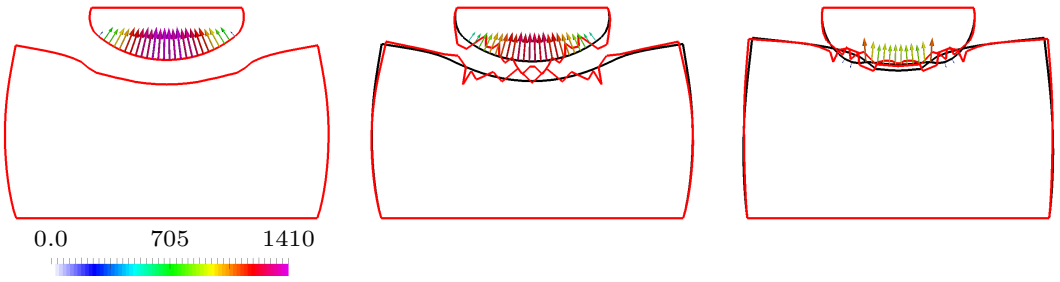
$$\frac{\partial}{\partial a}\underline{n}(t,a) = \frac{\sin(t)}{\|\hat{\underline{n}}(t,a)\|^3}\begin{pmatrix} -a\cos(t)\sin(t) \\ \|\hat{\underline{n}}(t,a)\|^2 - a^2\sin^2(t) \end{pmatrix} \tag{5.80}$$

and

$$\frac{\partial^2}{\partial a^2}\underline{n}(t,a) = \frac{\sin^2(t)}{\|\hat{\underline{n}}(t,a)\|^5}\begin{pmatrix} 3\cos(t)a^2\sin^2(t) - \cos(t)\|\hat{\underline{n}}(t,a)\|^2 \\ 3a^3\sin^3(t) - 3a\sin(t)\|\hat{\underline{n}}(t,a)\|^2 \end{pmatrix}, \tag{5.81}$$

respectively. The corresponding plots can be found in Figures 5.13b and 5.13c. The roots can immediately be identified at a multiple of $\pi/2$, i.e., $t \in \{0, \pi/2, 3\pi/2, \dots\}$. This makes sense, since a stretch or compression in x-direction does not change the normals at these angular positions. But the most interesting result of this study is that the ellipse with a equal to 0.5 shows much higher first and second order derivatives than a circle or a stretched ellipse with a equal to 2. Actually, the stretched ellipse shows only a very small magnitude with respect to the second order derivatives. This means that the unit normals change for the compressed ellipse much more rapidly in comparison to an already stretched ellipse. The position of the related maxima are also quite interesting and are highlighted by black straight lines in Figures 5.13b and 5.13c. For example., the highest first order derivative in the case of a perfect circle is reached for $t = \pi/4$. In general, it can be said that with an increasing a the maxima of the first and second order derivatives move from the intersection point between ellipse and the x-axis towards the intersection point with the y-axis.

Now, these results are transferred to the contact problems considered here. Actually, one example has already been close to the discussed ellipse, viz. the semicircular indentor. If the semicircular indentor takes the slave role, the related smooth unit normal field has very similar properties than the previously discussed analytical example. Furthermore, the following prerequisites must be satisfied to maximize the influence of the second order derivative: The shape of the semicircular indentor is in a compressed configuration and the Lagrange multiplier values should be unequal to zero. Furthermore, the distance between slave and master should be preferably high. These requirements lead to the following scenario: The Young's modulus of the slave side is reduced to $2{,}500$, such that it becomes equal to the master body, and a high compression is generated by using the initial maximal overlap of 5.0. This creates the desired shape of the slave body and leads to high values for the active Lagrange multipliers. After this starting position has been reached at the end of the first load step, the influence of the second order derivatives can be triggered by providing a sufficiently high gap and a quickly changing radial shape in the second load step. However, the decrease of the Lagrange multiplier values should not happen too quickly. The easiest way to achieve all these points is by moving the indentor upwards, i.e., by reducing the previously applied penetration. If this happens at the correct speed, i.e, not too



(a) #0 (b) #3 (c) #4

Figure 5.14.: Comparison between simulations including the second order derivatives of the unit smooth normal field —— and without these second order derivatives ——.

fast and not too slow (in this special case the penetration is reduced to a value of 3.0), the impact of the second order derivative can become tremendously high as shown in Figure 5.14.

However, in the first load step the simulation without the second order derivatives of the smooth normal field converges even one iteration faster, in 20 iterations, than the consistently linearized one. Only in the second step, the consistent variant reaches the solution in 20 iterations while the inconsistent one diverges.

In order to complete this section, it is to say that the second order derivatives of the smooth normal field are influenced by a variety of very different parameters. But, the presented detailed investigations undoubtedly reveal that they play a crucial role for the non-linear solver behavior. Especially, since in real world problems, i.e., for more complex geometries and loadings, such critical configurations may occur very localized even for quite small and more common load steps.

5.6.5. Convergence Rates of the Plain Modified System

In this section, the actual achieved convergence rates as well as the in Section 5.4.2 analytically predicted upper bound for the regularization parameter shall be verified. The discussion starts with the latter one and it is again noted that all computations use an initial regularization parameter of one, i.e., $c_{\mathrm{N}}^{\{0\}} = 1.0$. Figure 5.15a demonstratively shows the results for the wedge example of Section 5.6.1.1. The increase of the regularization parameter is indeed bounded from above and as shown by (5.72) this upper bound is increased by a rising $\beta_\theta^{c_{\mathrm{N}}}$ value. While the lowest $\beta_\theta^{c_{\mathrm{N}}}$ value leads to an upper bound around $7.27\mathrm{E}{+}02$, a $\beta_\theta^{c_{\mathrm{N}}}$ value of 0.9 leads to a much higher upper bound around $5.89\mathrm{E}{+}04$. Besides the different upper bounds, the shown curves in Figure 5.15a have a very similar characteristic. For example, all of them show a short plateau of constant regularization parameters over a short period of a few non-linear iterations before the value rises to the final upper bound. One reason for this behavior is the still changing active set distribution in the pre-asymtotic phase of the non-linear solution path. This can lead to a scenario where the previously estimated c_{N} value represents a sufficient guess to reach the necessary reduction of the linear model for a short sequence of successive iterations. Only close to the solution, as soon as all constraints are part of $\mathcal{A}_+$ and the model prediction becomes reliable, the fulfillment of the linear reduction demands again a higher regularization parameter, until finally a monotonic convergence to an upper bound can be observed as derived in Section 5.4.2. Since the last obser-

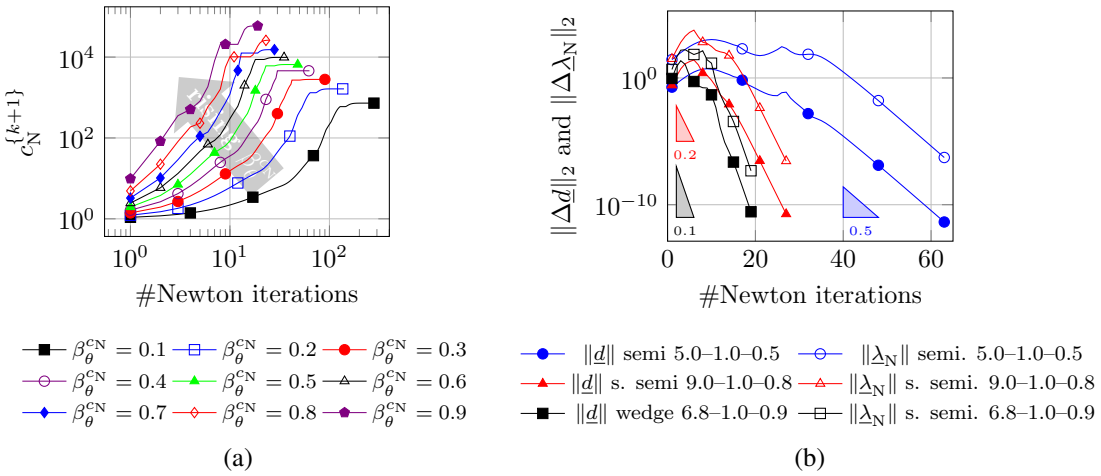
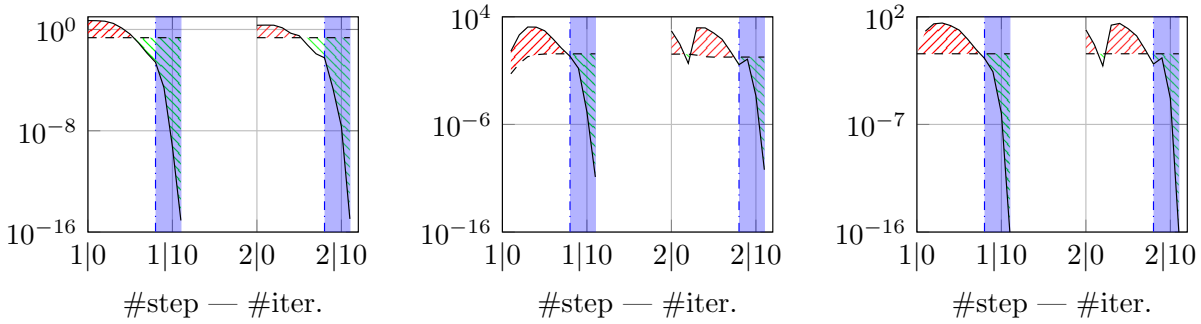


Figure 5.15.: In Figure 5.15a the influence on the upper bound for the regularization parameter c_N in case of a rising $\beta_{c_N}^{\theta}$ value is shown. Here, the wedge example with an initial penetration of 6.8 is used. In Figure 5.15b the decaying incremental norms for different examples are presented. The legend is supposed to be read as "initial penetration–$c_N^{\{0\}}$–$\beta_{c_N}^{\theta}$".

vation is also true for all the other 2-D examples, i.e., the upper bound is in the asymptotic phase consistently slightly underestimated, the convergence rates are bounded from below by (5.29) and thus Figure 5.15b shows for all considered examples a linear convergence rate with a slope close to $(1 - \beta_\theta^{c_N})$ in the asymptotic region, i.e., for an almost constant regularization parameter. This is in accordance with the predicted behavior in (5.29).

5.6.6. Effect of the Switching Condition

Within the scope of the current discussion, the opportunity is taken and the influence of the proposed switching conditions presented in Section 5.5 shall be also investigated in more detail. The basic idea of these switching conditions is to combine the benefits of the modified approach, viz. the higher robustness during the pre-asymptotic phase, with the fast local convergence of the classical consistently linearized system of equations. In total three different switching conditions have been proposed based on the current nodal gap, the relative residual of the active slave and master contributions and the angle between the incomplete weighted gap gradient and the structural potential gradient. To demonstrate the effectiveness the attention is exemplarily drawn to the soft semicircular example of Section 5.6.4. Thus, two load steps are considered. In the first step, the two bodies are pressed into each other leading to large initial penetration. In the follow-up step, the load is quickly reduced by an upward motion of the semicircular indentor, as shown in Figure 5.14. This leads to two very different developments of the considered conditions. To initiate a switch from the modified variant to the default system, all three conditions must be fulfilled simultaneously. This asymptotic phase is highlighted in blue in Figure 5.16. The dashed line in Figure 5.16a represents the right side of (5.74) with $\gamma_g = 0.5$. The smallest element edge length is around 0.4486. In Figure 5.16b the residual criterion defined in (5.76) is considered. Here, the $\mathrm{TOL}_{\mathrm{res}}$ has been set to $1.0\mathrm{e}{-3}$. However, the dashed line is also influenced by the gradient of the structural potential and, therefore, not constant. Finally, the tolerance for the angle



(a) gap (b) residual (c) angle

Figure 5.16.: Visualization of the three switching conditions presented in Section 5.5 for the example introduced in Section 5.6.4. In the red hatched areas the corresponding criterion is not fulfilled. In the green areas it is fulfilled and in the blue highlighted regions the method has been switched from the modified approach to the standard Lagrangian approach.

criterion, namely TOL_{φ}, in (5.75) has been set to $1.0\text{e}{-6}$. This is equivalent to an admissible angle below $0.08103\ \text{deg}$ or $1.414\text{e}{-3}\ \text{rad}$. The tolerances shown here worked also well for all the other tested examples. However, as already mentioned in Section 5.5 an adaptive strategy might be reasonable in some cases. For the considered first loading step, all considered criteria behave very predictively and show a monotonically decreasing trend after a first slight rise of the considered quantities. The total number of iterations could be reduced from 21 to 12. That is a reduction of almost 43%. On the other hand, the second unloading step is more challenging for the proposed switching conditions. As one might see in Figures 5.16b and 5.16c, the angle as well as the residual conditions show a sudden drop in the second Newton iteration, before they start to rise again. Thus, in this scenario the rough criterion considering the nodal gaps becomes crucial and avoids an unnecessary switch in a much too early iteration state. Actually, the switch in Newton iteration two would lead to divergence. In the second load step, the total number of non-linear iterations could be reduced from 20 to 12, i.e., about 40%.

5.6.7. Conditioning of the Tangential Stiffness Matrix

Finally, the attention is drawn to the condition number of the tangential stiffness matrix. These condition numbers play a crucial role whenever it comes to finer meshes or more complex models which ask for rising memory consumption. At some point it is no longer possible to solve the evolving linear system of equations directly and iterative linear solvers must be considered. Theory and practice show that penalty methods can lead to severe problems since the introduced penalty parameter influences beyond a specific threshold the condition number in such a way that the linear solver performance becomes increasingly worse, see Sections 3.2.1 and 3.2.3.2 for more information and references on this topic. One may have noticed that in the formulation presented here, as stated in Section 5.2.1, the modified system of equations becomes in the limit equal to the system of equations for the unmodified case, i.e., the standard Lagrangian approach. Furthermore, it has been said that the c_{N} value must rise sufficiently fast to achieve (super-) linear convergence. However, it can be expected that a extremely high c_{N} value will lead to an increasingly high condition number, similar to a traditional penalty approach, see also [23]

for a short analysis. Fortunately, it is possible to achieve fast convergence by simultaneously maintaining a comparatively low condition number of the system matrix. In order to prove this, the condition number estimates κ_∞ of the rather small sparse matrices will be directly computed with dGECON routine from the LAPACK distribution [4].

5.6.7.1. Semicircular Indentor

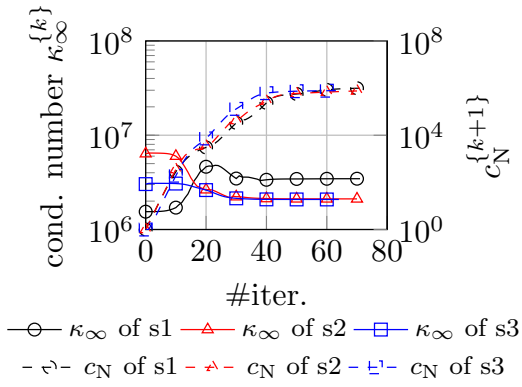
The discussion is started with the semicircular indentor example which has been already introduced in Figure 5.3 and Section 5.6.1.2. The only difference is that the mesh has been refined by a factor of two and both Young's moduli of the slave and master body are set to $25{,}000$. The considered three loading steps are:

1. The indentor is moved 4.0 units into the block.
2. The indentor is moved 2.0 units in the opposite direction.
3. The indentor travels by constant penetration 1.0 unit in x-direction.

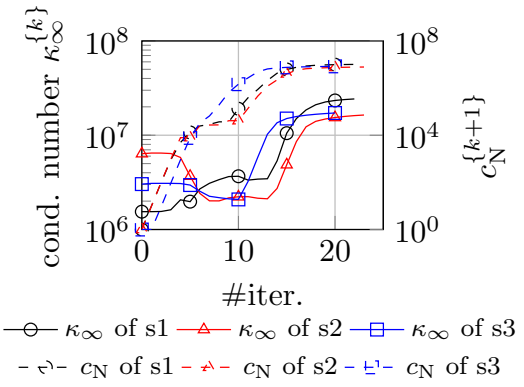
During these steps the condition number of the tangential system matrix will be affected by a variety of very different factors, such as the mesh distortion, the increasing and decreasing Lagrange multiplier values, the chosen material parameters, the chosen slave and master side, and the magnitudes of the different derivatives to name only a few. But it is also influenced by the actual magnitude of the c_N value and, therefore, indirectly by the chosen $\beta^\theta_{c_\mathrm{N}}$ value. The reader is encouraged to revisit Figure 5.15a: A high $\beta^\theta_{c_\mathrm{N}}$ value may lead to a faster convergence but also to a higher upper bound of the regularization parameter and thus also a higher condition number at least in the end of the load step, i.e., in the asymptotic phase.

To get a feeling for the different influences on the actual evolution of the condition number, the κ_∞ value has been computed for each non-linear iteration along to the three mentioned load steps. In Figure 5.17a the first results are visualized for the case that the correction parameter $\beta^\theta_{c_\mathrm{N}}$ is set to 0.45, i.e. very low, leading to a high number of non-linear iterations. But there is no verifiable bad impact on the condition number. Actually, rather the opposite is the case: The condition number noticeably decreases for high c_N values at the end of each load step. This can be noticed in the rise between the last iteration of load step 1 and the initial iteration of load step 2, where the c_N value is reset to 1.0. The same holds true between the end of step 2 and the beginning of step 3. In a next attempt, the same example has been recomputed with an increased correction parameter, i.e. $\beta^\theta_{c_\mathrm{N}} = 0.9$, see Figure 5.17b. In this case the rising c_N values undoubtedly start to increase the condition number. Compared to the condition number of the unloaded initial state, the value is increased up to a factor of 15.7. On the other hand, the necessary number of non-linear iterations is drastically reduced in contrast to Figure 5.17a. Leading to the first conclusion: a high $\beta^\theta_{c_\mathrm{N}}$ parameter has a bad impact on the condition number, since it inevitably leads to high conditioning in the asymptotic phase.

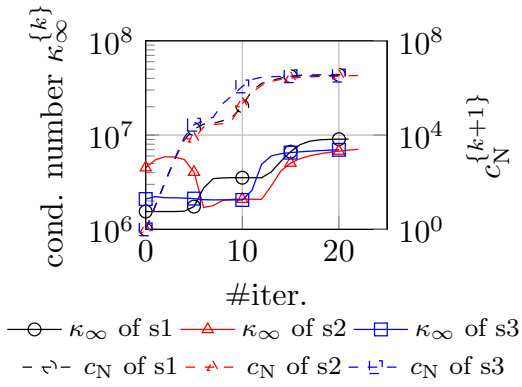
To obtain a more complete picture, it has been also tried to switch the slave and master side, such that the block becomes slave and the indentor master. The correction parameter $\beta^\theta_{c_\mathrm{N}}$ is kept constant at 0.9. The condition number still rises to the end of the load step, but interestingly the maximal reached values in each load step become smaller, see Figure 5.17c. Instead, of a factor of 15.7 a maximal increase by a factor of 5.86 is observable. That is remarkable since the maximal c_N value drops only slightly between the two cases. While in load step one a maximal value



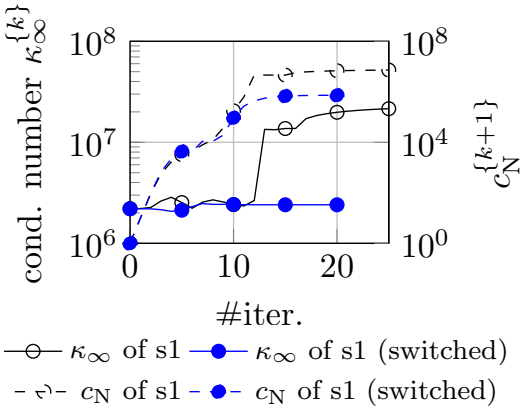
(a) E = 25,000, $\beta_\theta^{c_N} = 0.45$



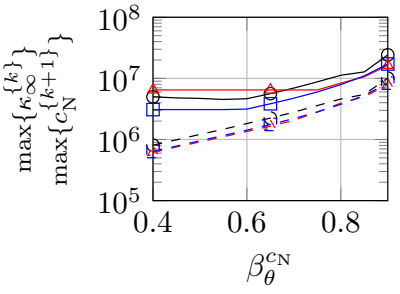
(b) E = 25,000, $\beta_\theta^{c_N} = 0.9$



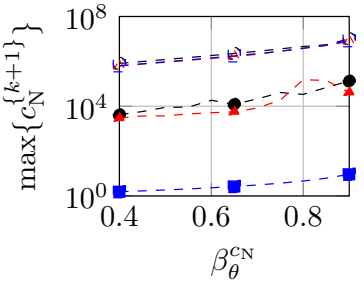
(c) E = 25,000, $\beta_\theta^{c_N} = 0.9$ (switched)



(d) $E^{[1]} = 25{,}000$, $E^{[2]} = 2{,}500$, $\beta_\theta^{c_N} = 0.9$



(e) E = 25,000, κ_∞ vs c_N



(f) E = 25,000, switching strategy (s)

Figure 5.17.: Investigation of the conditioning for the semicircular indentor example. In Figures 5.17a to 5.17d the results for a variety of different parameters and setting are presented. While in Figure 5.17e the coupling between the correction parameter $\beta_{c_N}^{\theta}$, the maximal arising c_N value and the conditioning is shown. Finally, Figure 5.17f demonstrates the beneficial effect of the switching strategy on the maximal arising regularization parameter values in each load step.

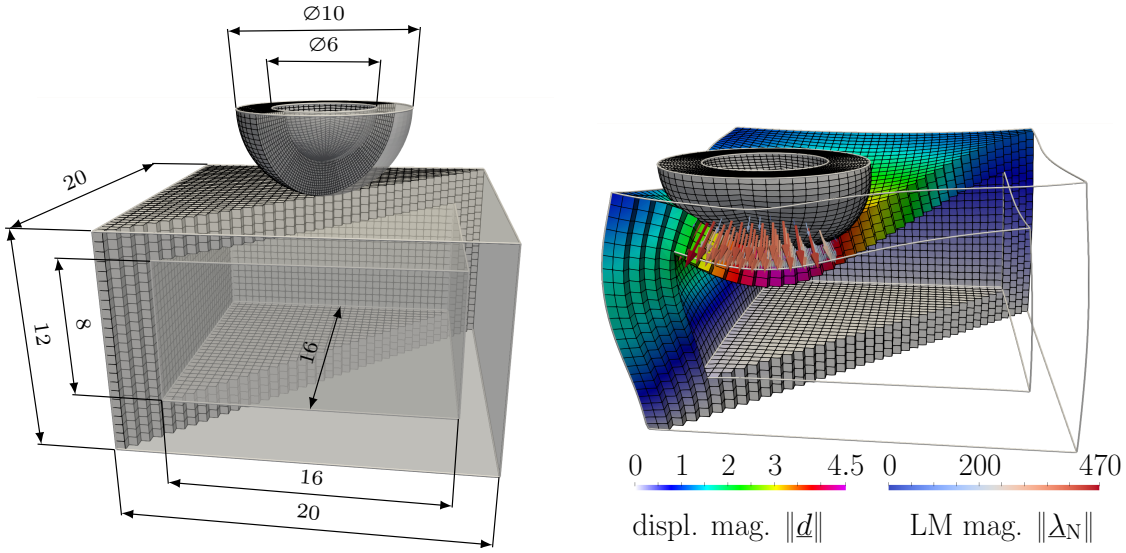
of $1.01\text{E}{+7}$ has been reached for the case shown in Figure 5.17b, the corresponding maximal value in Figure 5.17c is $9.06\text{E}{+6}$, i.e., only around 10% lower.

All of the so far discussed examples considered the same Young's modulus for both bodies. To demonstrate the impact of this choice, the simulation is repeated once more with almost the set-up used in Section 5.6.1.2. Only the actual penetration has been reduced from 5.0 to 4.0 units. As in the mentioned section, the slave and master side are interchangeable. The results are shown in Figure 5.17d, again with a $\beta^{\theta}_{c_{\text{N}}}$ equal to 0.9. The switch of the slave and master surfaces causes a severely different non-linear solver behavior for this case. Not only the condition number and the maximal c_{N} value drop but also the non-linear iteration numbers. The reason might be the very different intermediate smoothed slave normal fields.

In summary, it could be seen that in general a $\beta^{\theta}_{c_{\text{N}}}$ value of 0.9 makes the conditioning worse, while a $\beta^{\theta}_{c_{\text{N}}}$ of 0.45 seems to be even slightly beneficial. So the question might raise: Where is the break-even point? How long can the correction parameter be increased without worsening the linear solver behavior? At least for the first discussed example, i.e., equal Young's modulus for both bodies and indentor as slave, the answer is given in Figure 5.17e. The maximally reached condition number starts to rise at a $\beta^{\theta}_{c_{\text{N}}}$ value around 0.6. But, unfortunately, that is an intolerable low value since it would lead to high non-linear iteration numbers between 43 and 48 iterations per load step. Thus, the decrease of the $\beta^{\theta}_{c_{\text{N}}}$ value is not really a meaningful option. Instead, the already discussed possibility to switch to a different system of equations during the (modified) Newton scheme can be used to obtain a system of equations which is less influenced by a high c_{N} value. This leads to an even faster asymptotic convergence without the constraint that the regularization parameter must be increased any further. The drawback is the switch from a condensed system of equations to a saddle point system of equations. However, this is mainly a technical issue, since it demands for a different preconditioning for the linear system of equations. Furthermore, it might be also possible to switch to another condensed system, namely to the dual Lagrange multiplier formulation proposed in [215, 218, 279]. A very comprehensive discussion totally dedicated to this linear solver problematic can be found in [276].

This topic will be reconsidered in a moment. For now, the reader is kindly referred to Figure 5.17f where the beneficial effect of the switching strategy becomes very obvious. Here, the during each load step maximally reached c_{N} value of the switching strategy is compared with the values of the previously discussed condensed plain modified Newton method. As one can easily see, the maximally occurring c_{N} values are dramatically reduced. They stay even under the maximal value of the plain method corresponding to $\beta^{\theta}_{c_{\text{N}}}$ equal to 0.4 for the entire range of considered $\beta^{\theta}_{c_{\text{N}}}$ values. Thus, the given investigations have a clear result: Fast and robust convergence in combination with a low conditioning can be achieved by combining the modified Newton method and the consistently linearized system of equations.

Remark 5.8. Note that the condition numbers of the saddle point system of equations are not shown, since they would lead to an incorrect impression. In case of a saddle point system of equations, it makes no sense to consider the entire KKT system of equations to compute the condition number, since this would lead to very high values which do not really indicate much about the actual solvability of such systems. Instead, saddle-point systems asks for a special type of preconditioners which take advantage of the individual block structure (see e.g. Nocedal and Wright [204, Ch. 16] for more details).



(a) reference config.

(b) converged config. after first load step

Figure 5.18.: In Figure 5.18a the reference configuration of the sliding hemisphere on the hollow block including its dimensions and the mesh is presented. Figure 5.18b shows the deformed configuration for the first load step. The color bar white to pink represents the deformation of the hollow block, while the second color bar shows the magnitude of the Lagrange multiplier values living on the hollow block surface (slave).

5.6.7.2. Hollow Hemisphere on Hollow Block

So far the discussion has been restricted to observations and theoretical considerations, but it has not yet been proven that an iterative linear solver truly benefits from the presented switching strategy. For this purpose a more advanced test case presented in Figure 5.18 shall be investigated. Before the comprehensive examination can start, the setting is described in more detail: The actual geometrical configuration in the reference state can be seen in Figure 5.18a. Basically it is the 3-D version of the example considered in Section 5.6.3. The hollow block as well as the hemisphere have a Young's modulus equal to $2{,}500$ and a Poisson's ratio equal to 0.25. The example consists of 46,712 hexhedral linear finite elements with around $160{,}000$ primal degrees of freedom. No anti-locking strategy is applied. The considered load steps are:

1. The hemispherical indentor is moved equally by 5.0 units in the negative x-, y- and z-direction.
2. Subsequently, the penetration in z-direction is relaxed from 5.0 to 4.0.
3. The penetration is kept constant and sliding by 1.5 units in the positive x- and y-directions is initiated.
4. Finally, the penetration is again slightly relaxed to 3.5 units and the hemispherical indentor is moved to its final xy-position, viz. x equal to -2.5 and y equal to -3.5.

Thus, the last step combines sliding in x-direction with a simultaneous relaxation of the penetration. The final deformation state of the very first load step can be seen in Figure 5.18b, where besides the displacement field of the hollow block also the Lagrange multiplier field living on the block's slave surface are visualized.

Formulation as Condensed System

Since the presented method is currently based on the variationally incomplete approach from [131], the arising linear systems of equations become non-symmetric as soon as active contact contributions must be considered. Therefore, a suitable iterative solver, which can handle these non-symmetric systems, is chosen. In this example the *generalized minimal residual* (GMRES) algorithm is applied. However, in most cases the GMRES algorithm cannot directly be applied, since the bad conditioned system matrix would demand for a high iteration number and a high dimensional Krylov subspace [153, 232, 275]. Instead, a suitable preconditioner must be chosen. However, the correct choice of such a preconditioner is a research topic on its own and can become quickly very challenging.

In this work so-called algebraic multigrid preconditioners shall be considered. The foundation for the used linear solvers are the corresponding Trilinos packages [106, 145, 220]. Furthermore, the necessary adaptions for the algorithms to become usable for contact problems are entirely based on the work of [275–277]. Referring to the performance tips in [220], which state that *"it can be very challenging to find an appropriate set of multigrid parameters for a specific problem"*, the following shown set of parameters is by far not optimal and should be rather understood as a first attempt to solve the arising linear systems in an efficient way. Additionally, the linear system used here does not coincide with the system matrix arising from [215, 218, 279], nevertheless, the used linear solving strategies were specifically designed to fit these systems. Thus, not all developed methods could be successfully adapted in this work. Instead, the adaption is part of future work. First and foremost, the objective of this example is to demonstrate a beneficial effect of the described switching strategy.

The basic parameter set for the different preconditioners can be found in Table 5.2. In all cases the incomplete LU-factorization is used as a smoother for the fine and mediate levels, while a direct solver is applied to the coarsest level, such as KLU [63] or UMFPACK [62].

The condensed system during the pre-asymptotic phase is solved with a special implementation which is based on similar ideas as described in [282]. Details to the used permutations and much more can be found in [276]. Finally, the saddle-point system of equations during the aymptotic phase is preconditioned by a so-called *cheap semi-implicit method for pressure linked equations* (CheapSIMPLE). The basic idea is described in [211]. During this algorithm the inverses of the matrix blocks on the diagonal must be computed and exactly this computation is replaced by applying a fixed number of smoothing sweeps. The theory for this cheaper variant can be found for instance in [299]. All shown simulations have been performed on a dual socket Intel Xeon E5-2630 v3 node (with 2×8 cores) and 64 GByte RAM.

The correction parameter $\beta^{\theta}_{c_\mathrm{N}}$ is set to 0.8 for the following simulations and, additionally, only the condensed modified system of equations is considered in a first run. As expected: the behavior is very similar to the smaller 2-D examples of Section 5.6.7.1. In Figure 5.19 the GMRES iteration number stays quite constant at the beginning of each load step, till all of a sudden the necessary iterations heavily rise. In this example the increase is roughly by a factor of 2. However, with the gained knowledge of the previous condition number analysis the reason can be clearly traced back to increasing c_N values which are also shown in Figure 5.19 (see the dashed lines). Again, the effect on the conditioning is only detectable as soon as the c_N value passes some threshold.

	pure structure	cond. system (mN)	SP system (std)
SOLVER	GMRES		
solve tolerance	1.0E−6	1.0E−8	1.0E−9
convergence test	relative to initial residual		
max. Krylov subspace size	200 (no restart necessary)		
PRECONDITIONER	ML	MueLu (contact)	CheapSIMPLE/ML
fine/med. level smoother	incomplete LU factorization (ILU)		
fine/med. level damping	1.0	0.7	
coarse level smoother	Umfpack	KLU	
max. coarse level size	15,000		10,000
permutation	no	yes	no
aggregation type	uncoupled (UC)		

Table 5.2.: Linear solver parameters for the hollow block on hollow hemisphere example.

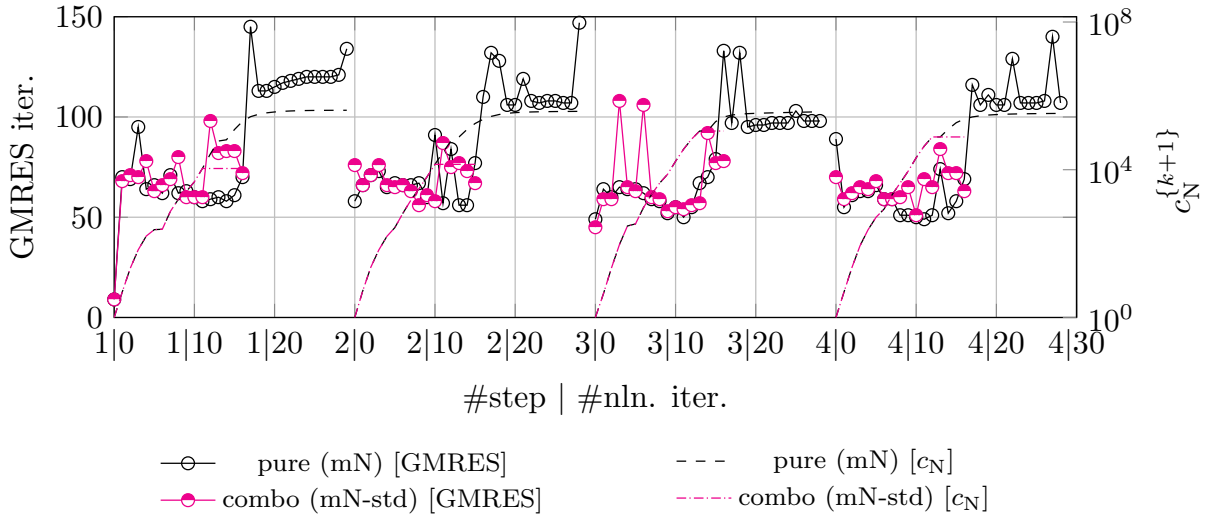


Figure 5.19.: Necessary GMRES iterations in each non-linear iteration of the four considered load steps. Comparison between the pure modified Newton approach (mN) and the switching approach (mN-std). Besides the GMRES iterations, also the rising c_N values are shown by dashed and dotted dashed lines.

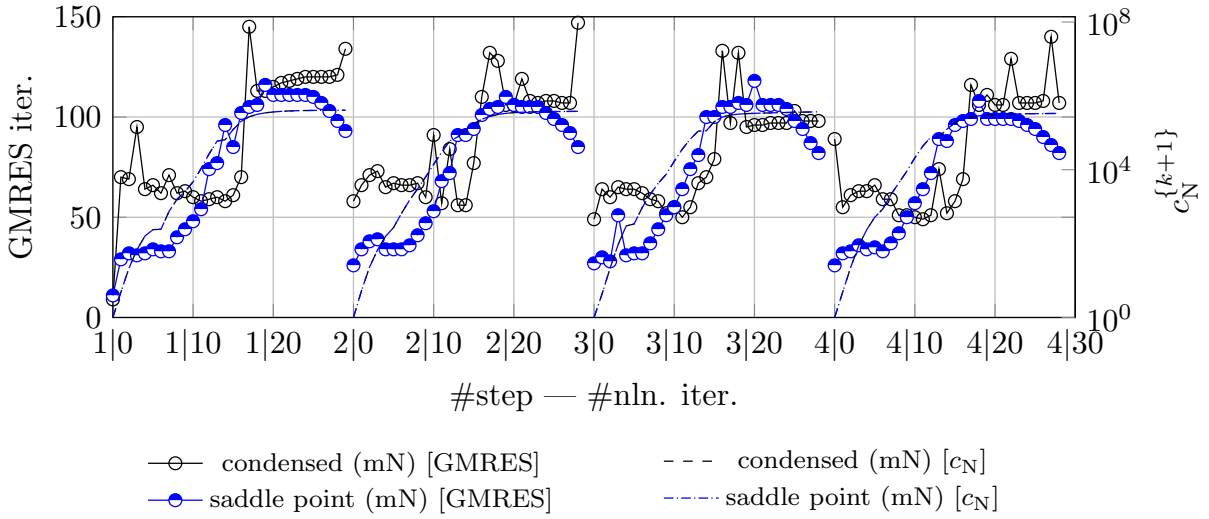


Figure 5.20.: Necessary GMRES iterations in each non-linear iteration of the four considered load steps. Comparison between the pure condensed modified Newton approach and pure saddle point modified Newton approach. Besides the GMRES iterations, also the rising c_N values are shown by dashed and dotted dashed lines.

In a second run, the switching strategy is applied. See Figure 5.19 for an overview. As expected the dramatic increase of the linear solver iterations can be prevented, even though the necessary solver switch causes a slight increase in iteration numbers. This effect can surely be resolved by a different set of linear solver parameters. Furthermore, Figure 5.19 shows especially in load step 2 some heavy outliers, which cannot totally be explained. Actually, the iteration numbers should coincide with the previous simulation in this pre-asymptotic phase. As one can see the c_N values of the SIR update are identical which underlines this point, but the probably slightly different parallel distribution seems to cause some unpredictable effects. However, the main message stays valid: The presented switching strategy combines the new modified Newton method and the consistently linearized approach from [131] in a meaningful way, such that the non-linear solver robustness could be successfully increased while the solvability of the arising linear systems of equations stays well controllable.

Formulation as Saddle Point System

A second possibility to resolve the bad conditioning has been already mentioned in Section 3.2.3.2 and Section 5.2.1: All of the previous investigations were based on the condensed system of equations. From a mathematical point of view the derived modified saddle point formulation, which is defined in (5.1a) and (5.2), and the condensed formulation coincide. However, for numerical iterative solution methods which deal with rounding errors and heavily rely on a low condition number to become competitive, this might be no longer true. Furthermore, the original modified saddle point system shows this nice and clear transition from the modified linear system of equations to the default system of equations for $c_N \to \infty$. This is an observation which has also been made by Gould [113] and led to the reformulation of the penalty system of equations as given in (3.59). Another advantage of dealing rather with a saddle point system instead of the condensed system of equations is the simple fact that there is already a well-suited method at hand to solve these systems, namely the mentioned CheapSIMPLE method, see also Table 5.2.

The obtained results are presented in Figure 5.20. A rough comparison with Figure 5.19 reveals that the saddle-point formulation is working better at the beginning and the end, i.e., for small and very high values of c_N. The first lower bound seems to lie somewhere near $c_\mathrm{N}^{\{k\}} = 1.0\mathrm{e}{+4}$ for this specific example. The second bound can be found near the actual solution, i.e., in the asymptotic regime when the $\{c_\mathrm{N}^{\{k\}}\}$ sequence has already reached its upper limit. Then, the GMRES iterations start to drop again in the last few Newton iterations. Especially, this drop at the end is a promising indicator that the linear solver performance can probably be improved by a better suited preconditioner. However, this has not been further investigated in this thesis.

5.6.8. Incomplete Versus Complete Variation

In the last example of this chapter the applicability of the modified approach to the variationally consistent and symmetric contact formulation as introduced in Chapter 4 shall be addressed. It must be firmly underlined that the entire derivation of the method holds also for the variationally consistent formulation and to some extent many of the used ideas become even more reliable, since there is no longer any difference between the operators $\nabla_{(\cdot)}(\cdot)$ and $\tilde{\nabla}_{(\cdot)}(\cdot)$. However, as already mentioned in Remark 5.2 there seem to be additionally cumbersome influences which are not revealed by the incomplete approach. For a comprehensive comparison of the two variational formulations the reader is referred to Section 4.4.

The example from Section 4.7.1 shall be reconsidered to demonstrate these differences. This simple example has the advantage that the need of discussing any further numerical integration artifacts is avoided. Precisely, exactly the same set-up is used, only the solving strategy is changed. Instead of the consistently linearized system, the modified system under consideration of the SIR correction scheme shall be applied. Furthermore, only the coarsest mesh as presented in Figure 4.2 is addressed.

First, the variationally incomplete and afterwards the complete formulation is applied to the problem. Each formulation is run in two configurations: Once with a $\beta_\Theta^{c_\mathrm{N}}$ value equal to 0.5 and once equal to 0.9. All simulations use an initial c_N value of 1.0. This stands in contrast to the results of Section 4.7.1 where c_N had been set to a constant value of 100 to avoid any cycling of the active set. The expected behavior is a stable but slow performance for the smaller value of $\beta_\Theta^{c_\mathrm{N}}$ and a fast but possible unstable performance for the higher value. This would be in accordance with the parameter study of Section 5.6.2. However, for this very simple example a stable result for both variants can be expected, since the initially set penetration of 1.0 is also solvable with the consistently linearized system (i.e. for the limit case $c_\mathrm{N} \to \infty$ in (5.2)). Thus far to the expectations.

Starting with the incomplete variational approach, all expectations are fulfilled. The algorithm takes 39 iterations for $\beta_\Theta^{c_\mathrm{N}} = 0.5$ and 17 iterations for $\beta_\Theta^{c_\mathrm{N}} = 0.9$. The final nodal displacements coincide with the results of Section 5.6.2 to at least 10 digits. Next, the attention is drawn to the variationally consistent formulation. If $\beta_\Theta^{c_\mathrm{N}}$ is set to 0.9 the expectations are satisfied once more and the solution is reached in only 15 iterations. Again, the result coincides to at least 10 digits with the displacement results of the consistently linearized approach. Also the sequence $\{c_\mathrm{N}^{\{k\}}\}$ stays bounded and at a first glance everything seems to line up with the expectations. However, if the $\beta_\Theta^{c_\mathrm{N}}$ value is reduced to 0.5 the solution procedure will start to diverge for the variationally consistent approach. The corresponding deformation states are shown in Figure 5.21a. It begins

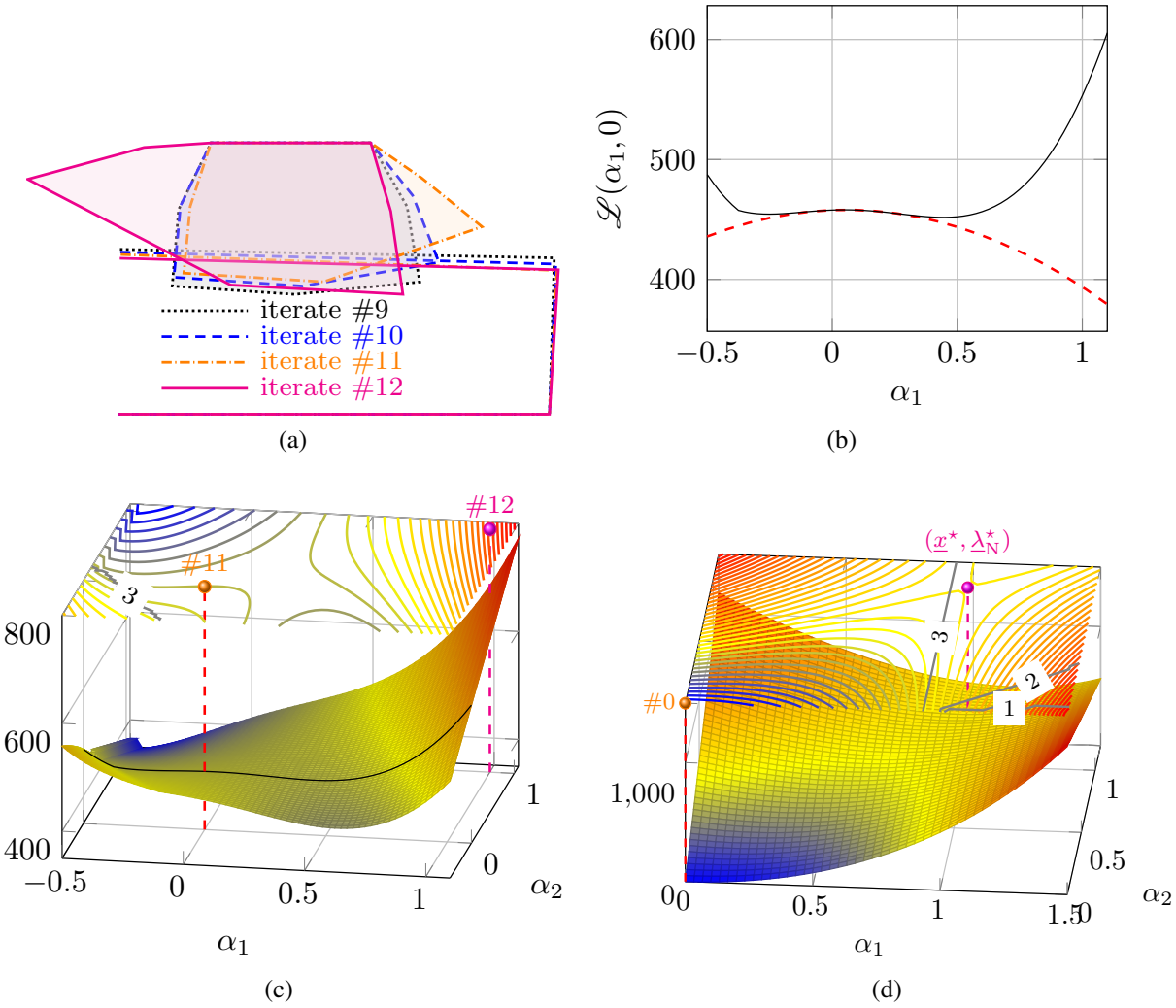


Figure 5.21.: In Figure 5.21a a selection of four deformation states for a variationally consistent simulation with $\beta_{\Theta}^{\mathrm{CN}} = 0.5$ are shown. In Figure 5.21b the related Lagrangian values are presented along the displacement search direction starting from iteration #11 together with the quadratic displacement model (see also the black line in Figure 5.21c). This reveals clearly the negative curvature around iteration #11. The surface plot of the Lagrangian under additional consideration of the varying Lagrange multiplier values between iteration #11 and #12 is shown in Figure 5.21c. Finally, Figure 5.21d shows a slice through the multidimensional Lagrangian function along the step from the initial predictor state to the final KKT-point which reveals nicely the saddle point structure.

in iteration $\#10$ where already a slight shift to the right side can be observed. In iteration $\#11$ the very right slave node slides extremely to the right, just to swing back in iteration $\#12$ to the other side. A similar behavior has been never apparent for the variationally inconsistent formulation. What is the cause for this strange solution path? To answer this question the attention is drawn to the Hessian of the Lagrangian with respect to the displacements, i.e., $\nabla^2_{\underline{d}\,\underline{d}}\mathscr{L}$. Exemplarily, the eigenvalues of $\nabla^2_{\underline{d}\,\underline{d}}\mathscr{L}^{\{11\}}$ are analyzed and it is revealed that the Hessian contains one negative eigenvalue, namely $\lambda^{\mathrm{H}}_{\min} \approx -37.84$. All remaining eigenvalues are positive, thus the matrix is indefinite. Notice that an indefinite Hessian is allowed in theory and is not a bad thing in first place. For instance, the Hessian might become negative because of the strongly positive curvature of the active weighted gap contributions. This might also explain the different behavior of the two variational approaches, since only the variationally consistent formulation contains the true second order derivatives of the active constraints.

The corresponding Lagrangian $\mathscr{L}(\alpha_1, \alpha_2) = \mathscr{L}(\underline{d}^{\{11\}} + \alpha_1 \Delta\underline{d}^{\{11\}}, \underline{\lambda}_{\mathrm{N}}^{\{11\}} + \alpha_2 \Delta\underline{\lambda}_{\mathrm{N}}^{\{11\}})$ for $\alpha_1, \alpha_2 \in [-0.5, 1.1]$ is shown in Figure 5.21c. Furthermore, Figure 5.21b demonstrates that the quadratic model fits well to the Lagrangian function, where the curvature in the current displacement direction is indeed negative.

Remark 5.9. The attentive reader might have noticed that there is a kink for $\alpha_1 < 0$ in the plotted Lagrangian functions presented in Figures 5.21b and 5.21c. The reason for this kink is a changing active set. A look at Figure 5.21a reveals that the gap of the very right slave node is positive, however, for $(\underline{d}^{\{11\}}, \underline{\lambda}_{\mathrm{N}}^{\{11\}})$ this node is still active due to the large Lagrange multiplier value. Nevertheless, if the gap is further increased while the Lagrange multiplier is decreased or is hardly increased, the node becomes inactive at some point. This is also marked by the gray line in the contour plot in Figure 5.21c. On the other hand, if the gap is further increased and the Lagrange multiplier is also sufficiently increased, the node will stay active. But, at some point the projections of the corresponding slave element will no longer hit the underlying master element. This phenomenon is also visible in Figure 5.21c: The unsteady part becomes obvious near the left edge by $\alpha_1 \to -0.5$.

However, it is pretty obvious that the computed step is not well suited: Instead of moving forward to the saddle point the step goes far beyond. The reason why the behavior of the modified approach gets suddenly so disastrous can be probably best explained under consideration of (5.8). The into the range-space of the constraint gradients mapped inverse Hessian matrix has also one negative eigenvalue and the applied regularization is not enough to remove it, since $c_{\mathrm{N}}^{\{11\}} = 6.1822\mathrm{E}{+}04$ is already too large. In fact a ten times smaller c_{N} value would be adequate to make the matrix on the left side of (5.8) sufficiently positive definite. However, if there is a negative definite matrix and another insufficient positive definite matrix is added, the resulting matrix is shifted closer to be singular, which makes things often even worse. Furthermore, it can be observed that the negative definite matrix in (5.8) results in Lagrange multiplier increments which point no longer in the opposite direction of the active constraints $\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}$ and this leads to a negative effect in (5.10) by increasing the displacement increments compared to the unmodified case instead of decreasing them (see also Section 5.2.2). The combination of all these effects seems to cause the undesirable behavior in Figure 5.21a. However, there is a way how this problem can be resolved by making the Hessian $\nabla^2_{\underline{d}\,\underline{d}}\mathscr{L}$ sufficiently positive definite, at least in the current search direction. A suitable algorithm for this task is given by Algorithm 6.3 and will

be presented in Section 6.6. By applying this algorithm (without considering any line-search), the method becomes stable again and the solution is reached in 38 iterations for the variationally consistent formulation and a $\beta_\Theta^{c_\mathrm{N}}$ equal to 0.5. Algorithm 6.3 can be also applied to the case $\beta_\Theta^{c_\mathrm{N}} = 0.9$, where the necessary iteration number rises slightly from 15 to 20. In contrast, if the additional regularization of the Hessian is applied to the incomplete approach, the results will not change. This underlines the initial hypothesis that the variationally inconsistent formulation is rather unaffected by this issue.

Remark 5.10. The discussed issue makes one wonder if the Hessian at the solution point $(\underline{d}^\star, \underline{\lambda}_\mathrm{N}^\star)$ is also indefinite and if so, what would be the consequences? A brief investigation reveals that the Hessian $\nabla^2_{\underline{d}\,\underline{d}}\mathscr{L}^\star$ stays indefinite and still contains one negative eigenvalue. However, two of the constraints are active at the solution $(\underline{d}^\star, \underline{\lambda}_\mathrm{N}^\star)$ and the associated gradients are linearly independent, i.e. $\mathrm{rank}(\nabla_{\underline{d}}\tilde{\underline{g}}_\mathrm{N}^{\mathcal{A}^\star}) = 2$. Consequently, the LICQ from Definition 3.1 holds. In addition, strict complementarity holds as well since $[\lambda_\mathrm{N}^\star]^i > 0 \; \forall i \in \mathcal{A}^\star$. Under these circumstances Theorem 3.5 can be reformulated as

$$\underline{\underline{Z}}^\mathrm{T} \nabla^2_{\underline{d}\,\underline{d}} \mathscr{L}^\star \underline{\underline{Z}} > 0, \tag{5.82}$$

i.e., the Hessian projected into the null-space of the active constraint gradients shall be positive definite, where $\underline{\underline{Z}} = \mathrm{null}([\nabla_{\underline{d}}\tilde{\underline{g}}_\mathrm{N}^{\mathcal{A}^\star}]^\mathrm{T})$ with $\underline{\underline{Z}} \in \mathbb{R}^{26\times 24}$ in this specific case. And, indeed, all eigenvalues of the projected Hessian are vastly positive, with the consequence that Theorem 3.5 is satisfied and the solution is an optimal point. For more information on (5.82) the reader is referred to Nocedal and Wright [204, Sec. 12.5, p. 337]. See also Figure 5.21d for a visualization of the Lagrangian around the solution point. At least in the presented 2-dimensional slice the saddle-point can be noticed very clearly.

5.7. Conclusion

In this chapter a new modification of the classical Newton's method has been proposed which can be used to solve non-linear contact problems under consideration of finite deformations. The presented algorithm has been designed on a strong mathematical foundation provided by the work of Bertsekas [23]. In this way, not only a significant improved non-linear solver behavior could be achieved, but also a reliable convergence analysis. Furthermore, the proposed algorithm uses a novel correction scheme and a sophisticated switching strategy which both help to successfully remove former drawbacks such as the dependency on an unknown regularization parameter c_N, the restricted (super-)linear convergence rate or the increasingly bad conditioning of the system matrix close to the solution. The provided profound theoretical and numerical analysis clearly indicate an advanced robustness and controllability.

However, there is still place for improvements: Currently, the proposed method has been only applied to quasi-static displacement controlled problems. While the treatment of dynamic problems will follow in the Sections 6.10.6 and 6.10.7, there is still the open question how to efficiently tackle contact problems considering Neumann loads. A key ingredient might be an improved choice of the initial regularization parameter $c_\mathrm{N}^{\{0\}}$. First attempts in this direction have shown promising results. Alternatively, it is possible to switch the modification off in case of

pure Neumann boundary conditions (see Section 6.10.2). Furthermore, the current method is in successive sliding load steps still unnecessarily bounded by the applied tangential predictor step. Therefore, a better performance in successive load steps seems achievable by improving the prediction method.

6. Line Search Filter Approach

The first two cornerstones on the way to a globally convergent contact algorithm have been presented in the Chapters 4 and 5 by constructing a robust, locally convergent computational contact method. The next important ingredient is the actual globalization procedure itself which will be considered now. Therefore, more sophisticated ideas from the numerical optimization literature shall be considered and thus help to improve the overall reliability of the proposed algorithms.

6.1. Motivation

The literature about numerical optimization methods offers a wide range of possibilities for incorporating a robust and globally convergent solution strategy into a numerical algorithm. For an overview the reader is kindly referred to Section 3.2 and the references therein. Here, a line search filter approach shall be considered. The used algorithm is closely related to the work of Wächter and Biegler [270, 271]. The filter method is based on ideas from multi-objective optimization, where, in the case of frictionless contact problems, the first objective is to minimize the structural energy, and the second objective is the fulfillment of the posed constraints with some emphasis on the latter one. By separating these goals instead of combining them into one merit-function (see Section 3.2.4), the algorithm becomes more flexible. The filter method will accept a trial point as soon as either the objective function could be decreased by a sufficient amount *or* the infeasibility measure has been sufficiently reduced. In this way the acceptability is less strict than the claimed reduction of a linear combination of both measures, such as it is the case for methods based on a merit function including some penalty parameter. A comparison of an interior-point line search filter algorithm and a penalty merit-function approach can be found, e.g., in Wächter and Biegler [272], whereas a comparison of a trust region and a filter trust region approach can be found in Gould and Toint [116], for instance. Unfortunately, this simple splitting idea is not enough to create a truly robust and globally convergent algorithm, such that additional ideas must be introduced from the constrained optimization literature and carried over to contact problems. It will be shown that the presented algorithms are able to overcome typical problems like cycling of the active set. In addition, the filter method provides a simple way to calculate minimal step-length estimates as long as there remain active constraints. If the step length drops below this barrier, it is very likely that the solution can not be improved in the current search direction. At this point a so-called *feasibility restoration* phase would come into play. Details on how to implement such a rescue strategy can be found, e.g., in Ulbrich et al. [263], Wächter and Biegler [272]. If the feasibility restoration phase should fail as well, such that the algorithm can not achieve the overall objective of finding a feasible local-minimum, the algorithm is supposed to stop at a local minimizer, or an at least stationary point with respect to some measure of infeasibility. In this case, there is a strong indication that the given problem seems at least locally

infeasible. It is to note here that more than that is hardly to obtain, since proving infeasibility is as difficult as finding a global minimizer and therefore beyond the capabilities of methods finding a local solution like those discussed here. Furthermore, it must be emphasized that the method presented here does *not* contain a feasibility restoration phase but a set of additional improvements which helps to avoid a stagnation of the line search filter method. In case of the studied computational contact problems, a drop below the minimal step length bound has always indicated a mistake at a different point in the algorithm, e.g., due to ill-posed boundary conditions or an insufficient mesh such that currently no feasibility restoration phase is needed. However, there might remain situations where such a feasibility restoration phase becomes necessary and, therefore, it should be considered in future work.

