Part B

Boundary-Coupled Problems

Preface to Part B

This group of projects deals with processes, in which different domains and their associated fields interact on their boundaries. Problems, in which different media such as flexible bodies and fluids interact, are considered. Moreover, domains with the same constitutive behavior are considered, in which fields of the same type interact on their surface by, e.g., contact involving friction.

The *Project Group B* is composed of the following projects:

- B1: Contact Problems in Machine Dynamics (1995–2003)
- B2: Transonic Flow on Flexible Airfoils (1995–1997)
- B4: Fluid-Structure-Interaction of Shells (1995–2006)
- B5: Acoustic and Hydroacoustic Radiation From Vibrating Solids (1995–2006)
- B6: Adaptive Multigrid Methods for Contact Problems with Friction (1996–2000)
- B7: Hybrid Multibody Simulation of Granular Media (2001–2006)
- B8: Contact Dynamics Treated by Multibody Systems and Multigrid Methods (2003–2006)

All projects of this group treat dynamic problems, in which spatial discretizations as well as time discretization play an important role. In Projects B4 and B5 interactions between elastic continua and compressible as well as incompressible fluids lead to complicated vibration problems; these come along with sound radiation in the acoustic case. Projects B7 and B8 belong to a class of problems, in which modeling of the boundary surface and the contact are in the focus.

B1: Contacts Problems in Machine Dynamics

This project (W. Schiehlen and W. Wendland) deals with the interaction between macroscopic motion of impacting machine elements and microscopic

deformation in the interface. Methods of multibody dynamics are well suited to describe macroscopic motion, while boundary element methods are well suited to describe microscopic deformation. The project aim is to describe the transient contact forces by boundary elements and to adopt multibody integration procedures for the high frequency process during impact. Rigid body motions in the interface modify boundary conditions time-dependently.

The theoretical modeling is validated by generic experiments from which the model parameters are identified.

B2: Transonic Flow on Flexible Airfoils

In this project (D. Dinkler and G. Wittum) the interaction between transonic flow and the vibration of flexible airfoils is analyzed. The flow in the structural nearfield is calculated by boundary layer theory with model refinement by separation.

Flutter analysis of the nonlinear-coupled systems leads to limit cycles with finite amplitudes. The stability of limit solutions is analyzed and the degrees of freedom of the nonlinear aero-elastic system are reduced to describe dominant modes of vibration and to ease parameter variations. Robust and adaptive multigrid methods are developed for the compressible flow of flexible structures.

The solution strategy of 2D and 3D Navier-Stokes equations is developed for unstructured grids with adaptive refinement for transonic flow. Local refinement is necessary to describe shocks and boundary layers. The coupling of the farfield with the stationary Euler equations are cast into a boundary integral equation providing the farfield boundary condition for the flow data. The result is imbedded in a global solution approach.

B4: Fluid-Structure Interaction of Shells

In B4 (E. Ramm and W. Wall) a robust and efficient numerical method to simulate the dynamics of thin nonlinear structures interacting within compressible flows is developed.

It relies on an advanced seven-parameter finite element shell formulation for the structure and a stabilized finite-element flow-solver for incompressible flows on deforming domains in an ALE formulation and it applies an iteratively staggered partitioned solution approach. Relaxation by means of the Aitken method or a steepest descent formulation accelerates the convergence. Weakly coupled schemes exhibit the so-called added mass instability which depends upon the mass density ratio. An analysis of this instability reveals why the destabilizing effect increases with an increased order of temporal accuracy of the flow solver. Thus iterations over the fields are required.

For long time fluid-structure interaction simulations special emphasis is put on the accuracy of the flow solver with respect to small time steps and

distorted meshes. The flow equations on a dynamically changing domain are formulated such that geometric conservation is inherently satisfied.

A free surface approach has been incorporated into the formulation where the update of the free surface geometry is included in the iteratively staggered solution scheme. Significant improvement of the efficiency could be achieved by a vectorization concept which optimizes the code to use a large portion of the actual peak performance.

The method is applicable to a broad spectrum of incompressible flows interacting with thin, flexible structures such as flow in pipes, accelerated liquid filled tanks, the aerodynamics of thin structures or physiological flows like blood flow.

B5: Acoustic and Hydroacoustic Radiation from Vibrating Solids

The Boundary Element Method (BEM) is widely used for the simulation of acoustic fields and elastic solids involving infinite domains. In the first phase of the Project B5 (L. Gaul), a symmetric hybrid BEM was developed, while in the second phase a Fast Multipole Multilevel (FMM) BEM was derived and implemented for the 3D Helmholtz equation. The key idea of the multipole method is to approximate the fundamental solution at some distance from the source point by a multipole series expansion. The boundary elements are combined in a cluster tree, allowing the efficient evaluation of the matrix-vector product. The resulting complexity of the algorithm is quasi-linear, and this, much faster than the traditional BEM. However, the application of the fast multipole BEM for the simulation of acoustic field-structure interaction problems poses now demands on the solver for the coupled systems: direct approaches are not applicable and iterative schemes suffer from ill-conditioning.

The research adopts the fast multipole BEM for 3D acoustics and extends its application to the simulation of acoustic-structure interaction. The fast multipole BEM is coupled to a finite element formulation for Kirchhoff plates. The coupling algorithm is of mortar type in which the interface pressure is interpolated as the Lagrange multiplier and this allows the coupling of non-matching grids. For the iterative solution of the resulting saddle point problem, an approximate Uzawa-type algorithm is proposed. The Generalized Minimal Residual Method (GMRES) is used for the outer iterations on the reduced equation for the Lagrange multiplier. The inner FEM and BEM subsystems are approximated with conjugate gradient and GMRES iterations, respectively. The efficiency of the solver is increased by pre-conditioning with an approximate inverse approach.