The remainder of this chapter is organized as follows: In Section 6.2 the basic idea of the filter method will be presented by introducing the underlying sufficient decrease criteria as well as by discussing a possible non-linear solution path visualized in the sub-space spanned by the two filter coordinates. Subsequently, the computation of the minimal step length estimates will be considered more deeply in Section 6.3. Afterwards, in Section 6.4, the attention is drawn to the Maratos effect (cf. Section 3.2.4) where one possible remedy will be studied in more detail: the *second order correction* (SOC) step. The SOC-step is constructed as an augmentation of the previously computed (insufficient) Newton step. The idea is that the combined step holds enough second order information to avoid the unnecessary step rejection. In Section 6.5 the steps of the used globalization algorithm will be presented. In the following sections some steps of the algorithm will be considered in more depth, starting with the applied correction of the linear system in Section 6.6. This correction is of major importance for a number of crucial scenarios such as the occurrence of structural or algorithmic instabilities. In this context, not only the used correction algorithm and the applied linear solver parameters are addressed, but the identification of invalid elements is also considered and a very general and reliable algorithm will be presented. In Section 6.7 further details of the globalization algorithm are addressed. Many of these smaller enhancements might seem unnecessarily complicated at a first glance, but without these small adaptions almost all of the results presented later would not have been achievable. Therein a number of necessary extensions are discussed such as a pre-testing approach to identify local mesh distortions which can not be detected by the global filter criteria. Another important point is the reinitialization of the filter in rare cumbersome situations or the scaling of the filter coordinates. In Section 6.8 an extension of the used filter coordinates to dynamic contact problems will be presented as well as the extension to EAS formulations. Then, before the numerical examples are addressed, some final practical remarks follow in Section 6.9. In Section 6.10 a number of challenging examples will be presented. Finally, a brief conclusion follows in Section 6.11.

6.2. Basic Idea of the Filter Method

The derivation of the line search filter method for finite deformation contact problems shall be begun by restating the problem in terms of an augmented Lagrangian formulation

$$\mathscr{L}_{c_\mathrm{N}}(\underline{x}, \underline{\lambda}_\mathrm{N}, \underline{\hat{s}}) = \mathscr{U}(\underline{x}) - \langle \underline{\lambda}_\mathrm{N}, \underline{\hat{g}}_\mathrm{N}(\underline{x}) - \underline{\hat{s}} \rangle_{\underline{\underline{A}}} + \frac{c_\mathrm{N}}{2} \| \underline{\hat{g}}_\mathrm{N}(\underline{x}) - \underline{\hat{s}} \|^2_{\underline{\underline{A}}}, \tag{6.1}$$

where the slack components $\hat{s}^i$ will be defined in Section 6.2.1. The consideration of (6.1) is commonly used as a technical step to define the coordinates of a filter-point $(\mathscr{L}, \Theta) \in \mathbb{R}^2$, where the first coordinate is the Lagrangian function value (2.59) as proposed by Ulbrich [264], Wächter and Biegler [271] and the second coordinate is an infeasibility measure $\Theta(\underline{x}, \underline{\hat{s}}) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ proportional to the last term in equation (6.1), which shall be defined as

$$\Theta(\underline{x}, \underline{\hat{s}}) = \|\underline{\hat{g}}_{\mathrm{N}}(\underline{x}) - \underline{\hat{s}}\|_{\underline{\underline{A}}^2} = \sqrt{\langle \underline{\tilde{g}}_{\mathrm{N}} - \underline{\underline{A}}\,\underline{\hat{s}}, \underline{\tilde{g}}_{\mathrm{N}} - \underline{\underline{A}}\,\underline{\hat{s}} \rangle}. \tag{6.2}$$

The choice of suitable slacks is again suspended to the next section. Alternatively, it is also possible to replace the first coordinate by the objective function value as exemplarily proposed by the original work of Fletcher et al. [98] or to define a linear model consisting of the objective function gradient and the complementarity condition to replace the first filter entry as proposed by Ulbrich et al. [263], even though this latter choice might have certain drawbacks since it would be based only on first order principles [246]. However, the usage of the Lagrangian can be favorable especially in combination with a Lagrange multiplier function (3.48), since the necessity for a *second order correction* (SOC) step might be circumvented. See Ulbrich [264] for a more detailed discussion on this more sophisticated avoidance strategy and Section 6.4 for the SOC approach considered here. Furthermore, since the definition of these coordinates is very flexible it is also possible to add more coordinates than just two. This approach has been suggested by Gould et al. [114], Gould and Toint [116] and is also followed in Milzarek and Ulbrich [196] for an unconstrained ℓ_1-regularized optimization problem. The definition of such additional filter point coordinates might become important if friction is considered in the future and, therefore, has been already kept in mind during the implementation of the framework.

6.2.1. Sufficient Reduction Criteria

In general, each of all the different globalization methods uses some kind of quality measure to assess a calculated trial step. This measure must fulfill some prerequisites, see Sections 3.1.2 and 3.2.4, which ensure a sufficient progress of an accepted trial step towards the solution. Typical examples are the comparison of the decrease of a model function to the actual achieved reduction as in trust region methods (see Conn et al. [53] or Section 3.1.2.2 for an overview), or the sufficient reduction of a merit function by enforcing the well-known (strong) Wolfe conditions (3.19a) or just the Armijo-rule (3.16). The latter one in combination with ideas originally proposed by Fletcher and Leyffer [96], Fletcher et al. [98] will be used here. A new trial point is going to be accepted, if it accomplishes either a sufficient decrease in the objective function value, i.e., the Lagrangian function defined in (2.59) and (4.10) , or if it reduces the infeasibility measure (6.2) by at least a factor proportional to the last accepted infeasibility measure value. The scaling factors for the infeasibility measure are denoted by $\gamma_f, \gamma_\Theta \in (0, 1/\sqrt{2})$ with respect to the two filter point coordinates, where the reason for the stated upper bound will be given in Section 6.7.6. In accordance with Wächter and Biegler [271], the sufficient reduction criteria are defined as

$$\Theta(\underline{x}(\alpha^{\{k,l\}}), \underline{\hat{s}}(\alpha^{\{k,l\}})) \leqslant (1 - \gamma_\Theta)\,\Theta(\underline{x}^{\{k\}}, \underline{\hat{s}}^{\{k\}}), \tag{6.3a}$$

or

$$\mathscr{L}(\underline{x}(\alpha^{\{k,l\}}), \underline{\lambda}_{\mathrm{N}}(\alpha^{\{k,l\}}), \underline{\hat{s}}(\alpha^{\{k,l\}})) \leqslant \mathscr{L}(\underline{x}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \underline{\hat{s}}^{\{k\}}) - \gamma_f \Theta(\underline{x}^{\{k\}}, \underline{\hat{s}}^{\{k\}}), \tag{6.3b}$$

where $(\underline{x}(\alpha^{\{k,l\}}), \underline{\lambda}_{\mathrm{N}}(\alpha^{\{k,l\}}))$ denotes a trial point evaluated for a step length $\alpha^{\{k,l\}} > 0$. The used index $k \in \mathbb{N}^+$ indicates the current Newton iteration, while the index $l \in \mathbb{N}^+$ indicates the current line search step during the Newton iteration k. As previously, all iteration superscripts are put in curly braces to avoid any confusion with other indices. In the following it is made use of the abbreviations

$$\begin{aligned} \Theta(\alpha^{\{k,l\}}) &:= \Theta(\underline{x}(\alpha^{\{k,l\}}), \underline{\hat{s}}(\alpha^{\{k,l\}})), \\ \mathscr{L}(\alpha^{\{k,l\}}) &:= \mathscr{L}(\underline{x}(\alpha^{\{k,l\}}), \underline{\lambda}_{\mathrm{N}}(\alpha^{\{k,l\}}), \underline{\hat{s}}(\alpha^{\{k,l\}})), \\ \mathscr{L}^{\{k\}} &:= \mathscr{L}(\underline{x}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \underline{\hat{s}}^{\{k\}}), \\ \text{and} \qquad \Theta^{\{k\}} &:= \Theta(\underline{x}^{\{k\}}, \underline{\hat{s}}^{\{k\}}), \end{aligned}$$

respectively. An alternative definition of the sufficient reduction criteria (6.3) which is also suitable for a set of (unconstrained) non-linear equations can be found in Gould and Toint [116].

In Wächter and Biegler [271] a second criterion is defined which is necessary since the defined acceptability checks (6.3) would allow a sequence of iterates $\{\underline{x}^{\{k\}}\}$ to be accepted only by reducing the infeasibility measure and never the objective function value. In the worst case such a behavior could lead to a convergence to a feasible but non-optimal solution point. To avoid such a scenario the following *$\mathscr{L}$-type switching condition* is proposed

$$m_{\mathscr{L}}(\alpha^{\{k,l\}}) < 0 \qquad \text{and} \qquad (-m_{\mathscr{L}}(\alpha^{\{k,l\}}))^{s_f} (\alpha^{\{k,l\}})^{1-s_f} > \nu_\Theta (\Theta^{\{k\}})^{s_\Theta}, \tag{6.4}$$

where $s_f \geqslant 1$, $s_\Theta > 1$ and $\nu_\Theta > 0$ are fixed constants. In equation (6.4) a model equation $m_{\mathscr{L}}(\alpha) : \mathbb{R} \to \mathbb{R}$ for the Lagrangian objective function is introduced which asks for further specification. This will follow in a moment. Whenever the switching condition (6.4) is fulfilled, the Armijo rule (3.16) is enforced

$$\mathscr{L}(\alpha^{\{k,l\}}) \leqslant \mathscr{L}^{\{k\}} + \eta_f m_{\mathscr{L}}(\alpha^{\{k,l\}}), \tag{6.5}$$

rather than the sufficient reduction criteria in (6.3). As soon as another trial step $l \in \{1, 2, \dots\}$ violates the switching condition, the algorithm switches back to the sufficient reduction criteria (6.3). The fixed scaling parameter $\eta_f \in (0,\ 0.5)$ in (6.5) is a well-known constant for the Armijo rule which has been already introduced in Section 3.1.2.

A detailed explanation, especially of the second part of the switching criterion (6.4) can be found in Wächter and Biegler [271]. The idea is that the switching criterion is only activated if the reduction of the objective function is not arbitrarily small compared to the current infeasiblity measure. Another important ingredient are the introduced fixed constants s_f and s_Θ. In the work by Wächter and Biegler [270], concerning the local convergence behavior, the authors propose that the relation $s_f > 2s_\Theta$ should hold. If this condition is satisfied, the switching condition (6.4)

close to a local solution will only become true, if a full (or second order corrected) step satisfies the Armijo rule. In accordance to the literature, a step $\alpha^{\{k,l\}}$ is called a $\mathscr{L}$-type step if it fulfills the $\mathscr{L}$-type switching condition and consequently a decrease solely due to the objective function value is demanded.

Now, the stated model equation $m_{\mathscr{L}}(\alpha)$ for the Lagrangian is addressed. In this thesis the Lagrangian function is modeled by a saddle-point model which is obtained by

$$\begin{aligned} m_{\mathscr{L}}(\alpha) =& \alpha \, \mathrm{D}\mathscr{L}((\underline{x}, \underline{\lambda}_{\mathrm{N}}, \underline{\hat{s}}); (\Delta\underline{d}, \Delta\underline{\lambda}_{\mathrm{N}}, \Delta\underline{\hat{s}})) \\ &+ \alpha^2 \, \Delta(\mathrm{D}\mathscr{L}((\underline{x}, \underline{\lambda}_{\mathrm{N}}, \underline{\hat{s}}); (\Delta\underline{d}, \Delta\underline{\lambda}_{\mathrm{N}}, \Delta\underline{\hat{s}})); (\Delta\underline{\lambda}_{\mathrm{N}})) \\ =& \alpha \, \{\langle \nabla_{\underline{x}} \mathscr{U}, \Delta\underline{d}\rangle - \langle \nabla_{\underline{x}} \underline{\tilde{g}}_{\mathrm{N}} \underline{\lambda}_{\mathrm{N}}, \Delta\underline{d}\rangle - \langle \underline{\hat{g}}_{\mathrm{N}} - \underline{\hat{s}}, \Delta\underline{\lambda}_{\mathrm{N}}\rangle_{\underline{\underline{A}}} + \langle \underline{\lambda}_{\mathrm{N}}, \Delta\underline{\hat{s}}\rangle_{\underline{\underline{A}}}\} \\ &- \alpha^2 \, \{\langle \Delta\underline{d}, \langle (\nabla_{\underline{x}} \underline{\tilde{g}}_{\mathrm{N}})^T, \Delta\underline{\lambda}_{\mathrm{N}}\rangle\rangle - \langle \Delta\underline{\hat{s}}, \Delta\underline{\lambda}_{\mathrm{N}}\rangle_{\underline{\underline{A}}}\}, \end{aligned} \tag{6.6}$$

where the dependency of the matrix $\underline{\underline{A}}$ on the displacements shall be ignored and the vector $\underline{\hat{s}} \in \mathbb{R}^m_+$ replaces the positive minimizer $\underline{s}^\star(\underline{x}, \underline{\lambda}_{\mathrm{N}})$, defined in equation (4.11), by the original set of independent, non-negative slack variables. This becomes only necessary during the line search and allows a smooth transition between inactive and active constraints as discussed in Gill et al. [107]. The search direction of the slack variables is obtained by

$$\begin{aligned} \underline{0} &= \{\underline{\underline{A}}(\underline{\hat{g}}_{\mathrm{N}} - \underline{\hat{s}})\}\Big|_k + \mathrm{D}((\underline{\underline{A}}(\underline{\hat{g}}_{\mathrm{N}} - \underline{\hat{s}})); (\Delta\underline{d}, \Delta\underline{\hat{s}}))\Big|_k \\ &= \left\{\underline{\tilde{g}}_{\mathrm{N}} + \nabla_{\underline{x}} \underline{\tilde{g}}_{\mathrm{N}} \Delta\underline{d} - \underline{\underline{A}}\,(\underline{\hat{s}} + \Delta\underline{\hat{s}})\right\}_k \end{aligned} \tag{6.7}$$

and finally

$$\Delta\underline{\hat{s}}^{\{k\}} = \left\{\underline{\hat{g}}_{\mathrm{N}} + \underline{\underline{A}}^{-1} \langle \nabla_{\underline{x}} \underline{\tilde{g}}_{\mathrm{N}}, \Delta\underline{d}\rangle - \underline{\hat{s}}\right\}_k . \tag{6.8}$$

This follows directly from the linear model of equation (3.42b) transferred to contact problems and evaluated at the last accepted iteration, where $(z^i)^2$ is substituted by $\hat{s}^i$. It can be seen in equation (6.7) that the term $\underline{\hat{s}} + \Delta\underline{\hat{s}}$ corresponds to the residual of the linear form for the weighted gap. One remaining parameter is the choice of the slack variable $\underline{\hat{s}}$ at the current iteration k. One possibility would again be the minimizer given in (4.11). The drawback of this choice would be the dependency of the function value on the non-physical regularization parameter c_{N}. Therefore, it is claimed that the objective function at the current iteration is independent of this parameter and is instead formulated only with respect to physically meaningful active contact contributions, viz.

$$\mathscr{L}^{\{k\}} = \mathscr{U}^{\{k\}} - \langle \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \underline{\tilde{g}}_{\mathrm{N}}^{\{k\}} - \underline{\underline{A}}^{\{k\}} \underline{\hat{s}}^{\{k\}} \rangle \overset{!}{=} \mathscr{U}^{\{k\}} - \langle \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}^{\{k\}}}, \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}^{\{k\}}} \rangle . \tag{6.9}$$

Solving for $\underline{\hat{s}}^{\{k\}}$ yields

$$
\begin{aligned}
\langle \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \tilde{\underline{g}}_{\mathrm{N}}^{\{k\}} \rangle - \langle \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \underline{\underline{A}}^{\{k\}} \bar{\underline{s}}^{\{k\}} \rangle &= \langle \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}^{\{k\}}}, \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}^{\{k\}}} \rangle, \\
\langle \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \underline{\underline{A}}^{\{k\}} \bar{\underline{s}}^{\{k\}} \rangle &= \langle \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}^{\{k\}}}, \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{I}^{\{k\}}} \rangle, \\
\Rightarrow \qquad \bar{\underline{s}}^{\{k\}} &= \begin{pmatrix} \underline{0} & \hat{\underline{g}}_{\mathrm{N}}^{\mathcal{I}^{\{k\}}} \end{pmatrix}^{\mathrm{T}},
\end{aligned} \tag{6.10}
$$

where it shall be assumed that the vector entries are sorted in such a way that all active contributions build up the first half and all inactive contributions the second half of the vector. Note that $\mathscr{L}(\underline{x}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \bar{\underline{s}}^{\{k\}}) \geqslant \mathscr{L}(\underline{x}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \underline{s}^{\star}(\underline{x}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}}))$ holds in general, since

$$
\mathscr{L}(\underline{x}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \bar{\underline{s}}^{\{k\}}) - \mathscr{L}(\underline{x}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}}, \underline{s}^{\star}(\underline{x}^{\{k\}}, \underline{\lambda}_{\mathrm{N}}^{\{k\}})) = \frac{1}{c_{\mathrm{N}}} \langle \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}^{\{k\}}}, \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}^{\{k\}}} \rangle_{\underline{\underline{A}}} \geqslant 0. \tag{6.11}
$$

This is a very important result, since exactly this relation is the key to avoid the acceptance of bad Newton steps. Otherwise, if the optimal slack variables $\hat{\underline{s}}^{\star}$ are used, it is possible that just an active set change causes a sufficient reduction of the Lagrangian function: A change from active to inactive can come along with a set of largely negative Lagrange multiplier values at the previously active nodes. These Lagrange multiplier values would cause an accordingly artificially high decrease of the Lagrangian function, completely dominating any increase due to, e.g., mesh distortion. The influence would be higher for smaller values of the regularization parameter c_{N} and therefore, the acceptance criterion would be unintentionally bound to the choice of this parameter. A similar observation suggests to use (6.10) for the infeasibility measure defined in (6.2) as well. In this case the use of $\hat{\underline{s}}^{\star}$ can cause a rejection of good Newton steps, since

$$
\Theta(\underline{x}, \hat{\underline{s}}^{\star}) \geqslant \Theta(\underline{x}, \bar{\underline{s}}) \quad \Leftrightarrow \quad \Theta^2(\underline{x}, \hat{\underline{s}}^{\star}) - \Theta^2(\underline{x}, \bar{\underline{s}}) = \frac{1}{c_{\mathrm{N}}^2} \langle \underline{\underline{A}}\, \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}}, \underline{\underline{A}}\, \underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} \rangle \geqslant 0, \tag{6.12}
$$

and, therefore, the inactive Lagrange multipliers would artificially increase the infeasibility measure value. Finally, it is to mention that the derived choice $\bar{\underline{s}}^{\{k\}}$ for the slack variables is during the pre-asymptotic phase no longer strictly positive since $\hat{\underline{g}}_{\mathrm{N}}^{i} > \frac{1}{c_{\mathrm{N}}} \lambda_{\mathrm{N}}^{i}$, $\forall i \in \mathcal{I}^{\{k\}}$ holds. Anyway, at the solution point, fulfilling (4.9), the positivity assumption is reobtained and the desired inequalities are enforced. The discussed choice of the slack variable is also in accordance with the literature, see e.g. Ulbrich [264].

In a next step, $\bar{\underline{s}}^{\{k\}}$ is inserted for $\hat{\underline{s}}^{\{k\}}$ into (6.8) and the equation is separated into its active and inactive part, yielding

$$
\Delta \bar{\underline{s}}^{\mathcal{A}} = \hat{\underline{g}}_{\mathrm{N}}^{\mathcal{A}} + (\underline{\underline{A}}^{\mathcal{A}})^{-1} \langle \nabla_{\underline{x}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, \Delta \underline{d} \rangle = 0, \tag{6.13a}
$$

$$
\Delta \bar{\underline{s}}^{\mathcal{I}} = (\underline{\underline{A}}^{\mathcal{I}})^{-1} \langle \nabla_{\underline{x}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{I}}, \Delta \underline{d} \rangle. \tag{6.13b}
$$

Note that the active part (6.13a) is equal to zero, due to the linearized demand in (3.54) which can be directly transferred to the contact problem. However, in case of a modified system such as the modified variant of Newton's method introduced in Chapter 5, this equation might no longer be satisfied. Next, (6.13) is inserted into the saddle point model (6.6) and it follows

$$m_{\mathscr{L}}(\alpha) = \alpha \left\{ \langle \nabla_{\underline{x}} \mathscr{W}, \Delta \underline{d} \rangle - \langle \nabla_{\underline{x}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}} \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}, \Delta \underline{d} \rangle - \langle \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}, \Delta \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \rangle \right\} - \alpha^2 \left\{ \langle \Delta \underline{d}, \nabla_{\underline{x}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}} \Delta \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \rangle \right\}. \tag{6.14}$$

An important observation is that there is no need to evaluate the gap or the corresponding gradient for any inactive constraint. This is crucial, since there are geometrical constellations where the evaluation of these quantities cannot be performed because of a missing feasible projection, or due to the chosen search radius of the given contact pair detection algorithm [288]. The latter one can not be chosen too large since otherwise a loss in efficiency would be the consequence. Thus, if the search radius is chosen in the correct manner, the gap evaluation of inactive contributions could easily fail.

6.2.2. Filter Definition

The proposed rule would still allow a sequence of iterations which fulfills criterion (6.3a) and (6.3b) on a rotating basis and, hence, could lead to an undesirable cycle of accepted iterates without any progress towards the solution. To avoid this, a taboo region in the half-plane $\{(\mathscr{L}, \Theta) \in \mathbb{R}^2 : \Theta > 0\}$ is proposed by Fletcher et al. [98]. This taboo region consists of a list of $(\mathscr{L}, \Theta)_p$ value pairs, called a filter-point. A new trial point is rejected by the filter as soon as both of its coordinates are larger than both coordinates of one of the previously added points in the list. More precisely, a trial point must fulfill at least one of the two conditions defined in (6.3) for each single filter point contained in the list to be accepted. According to Wächter and Biegler [271], the filter can also be defined as a set $\mathcal{F}^{\{k\}} \subseteq \mathbb{R} \times \mathbb{R}_+$ containing all prohibited coordinate pairs at iteration k. Consequently, a new trial point will be accepted by the filter, if the corresponding filter point is not part of the defined taboo region

$$\left(\mathscr{L}(\alpha^{\{k,l\}}), \Theta(\alpha^{\{k,l\}})\right) \notin \mathcal{F}^{\{k\}}. \tag{6.15}$$

There are different options how to initialize such a filter. For example, an empty filter $\mathcal{F}^{\{0\}} = \varnothing$ can be used, or a pre-defined upper-bound for the constraint violation $\mathcal{F}^{\{0\}} = \{(\mathscr{L}, \Theta) \in \mathbb{R}^2 \mid \Theta \geqslant \Theta_{\max}\}$ is possible, where $\Theta_{\max}$ is chosen such that $\Theta^{\{0\}} < \Theta_{\max}$ holds true. During the subsequent iterations the filter is for some iterates augmented by

$$\mathcal{F}^{\{k+1\}} := \mathcal{F}^{\{k\}} \cup \{(\mathscr{L}, \Theta) \in \mathbb{R}^2 \mid \mathscr{L} \geqslant \mathscr{L}^{\{k\}} - \gamma_f \Theta^{\{k\}} \text{ and } \Theta \geqslant (1 - \gamma_\Theta)\Theta^{\{k\}}\}. \tag{6.16}$$

If the filter is not supposed to be updated, the algorithm will stick to the previous one by setting $\mathcal{F}^{\{k+1\}} := \mathcal{F}^{\{k\}}$. This strategy provides that $\mathcal{F}^{\{k\}} \subseteq \mathcal{F}^{\{k+1\}}$ holds true for all k. One remaining open question is when to augment the filter? First, it must be considered that if the filter is augmented in each iteration, the algorithm might become too strict and too many "good" iterates will be rejected. Otherwise, if the algorithm is too loose, the problematic cycling can occur.

Fortunately, Wächter and Biegler propose also a meaningful rule for this case. First, feasible points are never added to the filter. Second, the filter will be augmented only for all non-$\mathscr{L}$-type iterates. If the switching condition (6.4) holds, the Armijo rule (6.5) must be satisfied and a

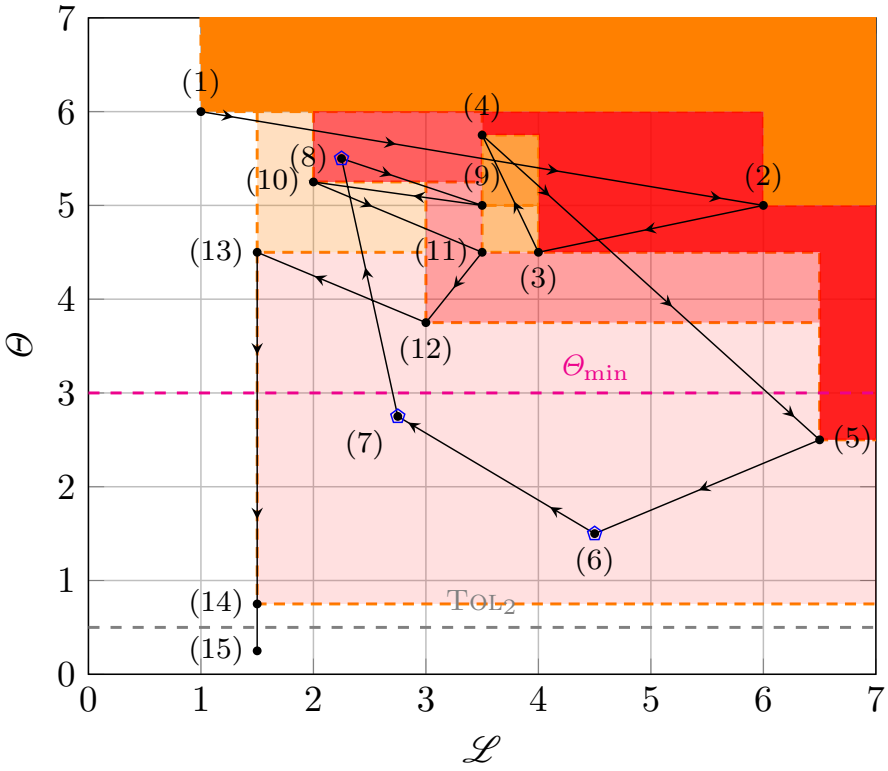


Figure 6.1.: Draft of a solution path in the filter sub-space spanned by the Lagrangian function value as first coordinate and the infeasibility measure as second coordinate.

monotone decreasing path with respect to the objective function values is followed. The reader is referred to Wächter and Biegler [271] for the proof that this augmentation strategy indeed avoids cycling.

Even though all of this sounds very intuitive, the filter draft in Figure 6.1 demonstrates that these simple rules can quickly lead to a quite complex solution path in the $(\mathscr{L}, \Theta)$-space. It is to emphasize that all these iterates would have been accepted by the filter without invoking any step length modification. This alone is already pretty impressive since most globalization methods demand some kind of monotonicity. With respect to the filter method there is also a taboo region which grows over the iterations as illustrated in Figure 6.1 by the growing red/orange area where a darker, less transparent color corresponds to an older part of the taboo region. However, the filter allows in each iteration a rise in the $\mathscr{L}$ or Θ value to a certain level. But, it does never allow a rise in both coordinates at once from one iteration to the next. This is prevented by the sufficient reduction conditions (6.3) or the Armijo rule (6.5) according to the result of the $\mathscr{L}$-type switching condition. Furthermore, it shall be assumed that all points marked by ⬠ represent $\mathscr{L}$-type steps, i.e., the filter set is not augmented. The explicit role of $\Theta_{\min}$ will be addressed in 6.7.2 and is of minor importance at the moment.

The sketch of a solution path presented in Figure 6.1 is now discussed in more detail, since a deep understanding of possible scenarios is essential for the later presented tweaks and adaptions of the classic filter method. In this example, the initial filter set is empty, i.e., $\mathcal{F}^{\{0\}} = \varnothing$. Now, the first trial point with coordinates $(1, 6)$ is accepted and added to the filter. The second trial point has the coordinates $(6, 5)$ and thus shows a rise in the objective function $\mathscr{L}$, but a decrease in

the infeasibility measure Θ what is enough to be accepted. Next, the trial point $(4, 4.5)$ follows, which shows a decrease in both coordinates compared to the previous iterate and is at least in one coordinates smaller than the first point. Consequently, it is not blocked by the filter and accepted by the sufficient decrease check. The same is true for the following iterations $4, 5, \ldots, 8$. All of them show a decrease in at least one coordinate compared to the previous iterate and none of them is blocked by the information stored in the filter. However, iteration 9 with the coordinates $(3.5, 5)$ is a special case: Firstly, in comparison to the previous iteration 8 it shows the demanded decrease in at least one coordinate. In this case the constraint violation Θ is getting smaller. But, if iteration 7 would not be marked as $\mathscr{L}$-type step, its coordinates $(2.75, 2.75)$ would be part of the filter set and would block iteration 9, since $2.75 < 3.5$ and $2.75 < 5$. Therefore, iteration 9 is only accepted by the filter due to the used $\mathscr{L}$-type switching strategy in conjunction with the augmentation strategy. The same holds true for the iterations 11 and 12. Both of them would be blocked by iteration 7. Afterwards the solution path shows a last slight rise in the constraint violation till it starts to drop to very small constraint violations and iteration 15 is already called feasible since its constraint violation coordinate is already below the specified tolerance TOL_2. Feasible points are not used to augment the filter set. Actually, this is a quite usual solution path which has been chosen to demonstrate that all ingredients of the filter method are important for the over-all performance. Fortunately, it is possible to find a set of parameters which works very well for a large variety of examples. This will be further addressed in Section 6.10.

6.3. Minimal Step Length Estimates

Even though the algorithm avoids the need for an increasing penalty parameter and defines a looser acceptability criterion than a classical line search algorithm based on a merit function [96], there is still the possibility that there is no admissible step-length in the current calculated search direction. In such a case an alternative fall-back strategy must be provided. But before any counteraction can be initialized, such a bad search direction must be reliably detected. This can be achieved by consideration of the defined acceptability and switching conditions. Each of them contains an inherent minimal step length estimate. It must be highlighted that the possibility to estimate a minimal step length is a big advantage of the proposed method since it avoids the definition of a heuristic stopping criterion such as a maximal allowed number of line search reductions, for example.

6.3.1. Sufficient Infeasibility Reduction

Beginning with (6.3a) an approximate value of the infeasibility measure (6.2) at the trial point can be obtained by a linear model

$$\Theta^{\{k\}} + m_\Theta(\alpha^{\{k,l\}}) \leqslant (1 - \gamma_\Theta)\Theta^{\{k\}}, \tag{6.17}$$

where the linear model of the infeasibility measure follows as

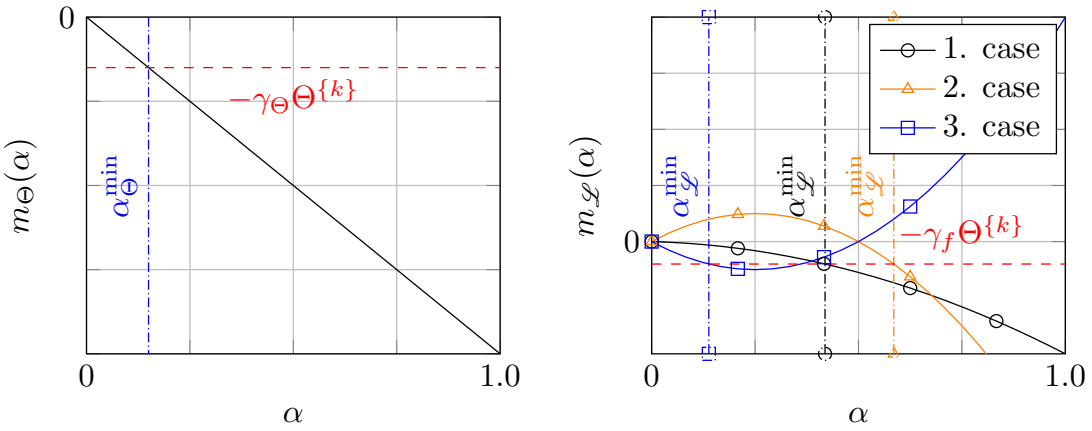


Figure 6.2.: Visualization of the minimal step length estimates for the sufficient decrease conditions. In Figure 6.2a the minimal step length estimate deduced from the linear infeasibility model is shown. In Figure 6.2b the three different cases for the quadratic Lagrangian model are illustrated.

$$\begin{aligned} m_\Theta(\alpha) =& \alpha \, \mathrm{D}(\Theta(\underline{x}, \hat{\underline{s}}); (\Delta\underline{x}, \Delta\hat{\underline{s}})) \\ =& \frac{\alpha}{\Theta} \left\{ \langle \tilde{\underline{g}}_{\mathrm{N}}, \langle \nabla_{\underline{x}} \tilde{\underline{g}}_{\mathrm{N}}, \Delta\underline{d} \rangle \rangle - \langle \hat{\underline{s}}, \langle \nabla_{\underline{x}} \tilde{\underline{g}}_{\mathrm{N}}, \Delta\underline{d} \rangle \rangle_{\underline{\underline{A}}} - \langle \tilde{\underline{g}}_{\mathrm{N}}, \Delta\hat{\underline{s}} \rangle_{\underline{\underline{A}}} + \langle \hat{\underline{s}}, \Delta\hat{\underline{s}} \rangle_{\underline{\underline{A}}^2} \right\}. \end{aligned} \tag{6.18}$$

By splitting (6.18) in an active and an inactive part and inserting (6.10), as well as (6.13), the following result is obtained

$$m_\Theta(\alpha) = \frac{\alpha}{\Theta} \left\{ \langle \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, \langle \nabla_{\underline{x}} \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}, \Delta\underline{d} \rangle \rangle \right\}, \tag{6.19}$$

see also the SIR correction scheme and (5.22). Under the prerequisite that $m'_\Theta(0) < 0$ holds, where the number of prime superscripts denotes the order of the considered derivative with respect to α, the equation (6.17) can be solved for the step length parameter, yielding

$$\alpha^{\{k,l\}} \geqslant \alpha_\Theta^{\min} = -\frac{\gamma_\Theta \, \Theta^{\{k\}}}{m'_\Theta(0)}. \tag{6.20}$$

Note that for the unmodified system of equations, (6.20) yields $\alpha_\Theta^{\min} = \gamma_\Theta$. A visualization of the stated relation is given in Figure 6.2a.

6.3.2. Sufficient Lagrangian Function Reduction

The second estimate is derived from equation (6.3b). Under consideration of the saddle-point model the mentioned criterion can be rewritten as

$$m'_{\mathscr{L}}(0) \, \alpha^{\{k,l\}} + \frac{1}{2} m''_{\mathscr{L}}(0) \, (\alpha^{\{k,l\}})^2 + \gamma_f \Theta^{\{k\}} \leqslant 0. \tag{6.21}$$

Again, a minimal step length estimate can only be obtained, if the current search direction is a descent direction for the considered criterion, i.e., $m_{\mathscr{L}}(\alpha) < 0$. In general there are three possible cases:

1. ***$\mathscr{L}$-reduction case:*** $m'_{\mathscr{L}}(0) < 0$ *and* $m''_{\mathscr{L}}(0) < 0$. The parabola opens downward. Since only descent directions are considered, the desired result is obtained by the positive root. There exists always a minimal step length estimate, but it is possible that it is larger than the default step length. The minimal step length estimate is obtained by

$$\alpha_{\mathscr{L}}^{\min} = \frac{-m'_{\mathscr{L}}(0) - \sqrt{(m'_{\mathscr{L}}(0))^2 - 2\, m''_{\mathscr{L}}(0)\gamma_f \Theta^{\{k\}}}}{m''_{\mathscr{L}}(0)}. \tag{6.22}$$

2. ***$\mathscr{L}$-reduction case:*** $m'_{\mathscr{L}}(0) > 0$ *and* $m''_{\mathscr{L}}(0) < 0$. In this case the linear model would predict no descent direction. However, the negative curvature may lead to a decrease of the objective function values as long as $m_{\mathscr{L}}(\alpha^{\{k,0\}}) < 0$ holds. In this case the relevant root is again defined by (6.22).

3. ***$\mathscr{L}$-reduction case:*** $m'_{\mathscr{L}}(0) < 0$ *and* $m''_{\mathscr{L}}(0) > 0$. In the last case the parabola defined in (6.21) opens upward. Here, the minimizer of the mixed second order 1-D model plays an important role. There exists no root, if the minimum value of the parabola is larger than $-\gamma_f \Theta^{\{k\}}$, i.e., there is a root only if

$$\frac{(m'_{\mathscr{L}}(0))^2}{2\, m''_{\mathscr{L}}(0)} \geqslant \gamma_f \Theta^{\{k\}}. \tag{6.23}$$

   Then two positive roots can be expected and the minimal step length estimate is obtained by the smaller one, which corresponds to (6.22). Note that there can be a reasonable root even if $m_{\mathscr{L}}(\alpha^{\{k,0\}}) > 0$ holds.

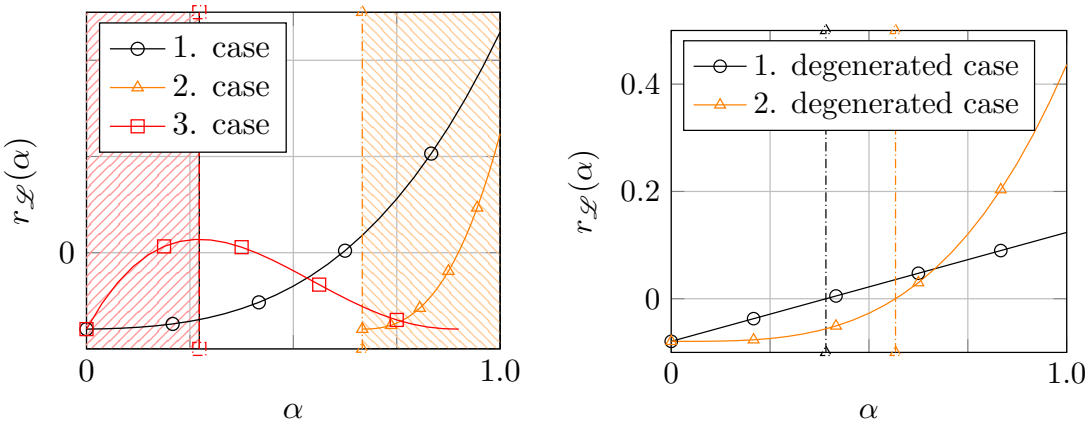
In summary the following results are obtained

$$\alpha_{\mathscr{L}}^{\min} = \min\{\alpha^{\{k,0\}},\ \frac{-m'_{\mathscr{L}}(0) - \sqrt{(m'_{\mathscr{L}}(0))^2 - 2\, m''_{\mathscr{L}}(0)\gamma_f \Theta^{\{k\}}}}{m''_{\mathscr{L}}(0)}\} \tag{6.24}$$

if either $m''_{\mathscr{L}}(0) \leqslant 0$ and $m_{\mathscr{L}}(\alpha^{\{k,0\}}) < 0$, or if $m''_{\mathscr{L}}(0) > 0$ and (6.23) holds. Otherwise there is no minimal step length estimate for the objective function criterion and the used linear model predicts that all trial steps will be rejected by criterion (6.3b). In this case $\alpha_{\mathscr{L}}^{\min}$ is set equal to the default step-length $\alpha^{\{k,0\}}$. See Figure 6.2b for a visualization of the different cases.

6.3.3. $\mathscr{L}$-type Condition

Finally, the last estimate is obtained under consideration of the $\mathscr{L}$-type switching condition (6.4). Since this function is in general non-linear due to the chosen exponential parameters s_f and s_Θ,



(a) (b)

Figure 6.3.: In Figure 6.3a the three general cases for the $\mathscr{L}$-type condition are visualized. While for the 1. case the desired estimate is initially searched in the entire domain $\alpha \in (0, 1]$, the cases 2 and 3 restrict the search domain to the related hatched regions. In Figure 6.3b the two meaningful degenerated cases together with their minimal step length estimates are presented. Note that $s_f = 2.3$ and $s_\Theta = 1.1$ have been chosen for all shown curves.

a local Newton approach shall be implemented to compute the desired estimate. Therefore, the local residual is defined as

$$r_{\mathscr{L}}(\alpha) = (-m'_{\mathscr{L}}(0) - \frac{\alpha}{2}\, m''_{\mathscr{L}}(0))^{s_f} \alpha - \nu_\Theta \Theta^{s_\Theta} \tag{6.25}$$

and its first derivative with respect to the step length parameter α as

$$r'_{\mathscr{L}}(\alpha) = (-m'_{\mathscr{L}}(0) - \frac{\alpha}{2}\, m''_{\mathscr{L}}(0))^{s_f}\, \frac{2m'_{\mathscr{L}}(0) + (1+s_f)\,\alpha\, m''_{\mathscr{L}}(0)}{2m'_{\mathscr{L}}(0) + \alpha m''_{\mathscr{L}}(0)}. \tag{6.26}$$

For now it shall be assumed that neither $m'_{\mathscr{L}}(0) = 0$, nor $m''_{\mathscr{L}}(0) = 0$ holds. These degenerated cases are addressed afterwards. Before the Newton scheme is entered, it must be guaranteed that a solution exists. Due to the non-linearity, there are three possible scenarios, which can all be treated in one algorithm under consideration of different lower and upper bounds.

1. ***$\mathscr{L}$-type case:*** $m'_{\mathscr{L}}(0) < 0$ *and* $m''_{\mathscr{L}}(0) < 0$. This is the easiest case. Here, $m_{\mathscr{L}}(\alpha) < 0$ holds for all $\alpha \in (0, 1]$ and, consequently, the first part of (6.4) is always fulfilled. The attention is drawn to the first derivative of the residual with respect to the step length parameter given in (6.26). This derivative has two possible roots

$$\hat{\alpha}^\star_{r,1} = -\frac{2m'_{\mathscr{L}}(0)}{(1+s_f)m''_{\mathscr{L}}(0)} \qquad \text{and} \qquad \hat{\alpha}^\star_{r,2} = -\frac{2m'_{\mathscr{L}}(0)}{m''_{\mathscr{L}}(0)}. \tag{6.27}$$

   Note that the second root denotes a limit case, since for a positive non-integer value of s_f the evaluation of (6.25) will fail on the left side of $\hat{\alpha}^\star_{r,2}$ and (6.26) will fail at $\hat{\alpha}^\star_{r,2}$ due to a division by zero. Actually, in the considered case $\hat{\alpha}^\star_{r,2} < \hat{\alpha}^\star_{r,1} < 0$ holds and thus both roots lie in the

negative half plane. Now, the evaluation at $\hat{\alpha}^{\star}_{r,1}$ of the second derivative of (6.25) with respect to α yields

$$r''_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,1}) = \frac{(1+s_f)^2\left(-\frac{m'_{\mathscr{L}}(0)s_f}{1+s_f}\right)^{s_f} m''_{\mathscr{L}}(0)}{2s_f m'_{\mathscr{L}}(0)} \tag{6.28}$$

and for the considered case $r''_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,1}) > 0$ holds, thus, $\hat{\alpha}^{\star}_{r,1}$ denotes a minimizer of (6.25) for the first case and the slope to the right side of $\hat{\alpha}^{\star}_{r,1}$ is always positive. Furthermore, as long as $\Theta \geqslant 0$ is satisfied, $r_{\mathscr{L}}(0) \leqslant 0$ must hold and it can be concluded: If $m'_{\mathscr{L}}(0) < 0$, $m''_{\mathscr{L}}(0) < 0$ and $r_{\mathscr{L}}(\alpha^{\{k,0\}}) > 0$ holds, the minimal step length estimate for the $\mathscr{L}$-type condition must lie between

$$\hat{\alpha}_{\text{low}} = 0 \qquad \text{and} \qquad \hat{\alpha}_{\text{up}} = \alpha^{\{k,0\}}. \tag{6.29}$$

These two values are then used as initial upper and lower bounds for the following Algorithm 6.1. Otherwise, if $r_{\mathscr{L}}(\alpha^{\{k,0\}}) \leqslant 0$ holds, the minimal step length estimate for the $\mathscr{L}$-type condition is set to the default step length, i.e. $\alpha_f^{\min} = \alpha^{\{k,0\}}$.

2. ***$\mathscr{L}$-type case:*** *$m'_{\mathscr{L}}(0) > 0$ and $m''_{\mathscr{L}}(0) < 0$.* Again, it shall be started under consideration of the first part of (6.4). The used model indicates a descent direction only if

$$\alpha > -\frac{2m'_{\mathscr{L}}(0)}{m''_{\mathscr{L}}(0)}. \tag{6.30}$$

This is equal to the second root $\hat{\alpha}^{\star}_{r,2}$ of (6.26) which is defined in (6.27). Furthermore, due to the changed assumptions, now $0 < \hat{\alpha}^{\star}_{r,1} < \hat{\alpha}^{\star}_{r,2}$ holds. Next, the sign of the slope of (6.25) right to $\hat{\alpha}^{\star}_{r,2}$ is obtained by

$$\lim_{\varepsilon\to 0^+} r'_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,2}+\varepsilon) = \lim_{\varepsilon\to 0^+} \frac{(-\frac{\varepsilon}{2}m''_{\mathscr{L}}(0))^{s_f}[-2s_f m'_{\mathscr{L}}(0) + \varepsilon(1+s_f)m''_{\mathscr{L}}(0)]}{\varepsilon m''_{\mathscr{L}}(0)} > 0. \tag{6.31}$$

Thus, it can be concluded: If $m'_{\mathscr{L}}(0) > 0$, $m''_{\mathscr{L}}(0) < 0$ and $r_{\mathscr{L}}(\alpha^{\{k,0\}}) > 0$ holds, the minimal step length estimate for the $\mathscr{L}$-type condition must lie between

$$\hat{\alpha}_{\text{low}} = \hat{\alpha}^{\star}_{r,2} = -\frac{2m'_{\mathscr{L}}(0)}{m''_{\mathscr{L}}(0)} \qquad \text{and} \qquad \hat{\alpha}_{\text{up}} = \alpha^{\{k,0\}}. \tag{6.32}$$

These values are used as initial upper and lower bounds in the 2. case. Again, if $r_{\mathscr{L}}(\alpha^{\{k,0\}}) \leqslant 0$ holds, the minimal step length estimate for the $\mathscr{L}$-type condition is set to the default step length, i.e. $\alpha_f^{\min} = \alpha^{\{k,0\}}$.

3. ***$\mathscr{L}$-type case:*** *$m'_{\mathscr{L}}(0) < 0$ and $m''_{\mathscr{L}}(0) > 0$.* This time $m_{\mathscr{L}}(\alpha) < 0$ holds if $\alpha \in (0, \hat{\alpha}^{\star}_{r,2})$ and, furthermore, $\hat{\alpha}^{\star}_{r,1} \in (0, \hat{\alpha}^{\star}_{r,2})$ holds as well. Now, with the additional demand that $r_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,1}) > 0$ is satisfied, it follows

$$\begin{aligned} r_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,1}) &= -\frac{2\left(-\frac{s_f m'_{\mathscr{L}}(0)}{1+s_f}\right) m'_{\mathscr{L}}(0)}{(1+s_f) m''_{\mathscr{L}}(0)} - \nu_{\Theta}\Theta^{s_{\Theta}} \\ &= -\frac{2[m'_{\mathscr{L}}(0)]^2}{(1+s_f)^3[m''_{\mathscr{L}}(0)]^2} r''_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,1}) - \nu_{\Theta}\Theta^{s_{\Theta}} > 0, \end{aligned} \tag{6.33}$$

where (6.28) has been inserted. This directly yields

$$r''_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,1}) < -\frac{(1+s_f)^3[m''_{\mathscr{L}}(0)]^2}{2[m'_{\mathscr{L}}(0)]^2}\nu_{\Theta}\Theta^{s_{\Theta}} < 0, \tag{6.34}$$

if $\Theta \neq 0$ and thus $\hat{\alpha}^{\star}_{r,1}$ denotes a maximizer in this case. Finally, it can be concluded: If $m'_{\mathscr{L}}(0) < 0$, $m''_{\mathscr{L}}(0) > 0$ and $r_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,1}) > 0$ holds, the minimal step length estimate for the $\mathscr{L}$-type condition must lie between

$$\hat{\alpha}_{\text{low}} = 0 \qquad \text{and} \qquad \hat{\alpha}_{\text{up}} = \hat{\alpha}^{\star}_{r,1} = -\frac{2m'_{\mathscr{L}}(0)}{(1+s_f)m''_{\mathscr{L}}(0)}. \tag{6.35}$$

These values are used as initial upper and lower bounds in the 3. case. If $r_{\mathscr{L}}(\hat{\alpha}^{\star}_{r,1}) \leqslant 0$ holds, the minimal step length estimate for the $\mathscr{L}$-type condition is set to the default step length, i.e. $\alpha_f^{\min} = \alpha^{\{k,0\}}$.

If one of these cases occurs and the related conditions for the existence of a solution are satisfied, the local Newton scheme must converge. If the conditions are not fulfilled the estimate is defined as the default step-length $\alpha_f^{\min} = \alpha^{\{k,0\}}$. The complete local safe-guarded Newton algorithm is provided in Algorithm 6.1. The stated safe-guarding strategy ensures that the local Newton scheme converges to a solution in the correct interval. The starting point is chosen as the secant approximation based on the lower and upper bound values. Note that the start value is always positive and part of the desired interval. See also Figure 6.3a for a representative visualization of the three discussed general cases. Note that due to the chosen exponents it is not possible to evaluate function (6.25) in the entire domain for the cases 2 and 3. However, the evaluation is completely uncritical inside the hatched regions. Furthermore, the used algorithm avoids a evaluation of (6.26) at the boundary positions such that these points are also uncritical for Algorithm 6.1.

As mentioned at the beginning this Newton approach is only necessary for the general case. If a degenerated case occurs, i.e., either $m'_{\mathscr{L}}(0)$ or $m''_{\mathscr{L}}(0)$ is equal to zero, the solution can be obtained without considering a Newton scheme:

1. ***Degenerated $\mathscr{L}$-type case:*** *$m'_{\mathscr{L}}(0) < 0$ and $m''_{\mathscr{L}}(0) = 0$.* Here, $m_{\mathscr{L}}(\alpha) < 0$ holds for all $\alpha \in (0, \alpha^{\{k,0\}}]$ and the estimate for the minimal step length follows as

Algorithm 6.1 MINIMAL STEP LENGTH FOR THE $\mathscr{L}$-TYPE CONDITION

Given. Newton stopping tolerance $\text{TOL} \in \mathbb{R}_+$.

0. *Initialize.* Set the iteration counter $j \leftarrow 0$; depending on the current $\mathscr{L}$-type case, set the lower $\hat{\alpha}_{\text{low}}$ and upper bound $\hat{\alpha}_{\text{up}}$ in accordance to the given definition in (6.29), (6.32) or (6.35); define a starting point $\hat{\alpha}^{\{j\}} = \frac{r_{\mathscr{L}}(\hat{\alpha}_{\text{low}})}{r_{\mathscr{L}}(\hat{\alpha}_{\text{low}}) - r_{\mathscr{L}}(\hat{\alpha}_{\text{up}})}$.
1. *Evaluation.* Evaluate the residual (6.25) and derivative (6.26) for the current iterate.
2. *Update bounds.* Update the lower and upper bound values. If $r_{\mathscr{L}}(\hat{\alpha}^{\{j\}}) < 0$ and $\hat{\alpha}^{\{j\}} > \hat{\alpha}_{\text{low}}$, set $\hat{\alpha}_{\text{low}} = \hat{\alpha}^{\{j\}}$, else if $r_{\mathscr{L}}(\hat{\alpha}^{\{j\}}) > 0$ and $\hat{\alpha}^{\{j\}} < \hat{\alpha}_{\text{up}}$, set $\hat{\alpha}_{\text{up}} = \hat{\alpha}^{\{j\}}$.
3. *Trial point.* Compute the Newton trial estimate

$$\bar{\hat{\alpha}}^{\{j\}} = \hat{\alpha}^{\{j\}} - \left.\frac{r_{\mathscr{L}}(\hat{\alpha})}{r'_{\mathscr{L}}(\hat{\alpha})}\right|_{\hat{\alpha}^{\{j\}}}. \tag{6.36}$$

4. *Safe-guarding.* Check the lower and upper bounds. If $\bar{\hat{\alpha}}^{\{j\}} < \hat{\alpha}_{\text{low}}$ or $\bar{\hat{\alpha}}^{\{j\}} > \hat{\alpha}_{\text{up}}$, set $\hat{\alpha}^{\{j+1\}} = 0.5\,(\hat{\alpha}_{\text{low}} + \hat{\alpha}_{\text{up}})$. Otherwise use the Newton iterate $\hat{\alpha}^{\{j+1\}} = \bar{\hat{\alpha}}^{\{j\}}$.
5. *Check convergence.* Increase the iteration counter $j \leftarrow j + 1$ and check the convergence criterion. If

$$|\hat{\alpha}^{\{j\}} - \hat{\alpha}^{\{j-1\}}| < \text{TOL} \cdot |\hat{\alpha}^{\{j\}}| \tag{6.37}$$

stop the local Newton scheme and set $\alpha_f^{\min} = \hat{\alpha}^{\{j\}}$, otherwise go to Step 1.

$$\alpha_f^{\min} = \frac{\nu_\Theta \Theta^{s_\Theta}}{(-m'_{\mathscr{L}}(0))^{s_f}}. \tag{6.38}$$

2. ***Degenerated $\mathscr{L}$-type case:*** *$m'_{\mathscr{L}}(0) = 0$ and $m''_{\mathscr{L}}(0) < 0$.* This case is more or less mainly of theoretical importance. However, $m_{\mathscr{L}}(\alpha) < 0$ still holds for all $\alpha \in (0, \alpha^{\{k,0\}}]$ and the estimate yields

$$\alpha_f^{\min} = \left[\frac{\nu_\Theta \Theta^{s_\Theta}}{(-\frac{1}{2}m''_{\mathscr{L}}(0))^{s_f}}\right]^{\frac{1}{1+s_f}}. \tag{6.39}$$

3. ***Degenerated $\mathscr{L}$-type case:*** *$m'_{\mathscr{L}}(0) > 0$ and $m''_{\mathscr{L}}(0) = 0$ or $m'_{\mathscr{L}}(0) = 0$ and $m''_{\mathscr{L}}(0) > 0$.* No solution can be obtained, and the minimal step length estimate is set to the default step $\alpha_f^{\min} = \alpha^{\{k,0\}}$.