The fundamentals of a Fuzzy BEM formulation for incorporating uncertainties of model data were derived.

B6: Adaptive Multigrid Methods for Contact Problems with Friction

In B6 (R. Kornhuber) a monotone multigrid method was developed for the Signorini problem. Convergence rates appeared as in the linear case.

Contact with friction was treated by a fix point iteration to determine the unknown normal stresses. In each fix point step a piecewise smooth variational problem has to be solved.

The extension of monotone multigrid methods requires a local Newton-linearization linked with a damping strategy. A local multilevel damping approach was developed and tested numerically with promising results.

B7: Hybrid Multibody Simulation of Granular Media

The aim of this project (P. Eberhard and S. Luding) is to simulate large systems like bulk solids, e.g., in silos, carriage plants or conveyers consisting of many particles. For granular matter simulations often the discrete element method (DEM) is used. Then, bodies are modeled as rigid bodies, and a global behavior of the physical properties is achieved by means of the contact forces. Here, in normal direction the Kelvin-Voigt model is used. For a stronger dissipation an extension of the model by Gonthier can be used. Another possibility is a model which can also incorporate adhesive behavior. For the contact forces in tangential direction three different models are implemented and compared. First, a regularized model is used where the contact force is dependent on the relative velocity of the bodies. Additionally, a model based on the elasto-plastic analogy as well as the Cundall-Strack model are used. Since the contact detection is very time consuming, sophisticated presorting algorithms are investigated, which search for the directly neighboring bodies. For collision detection, three different methods are implemented and compared, e.g., the ray crossing method, based on a geometrical approach, or the multipole method based on the calculation of the double layer potential, or the common plane method.

B8: Contact Dynamics Treated by Multibody Systems and Multigrid Methods

From the mathematical point of view, many important contributions to mortar discretization methods for two body frictional contact problems have been accomplished in the framework of the Project B8 (P. Eberhard and B. Wohlmuth). New optimal a priori error estimates have been established as well as efficient strategies for the numerical solution of such nonlinear problems. Iterative solvers based on multigrid techniques have been combined with the idea of semi-smooth Newton methods leading to primal-dual active set strategies. This approach leads to an efficient iterative solver, which can be

interpreted as a nonlinear multigrid method. It has been adjusted to energy-conserving time-integrators. Furthermore, an a posteriori error estimator for two-body contact problems has been developed which leads to a fine resolution of the solution at the contact zone. Since standard low-order finite-element formulations show volume locking for nearly incompressible materials a modified finite element formulation has been given to avoid these effects.

In the experimental part, we deal with the oblique impact of a rotating disc striking an elastic strip at low velocities. The experimental setup and methods are designed to (i) generate prescribed translational and rotational motions of and (ii) measure impact properties by a high-speed digital camera system with image processing techniques. The analysis of the measurements shows, also compared with computations, that higher strip flexibility leads to more energy transfer to structural vibrations and a smaller transition incidence angle from sticking to slipping, whilst higher disc initial normal velocity results in, also, more energy transfer to structural vibrations, but a larger transition incidence angle. A hybrid impact model for interpreting the measurements is developed, which can also provide some insight into effects of strip flexibility and disc initial motions.

General Research Achievements

B4: Efficient partitioned fluid-structure interaction scheme in the context of an Arbitrary Lagrangian Eulerian formulation
- stable flow solver on deforming domains using finite elements
- computational mesh dynamics
- stable staggered coupling schemes
- free surface approach included in staggered solution
- vectorization and code optimization

B5: Acoustic-structure interaction
- symmetric hybrid boundary element method
- fast multipole multilevel boundary elements (FMM)
- Mortar FEM-BEM coupling
- preconditioned inexact Uzawa algorithm
- Fuzzy BEM

B7: Multibody simulation of granular materials
- simulation of bulk solids consisting of many particles by means of discrete element method or molecular dynamics
- usage and comparison of different neighborhood search methods: Verlet method, linked cell method and linked linear list
- different collision detection methods: ray crossing, multipole method, common plane method
- consideration of elasticity and dissipation by means of Kelvin-Voigt model and extensions of this model, treatment of adhesion, cohesion

- tangential contact forces by means of Cundall-Strack model, elasto-plastic model and regularized friction models

B8: Contact and impact modeling
- new optimal a priori error estimates
- Newton-methods for frictional contact resulting in a multilevel primal-dual active set strategy
- new adaptive mesh refinement strategy and application to the dynamical case
- experimental setup of oblique contact
- effects of strip flexibility and disc initial motions
- hybrid impact model

Mathematical Models for the Sedimentation of Suspensions*

Stefan Berres[1], Raimund Bürger[2], and Wolfgang L. Wendland[1]

[1] Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
berres@mathematik.uni-stuttgart.de
wendland@mathematik.uni-stuttgart.de
[2] Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Casilla 160-C, Concepción, Chile
rburger@ing-mat.udec.cl

Summary. Mathematical models for sedimentation processes are needed in numerous industrial applications for the description, simulation, design and control of solid-liquid separation processes of suspensions. The first simple but complete model describing the settling of a monodisperse suspension of small rigid spheres is the kinematic sedimentation model due to Kynch [93], which leads to a scalar nonlinear conservation law. The extension of this model to flocculated suspensions, pressure filters, polydisperse suspensions and continuously operated clarifier-thickener units give rise to a variety of time-dependent partial differential equations with intriguing non-standard properties. These properties include strongly degenerate parabolic equations, free boundary problems, strongly coupled systems of conservation laws which may fail to be hyperbolic, and conservation laws with a discontinuous flux. This contribution gives an overview of the authors' research that has been devoted to the mathematical modeling of solid-liquid separation, the existence and uniqueness analysis of these equations, the design and convergence analysis of numerical schemes, and the application to engineering problems. Extensions to other applications and general contributions to mathematical analysis are also addressed.