For an example of the first two meaningful degenerated cases the reader is kindly referred to Figure 6.3b.

In summary, three different minimal trial step size estimates have been derived in the Sections 6.3.1 to 6.3.3. Thus, the final estimate is obtained by choosing the minimal value of all estimates

$$\alpha_{\min}^{\{k\}} = \gamma_\alpha \cdot \min\{\alpha_\Theta^{\min},\ \alpha_\mathscr{L}^{\min},\ \alpha_f^{\min}\}, \tag{6.40}$$

where $\gamma_\alpha \in (0, 1]$ is a safety-factor to account for the used assumptions and neglected higher order terms during the derivation of the estimates. Alternatively, it is possible to use even more sophisticated estimates and models. However, in contrast to Wächter and Biegler [271], where estimates for pure linear models can be found, here already more complex models are considered and a further improvement of the used models seems not necessary based on the gained experience. Furthermore, the generated effort is still acceptable, since only one dimensional functions are involved and the estimates are computed only once per Newton step. As soon as the trial step size is below $\alpha_{\min}^{\{k\}}$, a strong indication is given that there is no solution in the current search direction and either a feasiblity restoration phase must be activated or the simulation is stopped at a non-optimal point.

6.4. Second Order Correction

First, the reader is kindly referred to Section 3.2.4 and the therein provided discussion about the Maratos effect [186]. In the following one possible remedy for this problem shall be proposed which can avoid the unnecessary rejection of good steps. The so-called *second order correction* step, or short, *SOC* step. The idea is that the previously computed Newton step is augmented by an additional increment such that the new displacement iterate

$$\underline{d}^{\mathrm{SOC}} = \underline{d}^{\{k\}} + \Delta\underline{d}^{\{k\}} + \Delta\underline{d}^{\mathrm{SOC}} \tag{6.41}$$

can be easier accepted by the filter since it improves the entire step quality by reducing the error order. The main reason why the Maratos effect even occurs is given by an insufficient representation of the constraint curvatures in the applied linear model of the Newton or SQP framework, see e.g. (3.52b) or the active constraint rows in (4.18). This is revealed by a simple Taylor series expansion leading to

$$\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d} + \Delta\underline{d}) = \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}) + [\nabla_{\underline{d}}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d})]^{\mathrm{T}}\Delta\underline{d} + O(\|\Delta\underline{d}\|^2) \overset{(\dagger)}{=} O(\|\Delta\underline{d}\|^2), \tag{6.42}$$

where the iteration superscript k has been omitted and $(\dagger)$ holds under the assumption that the unmodified system of equations is solved. Furthermore, it shall be assumed for now that the active set did not change during the last Newton step, i.e. $\mathcal{A}(\underline{d}, \underline{\lambda}_{\mathrm{N}}) \equiv \mathcal{A}(\underline{d} + \Delta\underline{d}, \underline{\lambda}_{\mathrm{N}} + \Delta\underline{\lambda}_{\mathrm{N}})$ and that each constraint is at least twice continuously differentiable. In addition, it is to note that the entire SOC discussion is mainly a topic concerning local convergence and aims for avoidance of unnecessary rejections close to the solution by the filter, since in the pre-asymptotic phase the occurring errors are hardly quantifiable. Thus, in order to avoid this phenomenon close to the solution, this second order error must be reduced or *corrected*. For this task a large variety of different methods are available. See for example Conn et al. [53, Sec. 15.3.2.3] for a comprehensive overview. One quite intuitive option taken from this reference is to consider the following subproblem

$$\text{minimize} \qquad \frac{1}{2}\|\Delta \underline{d}^{\mathrm{SOC}}\|_{\underline{\underline{B}}}^2 \tag{6.43a}$$

$$\text{s. t.} \qquad \tilde{g}_{\mathrm{N}}^i(\underline{d}+\Delta\underline{d}) + [\Delta \underline{d}^{\mathrm{SOC}}]^{\mathrm{T}}\, \nabla_{\underline{d}} \tilde{g}_{\mathrm{N}}^i(\underline{d}+\Delta\underline{d}) = 0, \qquad \forall i \in \mathcal{A}. \tag{6.43b}$$

In words: The additional second order correction step shall reduce the active constraints evaluated at the trial iterate while minimizing the distance measured in a given matrix norm to this new auxiliary iterate. There are different possibilities how to choose the matrix $\underline{\underline{B}}$, for instance, if the identity is inserted the distance in (6.43a) is measured in the classical ℓ_2-norm. Under the assumption that the LICQ holds close to the solution it can be concluded that the solution of (6.43) yields

$$\begin{aligned}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}+\Delta\underline{d}+\Delta\underline{d}^{\mathrm{SOC}}) &= \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}+\Delta\underline{d}) + [\nabla_{\underline{d}}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}+\Delta\underline{d})]^{\mathrm{T}}\Delta\underline{d}^{\mathrm{SOC}} + O(\|\Delta\underline{d}^{\mathrm{SOC}}\|^2)\\ &= O(\|\Delta\underline{d}^{\mathrm{SOC}}\|^2)\end{aligned} \tag{6.44}$$

and, consequently, the constrained least-squares minimization implies that

$$\Delta\underline{d}^{\mathrm{SOC}} = O(\|\Delta\underline{d}\|^2) \qquad \Leftrightarrow \qquad \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}+\Delta\underline{d}+\Delta\underline{d}^{\mathrm{SOC}}) \overset{(6.44)}{=} O(\|\Delta\underline{d}\|^4) \tag{6.45}$$

holds. Therefore, with respect to (6.42) the desired improvement can be achieved. However, the shown optimization problem (6.43) asks for the evaluation of the weighted gap gradients at the trial point $\underline{d}+\Delta\underline{d}$ which can be pretty expensive, especially since there is always the possibility that the second order correction step might not lead to the desired reduction and must be rejected afterwards. Fortunately, the additional evaluation of the constraint gradients is not necessary, since (6.44) can be reformulated by applying the mean-value theorem to the gap gradients [53, 79] such that

$$\begin{aligned}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}+\Delta\underline{d}+\Delta\underline{d}^{\mathrm{SOC}}) =& \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}+\Delta\underline{d}) + [\nabla_{\underline{d}}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d})]^{\mathrm{T}}\Delta\underline{d}^{\mathrm{SOC}}\\ &+ O(\|\Delta\underline{d}^{\mathrm{SOC}}\|\max\{\|\Delta\underline{d}^{\mathrm{SOC}}\|, \|\Delta\underline{d}\|\}).\end{aligned} \tag{6.46}$$

Now, if the old gradient $\nabla_{\underline{d}}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d})$ is inserted into the least squares problem (6.43), the left relation in (6.45) still holds, only the right side must be adapted to

$$\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}+\Delta\underline{d}+\Delta\underline{d}^{\mathrm{SOC}}) \overset{(6.46)}{=} O(\|\Delta\underline{d}\|^3). \tag{6.47}$$

Thus, $\Delta\underline{d}^{\mathrm{SOC}}$ stays a valid second order correction step. The related system of equations which must be solved is given by

$$\begin{pmatrix} \underline{\underline{B}} & -\nabla_{\underline{x}}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}) \\ -[\nabla_{\underline{x}}\underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d})]^{\mathrm{T}} & \underline{\underline{0}} \end{pmatrix} \begin{pmatrix} \Delta\underline{d}^{\mathrm{SOC}} \\ \underline{\lambda}^{\mathrm{SOC}} \end{pmatrix} = \begin{pmatrix} \underline{0} \\ \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{d}+\Delta\underline{d}) \end{pmatrix}, \tag{6.48}$$

where $\underline{\lambda}^{\mathrm{SOC}}$ represents the Lagrange multiplier associated to the modified least squares problem (6.43). Thus, only the constraints at the new iterate $\underline{d} + \Delta\underline{d}$ must be evaluated. This happens anyway since they are needed for the acceptability check. Even though this would be a reliable option to compute a SOC step, the obtained system matrix in (6.48) is still quite different from the usual case and asks for a new matrix assembly. It would be favorable if the already assembled system matrix can be reused.

Coming back to the least squares approach in a moment, an alternative ansatz shall be presented now. Instead of searching for a correction of the already computed step based on a linear model for the constraints, it might be a better idea to replace the linear model simply by a quadratic model, thus, the new demand

$$\tilde{g}_{\mathrm{N}}^{i\{k\}} + [\nabla_{\underline{x}} \tilde{g}_{\mathrm{N}}^{i\{k\}}]^{\mathrm{T}} (\Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}}) + \frac{1}{2} (\Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}})^{\mathrm{T}} [\nabla^2_{\underline{d}\,\underline{d}} \tilde{g}_{\mathrm{N}}^{i\{k\}}] (\Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}}) = 0 \tag{6.49}$$

can be formulated for all active constraints $i \in \mathcal{A}^{\{k\}}$. The obvious drawback of such an attempt is that it explicitly asks for the Hessians of the active constraints, however, they are available since they are necessary for the evaluation of the Hessian $\nabla^2_{\underline{d}\,\underline{d}} \mathscr{L}$. Nevertheless, a direct insertion of (6.49) into the system of equations by replacing (3.52b) might lead to a much more difficult to solve linear system. Therefore, Nocedal and Wright [204, Sec. 18.3, p. 544] suggest another approximation based on the Taylor series expansion. By neglecting third order terms, the following estimate is deduced

$$\tilde{g}_{\mathrm{N}}^{i}(\underline{d} + \Delta\underline{d}) \approx \tilde{g}_{\mathrm{N}}^{i\{k\}} + [\nabla_{\underline{x}} \tilde{g}_{\mathrm{N}}^{i\{k\}}]^{\mathrm{T}} \Delta\underline{d} + \frac{1}{2} \Delta\underline{d}^{\mathrm{T}} \nabla^2_{\underline{d}\,\underline{d}} \tilde{g}_{\mathrm{N}}^{i\{k\}} \Delta\underline{d}. \tag{6.50}$$

Under the assumption that the original Newton step $\Delta\underline{d}$ and the augmented step $\Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}}$ are not too different, the second order term in (6.49) shall be approximated by

$$\begin{aligned} \frac{1}{2} (\Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}})^{\mathrm{T}} \nabla^2_{\underline{d}\,\underline{d}} \tilde{g}_{\mathrm{N}}^{i\{k\}} (\Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}}) &= \frac{1}{2} \Delta\underline{d}^{\mathrm{T}} \nabla^2_{\underline{d}\,\underline{d}} \tilde{g}_{\mathrm{N}}^{i\{k\}} \Delta\underline{d} \\ &\overset{(6.50)}{=} \tilde{g}_{\mathrm{N}}^{i}(\underline{d} + \Delta\underline{d}) - \tilde{g}_{\mathrm{N}}^{i\{k\}} - [\nabla_{\underline{x}} \tilde{g}_{\mathrm{N}}^{i\{k\}}]^{\mathrm{T}} \Delta\underline{d}. \end{aligned} \tag{6.51}$$

Next, this estimate is inserted into (6.49) and the new optimization problem is revealed

$$\underset{\Delta\underline{\bar{d}}^{\mathrm{SOC}} \in \mathbb{R}^n}{\text{minimize}} \quad [\nabla_{\underline{d}} \mathscr{U}^{\{k\}}]^{\mathrm{T}} (\Delta\underline{\bar{d}}^{\mathrm{SOC}}) + \frac{1}{2} (\Delta\underline{\bar{d}}^{\mathrm{SOC}})^{\mathrm{T}} \nabla^2_{\underline{d}\,\underline{d}} \mathscr{L}^{\{k\}} (\Delta\underline{\bar{d}}^{\mathrm{SOC}}) \tag{6.52}$$

$$\text{subject to} \quad \tilde{g}_{\mathrm{N}}^{i}(\underline{d} + \Delta\underline{d}) - [\nabla_{\underline{d}} \tilde{g}_{\mathrm{N}}^{i\{k\}}]^{\mathrm{T}} \Delta\underline{d} + [\nabla_{\underline{d}} \tilde{g}_{\mathrm{N}}^{i\{k\}}]^{\mathrm{T}} \Delta\underline{\bar{d}}^{\mathrm{SOC}} \geqslant 0, \qquad \forall i \in \mathcal{S}, \tag{6.53}$$

where $\Delta\underline{\bar{d}}^{\mathrm{SOC}} = \Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}}$ has been inserted to simplify the notation. This at hand the related system of equations follows to

$$\begin{pmatrix} \tilde{\nabla}^2_{\underline{d}\,\underline{d}} \mathscr{L}^{\{k\}} & -\tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\{k\}} \\ -[\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}]^{\mathrm{T}} & \underline{0} \end{pmatrix} \begin{pmatrix} \Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}} \\ \Delta\underline{\lambda}_{\mathrm{N}}^{\{k\}} + \Delta\underline{\lambda}_{\mathrm{N}}^{\mathrm{SOC}} \end{pmatrix} = \begin{pmatrix} -\nabla_{\underline{d}} \mathscr{U}^{\{k\}} + \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\{k\}} \underline{\lambda}_{\mathrm{N}}^{\{k\}} \\ \tilde{\underline{g}}_{\mathrm{N}}(\underline{d} + \Delta\underline{d}) - [\nabla_{\underline{d}} \tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}]^{\mathrm{T}} \Delta\underline{d} \end{pmatrix}, \tag{6.54}$$

where the Hessian of the Lagrangian as well as the gradients of the weighted gap have been replaced by their potentially slightly inconsistent counterparts introduced in Chapter 4 and only active contributions have been considered. The first row in (6.54) can be easily rewritten as

$$\begin{aligned}&\tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}^{\{k\}}\Delta\underline{d}^{\mathrm{SOC}} - \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\Delta\underline{\lambda}_{\mathrm{N}}^{\mathrm{SOC}}\\ =&-\nabla_{\underline{d}}\mathscr{W}^{\{k\}} - \tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}^{\{k\}}\Delta\underline{d} + \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}(\underline{\lambda}_{\mathrm{N}}^{\{k\}} + \Delta\underline{\lambda}_{\mathrm{N}}^{\{k\}}) \overset{(\dagger)}{=} \underline{0},\end{aligned} \tag{6.55}$$

where $(\dagger)$ follows since $(\Delta\underline{d}, \Delta\underline{\lambda}_{\mathrm{N}}^{\{k\}})$ fulfill the default linearized system of equations. Similarly, the second row in (6.54) can be reformulated, leading to the final system of equations

$$\begin{pmatrix}\tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}^{\{k\}} & -\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}\\ -[\tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\{k\}}]^{\mathrm{T}} & \underline{\underline{0}}\end{pmatrix}\begin{pmatrix}\Delta\underline{d}^{\mathrm{SOC}}\\ \Delta\underline{\lambda}_{\mathrm{N}}^{\mathrm{SOC}}\end{pmatrix} = \begin{pmatrix}\underline{0}\\ \tilde{\underline{g}}_{\mathrm{N}}(\underline{d}+\Delta\underline{d})\end{pmatrix}. \tag{6.56}$$

A short comparison between (6.48) and (6.56) instantly reveals the huge similarity between these two approaches. Thus, if the matrix $\underline{\underline{B}}$ is replaced by $\tilde{\nabla}^2_{\underline{d}\,\underline{d}}\mathscr{L}^{\{k\}}$ the systems would completely coincide in case of the variationally consistent approach. In the inconsistent case, system (6.56) represents a truncated variant of its consistent counterpart, which works often well in practice. Furthermore, the initial derivation via the least squares problem (6.43) confirms the assumption that the difference between $\Delta\underline{d} + \Delta\underline{d}^{\mathrm{SOC}}$ and $\Delta\underline{d}^{\mathrm{SOC}}$ can indeed be expected as being minimal, at least near the solution. System (6.56) is in the following considered as the *CheapSOC*-step. Another possibility is to do one more additional default step with the system of equations newly evaluated at $\underline{d} + \Delta\underline{d}$. Thus, two consecutive full Newton steps can also be used as a second-order correction step, see Wächter and Biegler [270] and Conn et al. [53, 15.3.2.3] for more information about this approach. This possibility is in the following called *FullSOC*-step and coincides with the idea of a watch-dog or non-monotone Newton method [45]. The default second order correction strategy in this thesis is that the CheapSOC step is always used, whenever the active set did not change between $\underline{d}$ and $\underline{d} + \Delta\underline{d}$, otherwise the FullSOC step is applied. Nevertheless, the CheapSOC step can theoretically also be applied if the active set changed. Further discussion on this topic will follow in Section 6.10.

6.5. Globalization Algorithm

At this point the full line search filter algorithm for contact problems can be presented. The algorithm is closely related to the algorithms proposed in Wächter and Biegler [270, 271, 272].

Algorithm 6.2 LINE SEARCH FILTER ALGORITHM

Given. Assume that a suitable pair $(\underline{x},\ \underline{\lambda}_{\mathrm{N}})_{\{0\}}$ is given, e.g., obtained from a predictor step; constants $\Theta_{\max} \in (\Theta^{\{0\}},\ \infty)$, $\Theta_{\min} > 0$, $\gamma_f, \gamma_\Theta \in (0,\ 1/\sqrt{2})$, $\gamma_\Theta^{\max} \in (0,\ 1)$, $\gamma_\alpha \in (0, 1]$; $s_\Theta > 1$; $s_f > 2\, s_\Theta$; $\eta_f \in (0,\ 0.5)$, $0 < \sigma_1 \leqslant \sigma_2 < 1$, $\{\mathrm{TOL}_i\} \in \mathbb{R}_+,\ \forall i \in \{1, \ldots, 4\}$.

0. *Initialize.* Initialize the filter to $\mathcal{F}^{\{k\}} := \{(\mathscr{L}, \Theta) \in \mathbb{R}^2 \mid \Theta \geqslant \Theta_{\max}\}$ or $\mathcal{F}^{\{k\}} := \varnothing$ and the Newton iteration counter $k \leftarrow 0$.

1. *Check convergence.* Stop if the pair $(\underline{x},\ \underline{\lambda}_{\mathrm{N}})_{\{k\}}$ fulfills the KKT conditions (4.9). The current implementation performs the following set of residual tests

$$\left\|\nabla_{\underline{x}}\mathscr{U}(\underline{x}) - \tilde{\nabla}_{\underline{x}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}(\underline{x})\ \underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}\right\|\Big|_{\{k\}} \leqslant \mathrm{TOL}_1 \qquad \text{and} \qquad \Theta^{\{k\}} \leqslant \mathrm{TOL}_2. \tag{6.57}$$

   In addition, it is required that the step size of the last accepted step is the default step length. Under this prerequisite, the demands

$$\|\underline{d}^{\{k\}} - \underline{d}^{\{k-1\}}\| \leqslant \mathrm{TOL}_3 \qquad \text{and} \qquad \|\underline{\lambda}_{\mathrm{N}}^{\{k\}} - \underline{\lambda}_{\mathrm{N}}^{\{k-1\}}\| \leqslant \mathrm{TOL}_4 \cdot \|\underline{\lambda}_{\mathrm{N}}^{\{k\}}\|. \tag{6.58}$$

   must be satisfied as well. Finally, the solution method is only denoted as converged if the active-set is converged as well, i.e., the active set $\mathcal{A}^{\{k\}}$ must be identical to $\mathcal{A}^{\{k-1\}}$.
2. *Compute search direction.* Evaluate the tangential stiffness matrix and compute the search directions $(\Delta\underline{d},\ \Delta\underline{\lambda}_{\mathrm{N}})$ from the linear system currently used. If the system of equations is too ill-conditioned, modify the system of equations according to Algorithm 6.3. Otherwise, determine the minimal step length estimate $\alpha_{\min}^{\{k\}}$ defined in (6.40) and initialize the pre-testing of Step 3.3 by calculating the necessary reference volumes, see also Section 6.7.1.
3. *Backtracking line search.*
   3.1. *Initialize line search.* Set $l \leftarrow 0$, and $\alpha^{\{k,l\}} = 1$ or alternatively use a user-defined default step length. Create a back-up of internally stored condensed variables with respect to the lastly accepted iteration. Furthermore, set the boolean flag isSOC to FALSE.
   3.2. *Trial point.* Compute the new trial point $\underline{d}(\alpha^{\{k,l\}}) = \underline{d}^{\{k\}} + \alpha^{\{k,l\}}\Delta\underline{d}^{\{k\}}$ and $\underline{\lambda}_{\mathrm{N}}(\alpha^{\{k,l\}}) = \underline{\lambda}_{\mathrm{N}}^{\{k\}} + \alpha^{\{k,l\}}\Delta\underline{\lambda}_{\mathrm{N}}^{\{k\}}$.
   3.3. *Pre-testing.* Perform a pre-testing, e.g., as described in Section 6.7.1, which is supposed to avoid a locally largely distorted mesh. If the pre-testing is successful, go to Step 3.4. Otherwise, if isSOC = TRUE, recover any internally stored condensed variables first, and go directly to Step 3.12.
   3.4. *Check acceptability to the filter.* Compute $\mathscr{L}(\alpha^{\{k,l\}})$ and $\Theta(\alpha^{\{k,l\}})$. If the trial pair is not a part of the current filter $\left(\mathscr{L}(\alpha^{\{k,l\}}),\ \Theta(\alpha^{\{k,l\}})\right) \notin \mathcal{F}^{\{k\}}$, go to Step 3.5. Otherwise update the filter-blocking criteria and go to Step 3.6.
   3.5. *Check the sufficient decrease criteria.*
      3.5.1. $\Theta^{\{k\}} \leqslant \Theta_{\min}$ *and switching condition* (6.4) *holds.* If the Armijo condition (6.5) is satisfied, go to Step 4. Otherwise go to Step 3.6.
      3.5.2. $\Theta^{\{k\}} > \Theta_{\min}$ *or switching condition is not fulfilled.* If one of the acceptance criteria defined in (6.3) holds, go to Step 4. Otherwise go to Step 3.6.
   3.6. *Initialize the SOC strategy.* If $l \neq 0$ go to Step 3.10. Else if isSOC = TRUE the current trial point has been already corrected, first recover any internally stored condensed variables and go to Step 3.12. Otherwise, set isSOC to TRUE and evaluate the SOC system. If the current trial point indicates a converged active set and there are infeasible constraints left, the *CheapSOC* system is considered, otherwise the *FullSOC* strategy is followed.
   3.7. *Compute the SOC step.* Solve the SOC system of equations under consideration of the same modification as used in Step 2. If the linear solver fails, recover internally stored condensed variables from the back-up state and go to Step 3.12.

3.8. *Replace the current trial pair by the second order corrected pair.* Replace the trial points from Step 3.2 by $\underline{d}(\alpha^{\{k,l\}}) = \underline{d}(\alpha^{\{k,l\}}) + \alpha^{\{k,l\}}\Delta\underline{d}^{\mathrm{SOC}}$ and $\underline{\lambda}_{\mathrm{N}}(\alpha^{\{k,l\}}) = \underline{\lambda}_{\mathrm{N}}(\alpha^{\{k,l\}}) + \alpha^{\{k,l\}}\Delta\underline{\lambda}_{\mathrm{N}}^{\mathrm{SOC}}$.

3.9. *Check acceptability of the SOC trial point.* Use the new second order corrected trial point pair and go to Step 3.3.

3.10. *Reinitialize the filter.* If the blocking criteria are fulfilled, the filter is reinitialized. Therefore, the maximally allowed constraint violation $\Theta_{\max}$ is reduced to $\Theta_{\max} \leftarrow \gamma_{\Theta}^{\max}\Theta_{\max}$ and the filter is reset to $\mathcal{F}^{\{k\}} := \{(\mathscr{L}, \Theta) \in \mathbb{R}^2 \mid \Theta \geqslant \Theta_{\max}\}$. Go to Step 3.12. Otherwise, if the current filter does not block good iterates, go to Step 3.11.

3.11. *Check for the minimal allowed step length.* If $\alpha^{\{k,l\}} < \alpha_{\min}^{\{k\}}$, the simulation will be stopped.

3.12. *Adapt the trial step size.* Choose $\alpha_{k,l+1} \in [\sigma_1\alpha^{\{k,l\}}, \sigma_2\alpha^{\{k,l\}}]$ and increase the counter $l \leftarrow l + 1$. Go back to Step 3.2.

4. *Accept the trial point.* Set $\underline{x}^{\{k+1\}} = \underline{x}(\alpha^{\{k,l\}})$ and $\underline{\lambda}_{\mathrm{N}}^{\{k+1\}} = \underline{\lambda}_{\mathrm{N}}(\alpha^{\{k,l\}})$.

5. *Augment the filter.* If k is not a $\mathscr{L}$-type step and the current point is not a feasible point, i.e. $\Theta^{\{k+1\}} > \mathrm{TOL}_2$, augment the filter using (6.16). Otherwise, $\mathcal{F}^{\{k+1\}} := \mathcal{F}^{\{k\}}$.

6. *Continue with next iteration.* Update the regularization parameter $c_{\mathrm{N}}^{\{k+1\}}$, increase the iteration counter $k \leftarrow k + 1$ and go to the convergence checks in Step 1.

6.6. Correction of the Linear System of Equations

The first crucial ingredient of Algorithm 6.2 is the treatment of non-positive definite systems of equations during the computation of the search direction in Step 2 of Algorithm 6.2. Theory and practice show that a singular or closely singular system often occurs in the neighborhood of instabilities such as buckling or snap-through scenarios. While a trust-region algorithm can naturally treat these systems due to the underlying idea of such methods, this is not true for a line search method. Instead, an algorithm is needed which reliably detects these cumbersome situations and provides a meaningful strategy to handle them without asking for too much additional computational effort. Especially, the solution of an Eigenvalue problem shall be avoided in the following. Instead, a set of heuristic criteria is presented which proved to be successful in practice. To reach this goal the following linear system of equations is considered throughout this chapter

$$\left.\begin{pmatrix} \tilde{\nabla}_{\underline{d}\,\underline{d}}\mathscr{L} + \omega\underline{\underline{M}} & -\tilde{\nabla}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}} & \underline{\underline{0}} \\ -(\nabla_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}})^{\mathrm{T}} & -\frac{1}{c_{\mathrm{N}}}\underline{\underline{A}}^{\mathcal{A}} & \underline{\underline{0}} \\ -\frac{2}{c_{\mathrm{N}}}\nabla_{\underline{d}}(\underline{\underline{A}}^{\mathcal{I}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}})^{\mathrm{T}} & \underline{\underline{0}} & -\frac{2}{c_{\mathrm{N}}}\underline{\underline{A}}^{\mathcal{I}} \end{pmatrix}\right|_{\{k\}} \begin{pmatrix} \Delta\underline{d} \\ \Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \\ \Delta\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} \end{pmatrix} = \left.\begin{pmatrix} -\nabla_{\underline{d}}\mathscr{U} + \tilde{\nabla}_{\underline{d}}\tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}} \\ \tilde{\underline{g}}_{\mathrm{N}}^{\mathcal{A}} \\ \frac{2}{c_{\mathrm{N}}}\underline{\underline{A}}^{\mathcal{I}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{I}} \end{pmatrix}\right|_{\{k\}}. \tag{6.59}$$

Fundamentally, this is the same system of equations as derived in Chapter 5. However, with one important difference: While the modification of the active constraint rows is, as already shown, very favorable in the pre-asymptotic phase, an additional modification of the top-left block has been introduced. This modification is well-known in the literature and becomes beneficial as soon as the projection of this matrix onto the null space of the active constraint gradients becomes positive semi-definite or even indefinite, see also the discussion in Section 5.6.8. The addition of

a positive-definite matrix $\underline{\underline{M}} \in \mathbb{R}^{n \times n}$ scaled by a positive scalar ω helps to shift the eigenvalues of the top-left matrix block to more positive values. In this thesis a simple identity matrix $\underline{\underline{I}} \in \mathbb{R}^{n \times n}$ is inserted for matrix $\underline{\underline{M}}$. This is a very common choice and has the additional advantage that the diagonal dominance will be increased by a sufficiently high regularization parameter ω. In fact, the active lower-right block can be interpreted in a very similar way: As long as the contact regularization parameter c_{N} is chosen small enough, this block helps to avoid any rank-deficiency of the constraint gradients. For more information the reader is kindly referred to Section 5.2.2. However, in contrast to the so-called *Inertia Correction* algorithm proposed in Wächter and Biegler [272, Sec. 3.1], the parameter c_{N} is defined by the SIR correction strategy proposed in Section 5.3.2. While in the mentioned reference [272] this block is only touched if the iteration matrix shows zero eigenvalues, it has become evident that the SIR update is crucial to create a meaningful sequence of iterates $\{\underline{d}^{\{k\}}\}$ and $\{\underline{\lambda}_{\mathrm{N}}^{\{k\}}\}$ which do not lead to a heavily distorted mesh for contact problems. In fact, it might be even necessary to reduce the c_{N} value again during the non-linear solution procedure, since otherwise the algorithm might not show the desired progress or even fails to fulfill the described sufficient decrease conditions of the filter method in some rare scenarios. This point will be reconsidered in Section 6.7.4. Furthermore, the shown saddle-point system is used instead of the possible condensed version due to the beneficial properties with regard to the conditioning of the iteration matrix. In this way, the following modifications of the system matrix will generally lead to a decreasing number of GMRES iterations if the regularization parameter ω is sufficiently high and c_{N} sufficiently low. However, to reach quadratic convergence, it is still favorable to set ω equal to zero whenever possible and to switch to the standard Lagrangian system as soon as the proposed switching conditions from 5.5 are satisfied. Otherwise, only a linear convergence can be expected. Firstly, the correction strategy of the parameter ω shall be discussed. Therefore, a modified variant of the Inertia Correction algorithm of Wächter and Biegler [272] is applied which is summarized in Algorithm 6.3.

The presented algorithm uses a heuristic acceptance test which works well in practice without being too restrictive. One important ingredient is the reliable detection of *invalid* or *bad* elements. While *invalid* means that the element Jacobian determinant becomes negative somewhere in the parametric domain, a *bad* element is identified by a fast changing element volume (or element area in 2-D). Therefore, first each element undergoes a validity check based on the trial displacement field $\underline{d}^{\{k\}} + \Delta\underline{d}$ which is described in Section 6.6.2. If the element is valid then the reference and current volumes of each element are computed and, subsequently, the ratio between previous and new trial volume is calculated. If this ratio is higher than $r_{\max}$ or lower than $r_{\min}$, then the *bad element counter* n_{bad} is increased.

This is one of the applied tests. However, Algorithm 6.3 is only concerned with the computation of a reliable search direction and a too restrictive set of criteria might lead to a slow convergence, since the correction algorithm might start to turn the search direction too far away from the original Newton direction. Even though, it should never be allowed to accept a step leading to invalid elements, this might be better handled afterwards by the line search capability of Algorithm 6.2.

Therefore, another acceptance test is added which consists of an upper bound for the current step length and a simple additional matrix check. The simple matrix check inlucdes a positive definiteness test claiming $\delta_\omega \geqslant \delta_\varepsilon > 0$, thus it is stronger than the wrapping positive definite-

Algorithm 6.3 CORRECTION OF THE ITERATION MATRIX

Given. The following constants $\omega_{\min} < \omega_0 < \omega_{\max}$, $0 < \kappa_\omega^- < 1 < \kappa_\omega^+ < \kappa_\omega^{++}$, $N_\omega > 0$, $r_{\min} < 1 < r_{\max}$, $s_{\text{last}} > 0$, $0 < \delta_\varepsilon \leqslant \delta_\omega$.

0. *Initialize.* The very first time entering the algorithm the variable ω_{last} is set to zero and the successive iteration counter n_ω is set to N_ω. Furthermore, each time the algorithm is entered, the correction counter $c_\omega \leftarrow 0$ will be reset.
1. *Solve the linear system of equations.* If $n_\omega \geqslant N_\omega$ try to solve the unmodified system of equations, i.e., $\omega = 0$ and c_{N} is chosen in accordance to the SIR-correction strategy. If the linear solver succeeds, go to Step 2. Otherwise, if either $n_\omega < N_\omega$ or the linear solver fails, go directly to Step 3.
2. *Check search direction.*
   2.1. *Compute element volumes.* Firstly, the reference element volumes $v_{\text{ref}}^{(e)}$ for all solid elements $e \in \mathcal{E}$ with respect to the previously accepted iterate $\underline{d}^{\{k\}}$ are computed. Afterwards, the current element volumes $v_{\text{curr}}^{(e)}$ with respect to $\underline{d}^{\{k\}} + \Delta\underline{d}$ are evaluated.
   2.2. *Identify bad elements.* Initialize the bad element counter n_{bad} to zero. If either
      a) the evaluation of the current element volumes fails due to a negative determinant of the deformation gradient or a negative determinant of the element Jacobian *or*
      b) the ratio $v_{\text{curr}}^{(e)}/v_{\text{ref}}^{(e)}$ is for any element $e \in \mathcal{E}$ greater than $r_{\max}$ or smaller than $r_{\min}$

      the bad element counter n_{bad} shall be increased by one for each identified element.
   2.3. *Compute default step length measure.* Compute the quadratic ℓ_2-norm of the current search direction, viz. $s_{\text{curr}} = \langle \Delta\underline{d}, \Delta\underline{d} \rangle$.
   2.4. *Acceptance test.* The first part of the acceptance test will be passed, if either
      a) there are no bad elements, i.e. $n_{\text{bad}} = 0$, *or*
      b) the step length decreases $s_{\text{curr}} \leqslant s_{\text{last}}$ and the top-left block of (6.59) fulfills $\langle \Delta\underline{d}, (\tilde{\nabla}^2_{\underline{d}\underline{d}}\mathscr{L} + \omega\underline{\underline{I}})\Delta\underline{d} \rangle > \delta_\omega s_{\text{curr}}$.

      As second part of the test, the matrix must stay sufficiently positive definite in the current search direction. This is enforced by the additional demand $\langle \Delta\underline{d}, (\tilde{\nabla}^2_{\underline{d}\underline{d}}\mathscr{L} + \omega\underline{\underline{I}})\Delta\underline{d} \rangle > \delta_\varepsilon s_{\text{curr}}$. If these tests are successfully passed, the current direction will be accepted. Then, set $s_{\text{last}} \leftarrow s_{\text{curr}}$ and go to Step 7. Otherwise, if there are bad elements and simultaneously either the step length increases or the (stronger) matrix test is violated or, alternatively, the direction is no descent direction, go to Step 3.
3. *Initialize or decrease the correction.* If $c_\omega > 0$ go to Step 5. Else if $\omega_{\text{last}} = 0$, set $\omega \leftarrow \omega_0$. Otherwise, set $\omega \leftarrow \max\{\omega_{\min},\ \kappa_\omega^- \omega_{\text{last}}\}$. Increase the correction counter $c_\omega \leftarrow c_\omega + 1$.
4. *Solve the modified system of equations.* Solve the modified system of equations. If the linear solver succeeds, go to Step 2. Otherwise, go to Step 5.
5. *Increase the correction.* If $\omega_{\text{last}} = 0$ set $\omega \leftarrow \kappa_\omega^{++}\omega$. Otherwise, set $\omega \leftarrow \kappa_\omega^+ \omega$. Increase the correction counter $c_\omega \leftarrow c_\omega + 1$.
6. *Check upper correction bound.* If $\omega > \omega_{\max}$, stop the simulation. Otherwise, go directly back to Step 4.
7. *Acceptance.* If $c_\omega \leqslant 1$, set $n_\omega \leftarrow n_\omega + 1$. Otherwise, reset $n_\omega \leftarrow 0$. Furthermore, if $c_\omega \neq 0$ set $\omega_{\text{last}} \leftarrow \omega$. Accept the computed search direction and use it for the following backtracking line search algorithm.

ness test, which will be addressed in a moment. This enhancement is reasonable without being unnecessarily restrictive since it can be only triggered if the violation comes along with a bad element counter greater than zero. Back to the first part of this second test, the quadratic ℓ_2-norm of the current step is computed and compared to the previous step length. If now the current step-length is smaller than the step length of the last applied search direction *and* the stronger positive definiteness test is passed as well, then the search direction is accepted even if some elements are still invalid or show a huge volume change. It is assumed that one of both criteria is fulfilled after a sufficient number of corrections. This can be expected, since the modified system of equations will lead to a smaller and smaller step-length for a sufficiently high value of ω. This is in accordance to the derivation leading to trust region (3.24) and Levenberg-Marquardt (3.23) schemes. Thus, the regularization parameter ω can be also interpreted as a representative estimate of the optimal trust-region Lagrange multiplier which is necessary to restrain the step length. An increasing trust region Lagrange multiplier value leads to a smaller and smaller trust-region. Similar ideas lead to the Levenberg-Marquardt method [174]. In fact, the step length test of Algorithm 6.3 proposed here is designed in such a way, that it takes advantage of these well-known strategies.

However, due to the used *or*-combination in Step 2.4, the step-length is not restrained if no bad elements have been detected. That means, under the assumption that the problem is well-posed and the Newton iteration does not lead to any bad elements, the algorithm is still able to increase the s_{last} value such that it is in a meaningful range. Otherwise, if the Newton iteration leads to invalid or bad elements and s_{last} has been chosen too large such that the matrix might no longer be positive definite in the computed search direction, the value of s_{last} will be adapted accordingly. Furthermore, due to the fact that it is very unlikely that s_{curr} is exactly equal to s_{last}, the algorithm will lead to a decreasing sequence each time bad elements occur. On the other hand, the value of s_{last} will stay bounded from above since the update in Step 2.4 is executed for each accepted search direction and the length of these search directions must be bounded as long as a progress to the solution can be achieved. In summary, the value of s_{last} will automatically become smaller and smaller as soon as the iteration sequence $\{\underline{d}^{\{k\}}\}$ starts to approach the solution point and can only become larger if no bad elements are detected and the upper left matrix block is sufficiently positive definite.

This last part is enforced by the surrounding positive definiteness test. Its main task is to detect a negative definite upper left block in (6.59). Otherwise, the algorithm might not modify the system of equations in cases where the matrix-block is negative definite but the bad element counter is equal to zero. Such a case would very likely lead to a stagnation of the entire solution procedure. An example can be found in Section 6.10.2. It can be concluded that the proposed algorithm has a well-designed self-correction property which works very well in practice (cf. Section 6.10).

Remark 6.1. At this point one important remark must be made concerning the surrounding positive definiteness tests in Step 2.4 of Algorithm 6.3. The example 5.6.8 has shown that there are scenarios in which the positive definiteness test used here will fail, but the projection of the gradients of the active constraints into the null space stays positive definite, as described in Remark 5.10. In such a case, the matrix modification might be unnecessary or even counter-productive in the sense of a higher number of Newton iterations. However, there are two points which must be kept in mind:

- Firstly, the example in Section 5.6.8 has shown that the method proposed in Chapter 5 benefits from a positive definite upper left block. Actually, a negative definite matrix in search direction might lead to a disastrous outcome. Furthermore, this is independent of the positive definite matrix property of the projected matrix. That means: If the modified Newton approach of Chapter 5 is applied, the convergence properties are rather improved by the positive definiteness test in Step 2.4 of Algorithm 6.3 and this is true for the variationally consistent as well as for the variationally inconsistent contact formulation.

- Secondly, it is to mention that the case of a negative curvature in search direction due to the influence of the second order derivatives of the active constraints could only be observed for the complete variational approach. In case of the incomplete variational approach, which is solely considered throughout this chapter, a negative definite system matrix seems rather connected to some kind of structural instability such as buckling or a snap-through scenarios. These instabilities need the matrix modification and are detected correctly by Algorithm 6.3.

To put it in a nutshell: Even though the true curvature of the underlying Lagrangian might be sufficiently positive definite in sense of (5.82), the line search method and contact formulation, i.e., the variationally inconsistent approach from Chapter 4, used here seem to benefit from the enforced stronger demand in Algorithm 6.3. This impression is also supported by various numerical experiments which will be presented in Section 6.10. The brief application to the complete formulation in Section 5.6.8 also revealed a stable but maybe slightly slower convergence behavior.

Now, the attention is drawn to some further details of Algorithm 6.3: Two counters c_ω and n_ω are introduced. The first one, namely the *correction counter* c_ω, is set to zero each time Algorithm 6.3 is entered and is incremented by one, each time the regularization parameter ω is updated in Step 3 or Step 5. This parameter has two main tasks: Firstly, it allows the identification of the first modification and initiates the decrease of the regularization parameter in Step 3. If the decrease is unsuccessful, the next and all subsequent modifications increase the regularization parameter until a new search direction is found which satisfies the acceptance tests. Beside this switching task between decrease and increase, the parameter c_ω has another purpose: It is used to detect a sequence of Newton iterations which were accepted with a decreasing regularization parameter. Therefore, each time acceptance is achieved and the correction counter is below or equal to 1, the second counter n_ω, the so-called *successive decrease counter*, is incremented by one. If an increase of the regularization parameter becomes necessary and thus $c_\omega > 1$ the successive decrease counter is reset. This counter is used to switch back to the unmodified system of equations in Step 1, but only when it seems promising. Otherwise, if the last increase of the regularization parameter was in the recent past, the solution attempt of the unmodified system of equations is skipped and the algorithm tries directly whether a slightly smaller regularization parameter might be successful in Step 3. In summary, the counter n_ω has been introduced to avoid unnecessary attempts to solve the unmodified system of equations whenever the recent history indicates that it is unlikely to be successful.

The remaining part of Algorithm 6.3 is quite similar to the inertia correction algorithm proposed by Wächter and Biegler [272]. Note that the update of the regularization parameter c_N is not part of Algorithm 6.3, but is instead moved to Step 6 of Algorithm 6.2. Some further details

on this topic will be given in Section 6.7.4. Before the discussion on Algorithm 6.2 is continued, the attention is drawn to the used linear solver parameters and the invalid element identification.

6.6.1. Linear Solver Parameters

This section is supposed to give a brief overview of the applied linear solver parameters which are considered in this chapter to solve the probably very large system of equations given in (6.59). However, if the problem size is small enough, a simple direct solver will be applied. Possible choices would be the UMFPACK [62] or the KLU [63] solver, for instance. In this thesis, the UMFPACK solver will be preferred.

However, if a larger problem is addressed, the choice of a suitable set of linear solver parameters can become far more complicated. A similar issue has already been discussed during the presentation in Section 5.6.7.2. However, the new system of equations (6.59) in combination with Algorithm 6.3 can be helpful to simplify this choice. With the applied regularization of the upper left matrix block in combination with the regularization of the lower right block under consideration of the SIR correction scheme, two methods are now accessible which increase the diagonal dominance of the system matrix whenever necessary. Therefore, the regularization of the upper left matrix block discussed here will be further increased whenever the linear solver fails in Step 1 or Step 4, and, indeed, this helps in many cumbersome situations to find a solution of the linear system without tweaking any parameters. To demonstrate this, all large examples in Section 6.10 will be solved with the same set of linear solver parameters. Furthermore, only small changes have been made compared to Table 5.2:

- The max. coarse level size is reduced to $5{,}000$ for all examples.
- In case of any active contact contributions solely the CheapSIMPLE preconditioner for the saddle point formulation will be considered. This is independent of the fact whether the modified contact system or the standard Lagrangian system are in use.
- The max. Krylov subspace size is increased to 300.

The remaining parameters in Table 5.2 stay unchanged . It must be emphasized that especially the smaller max. coarse level size makes it rather more difficult, but also computational less expensive, to find a solution of the linear systems. For further information concerning the linear solver performance the reader is referred to Section 6.10 and the graphs concerning the GMRES iterations therein.

6.6.2. Invalid Element Identification

The invalid element identification which is for example used in Step 2.2 a) of Algorithm 6.3, can be quite involved, since it might not always be straight forward to reliably detect all elements with a locally negative Jacobian determinant. However, it is important to detect them since otherwise the non-linear iteration might start to stagnate. Such a stagnation can occur if even one single element is overseen. This especially happens for very thin walled structures more often than one might think. To avoid this undesirable situation it is not enough to track the value of the

Jacobian or deformation gradient determinant at each evaluated Gauss point, instead the determinant is supposed to be positive in the entire feasible parametric domain. In the current implementation 4-noded quadrilateral and 8-noded hexahedral elements for 2- or 3-D, respectively, are considered. These elements are based on the classical linear Lagrange shape functions. Before the 3-D case is addressed, the attention is drawn to the easier 2-D case. Typically, the parametric space for the 2-D QUAD4 elements is defined as the domain $[-1,1] \times [-1,1]$. Since the following presentation is leaned on Johnen et al. [150, 151], the domain shall be redefined as $[0,1] \times [0,1]$. This does in no way affect the validation criteria and thus is just a technical step. The corresponding Lagrangian polynomials are then given by

$$L_1(\xi^1,\xi^2) = (1-\xi^1)(1-\xi^2), \qquad L_2(\xi^1,\xi^2) = \xi^1(1-\xi^2), \tag{6.60}$$

$$L_3(\xi^1,\xi^2) = \xi^1\xi^2, \qquad L_4(\xi^1,\xi^2) = (1-\xi^1)\xi^2. \tag{6.61}$$

$$\tag{6.62}$$

As one can easily reproduce, the Jacobian follows as

$$\begin{aligned} \underline{\underline{J}}^{(e)}(\xi^1,\xi^2) &= \begin{pmatrix} x^1{}_{,\xi^1} & x^1{}_{,\xi^2} \\ x^2{}_{,\xi^1} & x^2{}_{,\xi^2} \end{pmatrix} \\ &= \begin{pmatrix} \Delta[x^1]_{12}(1-\xi^2) + \Delta[x^1]_{43}\xi^2 & \Delta[x^1]_{14}(1-\xi^1) + \Delta[x^1]_{23}\xi^1 \\ \Delta[x^2]_{12}(1-\xi^2) + \Delta[x^2]_{43}\xi^2 & \Delta[x^2]_{14}(1-\xi^1) + \Delta[x^2]_{23}\xi^1 \end{pmatrix} \end{aligned} \tag{6.63}$$

and the associated determinant yields

$$\begin{aligned} \det(\underline{\underline{J}}^{(e)}(\xi^1,\xi^2)) =& L_1(\xi^1,\xi^2)\{\Delta[x^1]_{12}\,\Delta[x^2]_{14} - \Delta[x^2]_{12}\,\Delta[x^1]_{14}\} \\ &+ L_2(\xi^1,\xi^2)\{\Delta[x^1]_{12}\,\Delta[x^2]_{23} - \Delta[x^2]_{12}\,\Delta[x^1]_{23}\} \\ &+ L_3(\xi^1,\xi^2)\{\Delta[x^1]_{43}\,\Delta[x^2]_{14} - \Delta[x^2]_{43}\,\Delta[x^1]_{14}\} \\ &+ L_4(\xi^1,\xi^2)\{\Delta[x^1]_{43}\,\Delta[x^2]_{23} - \Delta[x^2]_{43}\,\Delta[x^1]_{23}\} \\ =& \sum_{i=1}^{4} L_i(\xi^1,\xi^2)\; J^i, \end{aligned} \tag{6.64}$$

where $\Delta[\cdot]_{ij}$ denotes $[\cdot]_i - [\cdot]_j$ and J^i represents one of the Jacobian determinant values evaluated at the four corners of the QUAD4 element. The representation in (6.64) clearly reveals that the value of the Jacobian determinant can be formulated as a bilinear interpolation of its values at the corners. Consequently, if the Jacobian determinant at the four corner nodes is greater than zero, then the Jacobian determinant in the entire element is greater than zero. This allows the definition of an easy validation criterion in 2-D: Evaluate the Jacobian determinant at the corners of the QUAD4 element. If all four values are positive, the entire element will be valid.

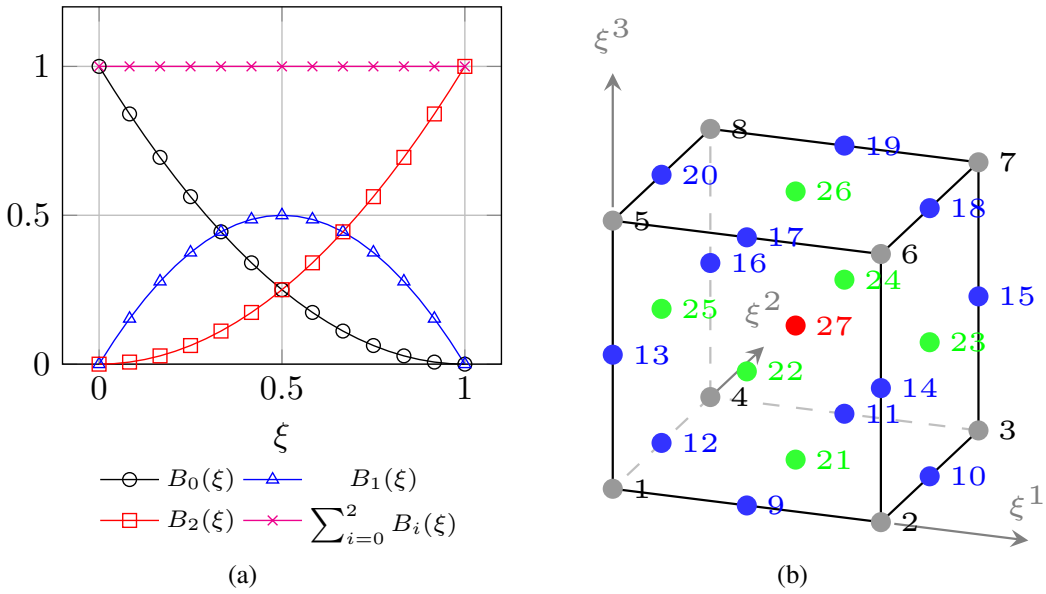
Unfortunately, the situation is much more complicated in 3-D. For a linear hexahedron, in the following called HEX8, the Jacobian determinant can be expressed as

$$\det(\underline{\underline{J}}^{(e)}(\xi^1,\xi^2,\xi^3)) = \varepsilon_{ijk}\; x^i{}_{,\xi^1}\; x^j{}_{,\xi^2}\; x^k{}_{,\xi^3}, \tag{6.65}$$

where ε_{ijk} is again the Levi-Civita symbol. Consequently, it is no longer a simple bilinear function, but instead a triquadratic one. Therefore, it is not enough to test only the corner points for validity. As stated in [147, 160], the positivity at the corner points can only be seen as a necessary condition. Furthermore, Knupp [160] has shown that it is also not enough to maintain positivity at the twelve edges of the HEX8 element. Therefore, he proposed that it might be sufficient to demand positivity on the six faces without giving a proof. However, this does not lead to a straight forward computation, since this new criterion asks for the minimization of a biquadratic function on each face subject to additional bound constraints given by the parametric domain. Since it shall be used as a fast and reliable pre-test in the filter method this seems unnecessarily complicated and to the best of the author's knowledge no usable algorithm for this minimization task has been published, yet. An alternative is the sub-division of the hexahedron into tetrahedra based on the eight corner nodes. This idea is exemplarily followed by [118, 147, 265, 268, 296]. These algorithms range from 8 to 64 computed tetrahedra. In Ushakova [266], the different approaches have been compared to each other and as a result it has been concluded that at least 58 tetrahedral volumes must be computed to obtain a sufficient criterion. This is quite a high number, and thus it might not be surprising that at the beginning a simpler remedy has been used in this thesis: The volume of the hexahedron in combination with the Jacobian determinant at the eight corners has been evaluated and checked for positivity. This is a approach which is also quite often followed in commercial packages [151, 161, 286]. If neither the volume calculation nor the evaluation at the corner nodes indicated an invalid element, the element has been declared valid. This worked in most cases very well, especially when it has been combined with an additional restriction of the volume change as proposed in Algorithms 6.2 and 6.3. However, a study of the related literature indicates that there are invalid configurations which might not be detected by such a simple approach. Fortunately, a further review has revealed that it is possible to improve the algorithm in an efficient way. The necessary steps have been implemented and shall be proposed in the following. Therefore, the spotlight is pointed onto the articles of Johnen et al. [150, 151]. In the first publication [150], a method has been proposed which is capable of reliably checking the validity of curved finite elements of any type. This is achieved by expanding the Jacobian determinant into the Bézier space. Due to the convex hull property of the Bézier expansion, the associated Bézier coefficients can be used to define an upper and lower bound for the minimum of the actual Jacobian determinant in the element. Subsequently, a *divide et impera* strategy is followed, which recursively divides the element into smaller subdivisions coming along with tighter and tighter bounds. This approach can be easily modified in such a way that it can be efficiently applied to the linear hexahedral element, since it is a special case of the more general framework discussed in [150]. This step has been done in [151] and will also be followed here. Furthermore, it is to note that this algorithm is to the best of the author's knowledge the only one which can truly be used as a reliable test since it works also if the minimum lies inside the element. Thus, it is not based on any unproven assumption.

Since a detailed description of the method can be found in [150, 151] only the most general steps shall be discussed here. Therefore, first, the necessary one-dimensional second order Bézier functions are stated

$$B_0(\xi) = (1-\xi)^2, \qquad B_1(\xi) = 2\xi(1-\xi), \qquad B_2(\xi) = (\xi)^2, \tag{6.66}$$



(a) (b)

Figure 6.4.: In Figure 6.4a the one-dimensional second order Bézier polynomials are visualized. Furthermore, the considered numbering of the sampling points for the HEX8 element is presented in Figure 6.4b. The color scheme is supposed to highlight the positioning more clearly: gray points denote corners, blue points are placed in the middle between two corner points on each edge, green points are positioned on the surface centers and the red point is located in the center of the HEX8 element.

for all $\xi \in [0, 1]$. A visualization can be found in Figure 6.4a. These at hand, their three-dimensional counterparts follow immediately as

$$B_{ijk}(\xi^1, \xi^2, \xi^3) = B_i(\xi^1)\, B_j(\xi^2)\, B_k(\xi^3), \tag{6.67}$$

where $0 \leqslant i, j, k \leqslant 2$. Next, the node numbering given in [151] is introduced by Figure 6.4b. Under consideration of this numbering the three indices of the B_{ijk} functions can also be assigned to the related point indices. For example, B_{000} becomes $\tilde{B}_1$ and B_{221} is identical to $\tilde{B}_{15}$. This said it is now possible to express the triquadratic Jacobian determinant (6.65) simply by

$$\det(\underline{\underline{J}}^{(e)}(\xi^1, \xi^2, \xi^3)) = \sum_{i=1}^{27} \tilde{b}^i\ \tilde{B}_i(\xi^1, \xi^2, \xi^3), \tag{6.68}$$

where $\tilde{b}^i$ are the coefficients or control values related to the corresponding Bézier basis function. Since the 27 Bézier functions are all positive over the entire parametric domain and due to the fact that the partition of unity holds, it can be concluded that the convex hull property holds and the coefficients can be used to define a lower bound of the Jacobian determinant value, viz.

$$\min_{i \in \{1,2,\ldots,27\}} \{\tilde{b}^i\} \leqslant \min_{\underline{\xi} \in [0,1]\times[0,1]} \{\det(\underline{\underline{J}}^{(e)}(\underline{\xi}))\}. \tag{6.69}$$

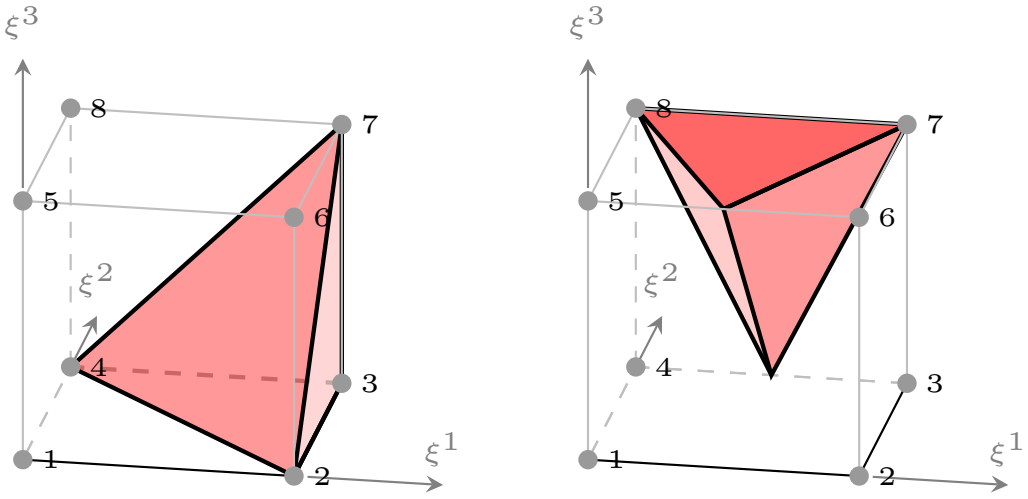


Figure 6.5.: In Figure 6.5a the tetrahedron related to the corner point 3 is given, while Figure 6.5b shows the tetrahedron belonging to the sampling point 19, i.e., the mid-point between the corners 7 and 8.