Keywords: Sedimentation, polydisperse suspensions, mathematical model, system of conservation laws, numerical simulation

1 Introduction

1.1 Problems of Sedimentation

The sedimentation of suspensions involves the mechanics, flow and transport properties of mixtures of fluids and solids, droplets or bubbles. Fundamental

* Research Project A2 "Sedimentation with Compression"

aspects of sedimentation and related solid-liquid separation processes such as filtration or centrifugation include properties of suspensions and emulsions (rheology, particle size and shape, particle-particle interaction, surface characteristics, yield stress, concentration, viscosity), individual particles (orientation and surfactants), and sediments and porous cakes (permeability, porosity and compressibility). These processes are of critical importance for solid-liquid separations in the chemical, mining, pulp and paper, wastewater, food, pharmaceutical, ceramic and other industries.

Mathematical models for these processes are obviously of theoretical as well as of practical importance. They have been widely suggested for more than half a century, starting with the celebrated paper by Kynch [93], who was the first to propose a partial differential equation (PDE), more precisely, a scalar, first-order nonlinear conservation law, together with initial and boundary conditions as a complete (albeit simple) model for the settling of a suspension. Kynch's model was capable to capture most phenomena ranging from the dilute limit to the packed sediment. To put the authors' research in the proper perspective, we emphasize that the available kinematic sedimentation model refers to a suspension of equal-sized hard spheres that settle in a column having neither inlets nor outlets.

Our research in the project A2 has been devoted to

- the mathematical modeling,
- the analysis of well-posedness (existence and uniqueness),
- the design and convergence analysis of numerical schemes, and
- the application to engineering problems

for several extensions of the classical kinematic model, including suspensions forming compressible sediments, filtration devices, continuously operated clarifier-thickener units, and polydisperse suspensions. Although all these extensions are based on engineering applications, they exhibit a variety of intriguing non-standard mathematical properties, such as nonlinear diffusion equations with strong type degeneracy, conservation laws with discontinuous flux, and strongly coupled systems of conservation laws with regions where hyperbolicity is lost.

Most of this research was not part of the original project proposal, but emerged in response to continuous interaction with colleagues from mathematics and engineering, public presentations, comments on published work, and the adaptation of ideas sketched in papers by others. It also turned out that the results could also be used for other applications, such as traffic models [35, 47, 48] and population balance models of ball wear in grinding mills [42]. In addition, some general contributions to mathematical analysis have been made in [30].

1.2 This Contribution

In Sect. 2, we briefly outline the kinematic sedimentation model, the original research problem, enumerate some of the questions stated more than a decade ago, and review the answers found during recent years.

A systematic review of the research performed forms the core of this paper, and is outlined in Sects. 3–6. Each of these sections deals with one particular mathematical or engineering aspect.

Of particular interest is the case of a monodisperse suspension with one single particle species. In Sect. 3 we focus on the one-dimensional flow of such mixtures in several devices, including clarifier-thickener units (Sect. 3.1) and pressure filtration (Sect. 3.2). We illustrate the non-standard features of each of these models; namely, a discontinuous flux and a diffusion term in the former and a free boundary in the latter.

Considering models representing such special cases as a given input, we summarize in the subsequent sections the mathematical analysis that has been applied to some of them. All models considered give rise to time-dependent partial differential equations that have certain non-standard features. In particular, we review results for the following types of equations: strongly degenerate, non-linear parabolic-hyperbolic convection-diffusion equations (Sect. 3.3); conservation laws, in part with discontinuous coefficients (Sect. 3.4); coupled systems of scalar equations of this kind with additional equations of motion (being similar to Navier-Stokes equations), (Sect. 3.5).

In Sect. 4 polydisperse suspensions are considered. In Sect. 4.1 a general mathematical model of suspensions is outlined, which beyond the kinematic approach features both an extension to particles with a size or density distribution (so-called polydisperse suspensions) and the formation of compressible sediments. The derivation of the model may seem fairly complex, but the final equations have the advantage that we may conveniently refer to all special models (say, for monodisperse or polydisperse suspensions, with or without sediment compressibility) as special cases. Thereafter, results for systems of conservation laws of mixed hyperbolic-elliptic type (Sect. 4.2) and quasilinear parabolic systems (Sect. 4.3) are reviewed.

In Sect. 5, we turn to the design and analysis of numerical schemes for the different models. These schemes include finite difference, front tracking and relaxation methods, and a very recent multiresolution technique for efficient computation of discontinuous solutions.

Some rather simple sedimentation and related models for simulations and applications have been implemented for practitioners in order to enable them to perform simulations with a user-friendly tool. This activity is briefly summarized in Sect. 6. However, all models require that certain concentration-dependent material-specific, empirical model functions are known. Clearly, this information has to be obtained from experimentation, which leads to the problem of parameter identification. Recent contributions on the numerical solution of this ill-posed inverse problem are equally reported in Sect. 6.

Finally, Sect. 6 presents a brief discussion of applications of the mathematical and numerical techniques generated while studying sedimentation models to related models in alternative applications (e.g. traffic flow and population balance models).

Selected open problems are addressed in Sect. 7.

2 Kinematic Model of Monodisperse Suspensions

Historically, the mathematical analysis of sedimentation processes began with the spatially one-dimensional kinematic sedimentation model due to Kynch [93], which leads to a scalar non-linear conservation law with a non-convex, smooth flux-density function. As is well known, solutions of such equations are discontinuous in general, and need to be defined as weak solutions. In addition, a selection criterion, i.e., a so-called entropy condition, needs to be postulated in order to single out the physically relevant among possibly several weak solutions. (A comprehensive treatment of the kinematic sedimentation model is given in the monograph *Sedimentation and Thickening* [58], while the articles [54, 64, 65] offer historical reviews on sedimentation and thickening research.)

We now restate the basic properties of the kinematic sedimentation model, which represent the state of knowledge at the beginning of our research. The (idealized) suspension is considered as a continuum and the sedimentation process is represented by the continuity equation of the solid phase:

$$\partial_t \phi + \partial_z f_{\mathrm{bk}}(\phi) = 0 \,, \quad 0 \le z \le L,\ t > 0 \,, \tag{1}$$

where ϕ is the local volume fraction of solids as a function of height z and time t, and $f_{\mathrm{bk}}(\phi) = \phi v_{\mathrm{s}}$ is the *Kynch batch flux density function*, where v_{s} is the solids phase velocity. The basic assumption is that the local solid-liquid relative velocity is a function of the solids volumetric concentration ϕ only, which for batch sedimentation in a closed column is equivalent to stating that $v_{\mathrm{s}} = v_{\mathrm{s}}(\phi)$. For the sedimentation of an initially homogeneous suspension of concentration ϕ_0, (1) is considered together with the initial condition

$$\phi(z,0) = \begin{cases} 0 & \text{for } z = L \,, \\ \phi_0 & \text{for } 0 < z < L \,, \\ \phi_{\max} & \text{for } z = 0 \,, \end{cases} \tag{2}$$

where it is assumed that the function f_{bk} satisfies $f_{\mathrm{bk}}(\phi) = 0$ for $\phi \le 0$ or $\phi \ge \phi_{\max}$ and $f_{\mathrm{bk}}(\phi) < 0$ for $0 < \phi < \phi_{\max}$, where $\phi_{\max}$ is the maximum solids concentration. The knowledge of f_{bk} is sufficient to determine $\phi = \phi(z,t)$ for a given initial concentration ϕ_0, and the solution can be constructed by the method of characteristics. To describe the batch settling velocities of particles in real suspensions of small particles, numerous material specific constitutive

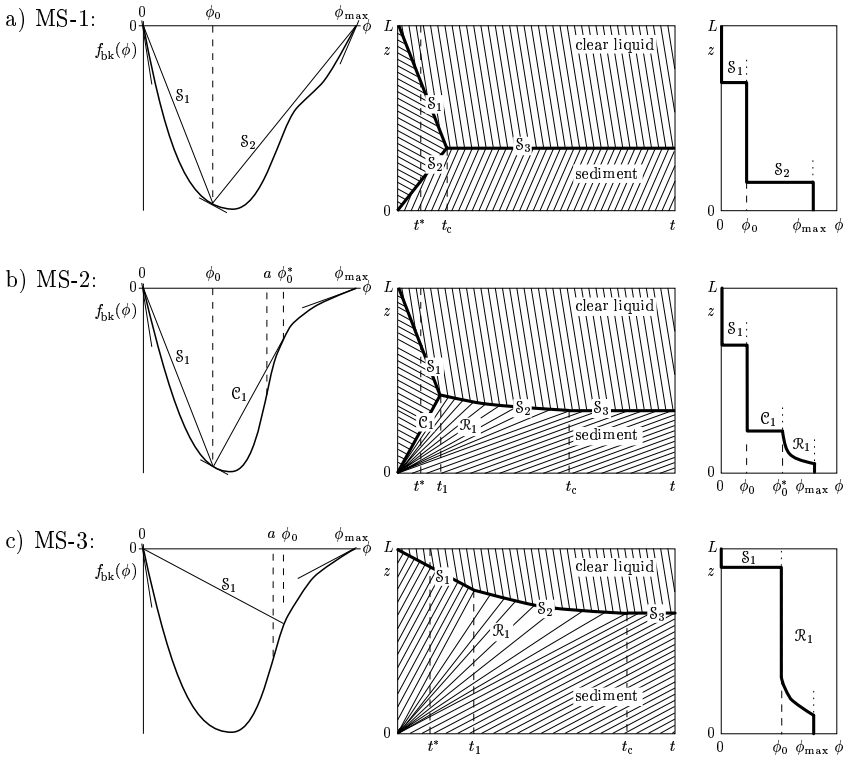


Fig. 1. Modes of sedimentation MS-1 to MS-3 [51]. From the left to the right, the flux plot, the settling plot showing characteristics and shock lines, and one concentration profile (for $t = t^*$) are shown for each mode. Chords in the flux plots and shocks in the settling plots having the same slopes are marked by the same symbols