Furthermore, the Bézier coefficients corresponding to the corner points of the HEX8 element coincide with the values of the Jacobian determinant at these positions. This fact becomes obvious by looking at (6.66). While the partition of unity holds in the entire domain of definition, only at the boundaries of the domain all Bézier functions are zero except for the one corresponding to the related coefficient. Thus, with respect to the numbering introduced in Figure 6.4b the first eight Bézier coefficients correspond to the eight values of the Jacobian determinant at these corners. Now, the minimum of these first eight coefficients can consequently be used to define an upper bound for the minimal determinant value in the hexahedron. Assuming that one of these first eight values is negative, it is proven that the considered HEX8 is invalid. On the other hand, if all 27 Bézier coefficients should be positive, it is proven that the entire HEX8 is valid. Finally, in the last case, when the first eight coefficients are positive and at least one of the remaining 19 coeffizients is negative, the current information is insufficient and a subdivision algorithm is initiated to tighten the bounds. The details of this subdivision algorithm are provided in [150] and will be discussed in a moment. Important to note is that such a strategy is proven to be successful, since the subdivision algorithm always stops and the bounds are expected to converge quadratically with the size of the subdomains. For more details the reader is referred to the literature on this topic [50, 173].

Now, the first question is how to obtain the 27 Bézier coefficients which are necessary for the algorithm. Therefore, following the derivation, it is easily possible to obtain a linear mapping between the Jacobian determinant values and the Bézier coefficients by

$$\begin{aligned} & \det(\underline{\underline{J}}^{(e)}(\underline{\xi}^i)) = \tilde{B}_k(\underline{\xi}^i)\tilde{b}^k, && \forall i,k \in \{1,2,\dots,27\}, \\ \Leftrightarrow \quad & \underline{j} = \underline{\underline{\tilde{B}}}\,\underline{\tilde{b}}, && \end{aligned} \tag{6.70}$$

where $\underline{\xi}^i$ denote the 27 parametric position vectors of the points defined in Figure 6.4b, e.g., $\underline{\xi}^1 = (0,0,0)^{\mathrm{T}}$ and $\underline{\xi}^{22} = (0.5,0,0.5)^{\mathrm{T}}$. Following (6.70), the $\tilde{B}_{ij}$ entries of the matrix $\underline{\underline{\tilde{B}}} \in \mathbb{R}^{27\times 27}$ are given by the function values $\tilde{B}_j(\underline{\xi}^i)$. Thus, if the vector $\underline{j}$ is known, i.e., if the values of the Jacobian determiant at the 27 points are known, then system (6.70) can be solved for the desired Bézier coefficients. Since this mapping has an important role in the up-coming subdivision algorithm, it makes sense to explicitly compute the inverse $\underline{\underline{T}} = \underline{\underline{\tilde{B}}}^{-1}$ once at the beginning of the simulation, or to take it from Johnen et al. [151]. This matrix $\underline{\underline{T}} \in \mathbb{R}^{27\times 27}$ maps the values of the Jacobian determinant evaluated at the 27 sampling points given in Figure 6.4b onto the 27 Bézier coefficients.

Moreover, Johnen et al. [151] nicely explain that there are only 20 independent Bézier coefficients as well as 20 independent entries in the vector $\underline{j}$. This can be used to construct a linear mapping from these 20 Jacobian determinant values onto the 27 Bézier coefficients, such that

$$\underline{\tilde{b}} = \underline{\underline{Q}}\, \underline{j}_{20}, \tag{6.71}$$

where the vector $\underline{j}_{20}$ holds the Jacobian determinant values of the first 20 sampling points. The actual definition of $\underline{\underline{Q}} \in \mathbb{R}^{27\times 20}$ can again be found in [151]. Note that this matrix $\underline{\underline{Q}}$ is also stored as a static member in the algorithm. Thus, neither the matrix $\underline{\underline{T}}$, nor the matrix $\underline{\underline{Q}}$ must be recomputed after the initial setup call. However, the first 20 entries of the vector $\underline{j}$ remain outstanding and are, therefore, considered next. The first eight values in the vector $\underline{j}$ correspond to the eight corners and are equal to the volumes of the tetrahedra constructed from the three edges of the respective corner scaled by a factor of six. For example, the entry three of the vector $\underline{j}$ follows as

$$j^3 = \det(\underline{\underline{J}}^{(e)}(1,1,0)) = 6\ \mathrm{vol}_\Delta(\underline{x}^3, \underline{x}^4, \underline{x}^2, \underline{x}^7). \tag{6.72}$$

In a similar way it is also possible to obtain the remaining 12 entries corresponding to the edges. Therefore, the corresponding tetrahedra can be constructed as exemplarily shown for the entry 19 which is defined by

$$j^{19} = \det(\underline{\underline{J}}^{(e)}(0.5,1,1)) = 6\ \mathrm{vol}_\Delta(\underline{x}^7, \underline{x}^8, \frac{1}{2}(\underline{x}^5 + \underline{x}^6), \frac{1}{2}(\underline{x}^3 + \underline{x}^4)). \tag{6.73}$$

A demonstrative visualization of the TET4 corresponding to (6.72) can be found in Figure 6.5a, while an illustration of the TET4 related to (6.73) can be seen in Figure 6.5b. These 20 volumes must be computed at the beginning of each validity check under consideration of the current nodal positions of the HEX8 element. In fact this is the only point in the algorithm where the current nodal coordinates are considered. Now, if either one of the 20 computed Jacobian determinant values is negative or if all the subsequently computed Bézier coefficients $\tilde{b}^i$ for $i \in \{9, 10, \ldots, 27\}$ following (6.71) are positive, the algorithm is finished. Only if all of the 20 Jacobian determinant values are positive and at least one of the remaining 19 Bézier coefficients are negative, the subdivision is started. This subdivision algorithm is based on the following relationship

$$j^{[q]}(\underline{\xi}^{[q]i}) = \tilde{B}^{[q]}{}_{k}[\underline{\xi}^{[q]i}]\ \tilde{b}^{[q]k} = \tilde{B}_{k}[\underline{\xi}(\underline{\xi}^{[q]i})]\ \tilde{b}^{k}, \tag{6.74}$$

where $\underline{\xi}^{[q]i}$ now denotes one of the 27 parametric position vectors in one of the eight subdomains $q \in \{1, 2, \ldots, 8\}$, i.e., similar to (6.70), $\underline{\xi}^{[q]1} = (0, 0, 0)^{\mathrm{T}}$ and $\underline{\xi}^{[q]22} = (0.5, 0, 0.5)^{\mathrm{T}}$ for example. Additionally, the linear mapping $\underline{\xi}(\underline{\xi}^{[q]}) = \underline{a}^{[q]} + (\underline{b}^{[q]} - \underline{a}^{[q]})\underline{\xi}^{[q]}$ is inserted, where $\underline{a}^{[q]}$ and $\underline{b}^{[q]}$ denote the lower left and upper right boundary point of the q^{th} subdomain cube. Thus, (6.74) yields

$$\tilde{b}^{[q]i} = T^{ij}\ \tilde{B}_{k}[\underline{a}^{[q]} + (\underline{b}^{[q]} - \underline{a}^{[q]})\underline{\xi}^{[q]}{}_{j}]\ \underline{\tilde{b}}^{k} \tag{6.75}$$

for $i, j, k \in \{1, 2, \ldots, 27\}$ and $q \in \{1, 2, \ldots, 8\}$ where the necessary inverse of the matrix $\tilde{B}^{[q]ik} = \tilde{B}^{[q]k}[\underline{\xi}^{[q]i}]$ has been replaced by the already computed matrix $\underline{\underline{T}}$. Furthermore, under the assumption that a previously boundary point pair $\underline{a}$ and $\underline{b}$ is given, the new eight sub-boundary pairs can be computed instantly and very inexpensively by

$$\begin{aligned}
&\underline{c} = \frac{\underline{a} + \underline{b}}{2}, && && && \\
&\underline{a}^{[1]} = \underline{a}, && \underline{b}^{[1]} = \underline{c}, && \underline{a}^{[2]} = (c_1, a_2, a_3)^{\mathrm{T}}, && \underline{b}^{[2]} = (b_1, c_2, c_3)^{\mathrm{T}}, \\
&\underline{a}^{[3]} = (c_1, c_2, a_3)^{\mathrm{T}}, && \underline{b}^{[3]} = (b_1, b_2, c_3)^{\mathrm{T}}, && \underline{a}^{[4]} = (a_1, c_2, a_3)^{\mathrm{T}}, && \underline{b}^{[4]} = (c_1, b_2, c_3)^{\mathrm{T}}, \\
&\underline{a}^{[5]} = (a_1, a_2, c_3)^{\mathrm{T}}, && \underline{b}^{[5]} = (c_1, c_2, b_3)^{\mathrm{T}}, && \underline{a}^{[6]} = (c_1, a_2, c_3)^{\mathrm{T}}, && \underline{b}^{[6]} = (b_1, c_2, b_3)^{\mathrm{T}}, \\
&\underline{a}^{[7]} = \underline{c}, && \underline{b}^{[7]} = \underline{b}, && \underline{a}^{[8]} = (a_1, c_2, c_3)^{\mathrm{T}}, && \underline{b}^{[8]} = (c_1, b_2, b_3)^{\mathrm{T}}.
\end{aligned} \tag{6.76}$$

Note that the initial boundary point pair is given by $\underline{a} = \underline{0}$ and $\underline{b} = (1, 1, 1)^{\mathrm{T}}$. Thus, the computational cost of the subdivision procedure lies in the evaluation of the 729 function values $\tilde{B}_{k}[\underline{\xi}(\underline{\xi}^{[q]i})]$ and the execution of the subsequent two matrix-vector products in (6.75).

At this point, it is possible to summarize the necessary steps in Algorithm 6.4 which is closely linked to the algorithms proposed in [150, 151]. Note that in Step 3 it is sufficient to check the last 19 Bézier coefficients, since the first eight have already been tested in Step 2. To verify the implementation, the unit tests proposed in [151] have been applied. In Figure 6.6 one heavily distorted element is shown, which would pass the old test, i.e., the Jacobian determinant values at the eight corners as well as at the Gauss point positions are all positive. In fact, if the minimum of the *scaled* Jacobian determinant at the eight corners is computed the value 0.647 is obtained. For more details how this scaled Jacobian determinant can be computed the interested reader is referred to Knupp [161]. Anyway, that's far away from being zero or critical. Thus, the previously used simple check would fail in this case, since as Figure 6.6b shows, the element is indeed invalid. It is to mention that this is detected by Algorithm 6.4 already in Step 2, since the volume of the TET4 associated to sampling point 14 in Figure 6.4b is negative. Thus, this example is not really challenging for the proposed method. Therefore, another test from [151] shall be discussed which is shown in Figure 6.7. Again the simple corner node and Gauss point check would tell us that the element is valid. However, Algorithm 6.4 reliably detects the location of a negative Jacobian determinant value by three successive subdivisions which are shown in Figure 6.7b.

Algorithm 6.4 ANALYSIS OF THE HEX8 JACOBIAN DETERMINANT

Given. The matrix $\underline{\underline{T}}$ as well as the matrix $\underline{\underline{Q}}$ are either given or must be constructed once at the beginning when the method is called the very first time. See Johnen et al. [151] for a detailed description.

1. *Compute TET4 volumes.* Compute the 20 TET4 volumes corresponding to the eight corners and the 12 edges of the hexahedron element. Store these volumes scaled by a factor of six in a vector $\underline{j}_{20}$.
2. *Check the volumes.* If at least one entry j_{20}^i for $i \in \{1, 2, \ldots, 20\}$ is negative or zero, return false and leave the algorithm.
3. *Compute the corresponding 27 Bézier coefficients.* Evaluate (6.71). If all entries in $\underline{\tilde{b}}$ are positive, return true and leave the algorithm.
4. *Subdivision.* Set the parent sub-domain boundaries to $\underline{a} = \underline{0}$ and $\underline{b} = (1, 1, 1)^{\mathrm{T}}$.
   4.1. *Create 8 sub-cubes.* Compute the sub-cube borders following (6.76).
   4.2. *Loop.* Loop over $q \in \{1, 2, \ldots, 8\}$.
      a) *Compute Bézier coefficients.* Evaluate $\tilde{B}_k[\underline{\xi}(\underline{\xi}^{[q]i})]$, $\forall i, k \in \{1, 2, \ldots, 27\}$ and use (6.75) to compute the associated Bézier coefficients.
      b) *Check Bézier coefficients.* If $\tilde{b}^{[q]i} \leqslant 0$ for any $i \in \{1, 2, \ldots, 8\}$, return false and leave the algorithm. Else if $\tilde{b}^{[q]i} > 0$ for all $i \in \{9, 8, \ldots, 27\}$, go to the next sub-cube, i.e. $q \leftarrow q + 1$, and go to Step 4.2a). If all coefficients are positive and no other sub-cube is left, return true.
      c) *Recursive subdivision.* Set $\underline{a} = \underline{a}^{[q]}$, $\underline{b} = \underline{b}^{[q]}$ and go to Step 4.1.

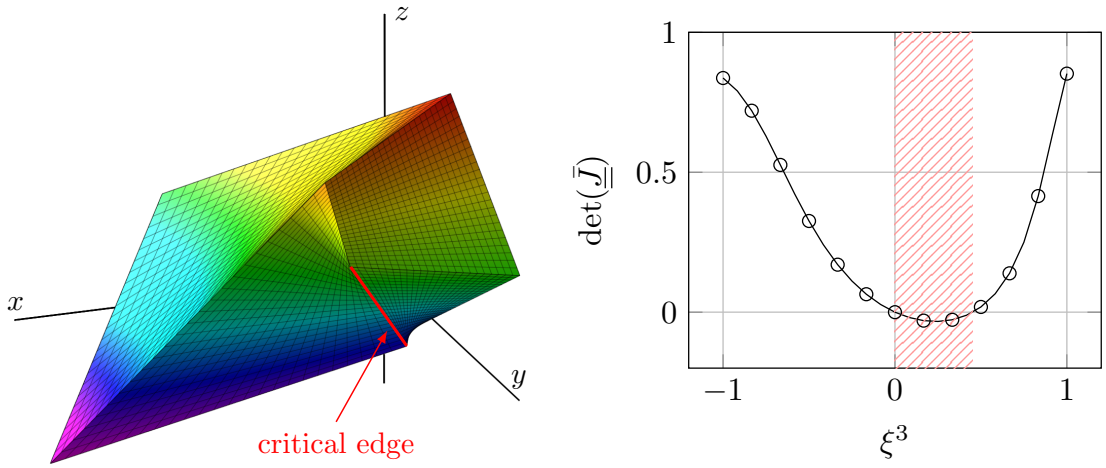
The seventh corner node of the smallest sub-cube has been correctly identified as invalid. Note that the subdivisions indicate that the Jacobian determinant values associated to the corners of the evaluated surrounding sub-cubes are all positive. Thus, the Jacobian determinant is only very locally invalid.

6.7. Further Details on the Globalization Algorithm

In the following more details of Algorithm 6.2 shall be discussed. Despite the fact that the following points might in theory not be strictly necessary, all of them represent crucial ingredients to obtain an algorithm which is practically applicable and successful. The different points are discussed hierarchically following the step order in Algorithm 6.2.

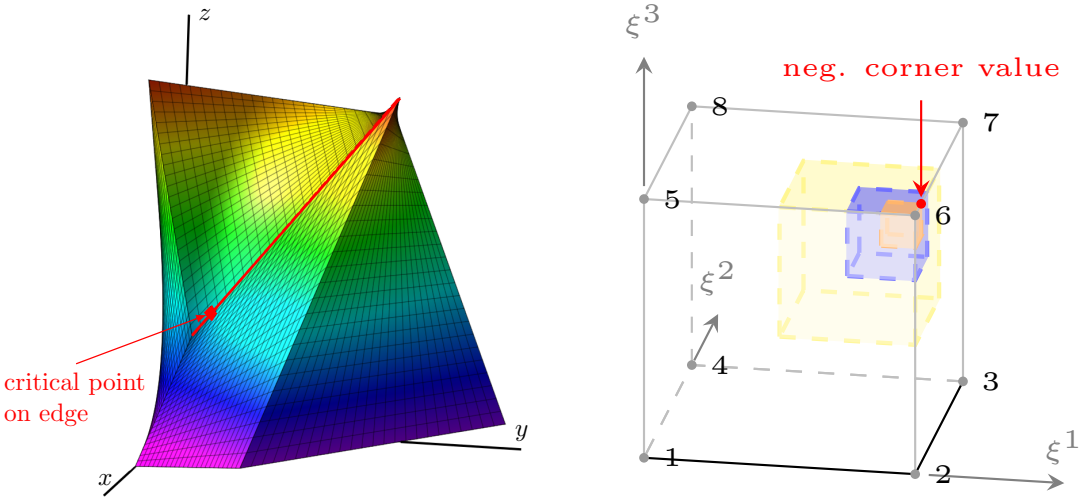
6.7.1. Pre-Testing

The attention is on Step 3.3 of Algorithm 6.2. Here, an optional pre-testing is inserted which is not part of the original line search filter algorithm proposed by Wächter and Biegler [270, 271, 272], but becomes important for discretized problems. The subsequent filter method completely relies on the fact that a bad iterate can be detected based on the objective function value and the used infeasibility measure. Unfortunately, the gained experience in the area of computational contact problems reveals that this is not always true. There exist Newton iterates which globally lead to a decreasing infeasibility violation or a decreasing structural energy such that they sat-



(a) (b)

Figure 6.6.: Figure 6.6a shows a heavily distorted element taken from Johnen et al. [151, Fig. 7]. In Figure 6.6b the associated graph of the scaled Jacobian determinant along a critical edge (highlighted in Figure 6.6a) is shown.



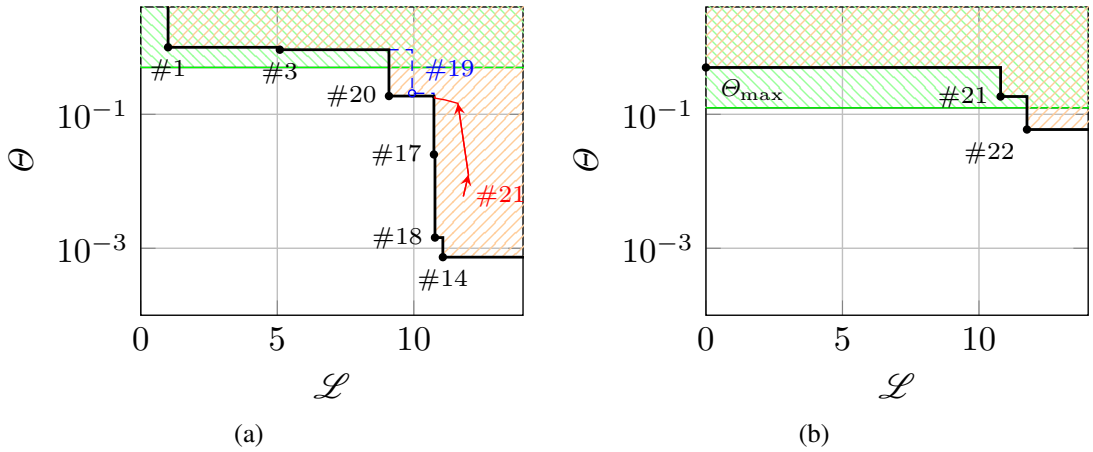
(a) (b)

Figure 6.7.: Figure 6.7a shows a heavily distorted element taken from Johnen et al. [151, Fig. 5]. In Figure 6.7b the result of the subdivision algorithm is shown. For this element three successive subdivisions must be executed before the computed bounds reliably indicate that the element is invalid. The corner node of the smallest sub-cube with a negative value is marked by ●.

isfy the acceptability criteria of the filter method, but come along with a very localized heavily distorted mesh. If such an iterate is accepted by the filter method, any further progress might be prohibited. Thus, the simulation stops, just because it might become very difficult or even impossible to resolve the heavily distorted geometrical configuration of a few or maybe only one element. Therefore, the already in Algorithm 6.3 applied bad element identification is placed in front of the filter method as a reliable pre-testing which identifies these rare unwanted cases. Now, since the search direction is kept constant and only the step length is further varied or, to say it clearer, potentially shortened, the demand is formulated that neither any solid element is allowed to have an invalid Jacobian determinant nor is any element allowed to show a drastically changing volume. If one of these criteria is violated, the current trial point will be rejected and the backtracking routine is directly initiated without even entering the filter method. Therefore, again a lower and an upper bound $0 < r_{\min}^{\text{pre}} < 1 < r_{\max}^{\text{pre}}$ for the volume ratios of reference to current volume with respect to each element must be defined. In general, these bounds can differ from the ones used in Algorithm 6.3. Then, if any ratio $v_{\text{curr}}^{(e)}/v_{\text{ref}}^{(e)}$ for $e \in \mathcal{E}$ lies outside the domain $[r_{\min}^{\text{pre}}, r_{\max}^{\text{pre}}]$, the current trial point will be rejected. A nice side-effect of this approach is that this volume testing is computationally much cheaper than a possible attempt of evaluating all structural and contact contributions. Consequently, this pre-testing can not only avoid a stagnation of the algorithm but might also save some time.

6.7.2. Bypassing of the $\mathscr{L}$-type Test

A small modification inherent in Algorithm 6.2, which has not been mentioned yet, is the additional condition during the check of the sufficient decrease criteria in Step 3.5. What has been added is the additional demand that not only the $\mathscr{L}$-type conditions formulated in (6.4) must be satisfied to fulfill the switching criterion but at the same time also the lastly accepted constraint violation is not allowed to be too big. Therefore, the so far unmentioned parameter $\Theta_{\min} > 0$ has been introduced (see also Figure 6.1). This adaption has been originally proposed by [272]. Even though this is only a very small change the consequences can be remarkable. In an earlier version of the algorithm this additional condition was missing. The result was that in some simulations the switching condition has been activated in a very early pre-asymptotic state. Now, an active $\mathscr{L}$-type condition demands more from the non-linear solution method than just to achieve a decrease in the constraint violation *or* the objective function value since the Armijo rule (6.5) must be satisfied. Actually, this can be a by far stricter condition which might lead to a very small step-length till acceptability is reached. Especially, as long as the underlying model is not able to reflect the actual objective function in a sufficient way. Exactly this is very likely during the pre-asymptotic phase. For a well-posed problem, it is almost always much easier to decrease the constraint violation than the objective function value in this early stage. The problem is that this small step-length will in turn indicate a repetitive satisfaction of the $\mathscr{L}$-type switching condition as long as the method does not succeed to leave this region. To put it in a nutshell, this might result in a very slow convergence even though the steps are all good and would be directly accepted by (6.3) without any step-length adaption. Due to the fact that generally no good accordance of the model and the actual objective function during the pre-asymptotic phase can be expected, it is a much better idea to skip the $\mathscr{L}$-type condition in this early stage. Therefore, $\Theta_{\min}$ is used. It is assumed that as soon as the constraint violation is below this value the model



(a) (b)

Figure 6.8.: In Figure 6.8a an example is shown, where between iteration #18 and #19 the constraint violation strongly rises while the objective function value is only slightly decreased. Subsequently, iteration #21 is completely blocked by historical information stored in the filter set. The problem can only be resolved by reinitializing the filter. The result is shown in Figure 6.8b. The green (upper left to lower right) hatched area marks the region defined by $\Theta(\alpha^{k,l}) \geqslant \gamma_{\Theta}^{\max}\Theta_{\max}$ before and after the reinitialization, respectively.

becomes trustworthy. Another side-effect is that the minimal step-length estimate for the $\mathscr{L}$-type condition described in Section 6.3.3 must not be evaluated during this pre-asymptotic iterates.

6.7.3. Reinitialization of the Filter

There is another problem which may arise during the application of the filter method. This time the filter might refuse to accept good iterates because of outdated historical information. This is something that is well-known in the literature and for instance mentioned by Wächter and Biegler [271, Remark 7] and later addressed in [272, Sec. 3.2]. The historical information might belong to a very different deformation state of the elastic bodies but, however, the associated filter point coordinates might be very similar to the coordinates of the currently considered filter point. To explain why this is actually not only an academic issue, let us consider the following scenario: The algorithm is on its way to the equilibrium state and during the previous Newton iterations the regularization parameter c_{N} has successively been updated based on the SIR-update. However, during the last iterations the active set drastically changed such that the number of active nodes and thus the active contact zone shrank to a very small region. Now, a bigger than expected change in the displacement field, for example induced by a slight structural instability, in combination with a rising contact force due to the high regularization parameter might lead to the undesired case that the two bodies come locally out of contact at the end of an accepted Newton iteration. Since the SIR update has led to a high c_{N} value, the active-set decision (4.12) mainly relies on the weighted gap value and, consequently, a small positive gap can be enough to locally set the previously active nodes inactive for the next Newton iteration. The consequence might be that the structures collapse in the next iteration due to the missing constraint and the new configuration indicates a large penetration but simultaneously comes along with a much smaller structural energy value, such that this new iteration might be again accepted by the filter. It is further to say that the previously discussed pre-testing might not detect such a case as long

as the collapsing is not attended by a heavily distorted mesh or large changes in the element volumes (see Section 6.7.1). A demonstrative example for such a scenario is shown in Figure 6.8a: As described the bodies come out of contact in Newton iteration $\#18$ and collapse in Newton iteration $\#19$ coming along with a much higher Θ value but also a lower $\mathscr{L}$ value. So one might think that this is no problem for the filter method, since nothing happened what can not be withdrawn. However, there are actual two severe problems with such a scenario which can even lead to complete failure of the algorithm:

- The first problem is that the current c_{N} value is probably much too high for the newly appeared large penetration state. This will be addressed in Section 6.7.4.
- The second problem are the filter points which have been already added to the filter during the filter augmentations in Step 5 of Algorithm 6.2. Actually, if as described this situation occurs only locally, i.e., the bodies collapse only at one contact point of many such that there are still other active contact zones left, then even the filter point referring to the collapse might be added to the filter (see Figure 6.8a). Especially, if the bypassing of the $\mathscr{L}$-type check is still activated due to the remaining constraints, and thus the last iterate can not be a $\mathscr{L}$-type and is also no feasible step. Now, after this collapsing scenario the algorithm tries to resolve the issue by pushing the bodies apart again, however, this will lead to a rising objective function coordinate of the associated filter point under almost all circumstances. And even though the constraint violation is supposed to be decreased compared to the previous iteration, there might be other filter points in the filter which show much smaller infeasibility values. These older points have the power to prohibit any further progress or recovery and, finally, the method will fail.

By looking at Figure 6.8a the situation seems a little bit different since the following iteration $\#20$ shows a decrease in the objective function and rather no change in the infeasibility measure. However, this originates from the mentioned too high c_{N} value which leads to a quite bad iterate. Thus, the c_{N} value is reset between Newton iteration $\#20$ and $\#21$ as described in Section 6.7.4. The scenario now becomes critical if points currently contained in the filter block the new iteration from acceptance. It is enough that one filter point is part of the set which combines a lower constraint violation value with a lower objective function value compared to the current iteration, independently from the fact that this new trial point might be a very good step in the sense that it tries to withdraw the previous unintended collapsing. Actually, exactly this is the case in Figure 6.8a. The good iteration $\#21$ is blocked by the filter points associated to the older Newton iterations $\#14$ and $\#18$. All belonging to deformation states before the collapsing occurred. In this special case, even the line search path indicated by the red line in Figure 6.8a is not able to resolve the issue and to leave the taboo-region of the filter.

Remark 6.2. One might expect that the line search path is approaching the previous filter point corresponding to iteration $\#20$ for very small step length values. The reason, why this is not the case here, will be given in Section 6.7.4.

The remedy implemented here is leaned on an idea proposed in Wächter and Biegler [272, Sec. 3.2], but it has been slightly extended. However, the basic strategy stays the same: The algorithm needs some counting mechanism such that it is able to detect such a situation. Therefore,

the filter method checks each time a trial point is rejected by the filter acceptability test in Step 3.4 of Algorithm 6.2, if the current trial filter point would have been accepted by the following sufficient decrease criteria in Step 3.5. Now, in the case that the filter rejected the trial point, but the sufficient decrease criteria, which are solely based on the previously accepted iterate, would have accepted the trial point, an additional condition is tested: If $\Theta(\alpha^{\{k,l\}}) < \gamma_\Theta^{\max}\Theta_{\max}$ holds, then a new *blocking point* is created and added to a so-called *blocking list*. The lastly mentioned condition has the important task to avoid an infinite number of reinitializations as will become clear in a moment. But first back to the blocking list: In the current implementation this blocking list contains two values per iteration. The first coordinate of such a blocking point is the current identification number of the Newton iteration. This value is used to identify successive blocking scenarios, i.e., the list will be cleared again if the next Newton iteration would not indicate a blocking scenario and is otherwise augmented. The second coordinate of a blocking point counts the successive number of blocks during the line search. In more details: Assuming that all blocking criteria are fulfilled, the first time the algorithm enters Step 3.4 during the current Newton iteration, it will be tested whether the list is continuous or not. If yes, then the new blocking point is added to the end of the list and the line search step counter is set to one. If the list is not continuous, it is cleared and a new list is initiated containing only the current point. Nevertheless, the second coordinate is initially set to one. If the algorithm enters now multiple times in one Newton iteration this check in Step 3.4 and fulfills each single time all blocking criteria, then the second coordinate associated to the current Newton iteration is incremented for each blocking line search step by one. Finally, the filter may be reinitialized in Step 3.10 either if the length of the list indicates that the last $n_{\text{newton}}^{\text{block}}$ Newton iterations have been blocked at least once by the filter or if the second coordinate of the lastly added blocking point is larger than $n_{\text{ls}}^{\text{block}}$, i.e., a pre-defined allowed number of consecutive blocking line search steps is exceeded. The filter is reset by setting $\mathcal{F}^{\{k\}} = \{(\mathscr{L}, \Theta) : \Theta \geqslant \gamma_\Theta^{\max}\Theta_{\max}\}$. Now, to avoid an infinite number of reinitializations, the $\Theta_{\max}$ is set after the reinitialization to $\Theta_{\max} \leftarrow \gamma_\Theta^{\max}\Theta_{\max}$. This algorithm showed to be very successful in practice in combination with the following adaption of the regularization parameter c_{N}. In Figure 6.8a the region $\Theta(\alpha^{k,l}) \geqslant \gamma_\Theta^{\max}\Theta_{\max}$ is represented by the green hatched area. As one can easily see the points corresponding to Newton iteration $\#21$ are all below this hatched area and are also below the taboo region defined by the lastly accepted iteration $\#20$. Thus, the entire red line search path fulfills the blocking criteria and, consequently, the second criterion based on $n_{\text{ls}}^{\text{block}}$ is activated. In this special case $n_{\text{ls}}^{\text{block}}$ is equal to 10 and the initial $\Theta_{\max}$ had been set to 2.0, while $\gamma_\Theta^{\max}$ had been chosen as 0.25. The result of the reinitialization can be seen in Figure 6.8b. The filter has now a clear upper bound for the infeasibility measure. Furthermore, the filter point corresponding to iteration $\#21$ has been added and, additionally, the filter point associated with the next Newton iteration is shown. The reinitialization was successful and allows progress to the equilibrium again. Otherwise, i.e. without the reinitialization, iteration $\#22$ would have been blocked as well.

6.7.4. Decrease of the Contact Regularization Parameter

In Section 5.3.2 the SIR update has been presented and represents a reliable method to obtain a new c_{N} value as long as the solution path is well-defined and no unforeseen situation occurs which would contradict the used linear infeasibility model. Therefore, it is a meaningful ap-

proach to use the SIR update as a monotone rule, i.e. $c_{\mathrm{N}}^{\{k+1\}} \geqslant c_{\mathrm{N}}^{\{k\}}$. Especially, since near the solution the upper bound of the c_{N} value is usually approached from beneath. However, the examples discussed in Section 5.6 showed that the correction parameter $\beta_{c_{\mathrm{N}}}^{\theta}$ should not be chosen too high, since otherwise a default Newton scheme might fail. Now, in contrast to Chapter 5, the examples considered here might show severe instabilities and as discussed beforehand in Section 6.7.3 it might happen that the filter method accepts a step which leads to a drastic rise of the infeasibility measure. In such a case it is quite likely that the previously computed c_{N} value might be a bad choice since it is based on older historical information. Thus, the new penetration state might ask for a smaller c_{N} value. But how is it possible to reliably detect these quite rare cases? Especially, since an unnecessary reduction of the regularization parameter can lead to a decreased convergence speed as Section 5.4 underlines.

The strategy used in this thesis is initiated by information collected over the last Newton iteration. Only if the last Newton iteration asked for a correction of the linear system of equations in Algorithm 6.3, indicated by $n_{\omega} > 0$ and the Newton step asked as well for a step length correction via the line search method, only then a possible reduction of the regularization is even considered in Step 7 of Algorithm 6.2. It is briefly to note that n_{ω} equal to zero can be used to identify the first part, since the mechanism in Algorithm 6.3 ensures that n_{ω} is only equal to zero if the unmodified system of equations is used. In this way a decrease of the c_{N} value is only applied if the algorithm was in trouble during the last Newton iteration, where $n_{\omega} > 0$ also indicates some instability. Now, the derivation of the c_{N} decrease is very close to the derivation of the c_{N} update derivation. The initial demand is just formulated differently, viz.

$$\Theta^{\{k\}} + \frac{1}{\Theta^{\{k\}}} \langle \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \nabla_{\underline{d}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}} \Big|_{\{k\}}^{\mathrm{T}} \Delta \underline{d} \rangle \overset{!}{\geqslant} (1 - \beta_{\Theta\mathrm{crit}}^{c_{\mathrm{N}}}) \Theta^{\{k\}}, \tag{6.77}$$

where the switched inequality sign is the important change. Thus, the linear model is used to define a bound for the maximally allowed reduction, where the new parameter $\beta_{\Theta\mathrm{crit}}^{c_{\mathrm{N}}}$ with $\beta_{\Theta}^{c_{\mathrm{N}}} < \beta_{\Theta\mathrm{crit}}^{c_{\mathrm{N}}} < 1$ is introduced. The collected experience during the parameter study in Section 5.6.2 indicates that a $\beta_{\Theta\mathrm{crit}}^{c_{\mathrm{N}}}$ value greater than 0.9 seems meaningful. In the example shown in Figure 6.8 the (increasing) correction parameter $\beta_{\Theta}^{c_{\mathrm{N}}}$ has been set to 0.8, while the critical (decreasing) correction parameter $\beta_{\Theta\mathrm{crit}}^{c_{\mathrm{N}}}$ had been set to 0.95. Now, following the same steps as in Section 5.3.2, the final c_{N} decrease updating rule is given by

$$c_{\mathrm{N}}^{\{k+1\}} = c_{\mathrm{N}}^{\{k\}} \min\{1, \frac{1}{1 - \beta_{\Theta\mathrm{crit}}^{c_{\mathrm{N}}}} (1 + \frac{\langle \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \nabla_{\underline{d}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}} \Big|_{\{k\}}^{\mathrm{T}} \Delta \underline{d} \rangle}{(\Theta^{\{k\}})^2})\} \tag{6.78}$$

as long as

$$\langle \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}}, \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}\{k\}} + \nabla_{\underline{d}} \underline{\tilde{g}}_{\mathrm{N}}^{\mathcal{A}} \Big|_{\{k\}}^{\mathrm{T}} \Delta \underline{d} \rangle > 0 \tag{6.79}$$

holds. Otherwise, $c_{\mathrm{N}}^{\{k+1\}} = c_{\mathrm{N}}^{\{k\}}$ is used. Note that (6.79) is theoretically supposed to be equal to zero in the limit case $c_{\mathrm{N}} \to \infty$. Thus, (6.79) is used as a safe-guarding condition to avoid

negative c_N values which seem to occur in some scenarios if the active set changes severely from one Newton iteration to the next.

The attention is now back on the example in Figure 6.8: The correction rule (6.78) is activated and applied in the end of Newton iteration $\#20$. Iteration $\#20$ fulfills the prerequisites since n_ω is equal to 4, i.e., the linear system had to be corrected four times and the step-length had to be reduced three times, such that (6.78) initiates a drop of the regularization parameter from $c_\mathrm{N}^{\{20\}} = 3.56\mathrm{E}{+}04$ to $c_\mathrm{N}^{\{21\}} = 1.88\mathrm{E}{+}03$. That is almost a factor 19 smaller. However, due to the described collapsing scenario it seems justified. As described in (6.10) and explained in the subsequent discussion: The used filter point coordinates $\mathscr{L}$ and Θ do not explicitly rely on the c_N value. This is a very important point since otherwise adaptions of the filter points already contained in the filter would become necessary, if the regularization parameter is touched during the Newton iterations. Nevertheless, a changing c_N value can still have an implicit impact on the filter point coordinates due to the active set decision (4.12). By keeping this active set decision in mind, it is possible to explain the different filter points in Figure 6.8: The drop in the $\mathscr{L}$ value as well as the rise in Θ in iteration $\#19$ has already been explained by the collapsing scenario. Now, the further drop of $\mathscr{L}$ from iteration $\#19$ to $\#20$ seems not logical at first glance. However, the origin can be traced back to the high c_N value. First of all, the c_N value is much too high for the large penetration state. This leads to a highly distorted mesh and non-physical high Lagrange multiplier values. This is actually very similar to the observations made in Chapter 5 with respect to the default standard Lagrangian system of equations. Now, due to the high c_N value, the chance is also high that at the end of iteration $\#20$ nodes with a negative averaged weighted gap value (i.e., showing a penetration and, therefore, indicating that they are active) and negative Lagrange multiplier values (i.e., contradictorily indicating that they are inactive) stay active. These nodes contribute to the active Lagrangian value (6.9) now in such a way, that they reduce the function value and are consequently the main reason for the smaller $\mathscr{L}$ value of iteration $\#20$ compared to $\#19$ in Figure 6.8a. On the other hand, since the linear system as well as the step-length had to be adapted in iteration $\#20$, the progress in terms of Θ are only marginal compared to $\#19$. When at the end of iteration $\#20$ the c_N value is now drastically reduced, such as in this case by a factor around 19, the former mentioned nodes with a negative Lagrange multiplier value and a negative averaged weighted gap value might change their status from active to inactive just because of this reduction of the regularization parameter. Actually, exactly this happens such that the cardinality of $\mathcal{A}$ changes from 1928 to 1886 without any change of the primal or dual variables. Since these nodes decreased the objective function value $\mathscr{L}$, it is no longer surprisingly that the path of the line search method ends somewhere on the right side of $\#20$. Furthermore, it also fits into the explanation that the Θ coordinate is slightly smaller since contributions of the previously active nodes are missing. To conclude the explanation it must be mentioned that a decrease of the c_N value can theoretically also lead to the opposite effect: a decrease of the objective function value. This would happen if previously inactive nodes joined the active set because of a high positive Lagrange multiplier value and small positive averaged weighted gap value. However, in this specific case the first effect is much stronger than the second one. Thus, all ingredients leading to Figure 6.8 are finally explained. This example demonstrates how complicated the interaction between the different parameters can actually be. Therefore, the importance of the modifications described here is hopefully emphasized. Many if not all of the numerical examples presented later would fail if only one of these modifications is left out.

6.7.5. Scaling of the Filter Coordinates

A scaling of the filter coordinates has already been applied in Figure 6.8 but was not mentioned so far. Such a re-scaling becomes necessary since the two coordinates defining a filter point are probably very differently scaled. While the first coordinate represents some energy or potential measure of all considered elastic bodies, the second coordinate is a pure geometric quantity representing the distance between the bodies. In the current implementation a very simple approach is followed: For all the presented examples a tangential predictor is used to obtain a meaningful initial deflection. The unscaled $\mathscr{L}$ value after this step is probably very low, especially if the contact detection is off during the predictor iteration $\#0$. Thus, it is a bad idea to use this value as a reference value for the scaling. Instead, the scaling is activated after this very first iteration. Consequently, the related scaling of the first coordinate is currently always defined as

$$\kappa_f = \frac{1}{|\mathscr{L}(\alpha^{\{1,0\}})|}. \tag{6.80}$$

However, this rule might be adapted if the very first trial point after the tangential predictor leads to a severely distorted mesh, such that $\mathscr{L}(\alpha^{\{1,0\}})$ might indicate a by far too large value. This case is excluded for now. The scaling of the second coordinate follows a quite similar approach. It is also skipped for iteration $\#0$. Furthermore, since it is possible that the follow-up iteration is a feasible iteration, an additional condition becomes necessary, viz.

$$\kappa_\Theta = \frac{1}{|\Theta(\alpha^{\{k,l\}})|} = \frac{1}{\Theta(\alpha^{\{k,l\}})} \qquad \text{if } k > 0 \text{ and } \Theta(\alpha^{\{k,l\}}) > \text{TOL}_2. \tag{6.81}$$

This scaling approach implicitly implies smaller changes to some of the already presented equations which shall be briefly addressed in the following. The first equation which slightly changes is (6.3b). It becomes

$$\kappa_f \mathscr{L}(\underline{x}(\alpha^{\{k,l\}}), \underline{\lambda}_\text{N}(\alpha^{\{k,l\}})) \leqslant \kappa_f \mathscr{L}(\underline{x}^{\{k\}}, \underline{\lambda}_\text{N}^{\{k\}}) - \gamma_f \kappa_\Theta \Theta(\underline{x}^{\{k\}}, \underline{\lambda}_\text{N}^{\{k\}}), \tag{6.82}$$

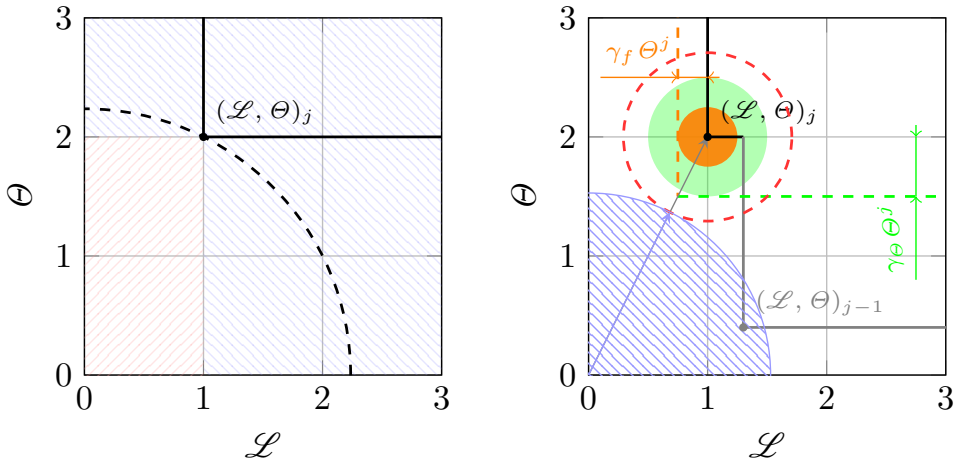
where especially the allowed margin benefits from the scaling by becoming more reliable. Note that the minimal step length estimation in Section 6.3.2 must be adapted accordingly. For instance, (6.21) becomes

$$m'_\mathscr{L}(0)\ \alpha^{\{k,l\}} + \frac{1}{2} m''_\mathscr{L}(0)\ (\alpha^{\{k,l\}})^2 + \gamma_f \frac{\kappa_\Theta}{\kappa_f} \Theta^{\{k\}} \leqslant 0. \tag{6.83}$$

Another point where the scaling plays a role is the $\mathscr{L}$-type switching condition (6.4). The second part changes to

$$(-m_\mathscr{L}(\alpha^{\{k,l\}}))^{s_f} (\alpha^{\{k,l\}})^{1-s_f} > \tilde{\nu}_\Theta (\Theta^{\{k\}})^{s_\Theta} \qquad \text{with } \tilde{\nu}_\Theta = \frac{\kappa_\Theta^{s_\Theta}}{\kappa_f^{s_f}} \nu_\Theta, \tag{6.84}$$

where all changes are hidden in a changed constant $\tilde{\nu}_\Theta$. This constant must be also inserted in the related equations for the minimal step length estimation in Section 6.3.3.



(a) (b)

Figure 6.9.: If Assumption 6.1 is fulfilled, only trial points with a smaller ℓ_2-norm than the already added filter points are able to strongly dominate one of these points. This is visualized in Figure 6.9a. However, only points in the red (bottom left to top right) hatched rectangle truly dominate the filter point l. Figure 6.9b shows an example for (6.86). This test can be used to speed-up the acceptability check, since as soon as a filter point l is found which fulfills (6.86), only the filter points with a smaller ℓ_2-norm must be considered by the point-to-point comparison.

6.7.6. Pre-Filtering

A minor implementation feature is a so-called *pre-filtering* approach which has been proposed by Gould and Toint [116] and is also applied here in Step 3.4 of Algorithm 6.2, but in a slightly different way. It has two distinct purposes: Firstly, it shall identify all points in the filter which are certainly not dominated by the new trial point. Secondly, it shall pre-identify all filter-points in the filter which would certainly accept the new trial point. To apply this pre-filtering technique the following assumption must hold:

> **AS 6.1.** All coordinates of the current trial point as well as of the filter points defining the filter are positive or equal to zero.

If Assumption 6.1 does not hold, the pre-filtering is skipped and a point-wise comparison is applied to identify dominated filter points and to test the acceptance to the filter. However, in the following it shall be assumed that Assumption 6.1 holds. Drawing the attention to the first goal: A filter point $(\mathscr{L}, \Theta)_j = (\mathscr{L}^j, \Theta^j) \in \mathcal{F}$ with $j = 1, 2, \ldots, |\mathcal{F}|$ is meant to be strongly dominated by a new trial point if all of its coordinates are larger than the coordinates of the current filter point, i.e.,

$$\mathscr{L}(\alpha^{\{k,l\}}) < \mathscr{L}^j \qquad \text{and} \qquad \Theta(\alpha^{\{k,l\}}) < \Theta^j. \tag{6.85}$$

If (6.85) holds, the old filter point j does no longer contribute any meaningful information, such that it can be removed. Now, while the strong dominance can only be checked by looking at the coordinates itself, it is possible to efficiently test the opposite, i.e., whether an already added filter point is certainly not dominated by a new trial point. Therefore, it is sufficient to look at

the ℓ_2-norm of such a point. If the ℓ_2-norm of the trial point is larger than the ℓ_2-norm of a point currently in the filter, the already added filter point can not be dominated by the new one. However, it is not true that a new trial point dominates all filter points in the filter, if it has a smaller ℓ_2-norm than these points. This fact is visualized in Figure 6.9a: All points with a ℓ_2-norm smaller than the filter point $(\mathscr{L},\ \Theta)_j = (1, 2)$ lie inside the dashed circle, but only points in the red hatched rectangle can strongly dominate the filter point with the index l. Note that it is also not possible if the ℓ_2-norm is replaced by the ℓ_∞-norm. To show this, a new trial point with the coordinates $(1.9, 1)$ shall be considered. Even though the trial point has a smaller ℓ_∞-norm as the point j in Figure 6.9a, since $1.9 < 2.0$, it still does not dominate the filter point j.

Now, to make this test efficient the filter points already in the filter set are stored in an ascending list with respect to the ℓ_2-norm of their entries. Next, the norm of a new trial filter point is compared to the norms of the already added filter points in the list and only if the norm of the trial point is smaller than the norm of one point in the filter, it is tested if this filter point and all subsequent points with a larger norm are potentially dominated by the new trial point. This is the first part of the pre-filtering approach. The second part is based on the inequality

$$\|(\mathscr{L}(\alpha^{\{k,l\}}),\ \Theta(\alpha^{\{k,l\}}))^{\mathrm{T}}\|_2 < \|(\mathscr{L},\Theta)_j^{\mathrm{T}}\|_2 - \sqrt{2}\ \max\{\gamma_f,\ \gamma_\Theta\}\ \Theta^j, \tag{6.86}$$

which can be easily derived under consideration of (6.3). To obtain (6.86), the ℓ_2-norm of the left and right side of (6.3) are taken, such that

$$\begin{aligned}\|(\mathscr{L}(\alpha^{\{k,l\}}),\ \Theta(\alpha^{\{k,l\}}))^{\mathrm{T}}\|_2 &< \|(\mathscr{L} - \gamma_f\Theta, (1-\gamma_\Theta)\Theta)_j^{\mathrm{T}}\|_2 \\ &\leqslant |\|(\mathscr{L},\Theta)_j^{\mathrm{T}}\|_2 - \|(\gamma_f,\gamma_\Theta)^{\mathrm{T}}\|_2\Theta^j| \\ &\leqslant \|(\mathscr{L},\Theta)_j^{\mathrm{T}}\|_2 - \sqrt{2}\ \max\{\gamma_f,\ \gamma_\Theta\}\ \Theta^j,\end{aligned} \tag{6.87}$$

where the last step becomes possible under the assumption that $\gamma_\Theta, \gamma_f < 1/\sqrt{2}$ holds. Now, still under the assumption that the algorithm loops over all filter points contained in the filter $\mathcal{F}$ in ascending order with respect to their ℓ_2-norms, the filter point j, which fulfills (6.86) for the first time, defines the lower bound of filter points which will certainly accept the new trial point. Therefore, only the first $j-1$ filter points must be tested by point-to-point comparison of their coordinates against the new trial point. The reader is kindly referred to Figure 6.9b, where this test is visualized. A look at this figure reveals that filter points with a smaller ℓ_2-norm are still able to block a new trial point lying in the blue shaded quadrant. See filter point $j-1$ for example.

A final remark: This pre-filtering approach might seem unnecessarily complicated due to the small number of only two coordinates, but it has been implemented with a possible extension of filter point coordinates in mind. One scenario are frictional contact problems. In addition, the framework allows to define many more meaningful coordinates based on the entries of the residual vector. For examples and possible applications the interested reader is referred to Gould and Toint [116] or Milzarek and Ulbrich [196].

6.8. Special Extensions for Structural Contact Problems

Up to this point, the description has been limited to a quasi-static frictionless contact formulation based on unmodified HEX8 finite elements. In this section two extensions shall be presented which are often necessary for real world applications: First, the treatment of dynamic contact problems with the line search filter method is addressed. Secondly, the EAS formulation will be considered and possible implications will be identified and discussed.

6.8.1. Dynamic Problems

To be more specific, the attention is on necessary adaptions for the treatment of dynamic structural problems under consideration of the Generalized-α time integration scheme. Other schemes are not considered. However, the adaptions would probably follow a quite similar pattern. The discussion starts where Section 4.6 has ended. By investigating the considered system of equations, where the focus lies on the residual (4.64a). To keep things simple any damping effects as well as the augmented regularization are excluded from the following discussion. As already mentioned, there exist at least two possibilities to evaluate the different terms at the respective mid-points $t_{n+1-\alpha_\mathrm{f}}$ and $t_{n+1-\alpha_\mathrm{m}}$, where the variant implemented here is based on the *trapezoidal rule*. The used residual is given in (4.65), which is also consistently linearized as presented in (4.68). Consequently, the solution is reached as soon as $\underline{r}_{\mathrm{cg}\alpha} = \underline{0}$ holds and the remaining KKT-conditions (4.67) are fulfilled as well. On the other hand, however, coming from a theoretical view point, the solution path is expected to follow Hamilton's principle, i.e.,

$$\delta \int_{t_n}^{t_{n+1}} \mathscr{U}(\underline{d}(t)) + \mathscr{V}_{\mathrm{ext}}(\underline{d}(t)) - \mathscr{K}(\underline{v}(t)) \mathrm{d}t = 0, \tag{6.88}$$

where the kinetic energy $\mathscr{K}$ has been defined in (2.36). In words: The total variation of the action integral shall vanish along the solution path through time and space. Following the presented derivation steps in Section 2.1.4 concerning the Lagrangian field theory, it is obvious that (4.65) indeed represents an approximation of (6.88) with all the issues mentioned in Section 4.6. However, with respect to the filter method discussed here one important question is still open: What is a meaningful representative for the first filter point coordinate? During the previous discussion the (quasi-static) Lagrangian functional has been used. This is no longer possible since it does not consider any time dependency and, furthermore, its derivative with respect to the displacements does not lead to (4.65). Consequently, it can not be used as a reliable objective function value. On the other hand, (6.88) would provide a theoretical possibility, but the involved time integral is not really suitable. Furthermore, the question remains how to choose the correct evaluation state to be in accordance with (4.65).

Therefore, (4.65) shall be directly considered and an auxiliary time dependent Lagrangian is constructed which delivers the Generalized-α residual when its derivatives with respect to the displacements are computed. This idea yields the scalar-valued *Generalized-α Lagrangian*, viz.

$$\mathscr{L}_{g\alpha} = \frac{\beta_{g\alpha}\,(\Delta t)^2}{2(1-\alpha_m)}\left[(1-\alpha_m)\underline{a}^{n+1} + \alpha_m \underline{a}^n\right]^T \underline{\underline{M}}\,\left[(1-\alpha_m)\underline{a}^{n+1} + \alpha_m \underline{a}^n\right] \tag{6.89a}$$

$$+ (1-\alpha_f)\mathscr{U}(\underline{d}^{n+1}) + \alpha_f(\underline{d}^{n+1})^T\,\nabla_{\underline{d}}\mathscr{U}|_{\underline{d}^n} \tag{6.89b}$$

$$+ (1-\alpha_f)\mathscr{V}_{ext}(\underline{d}^{n+1}) + \alpha_f(\underline{d}^{n+1})^T\,\nabla_{\underline{d}}\mathscr{V}_{ext}|_{\underline{d}^n} \tag{6.89c}$$

$$- (1-\alpha_f)\langle \underline{\tilde{g}}_N(\underline{d}^{n+1}), \underline{\lambda}_N^{n+1}\rangle - \alpha_f(\underline{d}^{n+1})^T\,\nabla_{\underline{d}}\underline{\tilde{g}}_N\Big|_{\underline{d}^n}\,\underline{\lambda}_N^n, \tag{6.89d}$$

where $\alpha_m, \alpha_f, \beta_{g\alpha}$ are the Generalized-α parameters and Δt denotes the discrete time step (see Section 2.4 for more information). The function (6.89) is one representative for a possible objective function. Actually, there exists an infinite number of suitable Lagrangians, e.g., the explicitly used displacement vector $(\underline{d}^{n+1})$ in front of the gradients, which depends on the previously converged state, can be also replaced by $(\underline{d}^{n+1} - \underline{d}^n)$. This would only lead to a constant shift of the objective function values throughout one time step and, therefore, it would only marginally influence the filter method. One point would be the used scaling factors (see Section 6.7.5). However, in this thesis solely (6.89) shall be considered. The proof that (6.89) is indeed a suitable objective function for (4.65) can be given by direct calculation. While the necessary derivatives of (6.89b) to (6.89d) obviously lead to the gradients in (4.65b) to (4.65d), the correctness of the first term can be validated under consideration of (2.84b). The consideration of a damping term seems possible in a similar straight forward manner.

6.8.2. Handling of Enhanced Assumed Strains

The idea of enhanced assumed strains has been briefly introduced in Section 2.3.4. It is one possible way to avoid certain kinds of locking, see Section 2.3.3. During the brief derivation it becomes obvious that crucial points are the condensation and post-processing steps described in (2.82) and (2.83). For a classical local Newton approach these steps are completely unproblematic. The algorithm will condensate the matrix and right hand side contributions on element level, afterwards the global system is assembled, the Dirichilet boundary conditions are applied and the system is solved. As soon as the solution vector containing $(\underline{d}, \underline{\lambda}_N)_{\{k\}}$ is updated, the recovery of the enhanced strain increment as well as the update of the enhanced strains in each element will be executed in a post-operation. Thus, everything stays simple.

Now, let us consider the line search case: As long as the step length is not modified the previous approach stays valid. However, if the step length is modified in Step 3.12, the procedure will slightly change. The computation of the trial point in Step 3.2 takes place in the same method as the default update for a classical Newton approach. The ingredient that changes is the step length which is now smaller than the default step length. This will become important for the enhanced strain update. Again, the post-operation in each element is executed, but this time, instead of recomputing the EAS increment and updating the internal stored strains, the routine

$$\underline{\tilde{\alpha}}_{eas}(\alpha^{\{k,l\}}) = \underline{\tilde{\alpha}}_{eas}(\alpha^{\{k,l-1\}}) + (\alpha^{\{k,l\}} - \alpha^{\{k,l-1\}})\,\Delta\underline{\tilde{\alpha}}_{eas}^{\{k\}}, \qquad \forall\, l \in \{1, 2, \dots\} \tag{6.90}$$

is applied, where $\underline{\tilde{\alpha}}_{eas}(\alpha^{\{k,l\}})$ denotes the enhanced trial strains for the current step length $\alpha^{\{k,l\}}$. The alternative would obviously be

$$\underline{\tilde{\alpha}}_{\text{eas}}(\alpha^{\{k,l\}}) = \underline{\tilde{\alpha}}_{\text{eas}}^{\{k-1\}} + \alpha^{\{k,l\}}\ \Delta\underline{\tilde{\alpha}}_{\text{eas}}^{\{k\}} \tag{6.91}$$

as it is done for the global state vectors. From a pure mathematical perspective there is no difference between (6.90) and (6.91). However, from a numerical point of view both approaches are different. First of all, the latter one is numerically more stable than the first one since it is hardly affected by any round-off or cancellation errors. But, (6.91) asks for the additional storage of the previous enhanced strain vector $\underline{\tilde{\alpha}}_{\text{eas}}(\alpha^{\{k-1\}})$ and since the element does not know if a line search step is expected or not, due to the fact that the decision is solely based on the comparison of current step length information $\alpha^{\{k,l\}}$ and $\alpha^{\{k,0\}}$, the old EAS state would have to be stored for each element in any scenario. To avoid this, (6.90) is considered, instead. As already mentioned, this approach might suffer from cancellation and round-off errors. This negative effect depends on many factors such as the ratio between $\underline{\tilde{\alpha}}_{\text{eas}}^{\{k-1\}}$ and $\Delta\underline{\tilde{\alpha}}_{\text{eas}}^{\{k\}}$ for example and has never become severe throughout any considered numerical experiment such that (6.90) is used for the line search trial point update of the internally stored enhanced strains.

Now, there is another point which is only necessary for Algorithm 6.2 and needs to be addressed here. The first hint for this additional adaptions can be found in Step 3.1: In this step the creation of a backup of internally stored variables is claimed. Why is this necessary? The answer is given by the second order correction step. Let us assume that the first trial point based on the default step is not accepted by the filter, i.e., the pre-testing was successful since otherwise the entire filter including the SOC step would have been directly skipped. In such a case the SOC is initialized in Step 3.6. If the SOC system could be successfully solved, the state variables are modified in Step 3.8, where the original trial point is augmented by the SOC direction (see Section 6.4 for more information). However, while the original state directions $(\Delta\underline{d}^{\{k\}}, \Delta\underline{\lambda}_{\text{N}}^{\{k\}})$ are stored in vectors different from $(\Delta\underline{d}^{\text{SOC}}, \Delta\underline{\lambda}_{\text{N}}^{\text{SOC}})$, this might not be automatically true for $\Delta\alpha_{\text{eas}}^{\{k\}}$ which are stored element-wise. Therefore, the back-up routine is called which stores a copy of $\underline{\tilde{\alpha}}_{\text{eas}}^{\{k-1\}}$ and a copy of $\Delta\underline{\tilde{\alpha}}_{\text{eas}}^{\{k\}}$. One might wonder since this is actually something which had been tried to be avoided by (6.90). However, in contrast to the EAS update, this back-up is really only evoked if a line search filter method is used. In all other cases no extra storage is needed.

Back to the SOC step: If the second order correction fails in Step 3.7, Step 3.3 or during the filter acceptability tests, the $\underline{\tilde{\alpha}}_{\text{eas}}(\alpha^{\{k,0\}})$ is replaced by the backup value of $\underline{\tilde{\alpha}}_{\text{eas}}^{\{k-1\}}$ and $\Delta\underline{\tilde{\alpha}}_{\text{eas}}^{\{k\}}$ is replaced by its stored backup counterpart, respectively. Since the algorithm jumps afterwards directly to the step length adaption, a final ingredient must be added to avoid an incorrect update in (6.90) for $l = 1$. If a recovery from the backup was necessary, the previous scalar valued step length parameter $\alpha^{\{k,0\}}$, which is stored in each element, is set to zero, thus the update for $l = 1$ will lead to the correct result. The presented way of handling additional EAS degrees of freedom will be applied to several examples in Section 6.10.

6.9. Final Practical Considerations

Finally, some concluding practical remarks follow before the numerical examples are considered. Therefore, numerical issues due to cancellation errors, the applied parameter sets or performance issues in parallel on high-performance computing (HPC) systems shall be discussed more deeply.

6.9.1. Numerical Issues

The treatment of numerical issues such as round-off errors must be discussed as one of the final implementation details. The loss of significant digits during floating point operations is a well-known issue which can become cumbersome close to solution points for the globalization strategy discussed here. If it is not addressed appropriately, it can happen that the fast local convergence is avoided by the line search method since the used checks, either based on the Lagrangian function value or the infeasibility measure, might indicate a slightly rising value even though the step is a very good step. Furthermore, if the origin are rather cancellation errors than influences of the Maratos effect, it is not guaranteed that a second order correction step resolves the problem. Fortunately, two very simple modifications help to avoid this problem completely:

- The first modification is already part of the line search algorithm and is based on the fact that feasible points shall not be used to augment the filter set. In case of a constrained optimization problem this can become significantly important. For the actual implementation a possible scaling as introduced in Section 6.7.5 should not be forgotten, i.e., a point is called feasible as soon as

$$\kappa_\Theta \Theta^{\{k\}} < \kappa_\Theta \mathrm{TOL}_2 \tag{6.92}$$

  holds.

- The first modification is able to avoid that the filter set becomes too restrictive. However, it can not avoid that good iterates are blocked by the sufficient decrease checks or the Armijo-rule in case of a $\mathscr{L}$-type step. Therefore, a second modification is applied based on the residual norm check. Throughout this thesis the structural force/gradient residual norm as presented in (6.57) is tested first for the default step length, i.e., $\alpha^{\{k,0\}}$, if this single convergence criterion is satisfied, the line search method is skipped and the filter method is not entered, instead, the step is directly accepted. This works very well in all considered cases what might be owed to the fact that a rather weak absolute value for TOL_1 about $1.0\mathrm{e}{-6}$ is used compared to the other tolerances. This strategy is recommended due to its simplicity and success in all performed experiments.

Other similar adaptions are based on the machine precision and can be found in Wächter and Biegler [272, Sec. 3.10], for instance.

6.9.2. Parameter Sets

Most if not all of the provided globalization algorithms rely on sets of user specified parameters. These sets are quite overwhelming at the beginning and might indicate that the provided methods can not be used in daily practice without being an expert on the topic. Fortunately, this is not true. In the following a set of meaningful parameter choices will be presented which have been used for almost all of the provided numerical examples. The only sub-set of parameters which might ask for adaptions are the ones concerning the filter reinitialization presented in Section 6.7.3.

	$\|\nabla_{\underline{d}}\mathscr{U} - \tilde{\nabla}_{\underline{d}}\tilde{g}_{\mathrm{N}}^{\mathcal{A}}\underline{\lambda}_{\mathrm{N}}^{\mathcal{A}}\|, i=1$	$\Theta, i=2$	$\|\Delta\underline{d}\|, i=3$	$\|\Delta\underline{\lambda}_{\mathrm{N}}\|, i=4$
norm-type	ℓ_2, absolute	ℓ_2, absolute	ℓ_2, absolute	ℓ_2, relative
TOL_i	1.0e−6	1.0e−10	1.0e−10	1.0e−10

Table 6.1.: Tolerances and norm types for the global convergence tests.

γ_f	γ_Θ	γ_α	s_f	s_Θ	SOC-type
1.0e−6	1.0e−6	0.05	2.3	1.1	automatic

Table 6.2.: Basic default parameters for Algorithm 6.2.

If the parameters differ from the default sets presented here, the changes will be named in the corresponding example section.

All tolerances for the global convergence tests are listed in Table 6.1. Note that these tolerances do not only specify when the solution is reached but also decide whether an iterate is feasible or not. Furthermore, TOL_1 is used to skip the filter method. See Section 6.9.1 for more information.

The basic default parameters for Algorithm 6.2 can be found in Table 6.2. These parameters are taken from [272] and work very well in practice. Note that the condition $s_f > 2s_\Theta$ is fulfilled by the proposed choice. This condition is important for the local convergence behavior as described in [270]. Another important parameter is $\Theta_{\min}$ which is used to by-pass the $\mathscr{L}$-type switching condition during the pre-asymptotic phase (see Section 6.7.2). This parameter is set to

$$\Theta_{\min} = 10^{-4}\max\{1.0, \Theta^{\{1\}}\}, \tag{6.93}$$

where $\Theta^{\{1\}}$ denotes the unscaled infeasibility violation in the first Newton iteration. If the infeasibility violation is zero in the first Newton iteration or the value of $\Theta^{\{1\}}$ is smaller than one, the value of $\Theta_{\min}$ will be set to 1.0e−4. The Armijo rule uses the typical parameters for a line search method based on Newton directions, i.e., the slope parameter c_1 is set to 1.0e−4 and the step length reduction parameter to $\beta = 0.5$, see also Algorithm 3.1.

Now, the parameters for the correction of the upper left block in (6.59) as part of Algorithm 6.3 are presented in Table 6.3. Most of these parameters have again been taken from [272]. Note that Algorithm 6.3 always starts with a reduction equal to κ_ω^- before the correction parameter is increased, i.e., the actual increase considered first is limited to $\kappa_\omega^-\kappa_\omega^+ = 8/3 \approx 2.667$.

The same volume change parameters as in Table 6.3 are also applied to the pre-testing proposed in 6.7.1, i.e., $r_{\min}^{\mathrm{pre}}$ is set to 0.5 and $r_{\max}^{\mathrm{pre}}$ to 2.0. For more information the reader is kindly referred to Section 6.7.1.

$r_{\min}$	$r_{\max}$	δ_ω	δ_ε	ω_0	$\omega_{\min}$	$\omega_{\max}$	κ_ω^-	κ_ω^+	κ_ω^{++}	N_ω
0.5	2.0	1.0	$10\cdot\varepsilon_{\mathrm{mach}}$	1.0e−4	1.0e−20	1.0e40	1/3	8	100	3

Table 6.3.: Default parameters for Algorithm 6.3. Note that the machine precision $\varepsilon_{\mathrm{mach}}$ depends on the used data type. Throughout this thesis all examples have been computed with double precision.

$\Theta_{\max}$ (scaled)	$\gamma_{\Theta}^{\max}$	$n_{\text{newton}}^{\text{block}}$	$n_{\text{ls}}^{\text{block}}$
2.0	0.25	4	7

Table 6.4.: Default parameter set for the reinitialization strategy proposed in Section 6.7.3.

Another set of open parameters are $\beta_{\Theta}^{c_{\text{N}}}$ as part of the constraint modification discussed in Chapter 5 and $\beta_{\Theta_{\text{crit}}}^{c_{\text{N}}}$ introduced in Section 6.7.4. For all the examples discussed in this chapter that will actually use the modified constraints approach, the parameter $\beta_{\Theta}^{c_{\text{N}}}$ has been set to 0.8, while $\beta_{\Theta_{\text{crit}}}^{c_{\text{N}}}$ has been set to 0.95.

Finally the last set of parameters is associated with the blocking criteria and the possible reinitialization of the filter set as discussed in Section 6.7.3. These are the only parameters which had to be modified for some of the examples. The default choice is presented in Table 6.4.

6.9.3. Parallel Redistribution

Reliability and robustness play an important role for the applicability of the presented methods, but it is also very important that the methods are efficient. A bottleneck for the methods considered here is clearly the evaluation of the derivatives associated to the contact terms. Especially, the necessary second order derivatives can be very expensive. Besides the fact that there is still plenty of room for improvements in the actual implementation of these evaluation procedures, there is the hope that a higher number of working horses represented by the used processors and enough physical memory can help to significantly reduce the computation time. It is obvious that a meaningful parallel treatment asks for a meaningful parallel distribution. This task can be accomplished with a number of well-established packages for the pure structural part. For example, in this thesis the load balancing and partitioning is achieved with the Zoltan toolkit, see Boman et al. [29] for an introduction. This approach works very well for the bulk material and thus it is often sufficient to find a meaningful distribution of the volume finite elements once at the beginning. The crux of this problem is to find a distribution which minimizes the necessary communication effort and redundancy of element evaluations, also known as ghosting, while simultaneously leading to a well balanced utilization of available resources.

However, in case of contact problems, it must be dealt with another problem that is directly related to the contact interaction among the contacting bodies. This redistribution issue has already been discussed in Popp [213, Sec. 4.6.1]. In the following, an algorithm will be presented which is based on this original work but contains important enhancements which are critical for the considered examples. But, before the new ingredients will be discussed, the basic idea shall be quickly recapitulated: First, a simple and naive distribution of the entire contact interface over all available processors would lead to a bad performance, since the specific differences between slave and master would stay unaccounted. Actually, the master elements have only a passive role when it comes to the evaluation of the contact contributions. In fact, they are mainly necessary for the projection and distance calculation within the (weighted) gap evaluation. The main load lies solely on the slave elements. Therefore, a meaningful processor distribution must take this into account. That is the first point. However, this issue can easily be solved at the very beginning, since the slave/master distribution does usually not change over a simulation (with the exception of self-contact [289]). Thus, this initial redistribution can be referred as a *static*

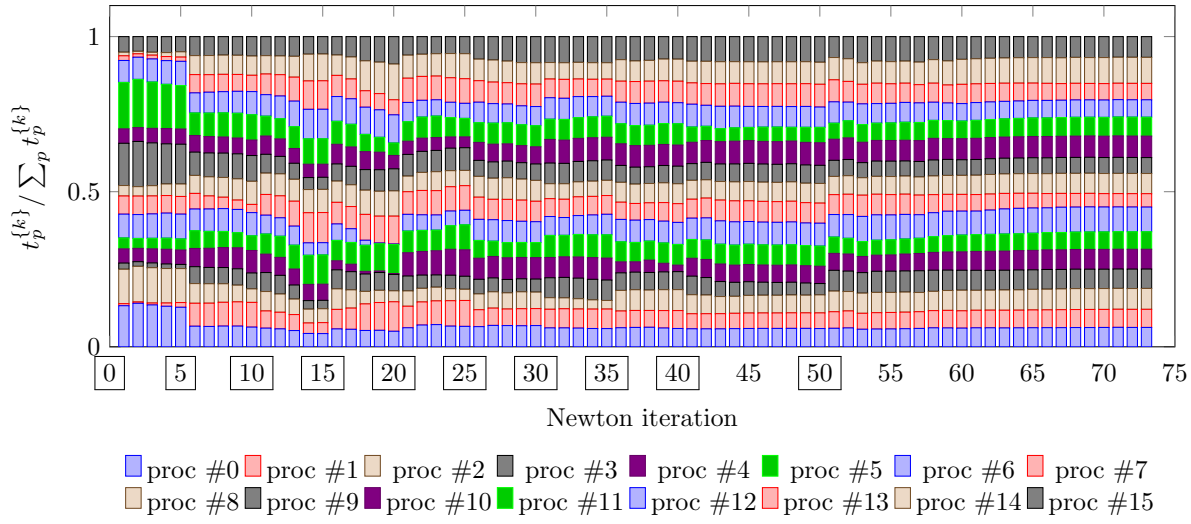
redistribution approach. The term redistribution is used since the interface distribution will generally differ from the parallel distribution of the underlying bulk elements, i.e., for the assembly of the final system matrix and right-hand-side vector additional communication is unavoidable. However, this communications can be performed very efficiently with the methods provided in the Epetra package as part of the Trilinos library [127].

Now, a simple uniform distribution of the processors over the slave side is still not sufficient to reach an efficient load balancing. The reason is that not all slave elements participate at a contact scenario. In fact, the slave and master elements define only the potential contact zone. Therefore, an efficient redistribution approach must also address this circumstance. This is the point at which the approach presented and applied in [213] and the strategy considered here start to diverge. In common is the fact that a *dynamic redistribution* will only take place if the ratio of the maximum and minimum processors time over all individual time measurements $t_p^{\{k\}}$ with $p \in \{0, 1, \ldots, N_\mathrm{p} - 1\}$, where N_p is the number of available processor cores, exceeds a pre-defined threshold. This can be summarized as

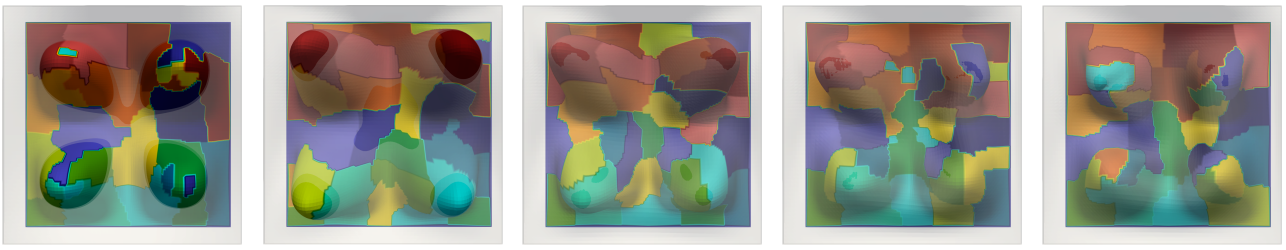
$$\frac{\max\limits_p \{t_p^{\{k\}}\}}{\min\limits_p \{t_p^{\{k\}}\}} > \mathcal{B}_\mathrm{proc}, \tag{6.94}$$

where $\mathcal{B}_\mathrm{proc} > 1$. Throughout this thesis the bound $\mathcal{B}_\mathrm{proc}$ has been set to 2.0, i.e., a redistribution is only initiated if there is a processor core which takes more than twice as long as another core, which indicates a clearly bad load balancing. However, it should be mentioned that the redistribution itself is also a quite expensive operation and, therefore, it should not be executed more often than absolutely necessary. As always, it is a matter of finding the best compromise. The difference between this thesis and [213] lies firstly in the measurement of the individual processor times $t_p^{\{k\}}$ in each Newton iteration k. While Popp [213] uses more of a global time measurement, a more locally restricted time measurement is used in this thesis. "Locally" refers hereby to the location in the source code. Here, only the accumulated time passed during the integration over the individual slave elements is considered. In this way it shall be ensured that the time measurements are reliable and are not disturbed by any internal communication calls among the processors. Secondly, the definition of the near and far field is severely different in this thesis. Near and far field relates to an additional split of the slave element sets. The near field contains all slave elements which actually contribute to the current evaluation, while the far field contains the remaining slave elements. For example, all elements without a valid projection. The definition of these two sets relied on the number of detected contact partners in the work of Popp [213] and was heavily based on the search algorithm and the therein set parameters [288]. This is surely a meaningful approach as long as the search radius is sufficiently small. However, in case of the examples discussed here, the search radius is often set to a quite large value since otherwise the contact detection for large initial penetration would fail. This has already been mentioned in Section 5.6 and is even more true for this chapter. Therefore, another strategy must be applied.

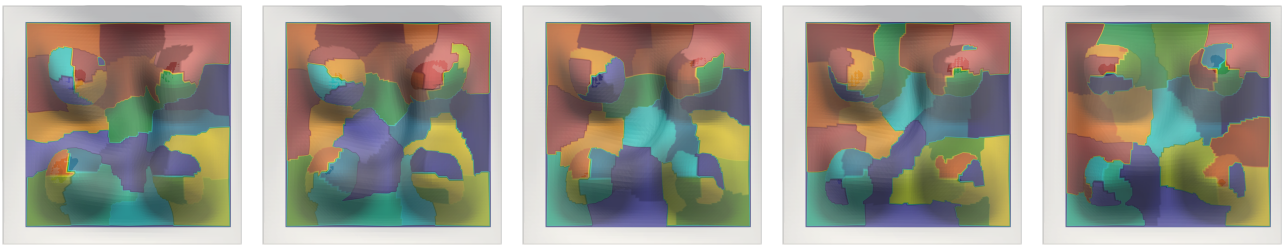
The novel simple idea is to measure the evaluation times of each individual slave element. Then the global maximum time spent for one single element is detected and all slave elements which spent at least a sufficient percentage of the maximum evaluation time are inserted into



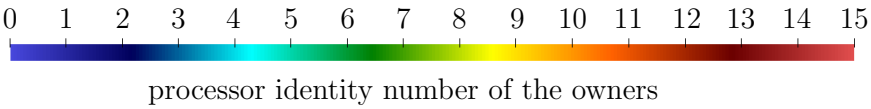
(a)



(b) after iteration #0 (c) after iteration #5 (d) after iteration #10 (e) after iteration #15 (f) after iteration #20



(g) after iteration #25 (h) after iteration #30 (i) after iteration #35 (j) after iteration #40 (k) after iteration #50



(l)

Figure 6.10.: Here, the effectiveness of the applied parallel redistribution approach is demonstrated for the upcoming example from Section 6.10.5. Figure 6.10a shows the relative evaluation times $t_p^{\{k\}} / \sum_p t_p^{\{k\}}$ of each participating processor $p \in \{0, 1, \ldots, 15\}$ with respect to the total evaluation time of the respective Newton iteration $k \in \{0, 1, \ldots, 73\}$. A parallel redistribution took place subsequently to each of the boxed iterations. The corresponding ownership distribution of the slave elements is presented in Figures 6.10b to 6.10k.

the near field set. All slave elements which stayed a shorter period in the integration methods will be moved to the far field. Afterwards, the individual sets, including the set of all master elements, are distributed evenly across all available processor cores. Thereby, each set is treated independently. In this way, the following sets can be defined

$$\mathcal{E}_{\mathrm{elb}}^{\mathrm{nf}} = \bigcup_{p} \{ e \in \mathcal{E}^{\mathcal{S}} \mid t_p^{(e)\{k\}} > \gamma_{\mathrm{proc}} \max_{p,e} \{ t_p^{(e)\{k\}} \} \}, \qquad \mathcal{E}_{\mathrm{elb}}^{\mathrm{ff}} = \mathcal{E}^{\mathcal{S}} \backslash \mathcal{E}_{\mathrm{elb}}^{\mathrm{nf}} \tag{6.95}$$

where elb is the abbreviation for element load balancing, while nf and ff denote near and far field, respectively. The set $\mathcal{E}^{\mathcal{S}}$ is simply the set of all slave elements. Just similar to the set of all slave nodes which is known to be defined as $\mathcal{S}$. The parameter γ_{proc} has been set constantly to 0.1 during all numerical examples, i.e., only slave elements which took less than 10% of the time with respect to the "slowest" slave element are moved to the far field. Furthermore, the original criterion based on the number of contact search partners is applied at the very beginning of a new load/time step, as well as whenever no valid evaluation times are at hand. Finally, an interval in terms of Newton iterations must be defined. After each of these intervals, which is set to 5 Newton iterations for example, (6.94) is checked and if the criterion is satisfied, the contact elements will be redistributed. The bad quality of the pure contact search based distribution in case of a very large search radius can be seen in Figure 6.10a, where the exemplary results for the example of Section 6.10.5 are presented. After every fifth Newton iteration, the evaluation times are checked. As a result, the evaluation times are far more evenly spread over the processors after iteration 5, for instance. However, the presented example undergoes very large deformations such that the parallel redistribution must take place several more times until a state of almost constant load distribution over all 16 cores is achieved (see Figure 6.10a). The last necessary redistribution takes place in iteration 50. The actual ownership of the individual slave elements after each redistribution can be seen in Figures 6.10b to 6.10k. At the end of the simulation only the tips of the sine waved membrane will be in contact. The gradual localization of the processors around these four tips is quite obvious in these figures.

The parallel redistribution approach presented here is not crucial for the simulation itself, but it is of major importance if the available resources shall be used as efficiently as possible. All of the following large examples, such that a parallel computation makes sense, will use the presented strategies.

6.10. Numerical Examples

In this section examples are provided which demonstrate the superior performance of the new globalization strategy. Furthermore, each of the following examples is supposed to focus on another feature of the algorithm such that the necessity of the introduced adaptions becomes clearer. All numerical examples rely on a self-implemented non-linear solver framework as part of Baci [274]. The framework on its own is based on the NOX package (see Heroux et al. [127]).

6.10.1. Pair of Plates

The first example demonstrates that the filter method is not unnecessarily restrictive. Therefore, an example is chosen which might seem simple at first glance but is hard to solve with the

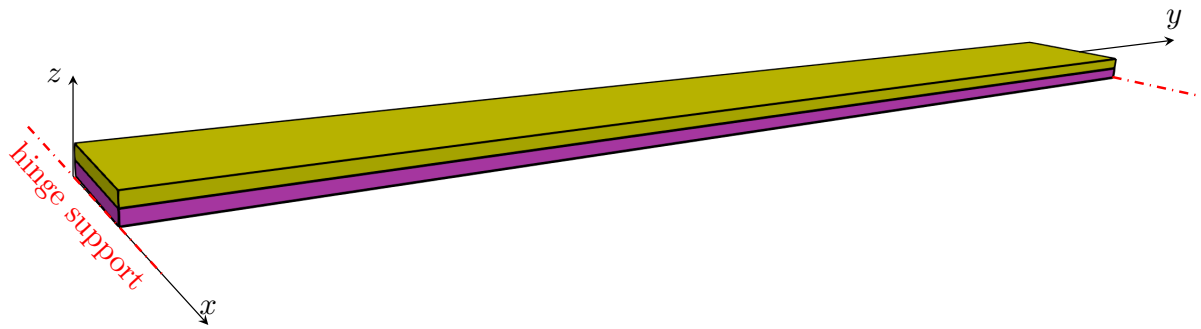
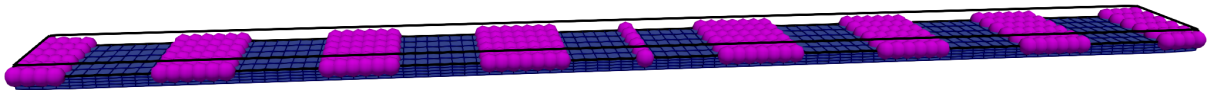


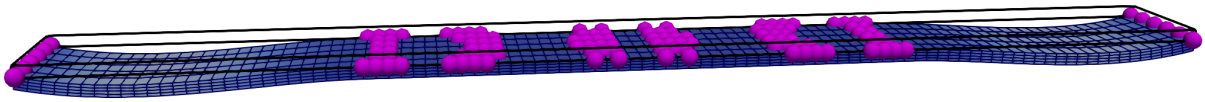
Figure 6.11.: Geometrical configuration of the pair of plates example. The red dash dotted lines indicate the hinge joints which are used to mount the lower plate. The example is taken from Miyamura et al. [197].

proposed active-set semi-smooth Newton method. The difficult part is the identification of the correct active-set. This issue has been already described by Miyamura et al. [197] where an interior point method, a semi-smooth Newton method as well as a newly proposed combination of both approaches have been compared to each other. They only consider linear elasticity and a node-to-node contact formulation, nevertheless, their results stay valid. The geometrical configuration is illustrated in Figure 6.11. The dimensions of the two plates are $1 \times 10 \times 0.1$ (width $\times$ length $\times$ thickness). Initially, both plates lie on top of each other as demonstrated in Figure 6.11. Furthermore, the lower plate is completely fixed on its two lower edges pointing in x-direction, i.e., also the displacements in x-direction are completely restrained. Consequently, only a rotational degree of freedom around the hinge support is possible. The Young's modulus is $E = 6.0$ and the Poisson's ratio is set to $\nu = 0.27$ [197]. Again, the simple coupled form of the compressible neo-Hookean material model is considered as defined in (2.26), see also [136, p. 247]. Now, a displacement controlled loading is applied. All degrees of freedom of nodes on the top surface are fixed in $x-$ and $y-$direction and are additionally applied with a prescribed shift of 0.1 units in negative z-direction. Finally, the contact condition needs to be specified: The top surface of the lower plate is the slave surface, while the bottom surface of the upper plate is the master surface. Note that this choice is not arbitrary. The reason is given in Remark 6.3. Another important point is that all slave nodes are initially declared as active. Otherwise, numerical round-off errors switch some nodes active and some inactive which would lead to a non-symmetric active set distribution in subsequent iterations due to the non-deterministic initial state. Now, with all these ingredients at hand the simulation can be started for two mesh options proposed by Miyamura et al. [197]. The first mesh variant considers 5 elements in x-direction, 100 in y-direction and 4 in z-direction. The second variant increases the element number in y-direction to 200. However, these choices lead to a bad aspect ratio of the HEX8 elements, such that the EAS-21 formulation is applied here as well, see also Andelfinger and Ramm [3] for more information.

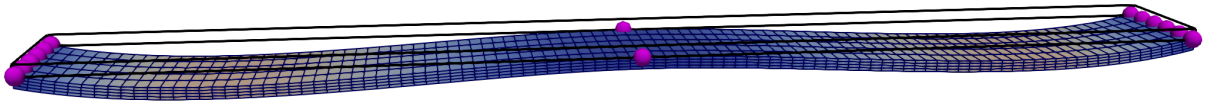
A selection of five iterations for the coarser mesh is presented in Figure 6.12. The first 38 iterations are necessary to find the correct active set. In these iterations, the active nodal distribution propagates somehow in waves towards the center while the lower plate starts to peel off from the top plate. This behavior is demonstrated in the Figures 6.12a to 6.12c. In iteration $\#38$, the correct active set is identified, but the lower plate still shows this sine-shaped deformation. Finally, in iteration $\#39$, the structure starts to move towards its final state. Note that iteration $\#39$ is the only one which activates the line search leading to a reduction of the step-length by



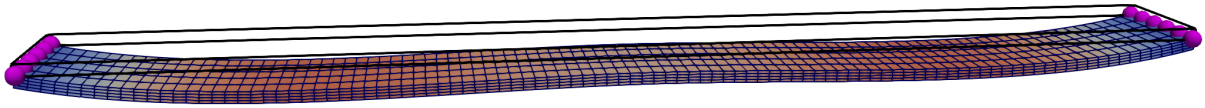
(a) $|\mathcal{A}^{\{0\}}|/|\mathcal{S}| = 282/606$ (predictor)



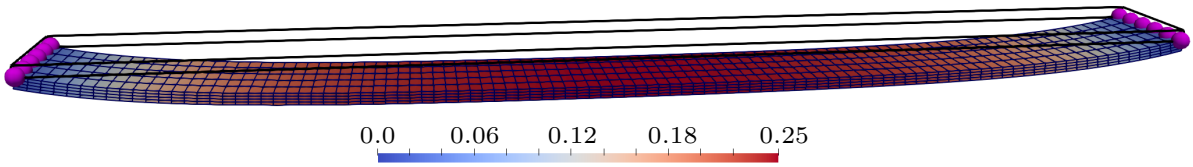
(b) $|\mathcal{A}^{\{14\}}|/|\mathcal{S}| = 124/606$



(c) $|\mathcal{A}^{\{37\}}|/|\mathcal{S}| = 14/606$



(d) $|\mathcal{A}^{\{39\}}|/|\mathcal{S}| = 12/606$



(e) $|\mathcal{A}^{\{46\}}|/|\mathcal{S}| = 12/606$ (converged)

Figure 6.12.: Figures 6.12a to 6.12d visualize four non-equilibrium displacement states corresponding to different Newton iterations. In Figure 6.12e the converged state is shown. The active nodes are highlighted as well. Note that the shown deformation state in Figure 6.12d already includes the line search correction.

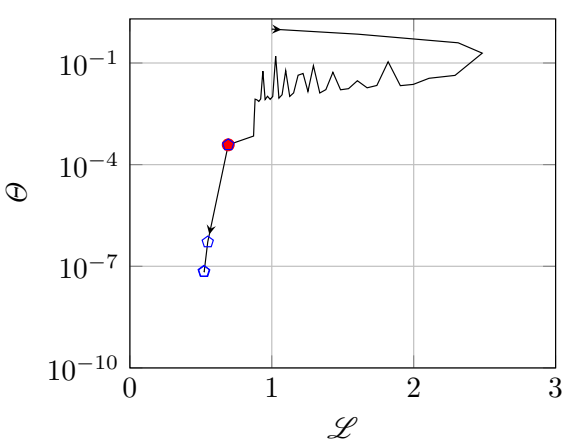
0.5. All previous and subsequent iterations are accepted by the filter method without modifying the step-length. This is a notable fact which would be hardly achievable with a merit function combining both goals. For example in Miyamura et al. [197], a simple merit function is used. However, to achieve the minimal reported iteration number, the minimal step length must be restricted to a lower limit of 0.2 for this example, i.e., the step length is chosen as 0.2 if the line search indicates a smaller step-length. Consequently, no monotonically decreasing solution path is followed with respect to the merit function. This heuristic adaption is not necessary in case of the filter method considered here.

To demonstrate this superior behavior, the solution paths in the filter space are visualized in Figure 6.13. Even though these paths look very scattered, all of them are valid in the sense of the filter approach. Three different settings have been investigated. In Figures 6.13a and 6.13b the coarser mesh is considered. Figure 6.13a belongs to a simulation where the SIR-update has been used during the initial pre-asymptotic phase. Interestingly, the switching condition is never satisfied during the peeling phase, i.e., from iteration 0 to 38. But right after this phase, at the beginning of iteration 39, the method switches to the standard Lagrangian formulation. This impressively underlines the meaning and importance of these switching conditions. Following this strategy the method manages to achieve the solution in 46 iterations. In contrast, if the standard Lagrangian system matrix with a constant c_{N} value equal to 1.0 is taken into account, the solution path presented in Figure 6.13b is followed. Again, only one line search adaption is necessary. Furthermore, the entire path indicates a monotonically decreasing first filter coordinate, i.e., the Lagrangian function value is reduced from iteration to iteration. The drawback is that it takes longer to reach the solution. In total, 60 Newton iterations are necessary.

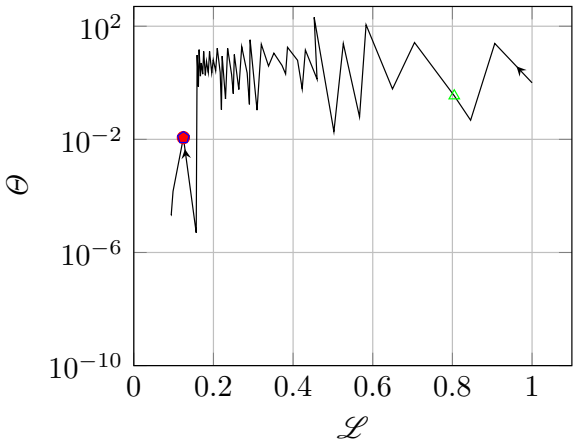
Finally, the finer mesh is considered. This time only the combined method has been applied. The method follows the solution path presented in Figures 6.13c and 6.13d. Note that except for one iteration, all the other points belong to a full Newton iteration. Thus, the in [197] already reported observation is verified that a finer mesh results in a slower convergence in case of the semi-smooth Newton method. Precisely, 78 iterations are necessary to reach the final converged state. This is very obvious in the detailed view given in Figure 6.13d where almost no progress with respect to the constraint norm can be obtained between iteration $\#11$ and $\#69$. Only the Lagrangian function value is slowly decreasing.

Furthermore, some points corresponding to special parts of Algorithm 6.2 are marked in Figure 6.13. For instance, in all settings except for Figure 6.13a one iteration asks for a second order correction (marked by $\triangle$) before becoming acceptable to the filter. Since these cases are part of the pre-asymptotic phase, the FullSOC approach from Section 6.4 is applied. In addition, the iterates which fulfill the $\mathscr{L}$-type condition are marked by $\pentagon$. Interestingly, the first iterate which satisfies (6.4) always asks for line search subsequently. However, this is not necessarily a bad thing, since the line search is always successful and helps to stick to a meaningful solution path.

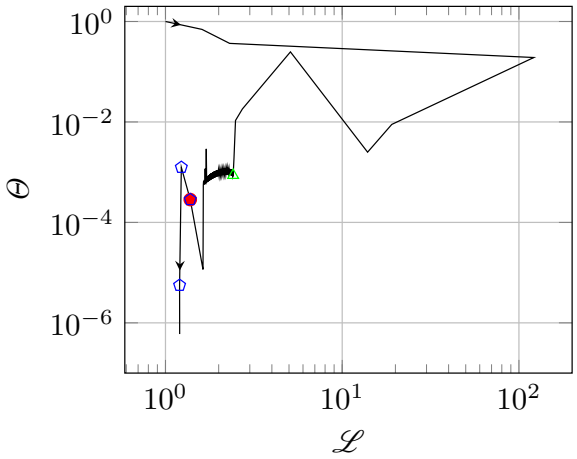
To conclude: This example has not been chosen to demonstrate that the filter method resolves issues inherent in the underlying semi-smooth Newton method, Thus, it is not surprising that the iteration number is such high. Nevertheless, it should demonstrate that the filter method stays reliable even in such cumbersome circumstances. The filter method is able to tell the user that the algorithm is still making progress to the solution even though the non-monotonic behavior of the norms and the active set might not immediately indicate it (see Figures 6.12 and 6.13). In addition, it also verifies that this progress is mathematically sufficient. This is exemplarily



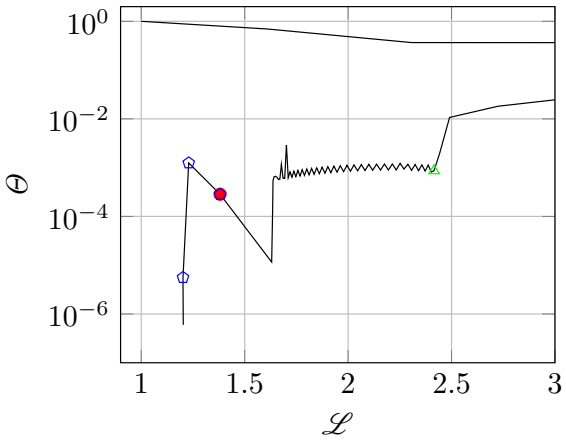
(a) $5 \times 100 \times 4$ (combo)



(b) $5 \times 100 \times 4$ (std)



(c) $5 \times 200 \times 4$ (combo)



(d) $5 \times 200 \times 4$ (combo), detailed view

Figure 6.13.: Solution paths for the pair of plates example in the filter sub-space. Figure 6.13a shows the solution path for the coarse mesh and the activated switching condition, i.e., the modified Newton approach is used for the pre-asymptotic phase while the strategy switches to the consistent linearization near the solution. This combined strategy is denoted by (*combo*). In Figure 6.13b the same configuration is solved with the consistently linearized standard strategy (*std*). In Figures 6.13c and 6.13d the (detailed) solution path for the fine mesh in conjunction with the combo strategy is presented.

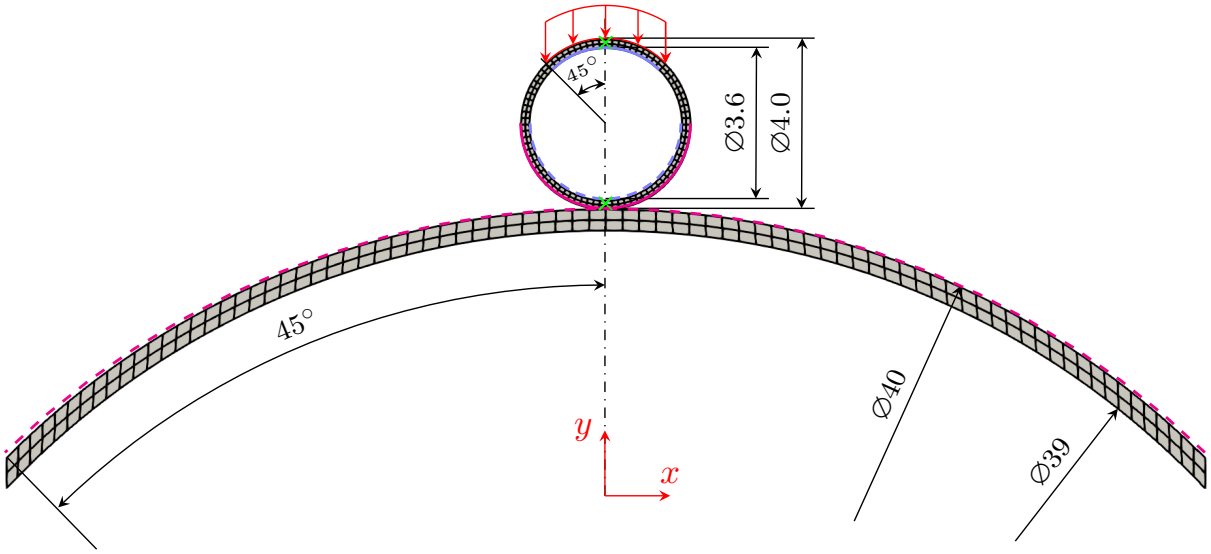


Figure 6.14.: Geometrical configuration of the snap-through example. The small circle is pressed via a Neumann load onto the large circular segment. The potential contact boundary segments are highlighted in two different colors where the slave side is represented by the solid and the master side by the dashed line, respectively.

enforced by the used margin in (6.3). This property alone makes the filter method truly valuable for any contact simulation.

Remark 6.3. It has been mentioned at the beginning of the problem description that the choice of slave and master side is not arbitrary. Indeed, a switch of slave and master side will lead to problems for one of the two meshes dependent on the actual interface definition. However, it is important to highlight that these issues are not based on the filter method but are again rather a problem of the applied element-wise numerical integration scheme. If the lower side of the top plate becomes slave and the ray-tracing projection is used, there are non-projectable parts of the boundary elements which might lead to convergence problems due to Gauss points which find a projection in one iteration and fail in the next. The locations of these critical Gauss points can be found close to the hinge supports (cf. Figure 6.11) and are directly linked to the inwards rotating upper edge of the lower plate. This is something which can easily be circumvented by the applied slave/master assignment.

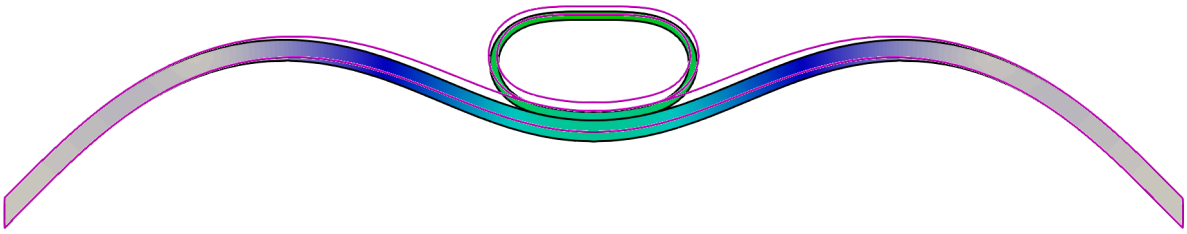
6.10.2. Snap-Through Buckling of Circular Structures

The next example demonstrates the ability of the proposed method to handle a snap-through instability. This is something which is mainly treated by Algorithm 6.3, since it asks for a reliable modification of the system matrix close to the instability. The considered initial geometric setup is presented in Figure 6.14. The example consists of a small circle which is pressed onto a larger circular segment. Therefore, a Neumann boundary condition is used on top of the small circle. Furthermore, the two marked nodes × on the symmetry line in the middle of the circle membrane are fixed in x-direction. The large circle segment is fixed on the two short cut edges left and right in x- and y-direction. Finally, the contact interfaces are defined between the light blue solid and dashed lines on the inner surface of the circle and between the solid and dashed lines on the outer

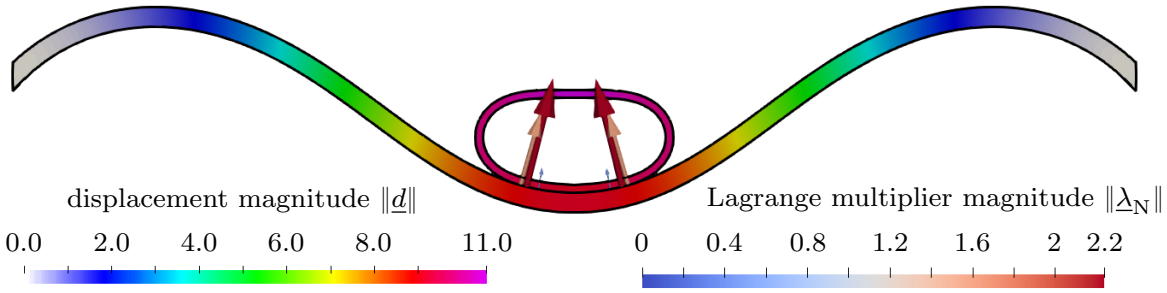
surfaces of the small and large circle (magenta color). As material law the already introduced logarithmic neo-Hookean material model (2.27) is considered. The Young's modulus is set to $1.0\mathrm{E}{+}3$ for both bodies and the Poisson's ratio to 0.25. Now, two load steps are reviewed to trigger the snap-through behavior. In the first load step the magnitude of the Neumann load is set to 0.4 and in the second load step it is doubled to 0.8. Since it is a pure Neumann loading and, consequently, no large penetrations are expected, the SIR-update is switched off for this simulation, i.e., there is no regularization term in the lower right block of (6.59). In addition, it is to highlight that the Neumann loading will lead to a negative total energy and Lagrangian value, i.e., the first coordinate of the filter points is expected to become negative which will automatically deactivate the pre-filtering introduced in Section 6.7.6.

As already mentioned: This example aims for Algorithm 6.3. Therefore, the main focus is on the correction of the upper left block in (6.59). In the first load step, the correction algorithm is triggered in iteration $\#5$ by the surrounding positive definiteness check of Step 2.4. The corresponding deformed (non-equilibrium) shape is presented in Figure 6.15a. It is obviously a configuration which is very close to the snap-through. The final equilibrium configuration is given in Figure 6.15b. Thus, the algorithm has the task to guide the two bodies through this locally unstable point to a new stable configuration. Therefore, the introduced positive definiteness check is performed and the unmodified system matrix (6.59) indicates a negative definite upper left block. Now, Algorithm 6.3 increases the correction factor ω until the acceptance check in Step 2.4 is successfully passed. This example takes 2 corrections. In the following linear solver attempts, a decreasing correction factor is sufficient, such that the counter n_ω is increased in each call during Step 7 of Algorithm 6.3, till n_ω exceeds N_ω. Afterwards, the unmodified system of equations is reconsidered. For this example this happens in iteration $\#10$ of load step $\#1$. Since the snap through point of the large circle segment has been successfully passed, no further corrections of the system matrix are necessary until the end of the first load step. The matrix corrections are shown in Figure 6.16a, while the entire safe-guarded solution path in the filter domain is given in Figure 6.16b. The converged configuration for load step $\#1$ is visualized in Figure 6.15b. In total, 20 Newton, 2 second order corrections steps (△) and 2 line search corrections (●) are necessary.

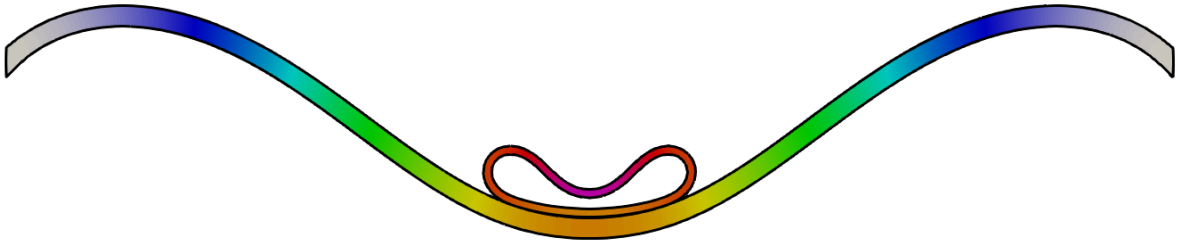
In the second load step the magnitude of the Neumann load acting on the small circle is further increased to 0.8. The Algorithm 6.3 asks again for a regularization in iteration $\#6$ and $\#7$ (see Figure 6.16a). But, in contrast to the previous load step, the correction is triggered by a bad element counter of 32 and a much larger step length compared to the previous iteration. Since the correction reduction in iteration $\#7$ leads to renewed failure, the regularization parameter is increased once more. Afterwards the critical point seems to be successfully passed, as from thereon a decreasing correction parameter is sufficient until the modification is finally switched off in iteration $\#11$. The reason for the necessary regularization is shown in Figure 6.15c. This time, the small circle snaps through. However, a look at the solution path in Figures 6.16c and 6.16d reveals an unexpected pattern. Further investigations point out that the found pattern is related to a scenario after the snap-through, roughly initiated in iteration $\#11$. The reason is that the small circle comes into self-contact in iteration $\#11$. This leads to the rising Θ coordinate in Figure 6.16d. Afterwards the active-set strategy has a tough time to identify the correct active set which comes along with a slow progress towards the solution (see iteration 13 to 24 in Figure 6.16d). The point which is remarkable here is that the shown solution path becomes only



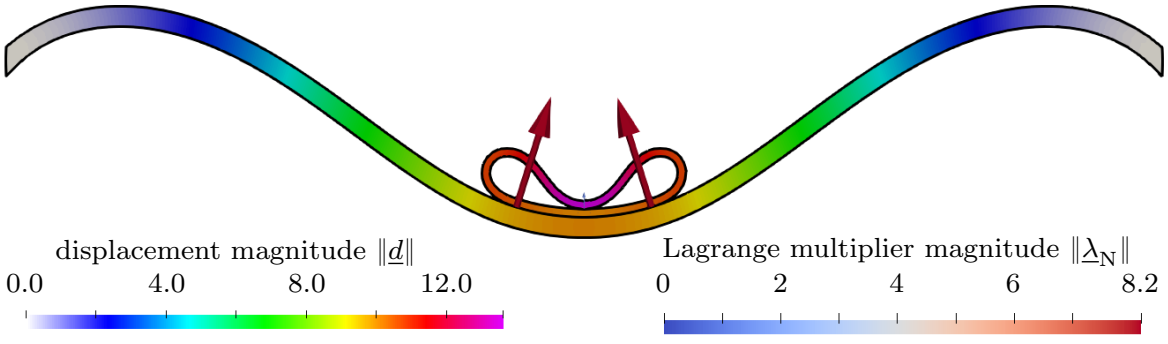
(a) load step #1, iteration #5



(b) load step #1, iteration #20 (converged)

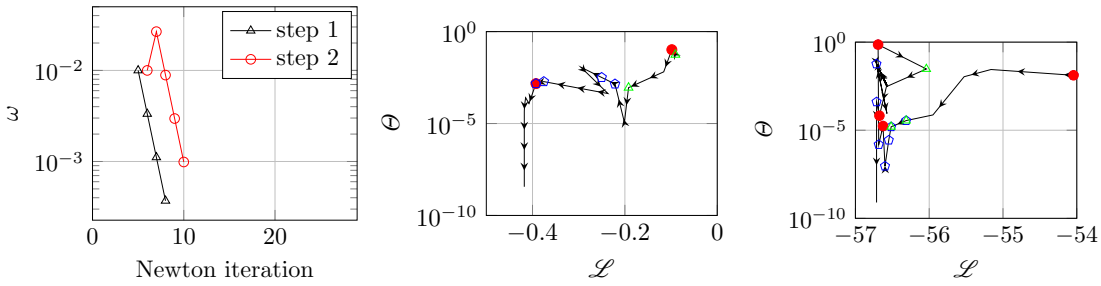


(c) load step #2, iteration #7

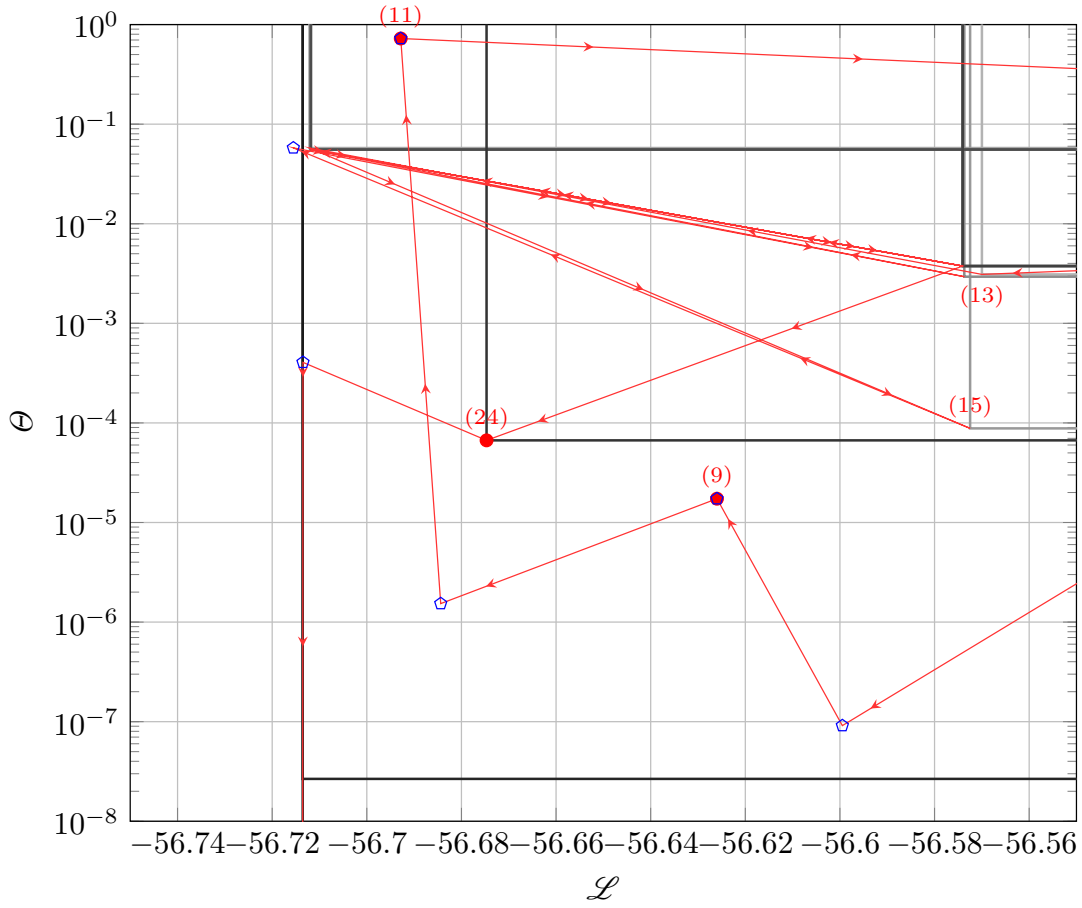


(d) load step #2, iteration #29 (converged)

Figure 6.15.: In Figure 6.15a the critical Newton iterate of load step #1 is shown. The pink solid lines represent the deformation state of the previous iteration #4. The applied matrix correction is just enough to push the intermediate state into the right direction to the final solution presented in Figure 6.15b. The second snap-through scenario, now concerning the small circle, is shown in Figure 6.15c. The final self-contacting solution state at the end of load step #2 can be found in Figure 6.15d.