equations for $v_s = v_s(\phi)$ or $f_{bk}(\phi) = \phi v_s(\phi)$ were proposed (see also [23, 77, 78]). The most frequently used is the equation due to Richardson and Zaki [107]:

$$f_{bk}(\phi) = \begin{cases} u_\infty \phi (1-\phi)^n & \text{for } 0 \le \phi \le 1 \,, \\ 0 & \text{for } \phi < 0 \text{ and } \phi > 1 \,, \end{cases} \qquad n > 1 \,, \tag{3}$$

where u_∞ is the Stokes velocity, i.e., the settling velocity of a single particle in an unbounded fluid. This equation has the inconvenience that the settling velocity becomes zero at the solids concentration $\phi = 1$, while experimentally this occurs at a maximum concentration ϕ_{max} between 0.6 and 0.7. Thus hindered settling functions that are designed to describe a specific suspension should be preferred whenever available.

To construct the solution of the initial value problem (1), (2), the method of characteristics is employed. This method is based on the propagation of $\phi_0(z_0)$, the initial value prescribed at $z = z_0$, at constant speed $f'_{bk}(\phi_0(z_0))$

in a z versus t diagram. These straight lines, the *characteristics*, might intersect, which makes solutions of (1) discontinuous in general. This is due to the nonlinearity of the flux density function f_{bk}. In fact, even for smooth initial data, a scalar conservation law with a nonlinear flux density function may produce discontinuous solutions. To outline some main properties of discontinuous solutions of (1), we consider the *Riemann problem*, where an initial function

$$\phi_0(z) = \begin{cases} \phi_0^+ & \text{for } z > 0 \,, \\ \phi_0^- & \text{for } z < 0 \end{cases} \tag{4}$$

consisting just of two constants is prescribed. Obviously, the initial-value problem (1), (2) consists of two adjacent Riemann problems producing two 'fans' of characteristics and discontinuities, which in this case start to interact after a finite time t_1. At discontinuities, (1) is not satisfied and is replaced by the *Rankine-Hugoniot condition*, which corresponds to the original conservation law and states that the local propagation velocity $\sigma(\phi^+, \phi^-)$ of a discontinuity between the solution values ϕ^+ above and ϕ^- below the discontinuity is given by

$$\sigma(\phi^+, \phi^-) = \frac{f_{\mathrm{bk}}(\phi^+) - f_{\mathrm{bk}}(\phi^-)}{\phi^+ - \phi^-} \,. \tag{5}$$

However, discontinuous solutions satisfying (1) at points of continuity and the Rankine-Hugoniot condition (5) at discontinuities are in general not unique. For this reason, an additional admissibility criterion is necessary to select the physically relevant discontinuous solution. One of these entropy criteria, which determine the unique weak solution and characterize irreversibility, is Oleĭnik's jump condition requiring that

$$\sigma(\phi, \phi^-) \geq \sigma(\phi^+, \phi^-) \geq \sigma(\phi, \phi^+) \quad \text{for all } \phi \text{ between } \phi^- \text{ and } \phi^+ \,. \tag{6}$$

This condition is satisfied if and only if, in an f_{bk} versus ϕ plot, the chord joining the points $(\phi^+, f_{\mathrm{bk}}(\phi^+))$ and $(\phi^-, f_{\mathrm{bk}}(\phi^-))$ remains above the graph of f_{bk} for $\phi^+ < \phi^-$ and below the graph for $\phi^+ > \phi^-$.

Discontinuities satisfying both (5) and (6) are called *shocks*. If, in addition,

$$f'_{\mathrm{bk}}(\phi^-) = \sigma(\phi^+, \phi^-) \quad \text{or} \quad f'_{\mathrm{bk}}(\phi^+) = \sigma(\phi^+, \phi^-) \,, \tag{7}$$

the shock is called a *contact discontinuity*. In that case the chord is tangent to the graph of f_{bk} in at least one of its endpoints. If we assume that $\phi_0^- < \phi_0^+$ and that $f'_{\mathrm{bk}}(\phi) > 0$ for $\phi_0^- \leq \phi \leq \phi_0^+$, it follows that no shock can be constructed between ϕ_0^- and ϕ_0^+. In that case, the Riemann problem (1), (4) has a continuous solution

$$\phi(z,t) = \begin{cases} \phi_0^+ & \text{for } z > f'_{\mathrm{bk}}(\phi_0^+)t \,, \\ (f'_{\mathrm{bk}})^{-1}(z/t) & \text{for } f'_{\mathrm{bk}}(\phi_0^-)t \leq z \leq f'_{\mathrm{bk}}(\phi_0^+)t \,, \\ \phi_0^- & \text{for } z < f'_{\mathrm{bk}}(\phi_0^-)t \,, \end{cases} \tag{8}$$

where $(f'_{\mathrm{bk}})^{-1}$ is the inverse of f'_{bk} restricted to the interval $[\phi_0^-, \phi_0^+]$. This solution is called a *rarefaction wave* and is the unique physically relevant weak solution of the Riemann problem. A piecewise continuous function satisfying the conservation law (1) at points of continuity and (5) and (6) at discontinuities is unique. For the problem of sedimentation of an initially homogeneous suspension, giving rise to two adjacent Riemann problems only, such a solution can be explicitly constructed by the method of characteristics. For example, for a flux density function f_{bk} with up to two inflection points, there are seven qualitatively different solutions, denoted according to [93] as *Modes of Sedimentation*. For the simplest case of a function f_{bk} with exactly one inflection point, three Modes of Sedimentation occur, see Fig. 1.