(a) matrix corrections (b) path of load step #1 (c) path of load step #2



(d) path of load step #2 (detailed view)

Figure 6.16.: The necessary system matrix corrections of the upper left block are shown in Figure 6.16a, while the solution paths of load step #1 and #2 in the respective filter sub-spaces are shown in Figures 6.16b and 6.16c. Additionally, a detailed view on the cumbersome part of load step #2 is given in Figure 6.16d.

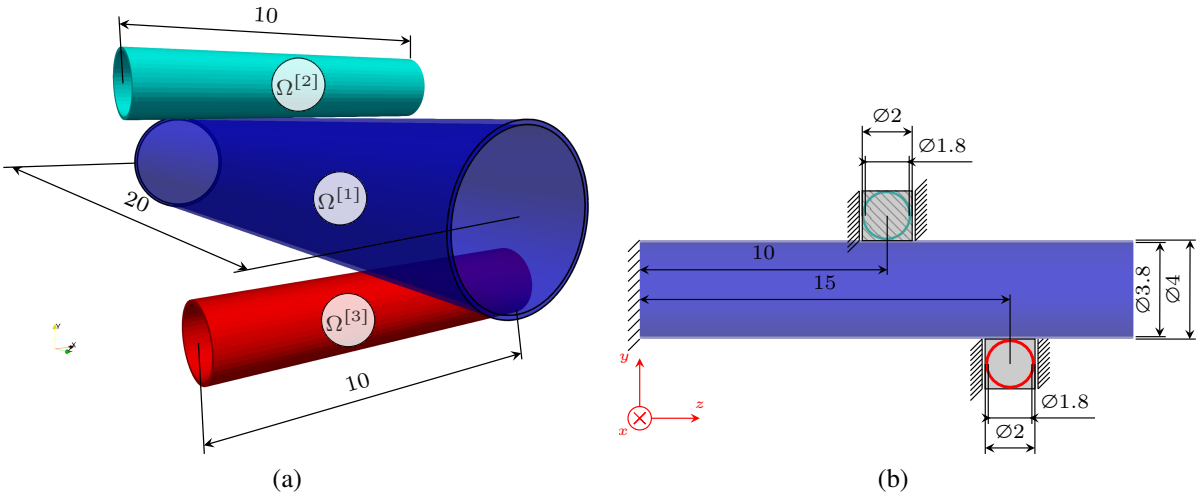


Figure 6.17.: Visualization of the clamped carbon fiber tube example. Herein, respective domains, boundary conditions, orientations and positions as well as dimensions of the three tubes are given.

possible due to the fact that iteration 5 to 11 are all $\mathscr{L}$-type iterations (⬠). Consequently, the corresponding points are *not* added to the filter as illustrated in Figure 6.16d. If these points were added to the filter, the points corresponding to iteration 9 to 10 would block many of the subsequent iterates and a reinitialization of the filter would become inevitable (see Section 6.7.3). This can be avoided by the presented strategy. Furthermore, the filter method does not intervene during the difficult active-set identification, i.e., the line search stays inactive. Only at the end of this cumbersome section the step length is halved once in iteration #24. In total 29 Newton and 3 second order correction steps (△) are necessary to reach the final solution shown in Figure 6.15d where only 4 steps ask for a line search correction (●).

6.10.3. Clamped Carbon Fiber Tube

Next, a more complex 3-D example is considered. The dimensioned geometry is presented in Figure 6.17. This time the focus lies, firstly, on the treatment of single bad elements which occur during the simulation. Secondly, the considered large tube will show again some kind of buckling. Thirdly, an anisotropic material model will be applied and, finally, the performance of iterative linear solvers will be reviewed. The discussion begins with a closer look at the used boundary conditions. The large blue tube oriented along the z-axis and corresponding to body $\Omega^{[1]}$ is fixed in all directions on the entire cut surface at $z = 0$. Now, the turquoise tube corresponding to $\Omega^{[2]}$ is fixed on the cut surface at $x = -5$ in x- and z-direction. In the third direction a prescribed motion will be applied. Quite similar Dirichlet boundary conditions hold also for $\Omega^{[3]}$. But this time the cut surface at $x = 5$, i.e., the opposing tube end, is considered. Again, a motion in y-direction will be applied. See Figure 6.17b for a visualization of the setting. The contact is detected on the outer side surfaces of the three cylinders, where the surface of $\Omega^{[1]}$ acts always as slave and the others as master. To avoid projection errors due to the large penetrations, the contact surface of $\Omega^{[1]}$ is split into two halves. The half pointing in positive y-direction is defined as slave for the contact between body $\Omega^{[1]}$ and $\Omega^{[2]}$, while the other half is considered

as the slave surface for the contact between body $\Omega^{[1]}$ and $\Omega^{[3]}$. Furthermore, the two smaller cylinders shall be moved in y-direction by a magnitude of 2.5 to the center line of the large tube.

Next, the used material laws are discussed. The two shorter tubes $\Omega^{[2]}$ and $\Omega^{[3]}$ use again a coupled neo-Hookean material law as defined in (2.26). The Young's modulus is defined as 172,000 and the Poisson's ratio as 0.25. This is motivated by aluminium alloy. Now, in case of the big tube $\Omega^{[1]}$, a more complicated material law is applied. The material law is an additive combination of (2.27) and

$$\Psi_{\text{tv}} = (\alpha_{\text{tv}} + \frac{\beta_{\text{tv}}}{2} \ln(I_3) + \gamma_{\text{tv}} \, (I_4 - 1))(I_4 - 1) - \frac{\alpha_{\text{tv}}}{2} \, (I_5 - 1), \tag{6.96}$$

where the first three invariants have already been introduced in (2.24), while the two missing pseudo invariants are given by

$$I_4 = \underline{a} \cdot \underline{\underline{C}} \cdot \underline{a}, \qquad I_5 = \underline{a} \cdot \underline{\underline{C}}^2 \cdot \underline{a}. \tag{6.97}$$

Here, the vector $\underline{a} \in \mathbb{R}^3$ denotes the considered fiber direction. The reader is kindly referred to the appropriate literature for more information, see e.g. Holzapfel [136] for a general introduction or Bonet and Burton [31] for a detailed derivation of (6.96). The complete material law contains five parameters and is consequently suitable to model a transversely isotropic material behavior. For this specific example the parameters are set to

$$\lambda_{\text{nH}} = 5736.552135, \qquad \mu_{\text{nH}} = 3454.231434, \qquad \alpha_{\text{tv}} = -2045.768566, \tag{6.98}$$

$$\beta_{\text{tv}} = -386.764504, \qquad \gamma_{\text{tv}} = 19{,}424.87514. \tag{6.99}$$

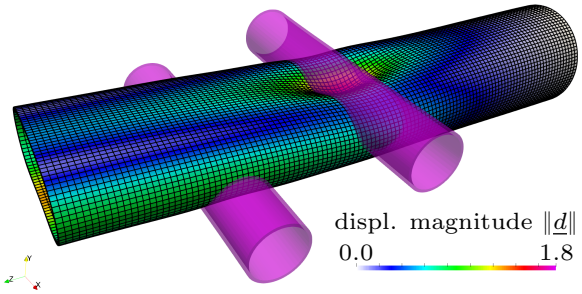
This choice corresponds to a carbon fiber reinforced plastic material with a PEEK matrix. The related more familiar parameters of the linear elastic regime are

$$E_{\parallel} = 172{,}000, \quad E_{\perp} = 10{,}000, \quad \nu_{\perp\parallel} = 0.27, \quad G_{\perp\parallel} = 5{,}500, \quad \nu_{\perp\perp} = 0.4475. \tag{6.100}$$

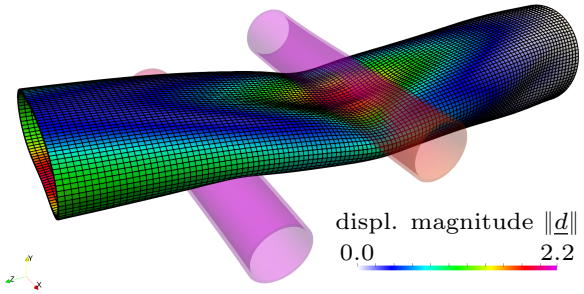
The derivation of these parameters is based on the demand that the non-linear material law coincides with the linear theory for the small strain regime. The reader is again referred to Bonet and Burton [31] for more information. The z-axis is chosen as the reference fiber direction.

The first tube $\Omega^{[1]}$ is subdivided into 120 elements in longitudinal direction and 96 elements in circumferential direction. The two smaller tubes $\Omega^{[2]}$ and $\Omega^{[3]}$ are subdivided into 60 elements in longitudinal and 48 elements in circumferential direction. In radial direction only one element layer is used, respectively. Consequently, the entire problem consists of $120 \cdot 96 + 2 \cdot (60 \cdot 48) = 17{,}280$ hexahedra elements, $34{,}944$ nodes and $104{,}832$ DOFs. Furthermore, the elements are enhanced by the EAS-21 formulation taken from Allgower and Georg [2] to avoid unwanted locking effects .

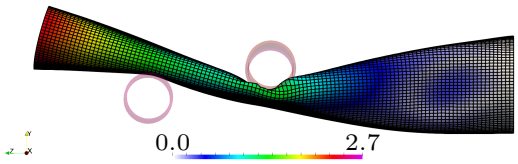
The considered boundary conditions in connection with the material laws and the element technology lead to a complicated problem. The proposed line search filter algorithm 6.2 takes in total 62 iterations to convergence. Thereby, many of the discussed potential issues occur and are



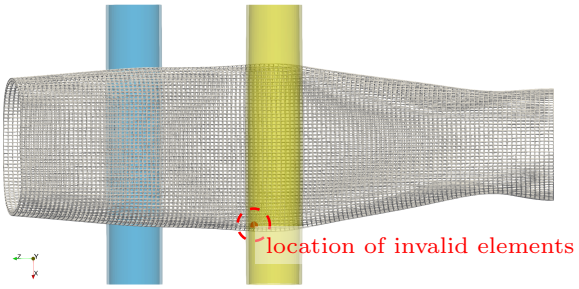
(a) iteration #3



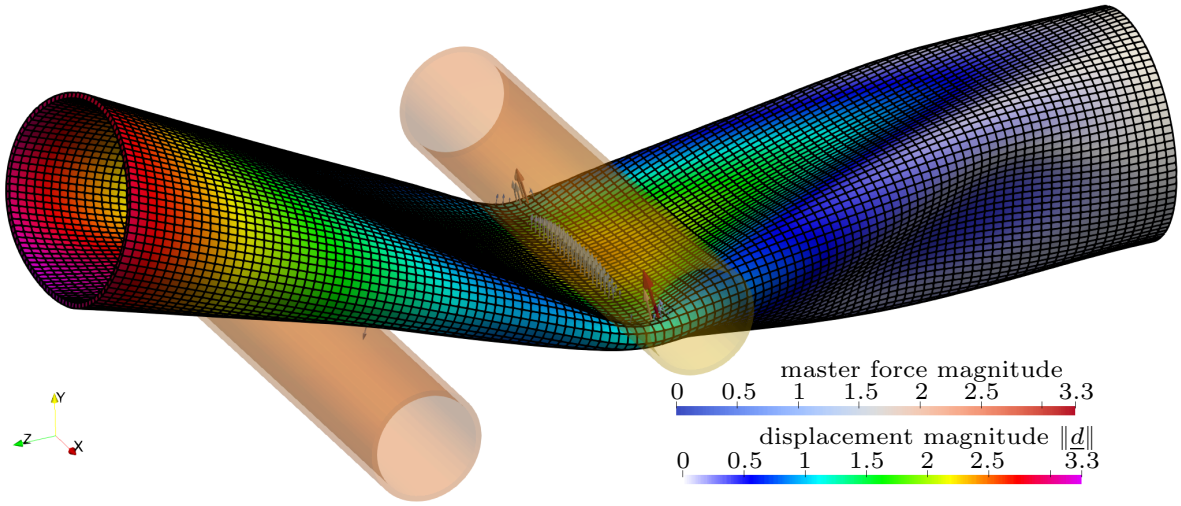
(b) iteration #8



(c) iteration #14

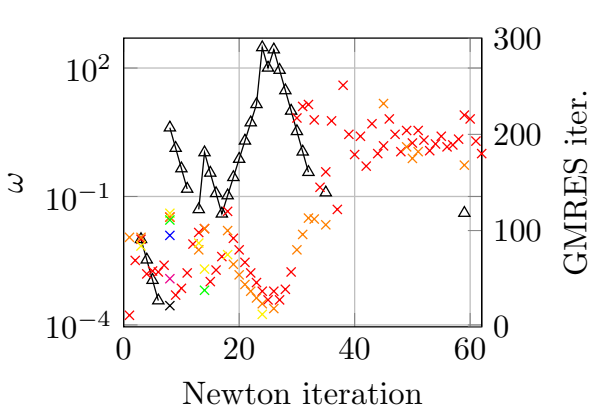


(d) iteration #18

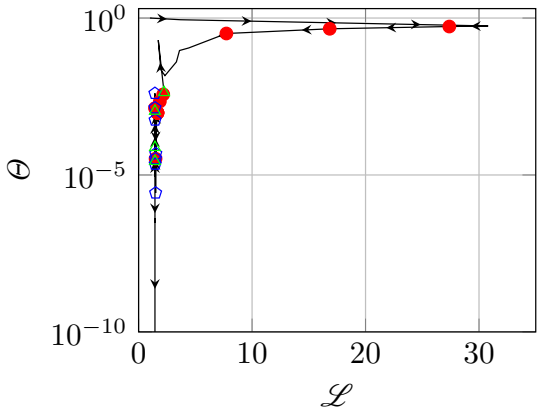


(e) iteration #62 (converged)

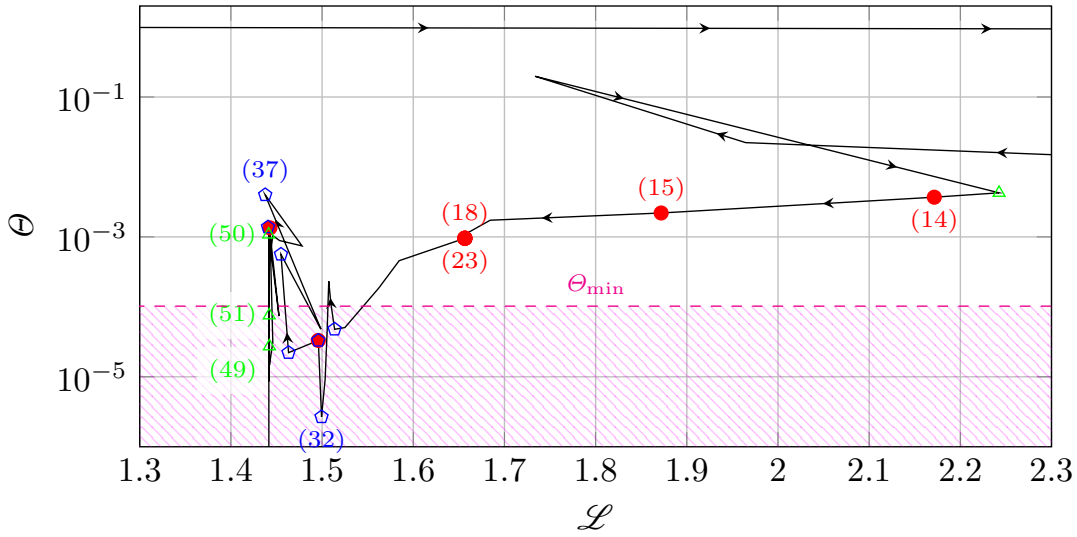
Figure 6.18.: A number of difficult intermediate deformation states occurring on the non-linear solution path are shown in Figures 6.18a to 6.18d. The difficulty arises from buckling phenomena and/or local mesh distortions. The final converged equilibrium state including the forces acting on the master bodies is shown in Figure 6.18e.



(a) matrix corrections



(b) solution path

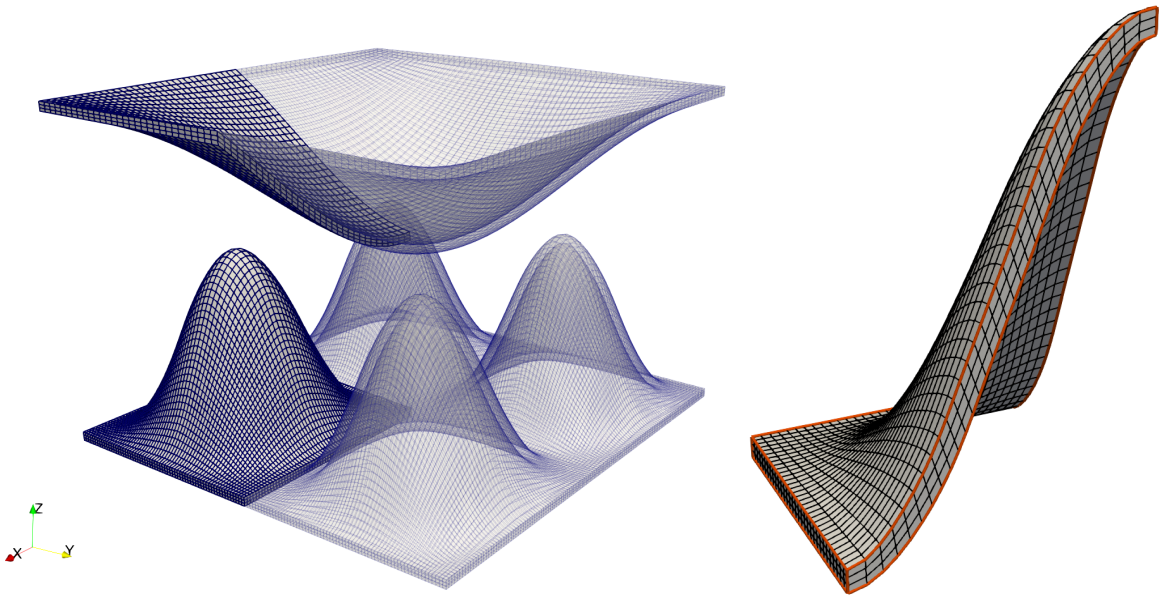


(c) detailed view on the solution path

Figure 6.19.: The necessary matrix correction factors ω as well as the necessary GMRES iteration numbers are shown in Figure 6.19a. The used color coding follows ×, ×, ×, ×, ×, ×, × in consecutive order corresponding to the index of the linear solver attempt taken in the respective Newton iteration. Note that not only Algorithm 6.3 might make an additional linear solver attempt necessary, but also a SOC step can ask for another solution of the linear system of equations. The non-linear solution path in the filter subspace is given in Figures 6.19b and 6.19c, where the almost complete stagnation between iteration #18 and #23 must be highlighted in Figure 6.19c. $\Theta_{\min}$ value as well as the therewith defined region $\{\Theta \in \mathbb{R}_+ \mid \Theta < \Theta_{\min}\}$ is shown in Figure 6.19c to underline the importance of this value for the $\mathscr{L}$-type step activation, see also ⬠.

resolved by the algorithm. A small collection of critical situations is presented in Figure 6.18. The associated solution path in the filter domain is given in Figures 6.19b and 6.19c, while the corrections and the necessary GMRES iterations are shown in Figure 6.19a.

The first critical situation occurs already in iteration $\#3$. Here, the tube corresponding to $\Omega^{[1]}$ starts to buckle for the first time as Figure 6.18a demonstrates. This comes along with a rising regularization parameter due to a negative definite matrix block (see Figure 6.19a). The next buckling occurs between $\Omega^{[1]}$ and $\Omega^{[2]}$ in iteration $\#8$ and is again attended by a negative definite upper left matrix block (see Figure 6.18b). This time, Algorithm 6.3 takes six iterations to reach the regularization ω which finally satisfies the acceptance criteria defined in Step 2.4 of the respective algorithm. Afterwards, a critical phase of the solution procedure starts. First, a heavy mesh distortion appears because of a negative definite system in iteration $\#14$ which initiates a decrease of the c_{N}-parameter at the end of iteration $\#15$. This decrease is guided by the strategy proposed in Section 6.7.4. However, the c_{N} parameter is only reduced very slightly by a factor equal to 0.978. This helps to temporally resolve the distortion. Unfortunately, a few iterations later the mesh gets locally distorted once more as visualized in Figure 6.18d. Actually, this time the mesh distortion leads to two invalid elements which are highlighted in red in this figure. These elements avoid any further progress for the next five iterations. The regularization parameter ω is successively increased, once, due to the invalid element counter which is equal to two and, secondly, due to the fact that the subsequent reduction of the regularization parameter ω does always increase the step length compared to the previous iteration. This is a great feature of Algorithm 6.3: As long as a rising regularization parameter and the line search are insufficient to resolve the distorted elements during an iteration, the parameter ω is probably increased once more in the subsequent attempt since ω is initially decreased in Step 3 and thus it is likely that s_{curr} becomes larger than s_{last}. This is also what finally helps to resolve the locally heavily distorted mesh. In iteration $\#24$ the regularization parameter ω reaches its maximum at a value of 302.72. Note that this cumbersome situation led almost to a complete stagnation of the entire globalization method between iteration 18 and 23. This is visible in Figure 6.19c. The step length has been reduced consecutively till a minimal value of $\alpha^{\{23,19\}} = 1.90735\mathrm{e}{-06}$ is reached before the rising ω value finally helps to overcome this point. This is a remarkable example of how one element out of $17{,}280$ has the power to cause a complete break-down of the algorithm. Therefore, it is very important to detect these elements as described in Section 6.6.2. Afterwards, the path pattern becomes more complex and hard to follow as demonstrated in Figure 6.19c. The origin can be again traced back to the correct active-set identification. In this specific case, the applied frictionless contact situation makes it more complicated: While tube $\Omega^{[2]}$ pushes the big tube $\Omega^{[1]}$ in positive x-direction, tube $\Omega^{[1]}$ creates a counter force and, consequently, a forth and back sliding in x-direction is initiated. The active-set finally converges in iteration $\#58$. Note that the last correction of the upper-left matrix block in iteration $\#59$ is not initiated by the acceptance tests in Step 2.4 but rather by a failing linear solver call in Step 1 of Algorithm 6.3. This underlines once more that the proposed algorithm can significantly improve the solvability of the underlying saddle-point system by increasing the diagonal dominance of the upper-left block (see the GMRES iteration numbers in Figure 6.19a). In addition it is to mention that at the end of iteration $\#58$ the switching criteria are fulfilled and the system matrix changes to the standard Lagrangian system which is necessary for a fast local convergence.



(a) (b)

Figure 6.20.: Visualization of the sine-shaped membranes defined by (6.101) and (6.102) including their thicknesses and their initial position which is specified in (6.105). Figure 6.20b shows a detailed view on a quarter of one sine hill to give a better impression of the used mesh.

6.10.4. Sine-Shaped Membranes

The following example demonstrates the necessity of the reinitialization routine proposed in Section 6.7.3. The filter reinitialization is something which should be avoided, however, under certain circumstances, it is not possible to reliably predict and resolve the undesired solution behavior without the risk of being too restrictive in other cases. Thus, the filter reinitialization is in some way a fall-back strategy when all the other strategies fail. Fortunately, this does not happen too often as the previous examples demonstrated. To trigger the reinitialization, two contacting membranes shall be considered. The mid-surface of the upper membrane corresponding to $\Omega^{[2]}$ (i.e. master body) is defined by

$$w(r,s) = -\frac{3}{8}(\frac{1}{5}r+3)\sin(\frac{\pi}{10}r+\pi)[1+\cos(\frac{\pi}{20}(r-s))], \tag{6.101}$$

which is intentionally chosen in such a way that it becomes non-symmetric. The mid-surface of the lower membrane corresponding to $\Omega^{[1]}$ (i.e. slave body) is given by

$$v(r,s) = [1+\sin(\frac{2\pi}{5}r+\frac{3\pi}{2})][1+\sin(\frac{2\pi}{5}s+\frac{3\pi}{2})], \tag{6.102}$$

and represents a sine-waved thin body. Due to its shape, it can be expected that multiple separated contact zones will occur between the two bodies. To obtain the top and bottom surfaces of these membranes, the associated normal fields must be computed. These are directly obtained by

$$\underline{N}_{(\bullet)}(r,s) = \frac{\underline{\tilde{N}}_{(\bullet)}(r,s)}{\|\underline{\tilde{N}}_{(\bullet)}(r,s)\|} \qquad \text{with} \qquad \underline{\tilde{N}}_{(\bullet)}(r,s) = \frac{\partial \underline{X}_{(\bullet)}(r,s)}{\partial r} \times \frac{\partial \underline{X}_{(\bullet)}(r,s)}{\partial s}, \tag{6.103}$$

where $\underline{X}_{(\bullet)}(r,s)$ is meant as a placeholder and is supposed to be replaced by

$$\underline{X}_w(r,s) = \begin{pmatrix} r & s & w(r,s) \end{pmatrix}^{\mathrm{T}} \qquad \text{or} \qquad \underline{X}_v(r,s) = \begin{pmatrix} r & s & v(r,s) \end{pmatrix}^{\mathrm{T}} \tag{6.104}$$

to obtain $\underline{N}_w(r,s)$ and $\underline{N}_v(r,s)$, respectively. Next, a thickness for each membrane is specified via $t = 0.2$ and, finally, the top and bottom surface pairs follow as

$$\underline{X}^{\pm}_{(\bullet)}(r,s) = \begin{pmatrix} r \pm \frac{t}{2} N^1_{(\bullet)}(r,s) \\ s \pm \frac{t}{2} N^2_{(\bullet)}(r,s) \\ (\bullet)(r,s) \pm \frac{t}{2} N^3_{(\bullet)}(r,s) + o_{(\bullet)} \end{pmatrix}, \tag{6.105}$$

where $(\bullet)$ is a again a placeholder for the respective identifier. The constant $o_{(\bullet)}$ is used to specify the initial offset between both membranes. In this example $o_w = 7.2$ and $o_v = 0$ are inserted. Furthermore, the domain is restricted to

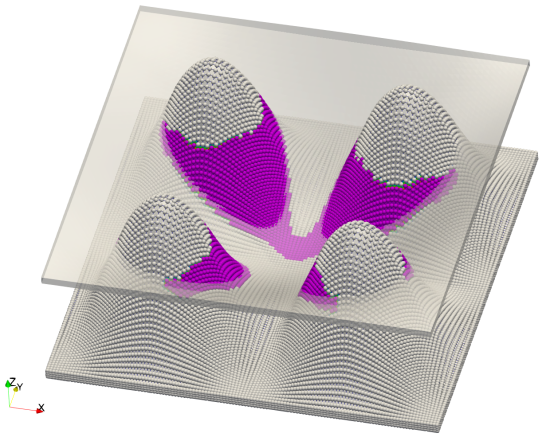
$$(r,s) \in \{(r,s) \in \mathbb{R} \times \mathbb{R} : [\underline{X}^{\pm}_{(\bullet)}]_1(r,s) \in [0,\ 10] \text{ and } [\underline{X}^{\pm}_{(\bullet)}]_2(r,s) \in [0,\ 10]\}, \tag{6.106}$$

i.e., restricted to $[0,\ 10] \times [0,\ 10]$ in the XY-plane. The geometrical reference configuration as well as the used mesh are presented in Figure 6.20. Once more, hexahedral elements are used. The subdivisions are as follows: In thickness direction three elements are considered, per quarter of a sine-shaped hill 32 elements in x- and y-direction are used as shown in Figure 6.20b and for the upper body corresponding to (6.101) 64 elements in x- and y- direction are inserted. In summary:

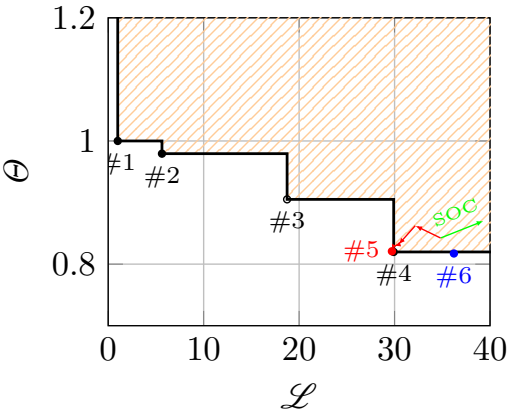
$$4 \cdot [4 \cdot (32 \cdot 32 \cdot 3)] + 64 \cdot 64 \cdot 3 = 61{,}440 \text{ elements}, \quad 83{,}464 \text{ nodes}, \quad 250{,}392 \text{ DOFs}.$$

This time no additional element technology is applied, even though it might be advisable to avoid possible shear locking due to the bad aspect ratio of some elements. Both bodies use once more (2.26) with $E^{[1]} = 2{,}500$, $E^{[2]} = 12{,}500$ and $\nu = 0.25$. Dirichlet boundary conditions are applied to all DOFs of nodes on the surrounding cut surfaces at $\underline{X}_{(\bullet)} \in \{X^1 \in \{0,\ 10\},\ X^2 \in [0,\ 10]\}$ as well as $\underline{X}_{(\bullet)} \in \{X^2 \in \{0,\ 10\},\ X^1 \in [0,\ 10]\}$ of both bodies. While body $\Omega^{[1]}$ stays fixed at these DOFs during the entire simulation, a prescribed motion in negative z-direction is applied to $\Omega^{[2]}$. The magnitude of this motion is set to 4.0 and is applied in one load step.

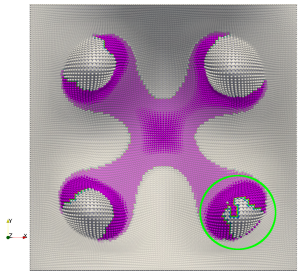
The resulting initial penetration state is presented in Figure 6.20a. All nodes highlighted in pink are set active by a negative averaged weighted gap value. A closer look reveals that some tips of the sine-shaped body $\Omega^{[1]}$ stay white. The reason is that an insufficient search parameter is chosen for the contact pair detection in this specific large penetration scenario [288]. This is done to demonstrate one source of failure which is completely unrelated to the filter method. A



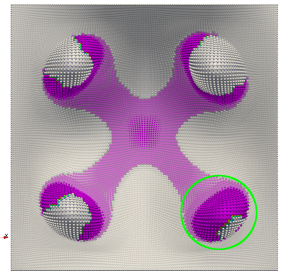
(a) iteration #0



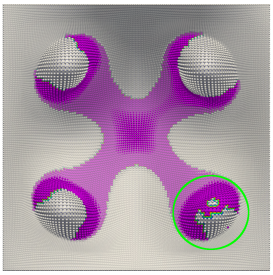
(b) filter domain



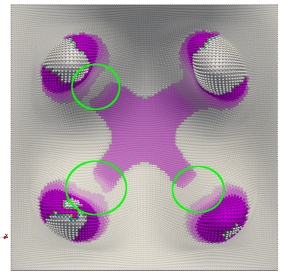
(c) iteration #4



(d) iteration #5/1



(e) iteration #5/4



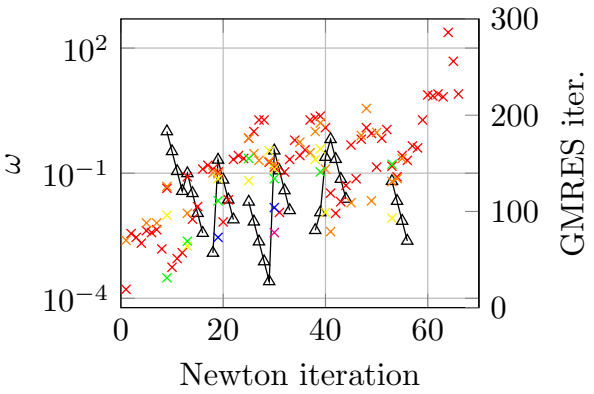
(f) iteration #6

Figure 6.21.: Active-set distribution during the initial phase of the sine-shaped membranes example. The Figures 6.21a to 6.21f show the impact of an insufficiently (i.e. here too small) chosen search parameter during contact pair detection. In such a case, it might happen that the initially missing and later progressively joining active nodes distort the filter measures and make them unreliable.

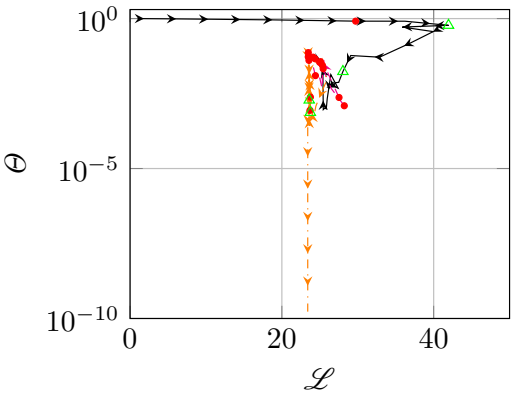
simple remedy would be to increase the according parameter, however, the search radius might be restricted by other considerations, such as efficiency or unwanted projection scenarios. In the following, the impact of this insufficiency on the filter method is discussed. Therefore, Figure 6.21 provides a detailed view of the situation. The insufficient search radius is less severe in the iterations 1 to 4. However, in iteration 5, the method is close to failure. This is obvious in Figure 6.21b. Even though the related displacement shows a clear progress into the right direction (see Figure 6.21d), the filter will block the iteration. Consequently, Algorithm 6.2 initiates a SOC step, however, the second order correction fails as well. Next, the line search begins and the algorithm finds an acceptable iteration point with a slightly smaller Lagrangian value after four backtracking steps (see Figures 6.21b, 6.21d and 6.21e). As it has been said at the outset, the problematic point is not the filter method but the provided truncated information. Although the solution proceeds in the right direction, the used infeasibility measure indicates a rising penetration since more and more nodes join the active-set which should have been already identified as active at the very beginning. This is highlighted in the Figures 6.21c to 6.21f. This circumstance becomes crucial in iteration $\#5$. Fortunately, iteration $\#6$ manages to leave this critical zone, since the gap between the two bodies starts to open at several other locations (see the green encircled regions in Figure 6.21f). This short interlude serves as a reminder that the origin of acceptance issues can be manifold and may be related to completely distinct parts of the used methods or algorithms.

Now, the attention is drawn to the filter reinitialization. In total the reinitialization is triggered twice for this example: Firstly, in iteration $\#27$ and, secondly, in iteration $\#39$. The solution path between iteration 1 and 27, between 27 and 39, as well as between 39 and 66 have been marked with different colors and line styles in Figures 6.22b and 6.22c. This highlights the fact that after a reinitialization, the previous filter points are erased from the filter and only one point solely based on the infeasibility measure acting as upper bound, as well as the current iterate itself are left. Therefore, the previous information loses its power to block new iterates. The related displacement field is given in Figure 6.23. Here, the sequence of iterations is shown which initiates the troublesome situation. In iteration $\#34$ everything looks still fine, in iteration $\#35$ already one line search step is necessary and the active set in the encircled region shows a severe change. Finally, in iteration $\#36$ the tip of this marked sine hill pops through the master body. As Figure 6.22c shows this leads to the scenario of a rising infeasibility measure while the Lagrangian value drops significantly at the same time, such that the new point is accepted without any modifications or corrections. The rest of the story has been already told in Section 6.7.3 and can be just transferred to this case. Also the first reinitialization in iteration 27 is caused by a similar circumstance. The related iterations $\#23$ and $\#24$ are marked in Figure 6.22c. However, even though the related path segments are very characteristic it would be false to say that a reinitialization becomes necessary each time when the infeasibility measure rises and the objective function value drops significantly. Many counter examples can be found in the already presented path graphs. Thus, the initial sentence of this section is underlined once more: It is not always possible to predict and prevent each and every cumbersome situation beforehand. Sometimes it is necessary to resolve it afterwards.

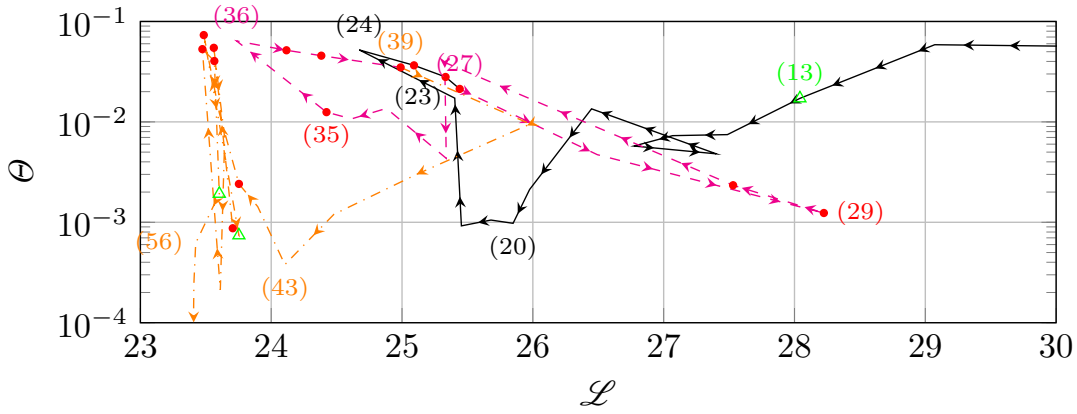
Furthermore, Figure 6.22a reveals again that the success of the filter method relies heavily on the proper regularization of the system matrix. Some kind of correction has been necessary almost throughout the entire simulation. Thereby, all of the regularizations have been initiated by a bad element counter greater than zero and an increasing step length. This strategy works



(a) matrix corrections

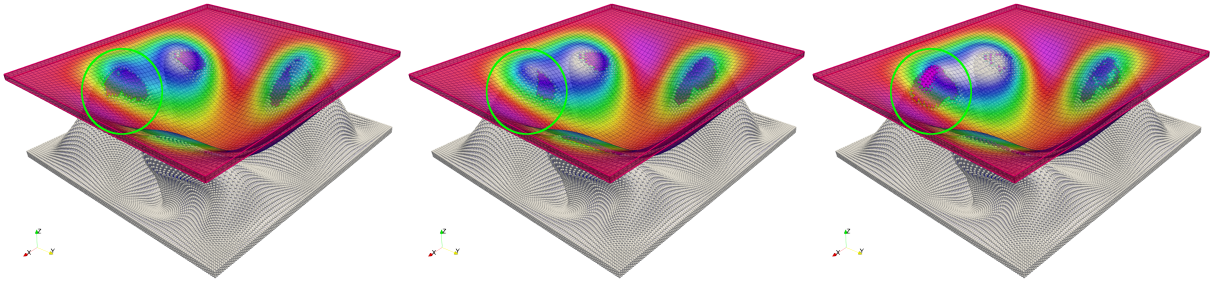


(b) solution path



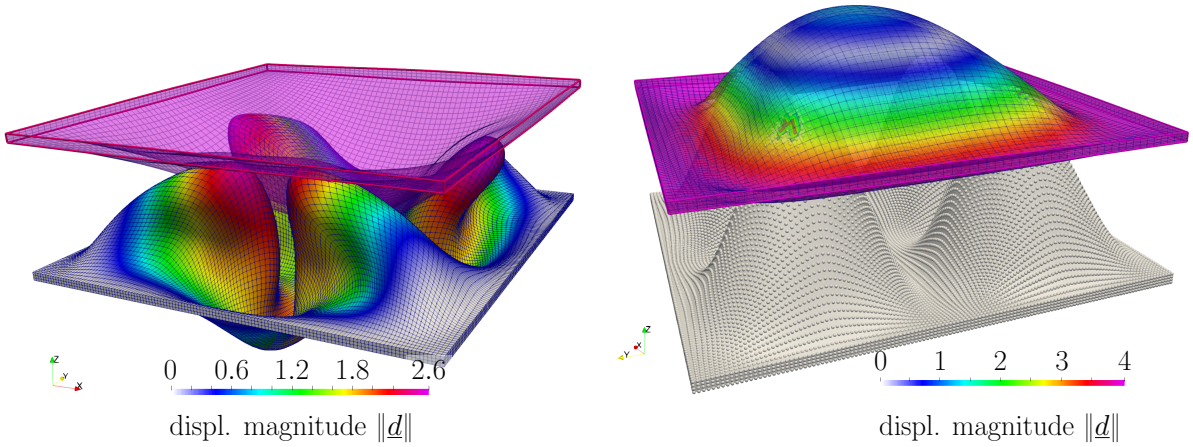
(c) detailed view on the solution path

Figure 6.22.: In Figure 6.22a the necessary matrix correction factor ω as well as the associated GMRES iteration numbers are shown for the first sine shaped membranes example. The used color coding follows the description provided in Figure 6.19. The complete solution path is presented in Figure 6.22b while Figure 6.22c shows only a smaller segment. The color and line style have been changed each time a reinitialization took place, i.e., after iteration #27 and after iteration #39.



(a) iteration #34 (b) iteration #35/1 (c) iteration #36

Figure 6.23.: A sequence of three accepted deformation states is shown which finally leads to a reinitialization in iteration #39.



(a) iteration #66 (converged)

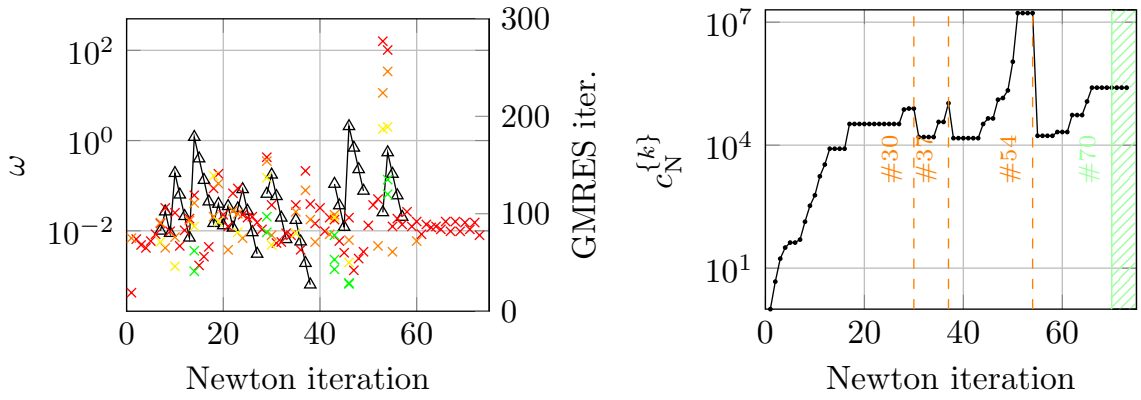
(b) iteration #73 (converged)

Figure 6.24.: Figure 6.24a shows the final deformation state for $E^{[1]} = 2{,}500$ and $E^{[2]} = 12{,}500$. The line search filter method proposed here takes 66 Newton iterations to find this solution. In Figure 6.24b a converged solution is presented which is revealed after 73 Newton iterations if the Young's modulus of the master body is reduced to $E^{[2]} = 4{,}500$.

really well and leads to a subsequent successful pre-testing in Step 3.3 of Algorithm 6.2 in many cases. Nevertheless, there are circumstances where pre-testing is still necessary. One example is iteration #30 where a local mesh distortion has been slowly developed over a few previously accepted Newton iterations and is finally safely resolved by the pre-testing strategy. Only in the very end, the correction of the system matrix becomes dispensable and allows a quadratic convergence near the solution. However, the back side of the missing regularization is a higher number of GMRES iterations. The sudden jump in GMRES iterations at iteration #64 in Figure 6.22a coincides again with the switching point to the standard Lagrangian saddle-point system of equations. The converged solution is reached at iteration #66. The associated final converged deformation state is presented in Figure 6.24a.

6.10.5. Sine-Shaped Membranes: Snap-Through

Now, the previous example is reconsidered with the only difference that the Young's modulus of body $\Omega^{[2]}$ is reduced. Instead of $12{,}500$, a value equal to $4{,}500$ is used. Thus, the stiffness gap between body $\Omega^{[1]}$ and $\Omega^{[2]}$ becomes severely smaller. This leads to a drastically changed system response as illustrated in Figure 6.24b. The reinitialization approach plays again an important role, but in contrast to the previous setting, another algorithmic feature becomes apparent: This time the decrease of the contact regularization parameter c_{N} introduced in Section 6.7.4 is triggered three times throughout the simulation, while the reinitialization is only executed once during iteration #55 after historic information in the filter blocked a new iterate in four consecutive Newton iterations. Note that the default parameter set for the reinitialization has been slightly changed in this example: The parameter $\gamma_{\Theta}^{\max}$ has been increased from 0.25 to 0.5 and the necessary number of consecutive Newton iterations has been also increased from 3 to 4 compared to Table 6.4. The development of the c_{N} parameter is illustrated in Figure 6.25b. The c_{N} regularization parameter is decreased at the end of iteration #30, #37 and #54. As described



(a) matrix corrections

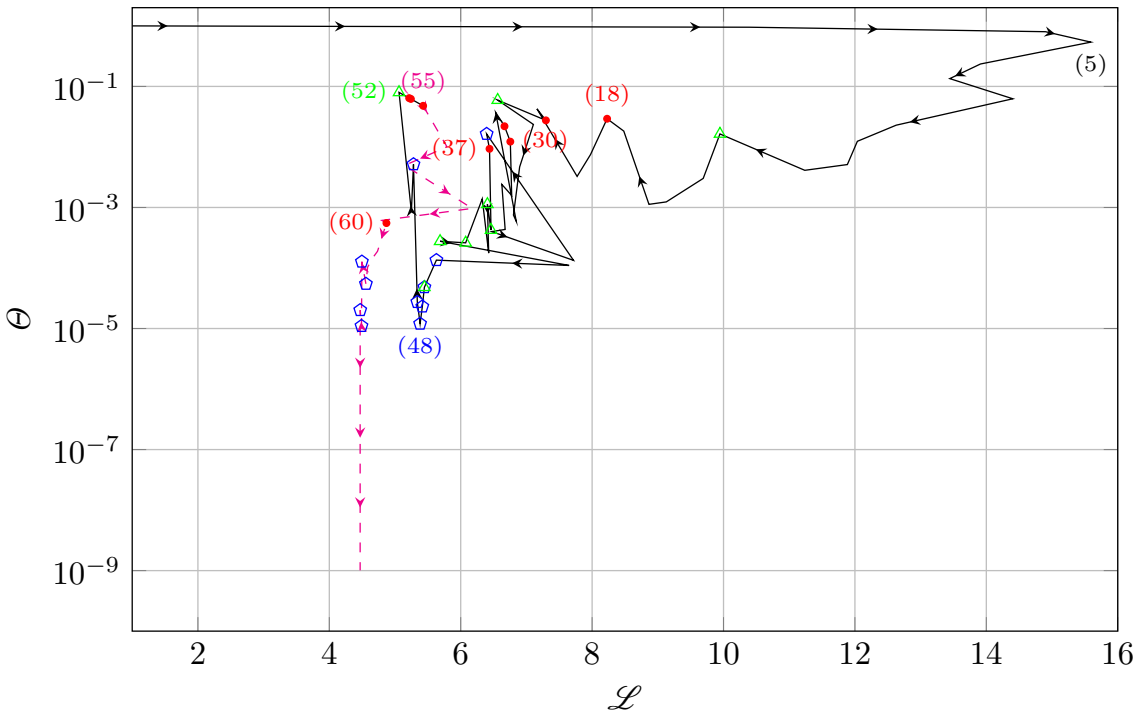
(b) development of c_{N}

Figure 6.25.: In Figure 6.25a the matrix correction parameter ω, which is computed by 6.3, is shown for the case that the sine shaped membranes example is solved with $E^{[2]} = 4{,}500$. The used color coding follows again the legend described in Figure 6.19. Figure 6.25b shows the development of the contact regularization parameter c_{N}. Especially, the decrease of this parameter under certain circumstances as described in 6.7.4 becomes visible.

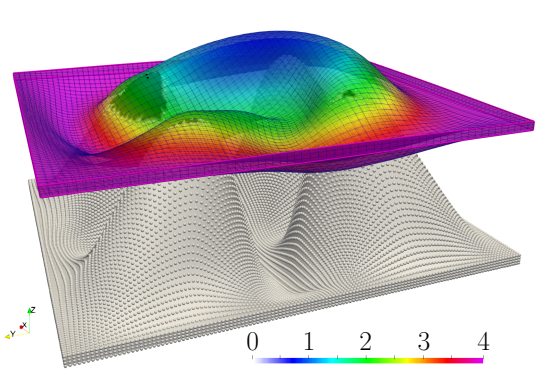
in Section 6.7.4 these decreases are initiated by a correction number unequal to zero as well as the necessity of line search step length reductions. If (6.77) additionally indicates a decrease, the parameter is reduced accordingly. However, as originally intended, the decrease of the regularization parameter takes also place during the troublesome situation where again the tip of one sine shaped hill pops through the master body. Fortunately, all the implemented strategies such as the regularization of the main displacement matrix block, the adapted regularization of the contact matrix block as well as the reinitialization strategy successfully manage to resolve the issue. Especially, the drop of the c_{N} parameter in the end of iteration $\#54$ is drastic and very important for the over-all performance. As soon as the algorithm finds its way back to a meaningful solution path, the c_{N} parameter starts to rise again in accordance to the SIR-update rule (see Figure 6.25). Finally, at the beginning of iteration $\#70$ the c_{N}-regularization is switched off and the asymptotic region is entered. This time the GMRES iterations do not rise during the switch as visualized in Figure 6.25 due to the different equilibrium configuration.

To the end of this example, some more general information about the actual solution path shall be given. In contrast to Section 6.10.4, the search parameter of the contact pair detection has been increased such that the problems displayed in Figure 6.21 could be successfully avoided. However, for the very large penetration at the beginning, probably even a larger search radius would be necessary to reach a level from which on any further increase does no longer change the initial active-set. Since the convergence seems not to be significantly influenced by the current choice, a further investigation has been prevented.

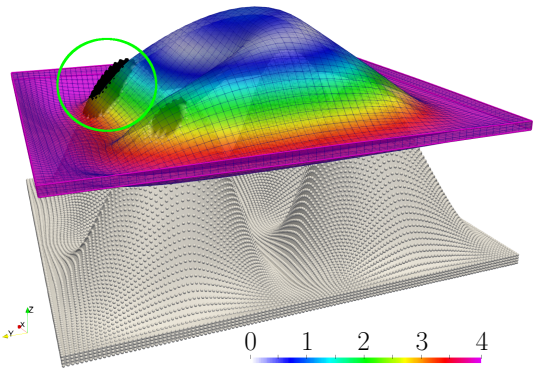
In total 73 Newton iterations and 8 successful second order correction steps were necessary to reach the equilibrium state shown in Figure 6.24b. The changed setup led to the circumstance that now the master body shows a much heavier deformation and the slave body is only deformed very little. Also the contact points are restricted to very small regions near the top of each sine hill. The related solution path is shown in Figure 6.26a. Furthermore, some critical deformation states are visualized as well in Figures 6.26b and 6.26c.



(a) solution path



(b) iteration #30



(c) iteration #52

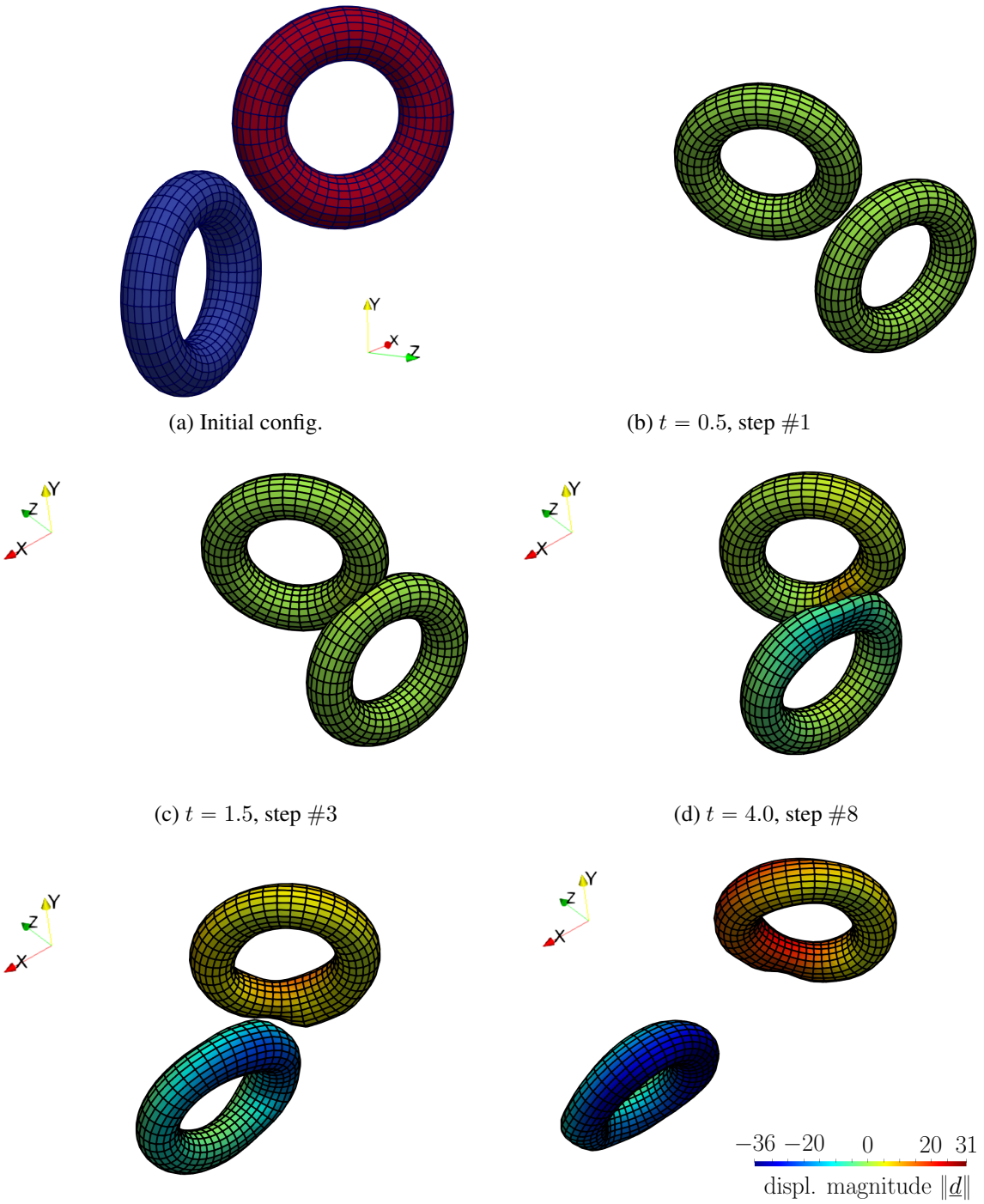
Figure 6.26.: Visualization of the solution path of the sine shaped membranes example with $E^{[2]} = 4{,}500$. The path after the reinitialization in iteration #55 is again plotted with a different color and line style. Furthermore, Figure 6.26b shows a deformation state which asks for a decrease of the c_{N} value and Figure 6.26c shows a sine hill popping through the master body which will lead to the reinitialization and drastic reduction of the c_{N} parameter later on.

In summary, the sine-shaped example proposed here is a great way to challenge the filter method and to trigger many of the exceptional cases, which often do not occur in simpler settings.