The formulation of admissibility conditions for more general discontinuous solutions (not necessarily piecewise differentiable ones) led to the concept of *entropy weak solutions*. Kružkov presented in [91] a general existence and uniqueness result.

In 1984, M.C. Bustos [57] embedded Kynch's theory into the state of the art of mathematical analysis. In a series of papers, summarized in Chap. 7 of [58], it was confirmed that the known solutions constructed in [83, 118] are indeed special cases of entropy weak solutions. Utilizing the method of characteristics, it was possible to extend the construction of modes of sedimentation to Kynch batch flux density functions with two and more inflection points.

In 1975 Petty [105] made an attempt to extend Kynch's theory to continuous sedimentation. The basic difference to batch settling in a cylindrical vessel of height L is that the upper end $z = L$ is identified with a feed inlet and the lower $z = 0$ with a discharge outlet. The vessel is fed continuously with feed suspension at the inlet (surface source) and discharged continuously through the outlet (surface sink). The overflow of clear liquid is not explicitly modeled. If $q = q(t)$ is defined as the volume flow rate of the mixture per unit area, which can be prescribed, then Kynch's equation for continuous sedimentation can be written as

$$\partial_t \phi + \partial_z \big(q(t)\phi + f_{\mathrm{bk}}(\phi)\big) = 0 \,. \tag{9}$$

Starting from Petty's model [105], Bustos, Concha and Wendland [59] studied a very simple model for continuous sedimentation, in which (9) is restricted to a space interval $[0, L]$ with Dirichlet boundary conditions at $z = 0$ and $z = L$. The problem is well posed if the boundary conditions are re-interpreted as set-valued entropy boundary conditions [61].

Experimental evidence demonstrated that while Kynch's theory accurately predicts the sedimentation behavior of suspensions of equally sized small rigid spherical particles, this is not the case for flocculent suspensions forming compressible sediments. For such mixtures, a kinematic model is no longer sufficient and one needs to take into account dynamic effects, in particular the concept of effective solid stress. One then obtains a strongly degenerate convection-diffusion equation. The supplement to (1) by an additional degenerating second-order diffusion term gives a suitable extension of Kynch's

theory [56]. The resulting equation (replacing (9)) can be stated as

$$\partial_t \phi + \partial_z \big(q(t)\phi + f_{\mathrm{bk}}(\phi) \big) = \partial_z^2 A(\phi) \ , \tag{10}$$

with

$$A(\phi) := \int_0^\phi a(s)\,\mathrm{d}s \ , \quad a(\phi) := -\frac{f_{\mathrm{bk}}(\phi)\sigma_{\mathrm{e}}(\phi)}{\Delta\varrho g \phi} \ , \tag{11}$$

where $\Delta\varrho > 0$ is the solid-fluid density difference, g is the acceleration of gravity, and σ_{e} is the effective solid stress function. It is assumed that the particles touch each other at a critical concentration $\phi_{\mathrm{c}} \in [0, \phi_{\max}]$, and that the effective solid stress σ_{e} and its derivative σ_{e}' satisfy

$$\sigma_{\mathrm{e}}(\phi) \ , \ \sigma_{\mathrm{e}}'(\phi) \begin{cases} = 0 & \text{for } \phi \le \phi_{\mathrm{c}} \ , \\ > 0 & \text{for } \phi > \phi_{\mathrm{c}} \ . \end{cases} \tag{12}$$

Under the assumptions (11) and (12), and when (3) is used, (10) is a second-order parabolic PDE that degenerates into first-order hyperbolic type for $\phi \le \phi_{\mathrm{c}}$. Since the type degeneracy occurs on a ϕ-interval of positive length, (10) is called strongly degenerate parabolic. The basic difficulty is that the location of the type-change interface, where $\phi = \phi_{\mathrm{c}}$, is not known beforehand.

Strongly degenerating parabolic equations like (10) were little understood at the time of the initial project proposal, and the main motivation of our research was to provide a well-posedness analysis for this equation, to design numerical methods for its solution, and to apply them to practical simulations.

Before formulating some specific problems during the initial stages of the project let us point out that (10) was restricted to a finite length interval $[0, L]$, corresponding to the so-called ideal continuous thickener. At $z = L$, it was presumed that the concentration can be explicitly prescribed as

$$\phi(L,t) = \phi_L(t) \ , \quad t \in (0, T] \ , \tag{13}$$

while at $z = 0$, the total flux is reduced to its convective part, i.e.

$$q(t)\phi + f_{\mathrm{bk}}(\phi) - \partial_z A(\phi) = q(t)\phi \quad \text{at} \quad z = 0 \ ,$$

which reduces to the boundary condition

$$f_{\mathrm{bk}}(\phi) - \partial_z A(\phi) = 0 \quad \text{at} \quad z = 0 \ . \tag{14}$$

It is furthermore assumed that an initial concentration is given:

$$\phi(z,0) = \phi_0(z) \ , \quad z \in [0, L] \ . \tag{15}$$

Equations (10) and (13)–(15) form an initial-boundary value problem (IBVP) for continuous sedimentation of a flocculated suspension, which forms compressible sediments; in short, this IBVP models *sedimentation with compression*. Note that it includes the problem of batch settling of a suspension whenever we set $q \equiv 0$ and $\phi_L \equiv 0$, and that the conventional Kynch analysis is included for $A \equiv 0$. This problem formed the starting point of our research, which was focused on the following questions:

a) Is it possible to extend existing analyses of initial-boundary value problems for conservation laws [2, 70] as well as early approaches to the existence and uniqueness analysis of strongly degenerate parabolic equations [116, 117, 120, 122] to provide well-posedness (existence and uniqueness) for the IBVP (10), (13)–(15)? How does the concept of entropy solutions look like here? In particular, how does the concept of entropy boundary conditions apply here?
b) Can one design an efficient numerical scheme for the IBVP and prove its convergence to an entropy solution?
c) Is it possible to use the results of (a) and (b) to formulate a control problem?
d) Can the model and its analysis be extended to several space dimensions?