6.10.6. Grazing Tori

The attention is now drawn to dynamic contact problems with all the restrictions mentioned in Section 4.6, i.e., the focus is not on energy conserving algorithms but instead on the robust and efficient treatment of contact dynamics. The first considered example consists of two tori of equal shape and material. The major and minor radii are 76 and 24, respectively. The Young's modulus and the Poisson's ratio are set to $E = 2{,}250$, $\nu = 0.3$ under consideration of (2.26), while the density is set to $\varrho_0 = 0.1$. The wall thickness is equal to 4.5. All this is in accordance with the tori example presented in Yang and Laursen [288]. In this first example the same coarse mesh as in the mentioned reference is used, thus, there are 3200 HEX8 elements, where 2 elements in thickness direction are used, 4800 nodes and $14{,}400$ DOFs. The mesh as well as the initial configuration is shown in Figure 6.27a. Furthermore, the EAS-21 formulation is applied to avoid any shear locking effects due to the thin walled structure and the high element aspect ratio [3]. The 800 outer surface elements of the blue torus represent the slave side. At $t = 0$ the center of this torus is at the origin and lies in the xy-plane. The red torus is moved 140 units in x and 140 units in y direction and is subsequently rotated about $45°$ around the y-axis as shown in the respective initial configuration in Figure 6.27a. The blue torus has an initial velocity of $\underline{v}(t = 0) = (30.0, 0.0, 23.0)^{\mathrm{T}}$. The time step size is chosen to $\Delta t = 0.5$ and the parameter ρ_∞ of the Generalized-α method is set to 0.75 (see Section 2.4). A selection of converged equilibrium states can be found in Figures 6.27b to 6.27f. As one can see: The two tori are only grazing each other. This is to demonstrate that the filter method can switch smoothly between a contact and a non-contact state, or to put it differently: It can switch smoothly between a constrained and an unconstrained optimization problem. The main ingredient which allows this switch is the $\mathscr{L}$-type switching condition (6.4). As soon as the constraints become feasible or inactive, i.e., $\Theta^{\{k\}} \approx 0$ holds, (6.4) is always fulfilled as long as the direction is a valid descent direction. However, this is guaranteed by the applied regularization of the system matrix in Algorithm 6.3. Here, the two tori are between $t_1 = 0.5$ and $t_{11} = 5.5$ in contact. Furthermore, this example includes some necessary linear system adaptions since the upper left block of (6.59) becomes multiple times either negative definite during the non-linear solution procedure or elements become invalid while the step length rises. This illustrates that the additional natural regularization by the mass matrix can be insufficient in the pre-asymptotic phase if the time step size is not sufficiently small (see (2.96)). In addition it is not only a problem during the contact phase, since time step $\#14$ asks for a regularization as well due to a negative definite system matrix. As demonstrated in Figure 6.28a the regularization term ω reaches its maximal value $26{,}666.67$ in time step $\#4$.

Furthermore, the simulation asks for a reinitialization in two time steps. At this point, it is again to note that the default parameters for the reinitialization have been slightly adapted for this example: The parameter $\gamma_\Theta^{\max}$ has been increased from 0.25 to 0.5 and the necessary number of consecutive Newton iterations has been also increased from 3 to 4 compared to Table 6.4. Furthermore, $\Theta_{\max}$ has been increased from 2.0 to 3.0. The reinitialization is triggered twice in time step 3, once the bound $n_{\mathrm{ls}}^{\mathrm{block}}$ is exceeded and, the second time, 4 Newton iterations in a row are blocked, thus, $n_{\mathrm{newton}}^{\mathrm{block}}$ initiates the reinitialization. The blocking behavior has the additional drawback that it is likely that a SOC step fails as well. This explains the high number



(a) Initial config.

(b) $t = 0.5$, step $\#1$

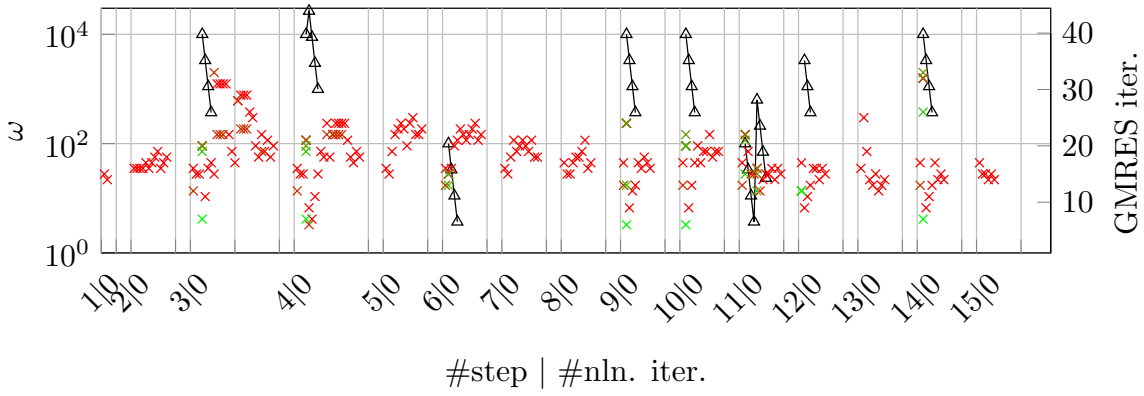
(c) $t = 1.5$, step $\#3$

(d) $t = 4.0$, step $\#8$

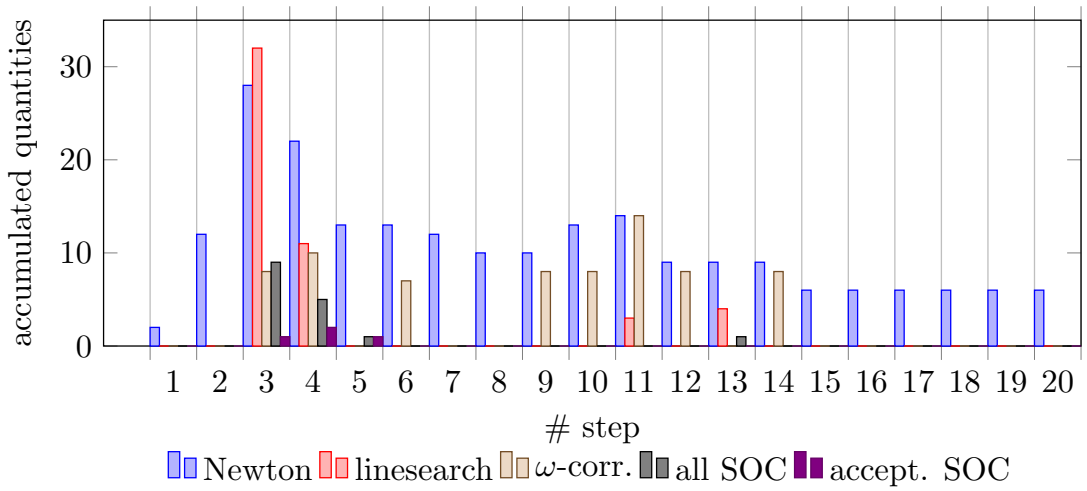
(e) $t = 6.5$, step $\#13$

(f) $t = 10.0$, step $\#20$

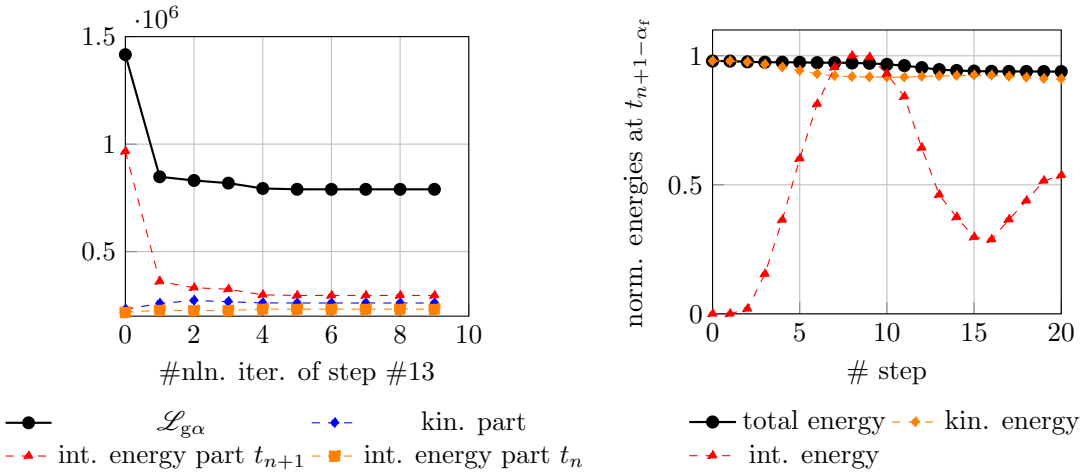
Figure 6.27.: Visualization of the grazing tori example. Figure 6.27a shows the initial positioning of the two tori ($\|\underline{d}\| = 0, t = 0$), while Figures 6.27b to 6.27f give impressions of different time steps. Note that the color coding in Figure 6.27a is only used to separate both tori more clearly, while the colors in the remaining images indicate the current displacement field.



(a) matrix corrections and GMRES iterations



(b) general statistics for the 20 considered time steps



(c) step #13

(d) total energy balance

Figure 6.28.: Figure 6.28a shows the necessary matrix corrections as well as the needed GMRES iterations for the first 15 time steps. The first number of the x-axis labels denotes the time step, while the following number represents the non-linear iteration counter, where zero is meant to indicate the start of a new time step. In Figure 6.28b interesting statistics are given about the non-liner solver behavior. Figure 6.28c shows the development of different contributions to (6.89) over time step #13, while Figure 6.28d presents the normalized energy balance for the the grazing tori example.

of unsuccessful SOC attempts in this time step. In addition, it is again triggered in time step $\#4$ also due to $n_{\text{newton}}^{\text{block}}$. A summary of all necessary Newton and line search iterations per time step, as well as the accumulated number of matrix corrections and the second order correction steps are shown in Figure 6.28b. Additionally, the number of successful, i.e., accepted SOC steps are presented as well. Finally, the attention is drawn to the newly introduced objective function (6.89) which becomes significantly important as soon as the two tori are getting out of contact. In Figure 6.28c the development of the different contributions to (6.89) are presented for time step $\#13$. This example demonstrates that not all contributions decrease monotonically. The kinetic incremental energy (6.89a) as well as the contributions (6.89b) related to the previous time step show also some temporary increase, while in this special case, the internal energy related to the current iteration is always decreasing. However, that is not in general true: Exemplarily, another pure structural load step (i.e., without contact contributions) between iteration 16 and 20 shows a different behavior. In these iterations the current internal energy decreases no longer monotonically. The explanation can be easily found in the global rise of the internal energy during these iterations as shown in Figure 6.28d. Nevertheless, the decisive point is that the sum of all these contributions is able to always follow a monotonically decreasing solution path as long as the non-linear solution method is controlled by the line search filter method proposed here.

Remark 6.4. The total energy drop in Figure 6.28d is small but still obvious for this example. The reasons for the drop are manifold and have been already discussed in Section 4.6. The rather small energy loss is caused by the only weak impact. In addition, the shown energies are scaled with respect to different quantities: While the internal energy is divided by its maximally reached value of $1.121\text{e}{+06}$ in time step $\#8$, the kinetic and total energy is scaled by the recursive value of the analytically calculated kinetic energy of the initially moving torus, viz.

$$\mathscr{K}(t=0) = \frac{1}{2}\varrho_0 V_0^{[1]} \|\underline{v}(t=0)\|^2 = 2.0982\text{e}{+7}, \tag{6.107}$$

where $V_0^{[1]} = 29{,}754\pi^2$. The initial offset between the analytical kinetic energy and the initial energy of the finite element model shown in Figure 6.28d can be traced back to the coarse mesh and the only approximative representation of the true volume by the used linear HEX8 elements.

6.10.7. Colliding Tori

As a final example the tori are considered once more but this time they shall collide directly and more widely with each other. The material parameters as well as the initial velocity field stay unchanged. The mesh is refined by a factor of two in all circumferential directions ending up with $12{,}800$ HEX8 elements, $19{,}200$ nodes and $57{,}600$ degrees of freedom. Again the EAS-21 formulation of Andelfinger and Ramm [3] is applied. The first torus is again placed at the origin, but now the second torus is moved 140 units in x- and z-direction and, subsequently, rotated by 45° around its new center point at $(140, 0, 140)^{\mathrm{T}}$. The time step size has been halved to 0.25, while ρ_∞ is again set to 0.75. The changed setup makes it necessary to detect contact not only between the outer surfaces of the two distinct tori but also among the inner surface of each of the two torus tubes itself. The initial configuration as well as a selection of converged time step

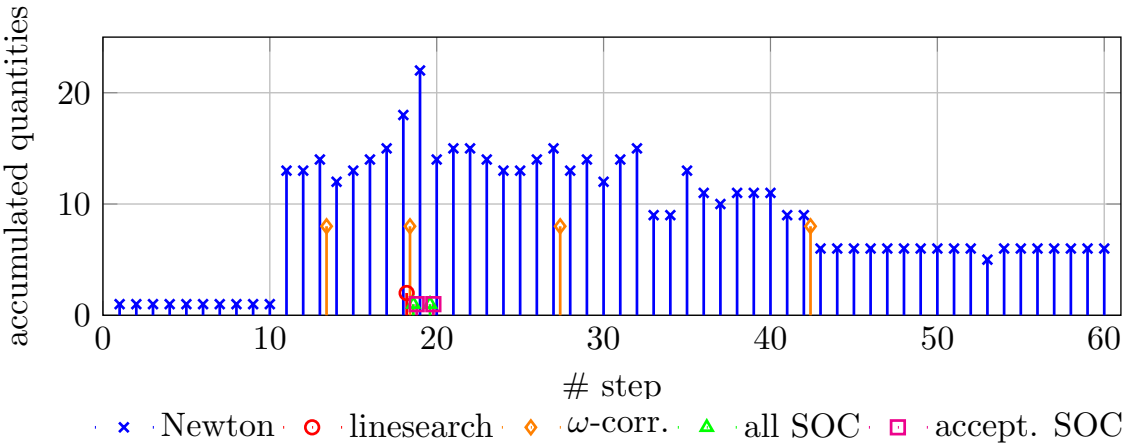


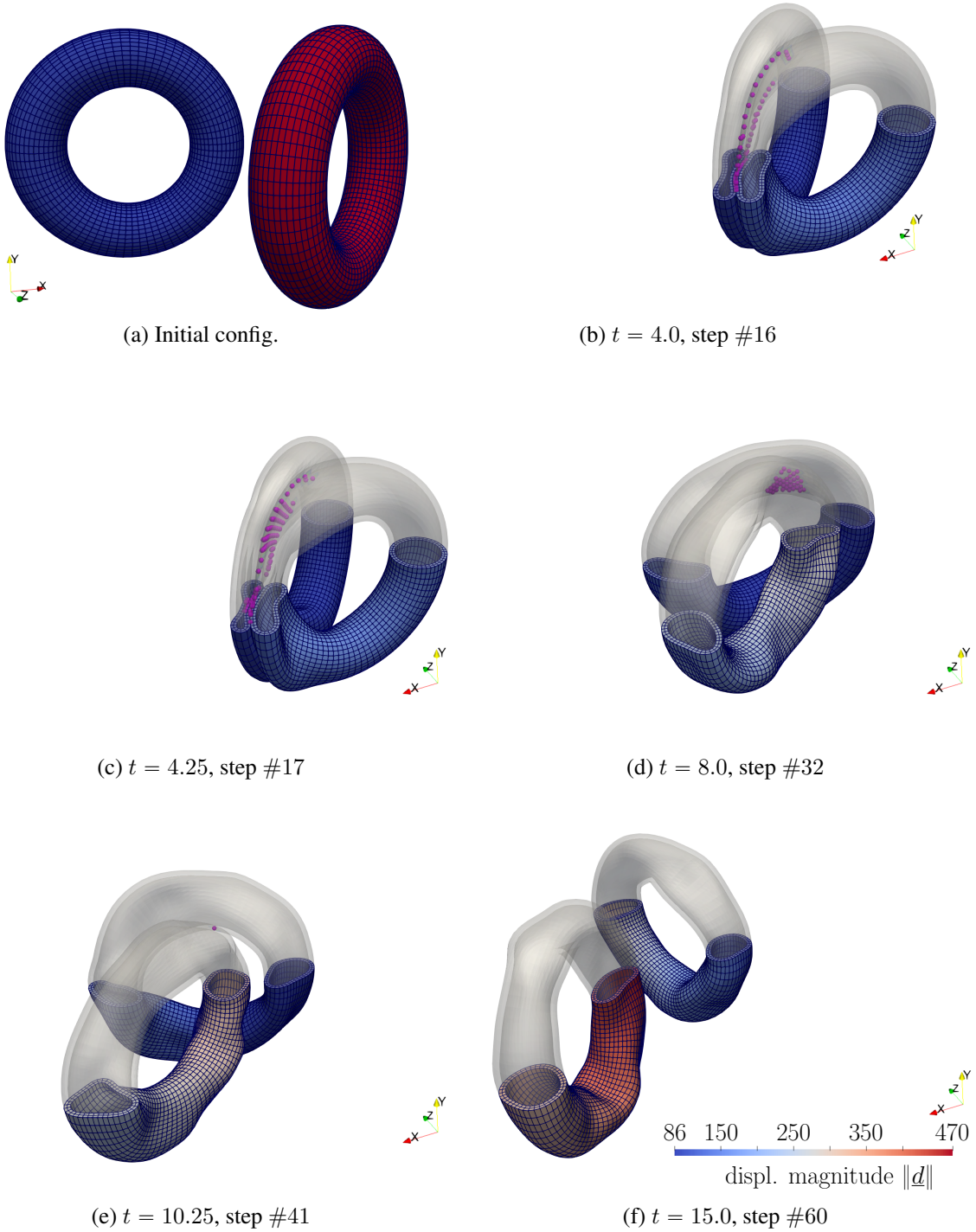
Figure 6.29.: Summary of different statistics about the impacting tori example.

solutions are presented in Figure 6.30. The self contact occurs for $t \in \{4.25, 4.75, 5.0\}$ and is exemplarily shown in Figure 6.30c.

A summary concerning the non-linear solver performance can be found in Figure 6.29. Even though the time step size has been halved, there are still time steps which need a matrix correction. However, the adaption becomes often necessary directly after the predictor step. This indicates that the performance might be further improved by a better predictor. The line search is only activated in load step 18 and the second order correction step helps to avoid a step length reduction in load step 18 and 19. In summary, this and the previous examples demonstrate impressively how well the line search filter method can handle a set of very different cumbersome situations which would fail otherwise.

6.11. Conclusion

A line search filter method for large deformation frictionless contact problems has been presented in this chapter. Thereby, a number of challenging issues have been revealed and comprehensively discussed. While the core functionality of the filter method considered here is largely inspired by [270–272], numerical experiments have indicated that further extensions and modifications are necessary to reach a robust and reliable globally convergent optimization algorithm for computational contact problems. One very important ingredient is the correction algorithm of the linear system of equations including the reliable identification of invalid elements, see Section 6.6, as well as the incorporation of the modified Newton method presented in Chapter 5. Algorithm 6.3 helps not only in case of the filter method, but it is also supportive in case of structural instabilities or whenever the iterative linear solver might face convergence problems. A wide variety of further important extensions has been summarized in Section 6.7. For example, since the method is applied to an already discretized problem, it is also important to tackle local issues which might not be noticeable on a global level. Therefore, pre-testing strategies have been introduced which help to identify locally distorted elements before the actual filter method is entered. Known issues which can be solved by bypassing of the $\mathscr{L}$-type test or the reinitialization of the filter have been addressed as well. These additional strategies are not al-



(a) Initial config.

(b) $t = 4.0$, step #16

(c) $t = 4.25$, step #17

(d) $t = 8.0$, step #32

(e) $t = 10.25$, step #41

(f) $t = 15.0$, step #60

Figure 6.30.: Visualization of the impacting tori example. Figure 6.30a shows the initial positioning of the two tori ($\|\underline{d}\| = 0, t = 0$), while Figures 6.30b to 6.30f give impressions of different time steps. The color coding follows again the same purpose as in Figure 6.27.

ways necessary, but under certain circumstances they might be the only way to a solution. This is for the implementation shown here especially true, since currently no feasibility restoration phase can be used as fallback strategy. This point needs to be addressed in the future.

At the beginning of this chapter it has been stated that for all considered examples the solution could be found without such a strategy and that all failing simulations failed because of other issues, e.g., badly chosen boundary conditions. However, there might be cases were a feasibility restoration phase would still be beneficial. For example, it might help to avoid the necessary adaption of the reinitialization parameters for certain examples and thus the usability of the method could be further improved. Examples for a meaningful feasibility restoration phase can be found in [263, 272], for instance.

Additionally, the successful treatment of advanced computational contact topics such as dynamic problems with focus on the Generalized-α method or the consideration of the EAS method have been addressed. Thereby, the numerical examples gave the impression that the correction of the linear system of equations can have a beneficial impact on the applicability of the EAS method in the presence of large compression, see also the stability issue described in Remark 2.5.

In summary, the presented line search filter method shows an impressive performance in all discussed numerical examples. The solvability of all the proposed simulations could be improved with the new line search filter method. For some examples it might have been possible to find a solution with a full Newton method by adapting the load/time step size, however, other examples which face structural instabilities are hardly solvable with classical pure Newton-Raphson approaches. Therefore, the new line search filter method can help to significantly improve the solvability of non-linear large scale (static or dynamic) contact problems by simultaneously generating only minimal overhead. Furthermore, the non-linear solution method does not become excessively restrictive due to the used acceptability checks, see exemplarily Section 6.10.1. Therefore, it seems promising to transfer the algorithms to other complex problems. An obvious candidate is the frictional contact problem, see Section 2.2.3 for a starting point.

7. Summary and Outlook

This thesis has dealt with the very important topic of finding ways to improve the robustness of existing computational contact simulation tools. Thereby, the focus has mainly lain on mortar-like contact formulation. The discretization methods of choice have been based on linear Lagrangian polynomials or non-linear rational B-splines. However, the specific discretization method itself has been of rather minor importance besides the single exception of the presented integration issue in Section 4.7.2 where the enhanced smoothness of the NURBS function can become handy to overcome certain issues.

But, before the most important achievements of this thesis are going to be recapitulated, the attention shall be drawn back to the beginning of this thesis. The entire story has begun with a brief introduction into the field of computational mechanics for large deformations. Already during these introductory words the idea has been to give a different point of view by shifting to a more mathematical instead of a pure engineering perspective. This had the important task to build and strenghten the bridge between both disciplines. Therefore, the derivation of the weak form has been discussed more comprehensively. Afterwards, the attention has been on contact mechanics and its special characteristics. However, the idea has always remained to provide a mathematical discussion which fits well with the constrained optimization literature. After a short interlude about frictional contact problems, which can be understood as a first foretaste for future developments, Chapter 2 ended with some further insights into spatial and temporal discretization methods. This part has been restricted to the essential key points which felt most important for this thesis.

Then, in the next Chapter 3, the discussion turned to the mathematical foundation of this work: the field of numerical (un-)constrained optimization. This is a vast and quickly developing research field, which is almost impossible to capture in its full extent. Therefore, this second introductory chapter summarized first important local and global techniques for the simpler case of unconstrained optimization, right before the focus had been shifted to the more complex field of constrained optimization. All in all, the provided content is still only scratching on the surface of all the therein mentioned locally and globally convergent methods and strategies. For a deeper insight, it is highly recommended to take a look at the referred literature.

Next, with these tools at hand, the main part of this thesis starts in Chapter 4. Therein, two novel contact formulations have been derived side by side which have much in common with the publications of Alart and Curnier [1], De Lorenzis et al. [65] and Pietrzak and Curnier [212]. However, the main difference is the combination of a mortar-based formulation together with a so-called ray-tracing normal field and a thereon based projection method. The ray tracing approach has been chosen in contrast to the often used closest-point projection. The applied smoothed ray-tracing approach can be also found in Popp [213] or Yang et al. [290], for instance. Actually, this method is often used in conjunction with a dual-mortar method for large deformation contact formulations. Despite the fact that the choice of the normal definition is of only minor importance for the actual final converged solution, it can be of essential importance

for the non-linear solver behavior. In summary, there exist many reasons why this thesis has chosen the ray-tracing approach. For example, the ray tracing, which is mainly defined on the slave side, can be beneficial when it comes to the actual implementation in a high performance computing framework which relies on an efficient parallel treatment of its individual ingredients. Another point is that this choice allowed to build directly on top of the work of Popp [213], Popp et al. [215, 216, 217, 218] and to utilize as many synergies as possible with the already existing mortar-based contact formulations in BACI (cf. Wall and Kronbichler [274]) at the Institute for Computational Mechanics.

The two approaches which have been developed in Chapter 4 contain firstly a symmetric and truly variationally consistent formulation and, secondly, a slightly variationally inconsistent variant, see also Hiermeier et al. [131]. The latter one has the advantage to be computationally much more efficient. Throughout the discussion and derivation, however, it could be shown that both approaches satisfy the balance of linear and angular momentum up to numerical precision. Especially, the balance of angular momentum is remarkable since this point is violated by many mortar-like formulations as comprehensively discussed in Section 4.5. Afterwards a comparison of the two noval methods with a well-established implementation summarized in Popp [213] is given. The mortar-like contact method proposed by Popp [213] is one of the candidates which, due to the applied variational approach, lack the conservation of angular momentum. Furthermore, one drawback of the variationally consistent formulation derived in this thesis is revealed, namely, the much stronger demand for a numerical exact integration. Certain remedies on this issue have been discussed as well. Finally, in the end of Chapter 4, another new discovery has been presented: Here, precisely in Section 4.7.4, an inherent instability of the variationally inconsistent as well as of the well-established formulation by [213] has been presented. To the author's knowledge this is the first time that the phenomenon has been discussed and explained. Additionally, the entire chapter comes a long with a comprehensive and detailed derivation of all necessary variations and linearizations, thus that both presented approaches fulfill the requirements for a locally quadratic convergence.

After this strongly contact related discussion, the attention has been drawn to one of the weaknesses of the previous chapter: the not always reliable convergence behavior, especially for large initial penetrations. In Chapter 5, a modified variant of Newton's method for constrained problems has been introduced. The foundation for this chapter can be found in Bertsekas [23]. However, many of the therein discussed steps represent completely novel extensions to the already existing literature. For example, the dynamic correction schemes for the regularization parameter c_{N} in Section 5.3. These schemes allow for the first time a much easier definition of the regularization parameter. Instead of choosing it in dependence on the problem dependent requirements, it is now possible to set the initial c_{N} value always to 1.0 and the associated dynamic correction increases its value on demand while always staying bounded. This alone is a significant improvement concerning usability. However, the modification of the linear system of equations allows also a much higher degree of initial penetration in case of displacement controlled problems. Furthermore, in combination with the mentioned dynamic c_{N} parameter correction, it has become possible to prove local convergence based on the ideas given in Bertsekas [23]. Furthermore, a novel switching strategy has been proposed which relies only on mathematically motivated conditions such as angles between certain gradient directions and relative residual measures rather than on problem dependent parameters. This stands in clear contrast to the ideas which can be found in other publications dealing with similar problems, see e.g. [294]. The new modified lin-

ear system of equations has additionally the advantage that the Lagrange multiplier can be quite easily condensed from the original saddle-point structure. This option has been also discussed in Chapter 5. Furthermore, the implications of such a procedure have been investigated under consideration of a detailed numerical study concerning the conditioning of the linear system. This study had the result that the original saddle-point system of equations might be beneficial as long as no special preconditioning strategy is available. The main part of the examples in Section 5.6 have put the focus on the variationally inconsistent formulation of Chapter 5. This stems mainly from the fact that the variationally consistent formulation suffers from the discussed numerical integration issues. However, the modification of the linear system presented in Chapter 5 does not depend on the inconsistencies introduced by the related contact formulation and in order to underline this, one more example has been added which considers the truly variationally consistent formulation. This example has revealed further interesting properties of the modified Newton method and has further expanded the knowledge about the differences of the two mortar-like contact formulations.

To this point all discussed methods and novel ideas have severely helped to improve the understanding of the mortar-like contact formulations and helped to improve the robustness of the methods, however, all of them are just locally convergent. In Chapter 6, therefore, the next logical step has been made towards a globally convergent method. For this thesis the so-called line search filter method has been chosen as one representative of a globally convergent constrained optimization strategy. The main algorithm presented in Chapter 6 has been largely inspired by the work of Wächter and Biegler [270, 271, 272]. However, the meaningful formulation of the filter coordinates as well as certain smaller enhancements in the calculation of the minimal step length estimates represent first smaller adjustments to the original algorithm. Furthermore, the consideration of a second order correction step has become necessary and had to be adapted to fit the requirements of the large-scale contact simulations considered here. Next, the inertia correction algorithm for the linear system of equations proposed in Wächter and Biegler [272] had to be significantly changed since certain information about the linear system of equations were just not at hand for the very huge parallel distributed linear systems of equations which have been solved within this thesis. Therefore, another strategy had to be developed which provided similar information about the solvability and quality of the search direction. Thereby, it has been decisive to detect and compensate a non-positive definite search direction since otherwise the following line search method would have been unable to find a sufficient step length to improve one of the filter measures. The novel linear system correction method contains further ingredients such as the reliable detection of invalid finite elements. This add-on is based on the work of Johnen et al. [150, 151] and represents another important contribution of this thesis when it comes to the treatment of globalization method for discretized highly non-linear (contact) problems. Additionally, it could be shown that the correction of the linear system is also beneficial for the general application of iterative linear solvers, such as GMRES, since the convergence of these solvers is no longer strictly bound to the correct initial parameter choice. Instead, the algorithm gains the power to heal itself. Furthermore, the invalid element detection has also been used as a pre-testing mechanism for the filter method. This is necessary, since numerical experiments have revealed that the typically used filter measures are not always able to reliably identify bad steps which lead to a very localized heavy mesh distortion. Besides this pre-testing approach many more small adjustments to the original line search filter method have been introduced. Some of them can also be found in the related literature such as the bypassing

of the $\mathscr{L}$-type switching test which is also mentioned in Wächter and Biegler [272], the reinitialization of the filter as discussed in [271, 272], or the pre-filtering which can be found in Gould and Toint [116] and Milzarek and Ulbrich [196]. These ideas have been only adapted for the contact problems considered here. However, there are also completely new ingredients such as the controlled decrease of the regularization parameter c_{N} which is discussed in Section 6.7.4, the scaling approach mentioned in Section 6.7.5, or the extensions of the filter measures for the treatment of dynamic contact problems which has been introduced in 6.8.1. Furthermore, the handling of enhanced assumed strains (EAS), which has been mentioned in Section 6.8.2, might seem to be only a small improvement, but is actually a very important contribution to the whole picture. When it comes to efficiency, the newly developed dynamic parallel redistribution must be mentioned as well.

All in all, the methods and algorithms discussed and presented in this thesis have proven to significantly improve the performance of classical non-linear solution approaches in numerous examples. Furthermore, the detailed discussions and provided background information have helped at many points to explain the observed behavior and to improve the understanding of the frictionless computational contact mechanics. Nevertheless, many points are still remaining on the agenda which ask for further investigations and extensions. Some of the most important open issues are summarized in the following in chronological order.

The numerical integration of the variationally consistent mortar-like contact formulation presented in Chapter 4 should be further investigated and resolved by a suitable extension of the segment-based integration approach (cf. Puso and Laursen [222], Wilking and Bischoff [278], Yang et al. [290]). Afterwards, it might be advisable to put more effort into the investigation of the inherent instability presented in Section 4.7.4. Furthermore, a stable, truly variationally consistent formulation would also allow a more distinct investigation of the modified Newton approach discussed in Chapter 5 as well as the line search filter method discussed in Chapter 6. The latter one has only been considered together with the variationally inconsistent formulation until now.

Another interesting research topic could be the extension of certain general ideas presented within this thesis to frictional contact problems. For example, it would be a great step forward if the dynamic correction of the c_{N} regularization parameter presented in Chapter 5 could also be carried over to the c_{τ} parameter of the frictional case, since the appropriate identification of this parameter is often much more difficult. Furthermore, also the application of the filter method could be of great interest. A possible suitable extension to a frictional contact formulation has been already presented in Section 2.2.3.

With respect to Chapter 5 and Chapter 6, it could be also beneficial to improve the used tangential predictor further, since some numerical experiments have become harder to solve just because of a bad initial displacement and/or Lagrange multiplier field. Furthermore, the modified Newton approach asks for further input to become applicable to problems with pure Neumann loads or mixed Neumann/Dirichlet boundary conditions.

Finally, the filter method itself can be improved by introducing a meaningful feasibility restoration phase. Even though all of the examples could be solved without a feasibility restoration phase, some of them asked for a special adaption of the default reinitialization parameters. A feasibility restoration might be the proper remedy here, since it would become only active as soon as the filter method parameterized with the default parameters given in Section 6.9.2 starts

failing. First ideas for a possible feasibility restoration phase can be found in Ulbrich et al. [263], Wächter and Biegler [272], for instance.

In conclusion, it can be said that the thesis has succeeded in achieving the initial goals. The presented methods help to create a more robust non-linear solver behavior. At the same time, however, it was also possible to explain in more detail inconsistencies in some formulations that had existed for some time and to point out possible solutions. The gap between engineering and mathematics could also be reduced again. This provides a promising basis for further improvements in the future.

A. Variation and Linearization of Basic Contact Terms

In this Appendix details to certain variations and linearizations introduced or mentioned in Chapter 4 are given. These additional details, together with the equations in the aforementioned chapter, should be enough to give a complete picture of the underlying mathematical relations necessary for the variationally complete and incomplete mortar-like contact formulations.

A.1. Variation of Basic Contact Terms

This section contains all remaining first order derivatives of Section 4.4.1. In the following, the element superscript will be omitted and the variation of the convective covariant base vectors follows first. The associated first order directional derivative yields

$$\begin{aligned} \mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[b]}{}_{k}) =& \frac{\partial}{\partial\xi^{[b]k}}\frac{\partial(X^{[b]i}+u^{[b]i})}{\partial u^{r}}\delta u^{r} + \frac{\partial\underline{\tau}^{[b]}{}_{k}}{\partial\xi^{[b]l}}\,\mathrm{D}_{\delta\underline{u}}(\xi^{[b]l}) \\ =& \frac{\partial(\delta\underline{u}^{[b]})}{\partial\xi^{[b]k}} + \frac{\partial\underline{\tau}^{[b]}{}_{k}}{\partial\xi^{[b]l}}\,\mathrm{D}_{\delta\underline{u}}(\xi^{[b]l}). \end{aligned} \tag{A.1}$$

The occurring directional derivative of the parameter space coordinate is equal to zero on the slave side. Further, the discretized form for $\underline{u}^{[b]}$ is inserted and

$$\frac{\partial(\delta\underline{u}^{[b]})}{\partial\xi^{[b]k}} = N^{[b]}{}_{i,\xi^{k}}\delta d^{[b]li}\underline{e}_{l} \tag{A.2}$$

is obtained. The next step is the calculation of the nodal averaged normal $\tilde{\underline{n}}^{[1]i}$, which is defined in (4.3) as the sum over the outward-pointing unit normals $\underline{n}^{[1](e)}$ of the adjacent elements evaluated at the parameter space coordinates of node i. Thus, the relation

$$\mathrm{D}_{\delta\underline{u}}(\tilde{\underline{n}}^{[1]}) = \sum_{e=1}^{\mathrm{N}_{\mathrm{e}}^{\mathrm{adj}}} \mathrm{D}_{\delta\underline{u}}(\underline{n}^{[1](e)}) \tag{A.3}$$

is received, where

$$\mathrm{D}_{\delta\underline{u}}(\underline{n}^{[1]}) = \mathrm{D}_{\delta\underline{u}}\left(\frac{\hat{\underline{n}}^{[1]}}{\|\hat{\underline{n}}^{[1]}\|}\right) = \frac{1}{\|\hat{\underline{n}}^{[1]}\|}(\underline{\underline{I}} - \underline{n}^{[1]}\otimes\underline{n}^{[1]})\,\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}), \tag{A.4}$$

$$\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}) = \mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{1}\times\underline{\tau}^{[1]}{}_{2}) = \mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{1})\times\underline{\tau}^{[1]}{}_{2} + \underline{\tau}^{[1]}{}_{1}\times\mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{2}). \tag{A.5}$$

The directional derivative of the element Jacobian determinant on the slave side follows as

$$\begin{aligned}\mathrm{D}_{\delta\underline{u}}(j^{[1]}) =& \mathrm{D}_{\delta\underline{u}}(\|\hat{\underline{n}}^{[1]}\|) = \underline{n}^{[1]} \cdot \mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}) \\ =& \frac{1}{j^{[1]}} \left\{ \left(\tau^{[1]}{}_{m2} \tau^{[1]}{}_{n1} - \tau^{[1]}{}_{m1} \tau^{[1]}{}_{n2} \right) \tau^{[1]m}{}_{2}\, \mathrm{D}_{\delta\underline{u}}(\tau^{[1]n}{}_{1}) \right. \\ & \left. + \left(\tau^{[1]}{}_{m1} \tau^{[1]}{}_{n2} - \tau^{[1]}{}_{m2} \tau^{[1]}{}_{n1} \right) \tau^{[1]m}{}_{1}\, \mathrm{D}_{\delta\underline{u}}(\tau^{[1]n}{}_{2}) \right\}. \end{aligned} \tag{A.6}$$

Finally, the projection rule from (2.54) is restated as

$$\underline{\chi}(\underline{x}(\xi^{[1]i}), \underline{x}(\xi^{[2]i}), \alpha_\chi) = \underline{x}(\xi^{[2]i}) - \underline{x}(\xi^{[1]i}) - \alpha_\chi \breve{\underline{n}}^{[1]}(\xi^{[1]i}) \tag{A.7}$$

and the directional derivative at the solution point is calculated for the derivation of the missing derivative of the master parametric coordinates. The directional derivative at the solution point $(\bar{\underline{x}}^{[2]}, \bar{\alpha}_\chi)$ with respect to the displacement degrees of freedom follows as

$$\begin{aligned} 0 =& \mathrm{D}_{\delta\underline{u}}(\underline{\chi}(\underline{x}(\xi^{[1]i}), \underline{x}(\bar{\xi}^{[2]i}), \bar{\alpha}_\chi)) \\ =& \delta\bar{\underline{u}}^{[2]} + \bar{\underline{\tau}}^{[2]}{}_{i}\, \mathrm{D}_{\delta\underline{u}}(\bar{\xi}^{[2]i}) - \delta\underline{u}^{[1]} - \mathrm{D}_{\delta\underline{u}}(\bar{\alpha}_\chi)\, \breve{\underline{n}}^{[1]} - \bar{\alpha}_\chi \mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]}). \end{aligned} \tag{A.8}$$

By inserting the directional derivative of the smooth normal field (4.28), (A.8) can be solved for the unknown directional derivatives of the master parameter space coordinates and the distance factor $\bar{\alpha}_\chi$:

$$\begin{pmatrix} \mathrm{D}_{\delta\underline{u}}(\bar{\xi}^{[2]1}) \\ \mathrm{D}_{\delta\underline{u}}(\bar{\xi}^{[2]2}) \\ \mathrm{D}_{\delta\underline{u}}(\bar{\alpha}_\chi) \end{pmatrix} = \bar{\underline{\underline{L}}}_{\chi}^{-1} \cdot \left(\delta\underline{u}^{[1]} + \bar{\alpha}_\chi \mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]}) - \delta\bar{\underline{u}}^{[2]} \right) \qquad \forall\ \delta\underline{u}, \tag{A.9}$$

where the matrix $\bar{\underline{\underline{L}}}_{\chi} \in \mathbb{R}^{3\times 3}$ is defined as

$$\bar{\underline{\underline{L}}}_{\chi} = \begin{pmatrix} \bar{\underline{\tau}}_1^{[2]} & \bar{\underline{\tau}}_2^{[2]} & -\breve{\underline{n}}^{[1]} \end{pmatrix}. \tag{A.10}$$

It is noted that the actual right and left hand sides are matrices of row dimension three and problem dependent column dimension.

A.2. Linearization of Basic Contact Terms

This section contains all remaining first and second order derivatives from Section 4.4.3. At the beginning the linearized first derivative of the virtual displacement shall be considered, i.e.

$$\mathrm{D}_{\Delta\underline{u}} \left(\frac{\partial(\delta\underline{u}^{[b]})}{\partial \xi^{[b]i}} \right) = \frac{\partial^2(\delta\underline{u}^{[b]})}{\partial \xi^{[b]i} \partial \xi^{[b]j}} \mathrm{D}_{\Delta\underline{u}}(\xi^{[b]j}). \tag{A.11}$$

The term will vanish for all slave contributions. Next, the linearization of the first order derivative of the convective element base vectors is derived as

$$\mathrm{D}_{\Delta\underline{u}}\left(\frac{\partial \underline{\tau}^{[b]}{}_{i}}{\partial \xi^{[b]k}}\right) = \mathrm{D}_{\Delta\underline{u}}\left(\frac{\partial^2 \underline{x}^{[b]}}{\partial \xi^{[b]i}\partial \xi^{[b]k}}\right) = \frac{\partial^2(\Delta\underline{u}^{[b]})}{\partial \xi^{[b]i}\partial \xi^{[b]k}} + \frac{\partial^2 \underline{\tau}^{[b]}{}_{i}}{\partial \xi^{[b]k}\partial \xi^{[b]l}}\mathrm{D}_{\Delta\underline{u}}(\xi^{[b]l}). \tag{A.12}$$

The second term is equal to zero for linear polynomials as well as for convective base vectors on the slave side. For the master side, the directional derivative of the projected coordinates, defined in (A.9), has to be taken into account.

The linearization of the total variation of the base vectors follows as

$$\begin{aligned}\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[b]}{}_{i})) =& \mathrm{D}_{\Delta\underline{u}}\left(\frac{\partial(\delta\underline{u}^{[b]})}{\partial \xi^{[b]i}}\right) + \mathrm{D}_{\Delta\underline{u}}\left(\frac{\partial \underline{\tau}^{[b]}{}_{i}}{\partial \xi^{[b]k}}\right)\ \mathrm{D}_{\delta\underline{u}}(\xi^{[b]k}) \\ &+ \frac{\partial \underline{\tau}^{[b]}{}_{i}}{\partial \xi^{[b]k}}\ \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\xi^{[b]k})).\end{aligned} \tag{A.13}$$

All terms are again equal to zero on the slave side. The last term depends on the linearization of the variation of the projection. The linearization of the variation of the element normal vector $\hat{\underline{n}}^{[1]}$ defined in (A.5) is computed by

$$\begin{aligned}\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]})) =&\ \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{1})) \times \underline{\tau}^{[1]}{}_{2} + \mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{1}) \times \mathrm{D}_{\Delta\underline{u}}(\underline{\tau}^{[1]}{}_{2}) \\ &+ \mathrm{D}_{\Delta\underline{u}}(\underline{\tau}^{[1]}{}_{1}) \times \mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{2}) + \underline{\tau}^{[1]}{}_{1} \times \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{2})) \\ =&\ \mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{1}) \times \mathrm{D}_{\Delta\underline{u}}(\underline{\tau}^{[1]}{}_{2}) + \mathrm{D}_{\Delta\underline{u}}(\underline{\tau}^{[1]}{}_{1}) \times \mathrm{D}_{\delta\underline{u}}(\underline{\tau}^{[1]}{}_{2}),\end{aligned} \tag{A.14}$$

where all zero terms have been removed in the last step. By consideration of (A.4), (A.5) and (A.14) the linearized form of the variation of the element Jacobian on the slave side is derived by

$$\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(j^{[1]})) = \mathrm{D}_{\Delta\underline{u}}(\underline{n}^{[1]}) \cdot \mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}) + \underline{n}^{[1]} \cdot \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]})). \tag{A.15}$$

The linearized variation of the smooth non-unit normal field is still missing. Therefore, the following equations are obtained

$$\begin{aligned}\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]})) =& N^{[1]}{}_{i}\frac{1}{\|\tilde{\underline{n}}^{[1]i}\|}\Big\{-\langle \breve{\underline{n}}^{[1]i},\ \mathrm{D}_{\Delta\underline{u}}(\tilde{\underline{n}}^{[1]i})\rangle\ \mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]i}) \\ &-[\mathrm{D}_{\Delta\underline{u}}(\breve{\underline{n}}^{[1]i}) \otimes \breve{\underline{n}}^{[1]i} + \breve{\underline{n}}^{[1]i} \otimes \mathrm{D}_{\Delta\underline{u}}(\breve{\underline{n}}^{[1]i})] \cdot \mathrm{D}_{\delta\underline{u}}(\tilde{\underline{n}}^{[1]i}) \\ &+[\underline{I} - \breve{\underline{n}}^{[1]i} \otimes \breve{\underline{n}}^{[1]i}] \cdot \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\tilde{\underline{n}}^{[1]i}))\Big\},\end{aligned} \tag{A.16}$$

where the directional derivative of the smooth and averaged normals are defined in (4.28) and (A.3), respectively, and

$$\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\tilde{\underline{n}}^{[1]})) = \sum_{e=1}^{\mathrm{N}_{\mathrm{e}}^{\mathrm{adj}}} \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\underline{n}^{[1](e)})), \tag{A.17}$$

$$\begin{aligned}\mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\underline{n}^{[1]})) =& \frac{1}{\|\hat{\underline{n}}^{[1]}\|}\Big\{-\langle \underline{n}^{[1]},\ \mathrm{D}_{\Delta\underline{u}}(\hat{\underline{n}}^{[1]})\rangle\, \mathrm{D}_{\delta\underline{u}}(\underline{n}^{[1]}) \\ & -[\mathrm{D}_{\Delta\underline{u}}(\underline{n}^{[1]}) \otimes \underline{n}^{[1]} + \underline{n}^{[1]} \otimes \mathrm{D}_{\Delta\underline{u}}(\underline{n}^{[1]})] \cdot \mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}) \\ & +[\underline{\underline{I}} - \underline{n}^{[1]} \otimes \underline{n}^{[1]}] \cdot \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\hat{\underline{n}}^{[1]}))\Big\}. \end{aligned} \tag{A.18}$$

The last two equations are supposed to be evaluated at the parameter space coordinates of the current node i with regard to the sum in (A.16). As a last step, the linearization of the variation of the projection at the solution point $(\bar{\underline{x}}^{[2]}, \bar{\alpha}_\chi)$ must be calculated. The directional derivative of (A.8) follows as

$$\begin{aligned} 0 =& \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\underline{\chi}(\underline{x}(\xi^{[1]i}), \underline{x}(\bar{\xi}^{[2]i}), \bar{\alpha}_\chi))) \\ =& \mathrm{D}_{\Delta\underline{u}}(\delta\underline{u}^{[2]}) + \mathrm{D}_{\Delta\underline{u}}(\bar{\underline{\tau}}^{[2]}{}_i)\mathrm{D}_{\delta\underline{u}}(\bar{\xi}^{[2]i}) + \bar{\underline{\tau}}^{[2]}{}_i \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\bar{\xi}^{[2]i})) - \mathrm{D}_{\Delta\underline{u}}(\delta\underline{u}^{[1]}) \\ & - \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\bar{\alpha}_\chi))\breve{\underline{n}}^{[1]} - \mathrm{D}_{\delta\underline{u}}(\bar{\alpha}_\chi)\mathrm{D}_{\Delta\underline{u}}(\breve{\underline{n}}^{[1]}) - \mathrm{D}_{\Delta\underline{u}}(\bar{\alpha}_\chi)\mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]}) - \bar{\alpha}_\chi \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]})) \end{aligned} \tag{A.19}$$

By inserting the second order directional derivatives of the smooth normal field (A.16) the derived set of linear equations (A.19) can be solved for the unknown second order directional derivatives and the following result is obtained:

$$\begin{pmatrix} \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\bar{\xi}^{[2]1})) \\ \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\bar{\xi}^{[2]2})) \\ \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\bar{\alpha}_\chi)) \end{pmatrix} = \bar{\underline{\underline{L}}}_\chi^{-1} \cdot \bar{\underline{r}}_\chi, \qquad \forall\ \delta\underline{u},\ \Delta\underline{u}, \tag{A.20}$$

$$\begin{aligned} \bar{\underline{r}}_\chi =& \mathrm{D}_{\delta\underline{u}}(\bar{\alpha}_\chi)\mathrm{D}_{\Delta\underline{u}}(\breve{\underline{n}}^{[1]}) + \mathrm{D}_{\Delta\underline{u}}(\bar{\alpha}_\chi)\mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]}) + \bar{\alpha}_\chi \mathrm{D}_{\Delta\underline{u}}(\mathrm{D}_{\delta\underline{u}}(\breve{\underline{n}}^{[1]})) - \mathrm{D}_{\Delta\underline{u}}(\delta\underline{u}^{[2]}) \\ & - \mathrm{D}_{\Delta\underline{u}}(\bar{\underline{\tau}}^{[2]}{}_i)\mathrm{D}_{\delta\underline{u}}(\bar{\xi}^{[2]i}), \end{aligned} \tag{A.21}$$

where the matrix $\bar{\underline{\underline{L}}}_\chi$ is defined in (A.10). In addition, the fact can be used that $\mathrm{D}_{\Delta\underline{u}}(\delta\underline{u}^{[1]})$ is equal to zero. It is to note once more that the left and right hand sides represent tensors of order three and only the first dimension is a-priori known to be equal to three. The remaining two dimensions are problem dependent.

B. Tangential Predictor for Large Sliding Steps

Since the predictor step is a crucial ingredient of the non-linear solution procedure discussed in this thesis, this topic will be explicitly taken up and explained in context of the modified system of equations introduced in Chapter 5. First of all, it is to emphasize that there are many ways to define a meaningful predictor and, therefore, only one possibility will be addressed. However, the algorithm presented here performs pretty well in all considered numerical studies. Before the discussion can be started two new sets $\check{\mathcal{D}}$ and $\mathcal{D}$ shall be introduced to distinguish between the unknown and the prescribed degrees of freedom (DoF). The first set contains all DoF which are affected by applied Dirichlet boundary conditions while the latter one contains all free and, therefore, unknown DoF. The cardinality of each set is denoted by $|\check{\mathcal{D}}| = \check{n}$ and $|\mathcal{D}| = n$, respectively. Note that the following assumptions shall hold:

AS B.1. All nodes with prescribed DoF shall be excluded from the slave set, i.e. $\{i \in \mathcal{S} \mid \nexists d^{[1]ij} = \check{d}^k : j \in \{1,2,3\} \wedge k \in \check{\mathcal{D}}\}$.

AS B.2. A static simulation is considered and only steady-state solutions shall be computed.

AS B.3. All applied Neumann conditions $\underline{f}_{\text{ext}}$ are independent from the deformation and, therefore, must not be considered during the linearization.

However, it is possible to relax all of these assumptions by applying the necessary modifications to the system of equations. The discussion starts at the beginning of a new load step. The most general case is that the Neumann load as well as the prescribed displacements are simultaneously changed. Let us assume that from step s to $s+1$ the following changes shall be applied

$$\begin{pmatrix} \underline{d}^{\{s+1,0\}} \\ \underline{\check{d}}^{\{s+1,0\}} \end{pmatrix} = \begin{pmatrix} \underline{d}^{\{s\}} \\ \underline{\check{d}}^{\{s\}} \end{pmatrix} + \begin{pmatrix} \underline{0} \\ \Delta\underline{\check{d}}^{\{s\}} \end{pmatrix}, \qquad \underline{f}_{\text{ext}}^{\{s+1,0\}} = \underline{f}_{\text{ext}}^{\{s\}} + \Delta\underline{f}_{\text{ext}}^{\{s\}}, \tag{B.1}$$

where $\Delta\underline{\check{d}}^{\{s\}} \in \mathbb{R}^{\check{n}}$ might be unequal to zero for some Dirichlet degrees of freedom and the external force increment $\Delta\underline{f}_{\text{ext}}^{\{s\}} \in \mathbb{R}^{n}$ might be as well unequal to zero for some Neumann degrees of freedom. Under these prerequisites, the following system of equations is obtained

$$\left.\begin{pmatrix} \underset{n\times n}{\tilde{\nabla}^2_{\underline{d}\underline{d}}\mathscr{L}} & \underset{n\times m^{\mathcal{A}}}{-\tilde{\nabla}_{\underline{d}}\underline{\tilde{g}}_{\text{N}}^{\mathcal{A}}} & \underline{\underline{0}} \\ \underset{m^{\mathcal{A}}\times n}{-[\nabla_{\underline{d}}\underline{\tilde{g}}_{\text{N}}^{\mathcal{A}}]^{\text{T}}} & \underset{m^{\mathcal{A}}\times m^{\mathcal{A}}}{-\frac{1}{c_{\text{N}}}\underline{\underline{A}}^{\mathcal{A}}} & \underline{\underline{0}} \\ \underline{\underline{0}} & \underline{\underline{0}} & \underset{m^{\mathcal{I}}\times m^{\mathcal{I}}}{\underline{\underline{I}}} \end{pmatrix}\right|_{\{s\}} \begin{pmatrix} \underset{n\times 1}{\Delta\underline{d}} \\ \underset{m^{\mathcal{A}}\times 1}{\Delta\underline{\lambda}_{\text{N}}^{\mathcal{A}}} \\ \underset{m^{\mathcal{I}}\times 1}{\Delta\underline{\lambda}_{\text{N}}^{\mathcal{I}}} \end{pmatrix} = \left.\begin{pmatrix} \underset{n\times 1}{-\tilde{\nabla}_{\underline{d}}\mathscr{L}} - \underset{n\times\check{n}}{[\nabla_{\check{d}}(\tilde{\nabla}_{\underline{d}}\mathscr{L})]^{\text{T}}}\underset{\check{n}\times 1}{\Delta\underline{\check{d}}} - \Delta\underline{f}_{\text{ext}} \\ \underset{m^{\mathcal{A}}\times 1}{\underline{\tilde{g}}_{\text{N}}^{\mathcal{A}}} + \underset{m^{\mathcal{A}}\times\check{n}}{[\nabla_{\check{d}}\underline{\tilde{g}}_{\text{N}}^{\mathcal{A}}]^{\text{T}}}\underset{\check{n}\times 1}{\Delta\underline{\check{d}}} \\ \underline{0} \end{pmatrix}\right|_{\{s\}}, \tag{B.2}$$

where the rows with respect to the variation of the inactive Lagrange multipliers have been simplified under the assumption that all inactive Lagrange multipliers are equal to zero at the beginning of a new load step. This follows directly from the applied algorithm. Again, it is possible to condense the Lagrange multipliers of the active contributions following (5.2) and (5.4), such that the active part of (B.2) yields

$$\tilde{\nabla}^2_{\underline{d}\,\underline{d}} \mathscr{L}_{c_\mathrm{N}} \, \Delta\underline{d} = -\{[\nabla_{\check{\underline{d}}}(\tilde{\nabla}_{\underline{d}} \mathscr{L})]^\mathrm{T} + c_\mathrm{N} \tilde{\nabla}_{\underline{d}} \tilde{\underline{g}}^{\mathcal{A}}_\mathrm{N} [\underline{\underline{A}}^{\mathcal{A}}]^{-1} [\nabla_{\check{\underline{d}}} \tilde{\underline{g}}^{\mathcal{A}}_\mathrm{N}]^\mathrm{T}\} \, \Delta\check{\underline{d}} - \Delta\underline{f}_{\mathrm{ext}}, \tag{B.3}$$

where $\tilde{\nabla}_{\underline{d}} \mathscr{L} - c_\mathrm{N} \hat{\underline{g}}^{\mathcal{A}}_\mathrm{N} = 0$ has been inserted, which follows under the assumption that the reached balance among the internal, external and contact forces at the end of the previous load step has been successfully reached. Note that the second part of the remaining contribution on the right hand side, which represents the tangentially approximated reaction due to the changing gap induced by the Dirichlet conditions, has only relevance in one scenario: If the master contact surface is moved by a prescribed displacement. In all other cases the influence vanishes as direct consequence of Assumption B.1: The interface integrals will not contribute to the corresponding columns in the tangential stiffness matrix . The tangential predictor introduced here is used in many of the numerical examples presented in this thesis.