During the course of the project, these questions have been dealt with in the following way.

a) Analyses of strongly degenerate parabolic equations available at the beginning of the authors' research included the papers by Volpert [116] and Volpert and Hudjaev [117] published in the late 1960s, and contributions by Wu Zhuoqun and his collaborators from the 1980s [120, 121, 122, 123]. The analysis of the IBVP (10), (13)–(15) was the topic of the second author's doctoral work [15], finished in 1996, which gave rise to the papers [52, 53]. The existence proof for a generalized weak solution satisfying an entropy condition (in short, entropy solution), characterized by Kružkov entropy functions, was based on the vanishing viscosity method. The uniqueness proof for this problem is based on the jump condition established in [123] and relies on the assumption of a smooth diffusion coefficient. A correct formulation of the boundary conditions leads to set-valued so-called entropy boundary conditions [52].
Unfortunately, most diffusion coefficients $a(\phi)$ for the sedimentation-consolidation model do not satisfy these smoothness assumptions. In [27] an improved version of the analysis of [53] is presented, and it is shown that the viscosity method also handles even discontinuous diffusion coefficients and does not lead to new singularities. A new result by Carrillo permits to prove the uniqueness of the generalized solution by Kružkov's "doubling of variables" technique.
The choice of the solution space BV, which formed the basis of the well-posedness analyses in [27, 53], turned out to be a severe restriction for the attempt to generalize the results of these papers to other initial-boundary value problems, including spatially multi-dimensional problems. In part inspired by [101], we utilized the more general concept of divergence-measure fields [63] for the analysis of a free boundary for a problem of pressure filtration [31], see also Sect. 3.3. It seems difficult to extend this approach to multi-dimensional degenerate parabolic equations with zero flux boundary conditions, but see [30] for a partial result.

Though entropy boundary conditions ensure well-posedness of the initial-boundary value problem (10), (13)–(15) [52], it turned out that these conditions are unphysical due to the violation of conservation of mass, and should be replaced by zero-flux or flux-type boundary conditions. For this reason, we analyzed in [27] in parallel a second IBVP, in which the boundary condition (13) is replaced by

$$q(t)\phi + f_{\mathrm{bk}}(\phi) - \partial_z A(\phi) = \Psi(t) \quad \text{at } z = L\,, \tag{16}$$

where $\Psi(t)$ is a boundary feed flux. Well-posedness of the IBVP (10), (14), (15), (16) is shown in [27]; this includes the zero-flux initial-boundary value problem for batch settling attained by setting $q \equiv 0$ and $\Psi \equiv 0$.
It should be emphasized that for the IBVP (10), (14), (15), (16) with flux-valued boundary conditions, the boundary conditions are almost always satisfied in a pointwise sense, which makes entropy boundary conditions unnecessary. Finally, let us comment that the question of boundary conditions experienced another turn in recent years. Since modeling continuous sedimentation through feed and discharge boundary conditions (as stipulated in the Petty-Bustos model [59, 61, 105]) is unphysical, feed and discharge mechanisms should be expressed by singular source terms and flux discontinuities, see Sect. 3.1.

b) After the well-posedness analysis had been completed [15, 52, 53], our efforts were directed to the development of numerical schemes for (10), (13)–(15) (and its variants). In the papers [17, 18, 28], operator-splitting finite difference schemes were used for numerical simulations of the problem, but they were still lacking a rigorous numerical analysis. The analysis of finite difference schemes for strongly degenerate parabolic equations was greatly advanced by the paper by Evje and Karlsen [72], who proved convergence of a monotone finite difference scheme to an entropy solution for the initial-boundary value problem of strongly degenerate convection-diffusion equations. Though this scheme could be easily adapted to the IBVP (10), (14), (15), (16), see [34], it was not obvious how to deal with boundary conditions. This problem was solved very recently in [24, 25], where convergence of explicit and semi-implicit monotone difference methods to an entropy solution is proved. See Sect. 5.1 for details.

c) Due to its involved nonlinear and strongly degenerate parabolic nature, the IBVP (10), (13)–(15) and its variants do not admit exact solution constructions. It seems therefore impossible to precisely predict the effect of control actions, such as changes of the feed flux, as is possible for the kinematic sedimentation model [60]. However, the effect of control actions, in particular changes between steady states, was simulated numerically in [17, 18, 27, 28], and later, for the clarifier-thickener model, in [22, 41, 45]. The problem of optimal control, for example with the aim to maximize the solids throughput or to minimize the fill-up time, has not yet been treated.

d) To overcome the conceptual one-dimensionality of the Kynch model, we formulated in [16, 55, 56] a general continuum mechanical theory of sedimentation-consolidation processes of suspensions of fine, flocculated particles in a viscous fluid. According to the theory of mixtures, the model formulation is based on the mass and linear momentum balances of the solid and the fluid components. Then, for closing the system, constitutive assumptions are introduced, and the equations are simplified following a dimensional analysis. The resulting model is a Stokes- or Navier-Stokes-like system of equations for the (incompressible) flow of the mixture, which is coupled to a strongly degenerate parabolic-hyperbolic quasilinear convection-diffusion equation for the local solids volume fraction. This model was later extended to polydisperse suspensions in [11]. In Sect. 4.1 we briefly outline the general derivation for the polydisperse case, and then refer to monodisperse suspensions as a special case.