Bibliography

[1] P. Alart, and A. Curnier, A mixed formulation for frictional contact problems prone to Newton like solution methods. *Computer Methods in Applied Mechanics and Engineering*, **92**, 353–375, 1991.

[2] E. L. Allgower, and K. Georg, *Introduction to numerical continuation methods* volume 45. SIAM, 2003.

[3] U. Andelfinger, and E. Ramm, EAS-elements for two-dimensional, three-dimensional, plate and shell structures and their equivalence to HR-elements. *International Journal for Numerical Methods in Engineering*, **36**, 1311–1337, 1993.

[4] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*. (3rd ed.). Philadelphia, PA: Society for Industrial and Applied Mathematics, 1999.

[5] F. Armero, and E. Petőcz, Formulation and analysis of conserving algorithms for frictionless dynamic contact/impact problems. *Computer Methods in Applied Mechanics and Engineering*, **158**, 269–300, 1998.

[6] F. Armero, and E. Petőcz, A new dissipative time-stepping algorithm for frictional contact problems: formulation and analysis. *Computer Methods in Applied Mechanics and Engineering*, **179**, 151–178, 1999.

[7] M. Arroyo, and M. Ortiz, Local maximum-entropy approximation schemes: a seamless bridge between finite elements and meshfree methods. *International journal for numerical methods in engineering*, **65**, 2167–2202, 2006.

[8] C. Audet, and J. E. Dennis Jr., A pattern search filter method for nonlinear programming without derivatives. *SIAM Journal on Optimization*, **14**, 980–1010, 2004.

[9] I. Babuška, The Finite Element Method with Lagrangian Multipliers. *Numerische Mathematik*, **20**, 179–192, 1973.

[10] I. Babuška, and M. Suri, On locking and robustness in the finite element method. *SIAM Journal on Numerical Analysis*, **29**, 1261–1293, 1992.

[11] J. M. Ball, Convexity Conditions and Existence Theorems in Nonlinear Elasticity. *Archive of Rational Mechanics and Analysis*, **63**, 337–403, 1977.

[12] K.-J. Bathe, *Finite element procedures*. Prentice Hall, 1996.

[13] K.-J. Bathe, and F. Brezzi, Stability of finite element mixed interpolations for contact problems. *Proceedings della Accademia Nazionale dei Lincei*, **12**, 167–183, 2001.

[14] K.-J. Bathe, and A. B. Chaudhary, A solution method for planar and axisymmetric contact problems. *International Journal for Numerical Methods in Engineering*, **21**, 65–88, 1985.

[15] F. B. Belgacem, The Mortar finite element method with Lagrange multipliers. *Numerische Mathematik*, **84**, 173–197, 1999.

[16] F. B. Belgacem, P. Hild, and P. Laborde, The Mortar Finite Element Method for Contact Problems. *Mathematical and Computer Modelling*, **28**, 263–271, 1998.

[17] T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, and P. Krysl, Meshless methods: an overview and recent developments. *Computer methods in applied mechanics and engineering*, **139**, 3–47, 1996.

[18] T. Belytschko, W. K. Liu, B. Moran, and K. Elkhodary, *Nonlinear finite elements for continua and structures*. John Wiley & Sons, 2013.

[19] M. Benzi, G. H. Golub, and J. Liesen, Numerical solution of saddle point problems. *Acta Numerica*, **14**, 1–137, 2005.

[20] C. Bernardi, Y. Maday, and A. T. Patera, Domain Decomposition by the Mortar Element Method. In H. G. Kaper, and M. Garbey (Eds.), *Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters* (pp. 269–286). Luwer Academic Publishers, 1993.

[21] A. Bertram, *Elasticity and plasticity of large deformations: An introduction*. Berlin-Heidelberg: Springer, 2014.

[22] D. P. Bertsekas, Combined primal-dual and penalty methods for constrained minimization. *SIAM J. Control*, **13**, 521–544, 1975.

[23] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 1996.

[24] M. C. Biggs, Constrained minimization using recursive equality quadratic programming. In F. A. Lootsma (Ed.), *Numerical Methods for Nonlinear Optimization*. New York: Academic press, 1972.

[25] M. C. Biggs, Constrained minimization using recursive quadratic programming. In L.-C. W. Dixon, and G. P. Szego (Eds.), *Towards Global Optimization*. North Holland, 1975.

[26] M. C. Biggs, On the convergence of some constrained minimization algorithms based on quadratic programming. *IMA Journal of Applied Mathematics*, **21**, 67–81, 1978.

[27] M. Bischoff, *Theorie und Numerik einer dreidimensionalen Schalenformulierung*. PhD–thesis University of Stuttgart, 1999.

[28] I. D. L. Bogle, and J. D. Perkins, A new sparsity preserving quasi-Newton update for solving nonlinear equations. *SIAM journal on scientific and statistical computing*, **11**, 621–630, 1990.

[29] E. G. Boman, Ü. V. Çatalyürek, C. Chevalier, and K. D. Devine, The Zoltan and Isorropia parallel toolkits for combinatorial scientific computing: Partitioning, ordering and coloring. *Scientific Programming*, **20**, 129–150, 2012.

[30] A. Bompadre, B. Schmidt, and M. Ortiz, Convergence analysis of meshfree approximation schemes. *SIAM Journal on Numerical Analysis*, **50**, 1344–1366, 2012.

[31] J. Bonet, and A. J. Burton, A simple orthotropic, transversely isotropic hyperelastic constitutive equation for large strain computations. *Computer Methods in Applied Mechanics and Engineering*, **162**, 151–164, 1998.

[32] J. Bonet, and R. D. Wood, *Nonlinear continuum mechanics for finite element analysis*. Cambridge University Press, 2008.

[33] N. Bonfils, N. Chevaugeon, and N. Moës, Treating volumetric inequality constraint in a continuum media with a coupled X-FEM/level-set strategy. *Computer Methods in Applied Mechanics and Engineering*, **205-208**, 16–28, 2012.

[34] S. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[35] S. C. Brenner, and L. R. Scott, *The mathematical theory of finite element methods* volume 15. Springer Science & Business Media, 2007.

[36] F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, **8**, 129–151, 1974.

[37] F. Brezzi, and M. Fortin, *Mixed and Hybrid Finite Element Methods*. (15th ed.). Springer-Verlag, 2012.

[38] C. G. Broyden, The convergence of an algorithm for solving sparse nonlinear systems. *Mathematics of Computation*, **25**, 285–285, 1971.

[39] S. Brunssen, F. Schmid, M. Schäfer, and B. Wohlmuth, A fast and robust iterative solver for nonlinear contact problems using a primal-dual active set strategy and algebraic multigrid. *International Journal for Numerical Methods in Engineering*, **69**, 524–543, 2007.

[40] E. Burman, P. Hansbo, and M. G. Larson, The penalty-free Nitsche method and nonconforming finite elements for the Signorini problem. *SIAM Journal on Numerical Analysis*, **55**, 2523–2539, 2017.

[41] J. D. Buys, *Dual Algorithms for Constrained Optimization Problems*. PhD–thesis Rijksuniversiteit Leiden, Holland, 1972.

[42] R. H. Byrd, M. Marazzi, and J. Nocedal, On the convergence of Newton iterations to non-stationary points. *Mathematical Programming*, **99**, 127–148, 2004.

[43] N. J. Carpenter, Lagrange constraints for transient finite element surface contact. *International Journal for Numerical Methods in Engineering*, **32**, 103–128, 1991.

[44] M. Ceze, and K. J. Fidkowski, Constrained pseudo-transient continuation. *International Journal for Numerical Methods in Engineering*, **102**, 1683–1703, 2015.

[45] R. M. Chamberlain, M. J. D. Powell, C. Lemarechal, and H. C. Pedersen, The watchdog technique for forcing convergence in algorithms for constrained optimization. In A. G. Buckley, and J.-L. Goffin (Eds.), *Algorithms for Constrained Minimization of Smooth Nonlinear Functions* (pp. 1–17). Berlin, Heidelberg: Springer Berlin Heidelberg, 1982.

[46] F. Chouly, P. Hild, and Y. Renard, Symmetric and non-symmetric variants of Nitsche's method for contact problems in elasticity: theory and numerical experiments. *Mathematics of Computation*, **84**, 1089–1112, 2015.

[47] P. Christensen, A. Klarbing, J. Pang, and N. Strömberg, Formulation and comparison of algorithms for frictional contact problems. *International Journal for Numerical Methods in Engineering*, **42**, 145–173, 1998.

[48] J. Chung, and G. M. Hulbert, A Time Integration Algorithm for Structural Dynamics With Improved Numerical Dissipation: The Generalized-α Method. *Journal of Applied Mechanics*, **60**, 371–375, 1993.

[49] F. Clarke, *Optimization and Nonsmooth Analysis* volume 5. Philadelphia: SIAM, 1990.

[50] E. Cohen, and L. L. Schumaker, Rates of convergence of control polygons. *Computer Aided Geometric Design*, **2**, 229–235, 1985.

[51] L. Collatz, *Numerische Behandlung von Differentialgleichungen*. Berlin-Gbttingen- Heidelberg: Springer-Verlag, 2013.

[52] A. R. Conn, and N. I. M. Gould, On the Location of Directions of Infinite Descent for Nonlinear Programming Algorithms. *SIAM Journal on Numerical Analysis*, **21**, 1162–1179, 1984.

[53] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust region methods*. SIAM, 2000.

[54] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*. Philadelphia, PA: SIAM, 2009.

[55] R. Courant, Variational methods for the solution of problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society*, **49**, 1–24, 1943.

[56] M. A. Crisfield, A fast incremental/iterative solution procedure that handles "snap-through". *Computers and Structures*, **13**, 55–62, 1981.

[57] M. A. Crisfield, An arc-length method including line searches and accelerations. *International Journal for Numerical Methods in Engineering*, **19**, 1269–1289, 1983.

[58] M. A. Crisfield, Re-visting the contact patch test. *International Journal for Numerical Methods in Engineering*, **48**, 435–449, 2000.

[59] A. Curnier, Q.-C. He, and A. Klarbring, Continuum mechanics modelling of large deformation contact with friction. In M. Raous, M. Jean, and J. J. Moreau (Eds.), *Contact Mechanics* (pp. 145–158). Boston, MA: Springer, 1995.

[60] W. Dahmen, and A. Reusken, *Numerik für Ingenieure und Naturwissenschaftler*. Berlin-Heidelberg: Springer, 2006.

[61] W. C. Davidon, Variable Metric Method for Minimization. *SIAM Journal on Optimization*, **1**, 1–17, 1991.

[62] T. A. Davis, Algorithm 832: UMFPACK V4.3—an unsymmetric-pattern multifrontal method. *ACM Transactions on Mathematical Software (TOMS)*, **30**, 196–199, 2004.

[63] T. A. Davis, and E. Palamadai Natarajan, Algorithm 907: KLU, a direct sparse solver for circuit simulation problems. *ACM Transactions on Mathematical Software (TOMS)*, **37**, 36, 2010.

[64] S. De, and K.-J. Bathe, The method of finite spheres with improved numerical integration. *Computers & Structures*, **79**, 2183–2196, 2001.

[65] L. De Lorenzis, P. Wriggers, and G. Zavarise, A mortar formulation for 3D large deformation contact using NURBS-based isogeometric analysis and the augmented Lagrangian method. *Computational Mechanics*, **49**, 1–20, 2012.

[66] E. A. De Souza Neto, D. Perić, M. Dutko, and D. R. Owen, Design of simple low order finite elements for large strain analysis of nearly incompressible solids. *International Journal of Solids and Structures*, **33**, 3277–3296, 1996.

[67] E. A. de Souza Neto, D. Perić, G. C. Huang, and D. R. Owen, Remarks on the stability of enhanced strain elements in finite elasticity and elastoplasticity. *Communications in Numerical Methods in Engineering*, **11**, 951–961, 1995.

[68] J. E. Dennis Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, 1983.

[69] P. Deuflhard, *Newton methods for nonlinear problems: affine invariance and adaptive algorithms* volume 35. Springer Science & Business Media, 2011.

[70] P. Deuflhard, R. Krause, and S. Ertel, A contact-stabilized Newmark method for dynamical contact problems. *International Journal for Numerical Methods in Engineering*, **73**, 1274–1290, 2008.

[71] R. Dimitri, L. De Lorenzis, M. A. Scott, P. Wriggers, R. L. Taylor, and G. Zavarise, Isogeometric large deformation frictionless contact using T-splines. *Computer Methods in Applied Mechanics and Engineering*, **269**, 394–414, 2014.

[72] J. Dolbow, and T. Belytschko, Numerical integration of the Galerkin weak form in meshfree methods. *Computational Mechanics*, **23**, 219–230, 1999.

[73] T. X. Duong, L. D. Lorenzis, and R. A. Sauer, A segmentation-free isogeometric extended mortar contact method. *Computational Mechanics*, **63**, 383–407, 2019.

[74] T. X. Duong, and R. A. Sauer, An accurate quadrature technique for the contact boundary in 3D finite element computations. *Computational Mechanics*, **55**, 145–166, 2015.

[75] A. Eisenträger, *Nonmonotone Line Search and Trust Region Methods for Optimization*. Technical Report Computing Laboratory, University of Oxford, 2007.

[76] N. El-Abbasi, and K. J. Bathe, Stability and patch test performance of contact discretizations and a new solution algorithm. *Computers and Structures*, **79**, 1473–1486, 2001.

[77] N. El-Abbasi, S. A. Meguid, and A. Czekanski, On the modelling of smooth contact surfaces using cubic splines. *International Journal for Numerical Methods in Engineering*, **50**, 953–967, 2001.

[78] J. B. Erway, and R. F. Marcia, Algorithm 943: MSS: MATLAB Software for L-BFGS trust-region subproblems for large-scale optimization. *arXiv preprint arXiv:1212.1525*, , 2012.

[79] F. Facchinei, and S. Lucidi, Quadratically and superlinearly convergent algorithms for the solution of inequality constrained minimization problems. *Journal of Optimization Theory and Applications*, **85**, 265–289, 1995.

[80] J. Fan, and J. Pan, Convergence properties of a self-adaptive Levenberg-Marquardt algorithm under local error bound condition. *Computational Optimization and Applications*, **34**, 47–62, 2006.

[81] J.-Y. Fan, A modified Levenberg-Marquandt Algorithm for singular system of nonlinear equations. *Journal of Computational Mathematics*, **21**, 625–636, 2003.

[82] P. Farah, M. Gitterle, W. Wall, and A. Popp, Computational wear and contact modeling for fretting analysis with isogeometric dual mortar methods. *Key Engineering Materials*, **681**, 1–18, 2016.

[83] P. Farah, A. Popp, and W. A. Wall, Segment-based vs. element-based integration for mortar methods in computational contact mechanics. *Computational Mechanics*, **55**, 209–228, 2015.

[84] P. Farah, and A.-T. Vuong, Volumetric coupling approaches for multi-physics simulations on non-matching meshes. *International Journal for Numerical Methods in Engineering*, **108**, 1550–1576, 2016.

[85] P. Farah, W. Wall, and A. Popp, A mortar finite element approach for point, line, and surface contact. *International Journal for Numerical Methods in Engineering*, **114**, 255–291, 2018.

[86] P. Farah, W. A. Wall, and A. Popp, An implicit finite wear contact formulation based on dual mortar methods. *International Journal for Numerical Methods in Engineering*, **111**, 325–353, 2017.

[87] P. W. Farah, *Mortar Methods for Computational Contact Mechanics Including Wear and General Volume Coupled Problems*. PhD–thesis Technical University of Munich, 2017.

[88] S. Fernández-Méndez, and A. Huerta, Imposing essential boundary conditions in mesh-free methods. *Computer methods in applied mechanics and engineering*, **193**, 1257–1275, 2004.

[89] A. V. Fiacco, and G. P. McCormick, *Nonlinear programming: sequential unconstrained minimization techniques*. New York, N. Y.: John Wiley & Sons, 1968.

[90] K. A. Fischer, and P. Wriggers, Frictionless 2D contact formulations for finite deformations based on the mortar method. *Computational Mechanics*, **36**, 226–244, 2005.

[91] B. Flemisch, and B. I. Wohlmuth, Stable Lagrange multipliers for quadrilateral meshes of curved interfaces in 3D. *Computer Methods in Applied Mechanics and Engineering*, **196**, 1589–1602, 2007.

[92] R. Fletcher, A Class of Methods for Nonlinear Programming with Termination and Convergence Properties. In J. Abadie (Ed.), *Integer and Nonlinear Programming* (pp. 157–173). North-Holland, Amsterdam, Holland, 1970.

[93] R. Fletcher, Practical methods of optimization, 2000.

[94] R. Fletcher, N. I. M. Gould, S. Leyffer, P. L. Toint, and A. Wächter, Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming. *SIAM Journal on Optimization*, **13**, 635–659, 2002.

[95] R. Fletcher, and S. Leyffer, *A bundle filter method for nonsmooth nonlinear*. Technical Report Department of Mathematics, University of Dundee Nethergate, Dundee, Scotland. NA/195, 1999.

[96] R. Fletcher, and S. Leyffer, Nonlinear programming without a penalty function. *Mathematical programming*, **91**, 239–269, 2002.

[97] R. Fletcher, and S. Leyffer, Filter-type Algorithms for Solving Systems of Algebraic Equations and Inequalities. In G. Di Pillo, and A. Murli (Eds.), *High Performance Algorithms and Software for Nonlinear Optimization* (pp. 265–284). Boston, MA: Springer US, 2003.

[98] R. Fletcher, S. Leyffer, and P. L. Toint, On the global convergence of a filter–SQP algorithm. *SIAM Journal on Optimization*, **13**, 44–59, 2002.

[99] K. R. Fowler, and C. T. Kelley, Pseudo-Transient Continuation for Nonsmooth Nonlinear Equations. *SIAM Journal on Numerical Analysis*, **43**, 1385–1406, 2005.

[100] A. Francavilla, and O. C. Zienkiewicz, A note on numerical computation of elastic contact problems. *International Journal for Numerical Methods in Engineering*, **9**, 913–924, 1975.

[101] D. Franke, A. Düster, V. Nübel, and E. Rank, A comparison of the h-, p-, hp-, and rp-version of the FEM for the solution of the 2D Hertzian contact problem. *Computational Mechanics*, **45**, 513–522, 2010.

[102] T.-P. Fries, and H. G. Matthies, *Classification and Overview of Meshfree Methods*. Technical Report Institut für Wissenschaftliches Rechnen. Braunschweig, 2004.

[103] C. W. Gear, *Numerical initial value problems in ordinary differential equations*. Englewood Cliffs, New Jersey, USA: Prentice-Hall, Inc., 1971.

[104] M. Gee, *Effiziente Lösungsstrategien in der nichtlinearen Schalenmechanik*. PhD–thesis Universität Stuttgart, 2004.

[105] M. W. Gee, C. T. Kelley, and R. B. Lehoucq, Pseudo-transient continuation for nonlinear transient elasticity. *International Journal for Numerical Methods in Engineering*, **78**, 1209–1219, 2009.

[106] M. W. Gee, C. M. Siefert, J. J. Hu, R. S. Tuminaro, and M. G. Sala, *ML 5.0 smoothed aggregation user's guide*. Technical Report Sandia National Laboratories. SAND2006-2649, 2006.

[107] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *Some Theoretical Properties of an Augmented Lagrangian Merit Function*. Technical Report DTIC Document, 1986.

[108] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization.*. Academic press, 1981.

[109] M. Gitterle, *A dual mortar formulation for finite deformation frictional contact problems including wear and thermal coupling*. PhD–thesis Technical University of Munich, 2012.

[110] M. Gitterle, A. Popp, M. W. Gee, and W. A. Wall, Finite deformation frictional mortar contact using a semi-smooth Newton method with consistent linearization. *International Journal for Numerical Methods in Engineering*, **84**, 543–571, 2010.

[111] T. Glad, and E. Polak, A multiplier method with automatic limitation of penalty growth. *Mathematical Programming*, **17**, 140–155, 1979.

[112] H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics*. (3rd ed.). Addison Wesley, 2001.

[113] N. I. M. Gould, On the Accurate Determination of Search Directions for Simple Differentiate Penalty Functions. *IMA Journal of Numerical Analysis*, **6**, 357–372, 1986.

[114] N. I. M. Gould, S. Leyffer, and P. L. Toint, A multidimensional filter algorithm for nonlinear equations and nonlinear least-squares. *SIAM Journal on Optimization*, **15**, 17–38, 2004.

[115] N. I. M. Gould, D. Orban, A. Sartenaer, and P. L. Toint, Superlinear convergence of primal-dual interior point algorithms for nonlinear programming. *SIAM Journal on Optimization*, **11**, 974–1002, 2001.

[116] N. I. M. Gould, and P. L. Toint, FILTRANE, a Fortran 95 filter-trust-region package for solving nonlinear least-squares and nonlinear feasibility problems. *ACM Transactions on Mathematical Software (TOMS)*, **33**, 3, 2007.

[117] H. Gouraud, Continuous shading of curved surfaces. *IEEE Transactions on Computers*, **100**, 623–629, 1971.

[118] J. Grandy, Conservative Remapping and Region Overlays by Intersecting Arbitrary Polyhedra. *Journal of Computational Physics*, **148**, 433–466, 1999.

[119] M. Graveleau, N. Chevaugeon, and N. Moës, The inequality level-set approach to handle contact: membrane case. *Advanced Modeling and Simulation in Engineering Sciences*, **2**, 2015.

[120] L. Grippo, F. Lampariello, and S. Lucidi, A nonmonotone line search technique for Newton's method. *SIAM journal on numerical analysis*, **23**, 707–716, 1986.

[121] Y. Guo, and J. S. Curtis, Discrete element method simulations for complex granular flows. *Annual Review of Fluid Mechanics*, **47**, 21–46, 2015.

[122] C. Hager, S. Hüeber, and B. I. Wohlmuth, A stable energy-conserving approach for frictional contact problems based on quadrature formulas. *International Journal for Numerical Methods in Engineering*, **73**, 205–225, 2008.

[123] C. Hager, and B. I. Wohlmuth, Analysis of a space-time discretization for dynamic elasticity problems based on mass-free surface elements. *SIAM J. Numer. Anal.*, **47**, 1863–1885, 2009.

[124] C. Hager, and B. I. Wohlmuth, Nonlinear complementarity functions for plasticity problems with frictional contact. *Computer Methods in Applied Mechanics and Engineering*, **198**, 3411–3427, 2009.

[125] S. P. Han, A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications*, **22**, 297–309, 1977.

[126] H.-B. Hellweg, and M. Crisfield, A new arc-length method for handling sharp snap-backs. *Computers and Structures*, **66**, 705–709, 1998.

[127] M. A. Heroux, R. A. Bartlett, V. E. Howle, R. J. Hoekstra, J. J. Hu, T. G. Kolda, R. B. Lehoucq, K. R. Long, R. P. Pawlowski, E. T. Phipps, A. G. Salinger, H. K. Thornquist, R. S. Tuminaro, J. M. Willenbring, A. Williams, and K. S. Stanley, An overview of the Trilinos project. *ACM Trans. Math. Softw.*, **31**, 397–423, 2005.

[128] J. Herskovits, and S. R. Mazorche, A feasible directions algorithm for nonlinear complementarity problems and applications in mechanics. *Structural and Multidisciplinary Optimization*, **37**, 435–446, 2009.

[129] C. Hesch, and P. Betsch, A mortar method for energy-momentum conserving schemes in frictionless dynamic contact problems. *International Journal for Numerical Methods in Engineering*, **77**, 1468–1500, 2009.

[130] C. Hesch, and P. Betsch, Transient three-dimensional contact problems: mortar method. Mixed methods and conserving integration. *Computational Mechanics*, **48**, 461–475, 2011.

[131] M. Hiermeier, W. A. Wall, and A. Popp, A truly variationally consistent and symmetric mortar-based contact formulation for finite deformation solid mechanics. *Computer Methods in Applied Mechanics and Engineering*, **342**, 532–560, 2018.

[132] N. J. Higham, *Accuracy and stability of Numerical Algorithms* volume 80. Manchester: SIAM, 2002.

[133] P. Hild, Numerical implementation of two nonconforming finite element methods for unilateral contact. *Computer Methods in Applied Mechanics and Engineering*, **184**, 99–123, 2000.

[134] M. Hintermüller, K. Ito, and K. Kunisch, The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, **13**, 865–888, 2002.

[135] M. Hofer, *Ein Level-Set Ansatz unter Nebenbedingungen für Kontaktprobleme mit großen Deformationen*. Master's thesis Technical University of Munich Institute for Computational Mechanics. Supervised by Michael Hiermeier, 2016.

[136] G. A. Holzapfel, *Nonlinear solid mechanics: a continuum approach for engineering*. Wiley, 2000.

[137] B. Hübner, E. Walhorn, and D. Dinkler, A monolithic approach to fluid-structure interaction using space-time finite elements. *Computer Methods in Applied Mechanics and Engineering*, **193**, 2087–2104, 2004.

[138] S. Hüeber, *Discretization techniques and efficient algorithms for contact problems*. PhD–thesis University of Stuttgart, 2008.

[139] S. Hüeber, and B. I. Wohlmuth, A primal-dual active set strategy for non-linear multibody contact problems. *Computer Methods in Applied Mechanics and Engineering*, **194**, 3147–3166, 2005.

[140] T. J. Hughes, *The finite element method: linear static and dynamic finite element analysis*. Courier Corporation, 2012.

[141] T. J. Hughes, J. A. Cottrell, and Y. Bazilevs, Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Computer Methods in Applied Mechanics and Engineering*, **194**, 4135–4195, 2005.

[142] T. J. Hughes, and G. M. Hulbert, Space-time finite element methods for elastodynamics: Formulations and error estimates. *Computer Methods in Applied Mechanics and Engineering*, **66**, 339–363, 1988.

[143] T. J. R. Hughes, R. L. Taylor, J. L. Sackman, A. Curnier, and W. Kanoknukulchai, A finite element method for a class of contact-impact problems. *Computer Methods in Applied Mechanics and Engineering*, **8**, 249–276, 1976.

[144] G. M. Hulbert, Time finite element methods for structural dynamics. *International Journal for Numerical Methods in Engineering*, **33**, 307–331, 1992.

[145] S. A. Hutchinson, J. N. Shadid, and R. S. Tuminaro, *Aztec user's guide: Version 1.1*, 1995.

[146] B. M. Irons, and R. C. Tuck, A version of the Aitken accelerator for computer iteration. *International Journal for Numerical Methods in Engineering*, **1**, 275–277, 1969.

[147] S. A. Ivanenko, Harmonic Mappings. In *Handbook of grid generation* chapter 8. (p. 43). CRC Press Boca Raton, 1999.

[148] P. S. Jensen, Finite difference techniques for variable grids. *Computer & Structures*, **2**, 17–29, 1972.

[149] S. Jin, R. R. Lewis, and D. West, A comparison of algorithms for vertex normal computation. *The Visual Computer*, **21**, 71–82, 2005.

[150] A. Johnen, J. F. Remacle, and C. Geuzaine, Geometrical validity of curvilinear finite elements. *Journal of Computational Physics*, **233**, 359–372, 2013.

[151] A. Johnen, J. C. Weill, and J. F. Remacle, Robust and efficient validation of the linear hexahedral element. *Procedia Engineering*, **203**, 271–283, 2017.

[152] C. Kane, E. A. Repetto, M. Ortiz, and J. E. Marsden, Finite element analysis of nonsmooth contact. *Computer Methods in Applied Mechanics and Engineering*, **180**, 1–26, 1999.

[153] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*. SIAM, 1995.

[154] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*. Philadelphia, PA: SIAM, 2003.

[155] C. T. Kelley, and D. E. Keyes, Convergence Analysis of Pseudo-Transient Continuation. *SIAM journal on numerical analysis*, **35**, 508–523, 1998.

[156] C. T. Kelley, L.-Z. Liao, L. Qi, M. T. Chu, J. P. Reese, and C. Winton, Projected Pseudo-transient Continuation. *SIAM journal on numerical analysis*, **46**, 3071–3083, 2008.

[157] N. Kikuchi, and J. T. Oden, *Contact problems in elasticity: a study of variational inequalities and finite element methods*. SIAM, 1988.

[158] S. Klinkel, and W. Wagner, A Geometrical Non-Linear Brick Element Based on the EAS-Method. *International Journal for Numerical Methods in Engineering*, **40**, 4529–4545, 1997.

[159] G. Kloosterman, R. M. Van Damme, A. H. Van Den Boogaard, and J. Huetink, A geometrical-based contact algorithm using a barrier method. *International Journal for Numerical Methods in Engineering*, **51**, 865–882, 2001.

[160] P. M. Knupp, On the invertibility of the isoparametric map. *Computer Methods in Applied Mechanics and Engineering*, **78**, 313–329, 1990.

[161] P. M. Knupp, Achieving finite element mesh quality via optimization of the Jacobian matrix norm and associated quantities. Part II - A framework for volume mesh optimization and the condition number of the Jacobian matrix. *International Journal for Numerical Methods in Engineering*, **48**, 1165–1185, 2000.

[162] T. G. Kolda, R. M. Lewis, and V. Torczon, Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods. *SIAM Review*, **45**, 385–482, 2003.

[163] J. S. Koo, and B. M. Kwak, Post-buckling analysis with frictional contacts combining complementarity relations and an arc-length method. *International Journal for Numerical Methods in Engineering*, **39**, 1161–1180, 1996.

[164] F. Koschnick, *Geometrische Locking-Effekte bei Finiten Elementen und ein allgemeines Konzept zu ihrer Vermeidung*. PhD–thesis Technical University of Munich, 2004.

[165] M. Křížek, L. Liu, and P. Neittaanmäki, Post-processing of Gauss–Seidel iterations. *Numerical linear algebra with applications*, **6**, 147–156, 1999.

[166] R. Kučera, J. MacHalová, H. Netuka, and P. Ženčák, An interior-point algorithm for the minimization arising from 3D contact problems with friction. *Optimization Methods and Software*, **28**, 1195–1217, 2013.

[167] D. Kuhl, and E. Ramm, Generalized Energy-Momentum Method for non-linear adaptive shell dynamics. *Computer Methods in Applied Mechanics and Engineering*, **178**, 343–366, 1999.

[168] U. Küttler, *Effiziente Lösungsverfahren für Fluid-Struktur-Interaktions-Probleme*. PhD–thesis Technical University Munich, 2009.

[169] U. Küttler, and W. A. Wall, Fixed-point fluid-structure interaction solvers with dynamic relaxation. *Computational Mechanics*, **43**, 61–72, 2008.

[170] T. A. Laursen, *Computational contact and impact mechanics*. Berlin-Heidelberg: Springer-Verlag, 2002.

[171] T. A. Laursen, and V. Chawla, Design of energy conserving algorithms for frictionless dynamic contact problems. *International Journal for Numerical Methods in Engineering*, **40**, 863–886, 1997.

[172] T. A. Laursen, and G. R. Love, Improved implicit integrators for transient impact problems - geometric admissibility within the conserving framework. *International Journal for Numerical Methods in Engineering*, **53**, 245–274, 2002.

[173] R. Leroy, *Certificates of positivity in the simplicial Bernstein basis*. Technical Report Institut de Recherche Mathématique de Rennes. URL: `https://hal.archives-ouvertes.fr/hal-00589945`, 2011.

[174] K. Levenberg, A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics*, **2**, 164–168, 1944.

[175] A. S. Lewis, and M. L. Overton, Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, **141**, 135–163, 2013.

[176] D. Li, and M. Fukushima, A globally and superlinearly convergent Gauss-Newton-Based BFGS method for symmetric nonlinear equations. *SIAM journal on numerical analysis*, **37**, 152–172, 1999.

[177] T. Liszka, and J. Orkisz, The finite difference method at arbitrary irregular grids and its application in applied mechanics. *Computers and Structures*, **11**, 83–95, 1980.

[178] D. C. Liu, and J. Nocedal, On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, **45**, 503–528, 1989.

[179] M. Liu, and G. Liu, Smoothed Particle Hydrodynamics (SPH): an overview and recent developments. *Archives of computational methods in engineering*, **17**, 25–76, 2010.

[180] V. Lubarda, *Elastoplasticity Theory*. (1st ed.). Boca Raton: CRC Press, 2001.

[181] J. Lubliner, *Plasticity theory*. Mineola, New York: Dover Publications, 2006.

[182] S. Lucidi, New Results on a Continuously Differentiable Exact Penalty Function. *SIAM Journal on Optimization*, **2**, 558–574, 1992.

[183] A. J. Macleod, Acceleration of vector sequences by multi-dimensional $\Delta 2$ methods. *Communications in Applied Numerical Methods*, **2**, 385–392, 1986.

[184] Y. Maday, C. Mavriplis, and A. T. Patera, Nonconforming Mortar Element Methods: Application to Spectral Discretizations. *Domain Decomposition Methods*, (pp. 392–418), 1989.

[185] D. S. Malkus, and T. J. Hughes, Mixed finite element methods - Reduced and selective integration techniques: A unification of concepts. *Computer Methods in Applied Mechanics and Engineering*, **15**, 63–81, 1978.

[186] N. Maratos, *Exact penalty function algorithms for finite dimensional and control optimization problems*. PhD–thesis Imerpial College of Science and Technology, University of London, 1978.

[187] D. W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.*, **11**, 431–441, 1963.

[188] J. E. Marsden, T. J. R. Hughes, and D. E. Carlson, *Mathematical Foundations of Elasticity*. New York: Dover Publications, 1984.

[189] J. M. Martinez, Practical quasi-Newton methods for solving nonlinear systems. *Journal of Computational and Applied Mathematics*, **124**, 97–121, 2000.

[190] J. M. Martinez, and M. C. Zambaldi, An inverse column-updating method for solving large-scale nonlinear systems of equations. *Optimization Methods & Software*, **1**, 129–140, 1992.

[191] E. Marwil, Convergence results for Schubert's method for solving sparse nonlinear equations. *SIAM journal on numerical analysis*, **16**, 588–605, 1979.

[192] T. W. McDevitt, and T. A. Laursen, A mortar-finite element formulation for frictional contact problems. *International Journal for Numerical Methods in Engineering*, **48**, 1525–1547, 2000.

[193] C. Meier, R. Weissbach, J. Weinberg, W. A. Wall, and A. J. Hart, Modeling and characterization of cohesion in fine metal powders with a focus on additive manufacturing process simulations. *Powder technology*, **343**, 855–866, 2019.

[194] D. Millán, A. Rosolen, and M. Arroyo, Thin shell analysis from scattered points with maximum-entropy approximants. *International Journal for Numerical Methods in Engineering*, **85**, 723–751, 2011.

[195] D. Millán, A. Rosolen, and M. Arroyo, Nonlinear manifold learning for meshfree finite deformation thin-shell analysis. *International Journal for Numerical Methods in Engineering*, **93**, 685–713, 2013.

[196] A. Milzarek, and M. Ulbrich, A Semismooth Newton Method with Multidimensional Filter Globalization for l_1-Optimization. *SIAM Journal on Optimization*, **24**, 298–333, 2014.

[197] T. Miyamura, Y. Kanno, and M. Ohsaki, Combined interior-point method and semismooth Newton method for frictionless contact problems. *International Journal for Numerical Methods in Engineering*, **81**, 701–727, 2010.

[198] J. M. Modisette, *An Automated Reliable Method for Two-Dimensional Reynolds-Averaged Navier-Stokes Simulations*. PhD–thesis Massachusetts Institute of Technology, 2011.

[199] N. Moës, E. Béchet, and M. Tourbier, Imposing Dirichlet boundary conditions in the extended finite element method. *International Journal for Numerical Methods in Engineering*, **67**, 1641–1669, 2006.

[200] A. Munjiza, D. Owen, and N. Bicanic, A combined finite-discrete element method in transient dynamics of fracturing solids. *Engineering computations*, **12**, 145–174, 1995.

[201] J. A. Nelder, and R. Mead, A Simplex Method for Function Minimization. *The Computer Journal*, **7**, 308–313, 1965.

[202] D. Neto, M. Oliveira, L. Menezes, and J. Alves, A contact smoothing method for arbitrary surface meshes using Nagata patches. *Computer Methods in Applied Mechanics and Engineering*, **299**, 283–315, 2016.

[203] D. Neto, M. C. Oliveira, L. F. Menezes, and J. L. Alves, Improving Nagata patch interpolation applied for tool surface description in sheet metal forming simulation. *Computer-Aided Design*, **45**, 639–656, 2013.

[204] J. Nocedal, and S. Wright, *Numerical Optimization*. (2nd ed.). New York, NY: Springer-Verlag, 2006.

[205] J. Nocedal, and Y. Yuan, Combining trust region and line search techniques. *Advances in Nonlinear Programming*, (pp. 153–175), 1998.

[206] J. T. Oden, *Finite elements of nonlinear continua*. Courier Corporation, 2006.

[207] J. T. Oden, and S. J. Kim, Interior penalty methods for finite element approximations of the Signorini problem in elastostatics. *Computers and Mathematics with Applications*, **8**, 35–56, 1982.

[208] R. W. Ogden, *Non-Linear Elastic Deformations*. Mineola, New York: Dover Publications, 1983.

[209] S. Osher, and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces* volume 153, 2003.

[210] Y. Otoguro, K. Takizawa, and T. E. Tezduyar, A General-Purpose NURBS Mesh Generation Method for Complex Geometries. In *Frontiers in Computational Fluid-Structure Interaction and Flow Simulation. Modeling and Simulation in Science, Engineering and Technology.* (pp. 399–434). Birkhäuser, Cham, 2018.

[211] S. V. Patankar, and D. B. Spalding, A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows. In *Numerical Prediction of Flow, Heat Transfer, Turbulence and Combustion* (pp. 54–73). Elsevier, 1983.

[212] G. Pietrzak, and A. Curnier, Large deformation frictional contact mechanics: continuum formulation and augmented Lagrangian treatment. *Computer Methods in Applied Mechanics and Engineering*, **177**, 351–381, 1999.

[213] A. Popp, *Mortar Methods for Computational Contact Mechanics and General Interface Problems*. PhD–thesis Technical University of Munich, 2012.

[214] A. Popp, State-of-the-Art Computational Methods for Finite Deformation Contact Modeling of Solids and Structures. In *Contact Modeling for Solids and Particles* (pp. 1–86). Springer, 2018.

[215] A. Popp, M. W. Gee, and W. A. Wall, A finite deformation mortar contact formulation using a primal-dual active set strategy. *International Journal for Numerical Methods in Engineering*, **79**, 1354–1391, 2009.

[216] A. Popp, M. Gitterle, M. W. Gee, and W. A. Wall, A dual mortar approach for 3D finite deformation contact with consistent linearization. *International Journal for Numerical Methods in Engineering*, **83**, 1428–1465, 2010.

[217] A. Popp, A. Seitz, M. W. Gee, and W. A. Wall, Improved robustness and consistency of 3D contact algorithms based on a dual mortar approach. *Computer Methods in Applied Mechanics and Engineering*, **264**, 67–80, 2013.

[218] A. Popp, B. I. Wohlmuth, M. W. Gee, and W. A. Wall, Dual quadratic mortar finite element methods for 3D finite deformation contact. *SIAM Journal on Scientific Computing*, **34**, B421–B446, 2012.

[219] M. J. D. Powell, Convergence properties of algorithms for nonlinear optimization. *SIAM Review*, **28**, 487–500, 1984.

[220] A. Prokopenko, J. J. Hu, T. A. Wiesner, C. M. Siefert, and R. S. Tuminaro, *MueLu user's guide 1.0*. Technical Report Sandia National Laboratories. SAND2014-18874, 2014.

[221] M. A. Puso, and T. A. Laursen, A 3D contact smoothing method using Gregory patches. *International Journal for Numerical Methods in Engineering*, **54**, 1161–1194, 2002.

[222] M. A. Puso, and T. A. Laursen, A mortar segment-to-segment contact method for large deformation solid mechanics. *Computer Methods in Applied Mechanics and Engineering*, **193**, 601–629, 2004.

[223] M. A. Puso, T. A. Laursen, and J. Solberg, A segment-to-segment mortar contact method for quadratic elements and large deformations. *Computer Methods in Applied Mechanics and Engineering*, **197**, 555–566, 2008.

[224] E. Ramm, Strategies for tracing the nonlinear response near limit points. In W. Wunderlich, E. Stein, and K.-J. Bathe (Eds.), *Nonlinear finite element analysis in structural mechanics* (pp. 63–89). Berlin, Heidelberg: Springer, 1981.

[225] A. D. Rauch, A.-T. Vuong, L. Yoshihara, and W. A. Wall, A coupled approach for fluid saturated poroelastic media and immersed solids for modeling cell-tissue interactions. *International Journal for Numerical Methods in Biomedical Engineering*, (p. e3139), 2018.

[226] A. Rieger, O. Scherf, and P. Wriggers, Adaptive Methods for Contact Problems. In E. Stein (Ed.), *Error-controlled adaptive finite elements in solid mechanics* (pp. 147–179). John Wiley & Sons, 2003.

[227] L. M. Rios, and N. V. Sahinidis, Derivative-free optimization: A review of algorithms and comparison of software implementations. *Journal of Global Optimization*, **56**, 1247–1293, 2013.

[228] S. M. Robinson, Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear-programming algorithms. *Mathematical Programming*, **7**, 1–16, 1974.

[229] R. T. Rockafellar, The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and applications*, **12**, 555–562, 1973.

[230] R. T. Rockafellar, Augmented Lagrange Multiplier Functions and Duality in Nonconvex Programming. *SIAM Journal on Control*, **12**, 268–285, 1974.

[231] H. H. Rosenbrock, Some general implicit processes for the numerical solution of differential equations. *The Computer Journal*, **5**, 329–330, 963.

[232] Y. Saad, and M. H. Schultz, GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM journal on scientific and statistical computing*, **7**, 856–869, 1986.

[233] A. A. Samarskii, *The theory of difference schemes*. CRC Press, 2001.

[234] R. A. Sauer, and L. De Lorenzis, A computational contact formulation based on surface potentials. *Computer Methods in Applied Mechanics and Engineering*, **253**, 369–395, 2013.

[235] R. A. Sauer, and P. Wriggers, Formulation and analysis of a three-dimensional finite element implementation for adhesive contact at the nanoscale. *Computer Methods in Applied Mechanics and Engineering*, **198**, 3871–3883, 2009.

[236] R. B. Schnabel, and P. D. Frank, Tensor methods for nonlinear equations. *SIAM Journal on Numerical Analysis*, **21**, 815–843, 1984.

[237] B. Schott, *Stabilized Cut Finite Element Methods for Complex Interface Coupled Flow Problems*. PhD–thesis Technical University Munich, 2016.

[238] L. K. Schubert, Modification of a quasi-Newton method for nonlinear equations with sparse Jacobian. *Mathematics of Computation*, **24**, 27–30, 1970.

[239] L. Schulze, *Lagrange-Multiplikator-Funktionen für Mortar-Kontaktformulierungen unter Berücksichtigung großer Deformationen*. Bachelor-thesis Technical University of Munich. Supervised by Michael Hiermeier, 2017.

[240] A. Seitz, *Computational Methods for Thermo-Elasto-Plastic Contact*. PhD–thesis Technical University Munich, 2019.

[241] A. Seitz, P. Farah, J. Kremheller, B. I. Wohlmuth, W. A. Wall, and A. Popp, Isogeometric dual mortar methods for computational contact mechanics. *Computer Methods in Applied Mechanics and Engineering*, **301**, 259–280, 2016.

[242] A. Seitz, A. Popp, and W. A. Wall, A semi-smooth Newton method for orthotropic plasticity and frictional contact at finite strains. *Computer Methods in Applied Mechanics and Engineering*, **285**, 228–254, 2015.

[243] A. Seitz, W. A. Wall, and A. Popp, A computational approach for thermo-elasto-plastic frictional contact based on a monolithic formulation using non-smooth nonlinear complementarity functions. *Advanced Modeling and Simulation in Engineering Sciences*, **5**, 5, 2018.

[244] A. Seitz, W. A. Wall, and A. Popp, Nitsche's method for finite deformation thermomechanical contact problems. *Computational Mechanics*, **63**, 1091–1110, 2019.

[245] P. Seshaiyer, and M. Suri, hp submeshing via non-conforming finite element methods. *Computer Methods in Applied Mechanics and Engineering*, **189**, 1011–1030, 2000.

[246] R. Silva, M. Ulbrich, S. Ulbrich, and L. N. Vicente, *A globally convergent primal-dual interior-point filter method for nonlinear programming: new filter optimality measures and computational results*. Technical Report Centro de Matemática da Universidade de Coimbra, 2008.

[247] J. C. Simo, and F. Armero, Geometrically non-linear enhanced strain mixed methods and the method of incompatible modes. *International Journal for Numerical Methods in Engineering*, **33**, 1413–1449, 1992.

[248] J. C. Simo, and T. J. R. Hughes, On the Variational Foundations of Assumed Strain Methods. *Journal of Applied Mechanics*, **53**, 51, 1986.

[249] J. C. Simo, and M. S. Rifai, A class of mixed assumed strain methods and the method of incompatible modes. *International Journal for Numerical Methods in Engineering*, **29**, 1595–1638, 1990.

[250] J. C. Simo, and N. Tarnow, The discrete energy-momentum method. Conserving algorithms for nonlinear elastodynamics. *ZAMP Zeitschrift für angewandte Mathematik und Physik*, **43**, 757–792, 1992.

[251] J. C. Simo, P. Wriggers, K. H. Schweizerhof, and R. L. Taylor, Finite deformation post-buckling analysis involving inelasticity and contact constraints. *International Journal for Numerical Methods in Engineering*, **23**, 779–800, 1986.

[252] S. Sitzmann, *Robust Algorithms for Contact Problems with Constitutive Contact Laws*. PhD–thesis Friedrich-Alexander-University, 2016.

[253] W. Spendley, G. Hext, F. H. Technometrics, and U. 1962, Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics*, **4**, 441–461, 1962.

[254] G. Stadler, Path-following and augmented Lagrangian methods for contact problems in linear elasticity. *Journal of Computational and Applied Mathematics*, **203**, 533–547, 2007.

[255] S. Stupkiewicz, Finite Wear and Soft Elasto-Hydrodynamic Lubrication: Beyond the Classical Frictional Contact of Soft Solids. In *Contact Modeling for Solids and Particles* (pp. 125–176). Springer, 2018.

[256] N. Sukumar, and R. Wright, Overview and construction of meshfree basis functions: from moving least squares to entropy approximants. *International Journal for Numerical Methods in Engineering*, **70**, 181–205, 2007.

[257] D. Sun, A regularization Newton method for solving nonlinear complementarity problems. *Applied Mathematics and Optimization*, **40**, 315–339, 1999.

[258] G. Tanoh, Y. Renard, and D. Noll, Computational experience with an interior point algorithm for large scale contact problems. *Optimization Online*, **10**, 2004.

[259] R. A. Tapia, Diagonalized multiplier methods and quasi-Newton methods for constrained optimization. *Journal of Optimization Theory and Applications*, **22**, 135–194, 1977.

[260] R. L. Taylor, and P. Papodopoulos, On a patch test for contact problems in two dimensions. In P. Wriggers, and W. Wagner (Eds.), *Nonlinear Computational Mechanics* (pp. 690–702). Berlin: Springer, 1991.

[261] I. Temizer, M. Abdalla, and Z. Gürdal, An interior point method for isogeometric contact. *Computer Methods in Applied Mechanics and Engineering*, **276**, 589–611, 2014.

[262] M. Tur, E. Giner, F. Fuenmayor, and P. Wriggers, 2D contact smooth formulation based on the mortar method. *Computer Methods in Applied Mechanics and Engineering*, **247**, 1–14, 2012.

[263] M. Ulbrich, S. Ulbrich, and L. N. Vicente, A globally convergent primal-dual interior-point filter method for nonlinear programming. *Mathematical Programming*, **100**, 379–410, 2004.

[264] S. Ulbrich, On the superlinear local convergence of a filter-SQP method. *Mathematical Programming*, **100**, 217–245, 2004.

[265] O. V. Ushakova, Conditions of nondegeneracy of three-dimensional cells. A formula of a volume of cells. *SIAM Journal on Scientific Computing*, **23**, 1274–1290, 2001.

[266] O. V. Ushakova, Nondegeneracy tests for hexahedral cells. *Computer Methods in Applied Mechanics and Engineering*, **200**, 1649–1658, 2011.

[267] H. Uzawa, 10 iterative methods for concave programming. *Studies in Linear and Non-Linear Programming*, **2**, 154, 1968.

[268] S. Vavasis, A Bernstein-Bezier Sufficient Condition for Invertibility of Polynomial Mapping Functions (Draft). *arXiv preprint cs/0308021*, (pp. 1–12), 2003.

[269] L. Vu-Quoc, and X. G. Tan, Optimal solid shells for non-linear analyses of multilayer composites. I. Statics. *Computer Methods in Applied Mechanics and Engineering*, **192**, 975–1016, 2003.

[270] A. Wächter, and L. T. Biegler, Line search filter methods for nonlinear programming: Local convergence. *SIAM Journal on Optimization*, **16**, 32–48, 2005.

[271] A. Wächter, and L. T. Biegler, Line search filter methods for nonlinear programming: Motivation and global convergence. *SIAM Journal on Optimization*, **16**, 1–31, 2005.

[272] A. Wächter, and L. T. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, **106**, 25–57, 2006.

[273] W. A. Wall, M. Bischoff, and E. Ramm, A deformation dependent stabilization technique, exemplified by EAS elements at large strains. *Computer Methods in Applied Mechanics and Engineering*, **188**, 859–871, 2000.

[274] W. A. Wall, and M. Kronbichler, *BACI: A multiphysics simulation environment*. Technical Report Technical University of Munich, Institute for Computational Mechanics, 2018.

[275] T. Wiesner, *Fexible Aggregation-based Algebraic Multigrid Methods for Contact and Flow Problems*. PhD–thesis Technical University of Munich, 2015.

[276] T. A. Wiesner, A. Popp, M. W. Gee, and W. A. Wall, Algebraic multigrid methods for dual mortar finite element formulations in contact mechanics. *International Journal for Numerical Methods in Engineering*, (pp. 399–430), 2017.

[277] T. A. Wiesner, R. S. Tuminaro, W. A. Wall, and M. W. Gee, Multigrid transfers for non-symmetric systems based on Schur complements and Galerkin projections. *Numerical Linear Algebra with Applications*, **21**, 415–438, 2014.

[278] C. Wilking, and M. Bischoff, Alternative integration algorithms for three-dimensional mortar contact. *Computational Mechanics*, **59**, 203–218, 2017.

[279] B. Wohlmuth, Variationally consistent discretization schemes and numerical algorithms for contact problems. *Acta Numerica*, **20**, 569–734, 2011.

[280] B. I. Wohlmuth, A Mortar Finite Element Method Using Dual Spaces for the Lagrange Multiplier. *SIAM Journal on Numerical Analysis*, **38**, 989–1012, 2000.

[281] B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*. Berlin-Heidelberg: Springer, 2001.

[282] B. I. Wohlmuth, and R. H. Krause, A multigrid method based on the unconstrained product space for mortar finite element discretizations. *SIAM journal on numerical analysis*, **39**, 192–213, 2001.

[283] P. Wriggers, *Computational contact mechanics*. Berlin-Heidelberg: Springer-Verlag, 2006.

[284] P. Wriggers, Advanced discretization methods for contact mechanics. In *Contact Modeling for Solids and Particles* (pp. 87–123). Springer, 2018.

[285] P. Wriggers, and S. Reese, A note on enhanced strain methods for large deformations. *Computer Methods in Applied Mechanics and Engineering*, **135**, 201–209, 1996.

[286] S. Yamakawa, and K. Shimada, Fully-automated hex-dominant mesh generation with directionality control via packing rectangular solid cells. *International Journal for Numerical Methods in Engineering*, **57**, 2099–2129, 2003.

[287] N. Yamashita, and M. Fukushima, On the Rate of Convergence of the Levenberg-Marquardt Method. In *Topics in numerical analysis* (pp. 239–249). Vienna: Springer, 2001.

[288] B. Yang, and T. A. Laursen, A contact searching algorithm including bounding volume trees applied to finite sliding mortar formulations. *Computational Mechanics*, **41**, 189–205, 2008.

[289] B. Yang, and T. A. Laursen, A large deformation mortar formulation of self contact with finite sliding. *Computer Methods in Applied Mechanics and Engineering*, **197**, 756–772, 2008.

[290] B. Yang, T. A. Laursen, and X. Meng, Two dimensional contact methods for large deformation frictional sliding. *International Journal for Numerical Methods in Engineering*, **62**, 1183–1225, 2005.

[291] J. Youett, O. Sander, and R. Kornhuber, A globally convergent filter-trust-region method for large deformation contact problems. *SIAM Journal on Scientific Computing*, **41**, B114–B138, 2019.

[292] J. W. Youett, *Dynamic large deformation contact problems and applications in virtual medicine*. PhD–thesis Freie Universität Berlin, 2015.

[293] G. Zavarise, and L. De Lorenzis, The node-to-segment algorithm for 2D frictionless contact: Classical formulation and special cases. *Computer Methods in Applied Mechanics and Engineering*, **198**, 3428–3451, 2009.

[294] G. Zavarise, L. De Lorenzis, and R. L. Taylor, A non-consistent start-up procedure for contact problems with large load-steps. *Computer Methods in Applied Mechanics and Engineering*, **205**, 91–109, 2012.

[295] H. Zhang, and W. W. Hager, A Nonmonotone Line Search Technique and Its Application to Unconstrained Optimization. *SIAM Journal on Optimization*, **14**, 1043–1056, 2004.

[296] S. Zhang, Subtetrahedral test for the positive Jacobian of hexahedral elements. *Preprint available at http://www.math.udel.edu/˜szhang/research/p/subtettest.pdf*, , 2005.

[297] O. C. Zienkiewicz, and R. L. Taylor, *The finite element method for solid and structural mechanics*. (6th ed.). Elsevier Butterworth-Heinemann, 2005.

[298] O. C. Zienkiewicz, R. L. Taylor, and J. M. Too, Reduced integration technique in general analysis of plates and shells. *International Journal for Numerical Methods in Engineering*, **3**, 275–290, 1971.

[299] W. Zulehner, A class of smoothers for saddle point problems. *Computing*, **65**, 227–246, 2000.

Verzeichnis der betreuten Studienarbeiten

Im Rahmen dieser Dissertation entstanden am Lehrstuhl für Numerische Mechanik (LNM) in den Jahren von 2013 bis 2018 unter wesentlicher wissenschaftlicher, fachlicher und inhaltlicher Anleitung des Autors die im Folgenden aufgeführten studentischen Arbeiten. Der Autor dankt allen Studierenden für Ihr Engagement bei der Unterstützung dieser wissenschaftlichen Arbeit.

Studierende(r)	**Studienarbeit**
Michael Kölbl	*An Arc-Length Method for Finite Deformation Problems: Static Analysis*, Bachelorarbeit, 2015.
Waschriporn Ampunant	*Analysis of a modified Pseudo Transient Continuation Method*, Semesterarbeit, 2016.
Michael Hofer	*The Inequality Level-Set Approach for Finite Deformation Contact Problems*, Masterarbeit, 2016.
Lennart Schulze	*Lagrange Multiplier Functions for Mortar-like Contact Formulations under Consideration of Large Deformations*, Bachelorarbeit, 2017.