3 Solid-Liquid Separation

In this section let us consider one space dimension, for which (34) is the governing equation, and focus on a monodisperse suspension. The governing equation is (10), where the diffusion function $A(\cdot)$ is given by (11) and (12). If reduced to one space dimension only, the mixture flow is completely determined by boundary conditions. The non-standard property of this equation is the solution-dependent parabolic-hyperbolic type degeneracy.

This spatially one-dimensional model leads to initial or initial-boundary value problems for strongly degenerating convection-diffusion equations. Similar models are also obtained for the centrifugation [19] and pressure filtration [21] of flocculated suspensions. The latter case involves a free boundary problem. These models are illustrated in [18, 19, 21] by numerical simulations, in part taking into account published experimental data. A summarizing theory of the solid-liquid separation of suspensions can be found in [81]. This model can be extended to vessels with varying cross-sectional area [26], which opens new design elements for industrial applications.

3.1 Clarifier-Thickener Models

Let us for a moment assume that $A \equiv 0$. The one-dimensional sedimentation model by Kynch [93] arises as a special case of the general sedimentation-consolidation theory. For the continuous settling of an ideal suspension it leads to the conservation law

$$\partial_t u + \partial_x \big(q(x,t)u + h(u) \big) = 0 \,, \tag{17}$$

where the solids concentration u varies with depth x and time t, and $q(x,t)$ is the local mixture velocity. To be consistent with quoted publications, the symbol ϕ for the local solids concentration is replaced by u and $f_{\mathrm{bk}}(\phi)$ by $-h(u)$.

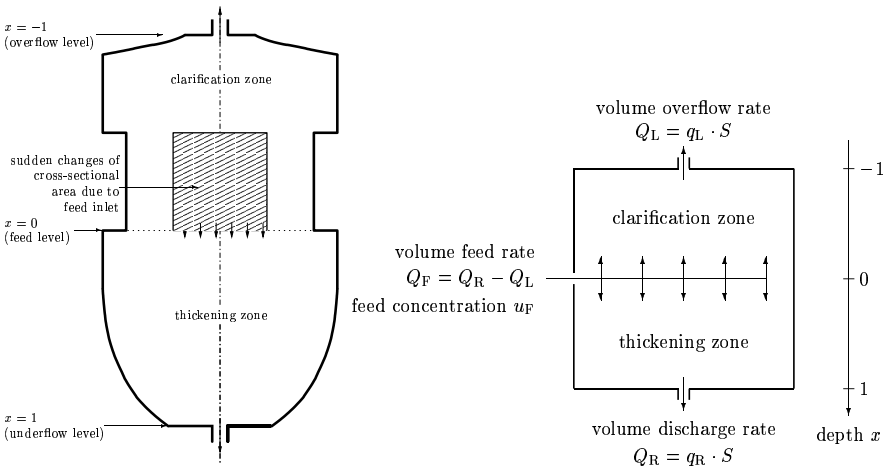


Fig. 2. **(a)** Clarifier-thickener with variable container cross section [39], **(b)** ideal clarifier-thickener with constant cross section [41]

The simple model for continuous sedimentation of [59] reduces (17) to an interval, say $x \in [0, L]$, corresponding to a cylindrical container. At the top $x = 0$ there is an inflow and at the bottom $x = L$ there is a discharge for the thickened sediment. It is assumed that the thickener is fed continuously by the inflow and discharged continuously through the plughole. The mixture velocity $q = q_R(t)$ is a function of time only, where q_R can be controlled at the bottom discharge. In [59] (17) is analyzed together with Dirichlet boundary conditions, where so-called entropy boundary conditions are used in order to show existence and uniqueness of entropy solutions. This model, which can be traced back to Petty [105], has some major disadvantages like the lack of a global conservation principle. In addition, it has been recognized that Dirichlet boundary conditions are physically unrealistic, even though they lead to a mathematically well-posed problem. They should be replaced by the change of the transport flux $q(x,t)u$ and the composite flux $q(x,t)u + h(u)$, which leads to a pure initial value problem.

Moreover, the feed suspension should be fed *between* the discharge openings at the sediment discharge at the bottom and the overflow at the top of the container. Then the one-dimensional modeling leads to an upward-directed mixture velocity $q_L \leq 0$ above and a downward-directed velocity $q_R \geq 0$ below the feed level. The feed source itself is described by a singular source term. Such "clarifier-thickeners" have been proposed by several authors including [62, 95] and have been analyzed thoroughly in particular by Diehl (see [68, 69] and the references therein). To sketch the mathematical models we consider a clarifier-thickener with a (generally discontinuously) varying cross-sectional area $0 < S_{\min} \leq S(x) \leq S_{\max}$, see Fig. 2 (a). The model is then given by the

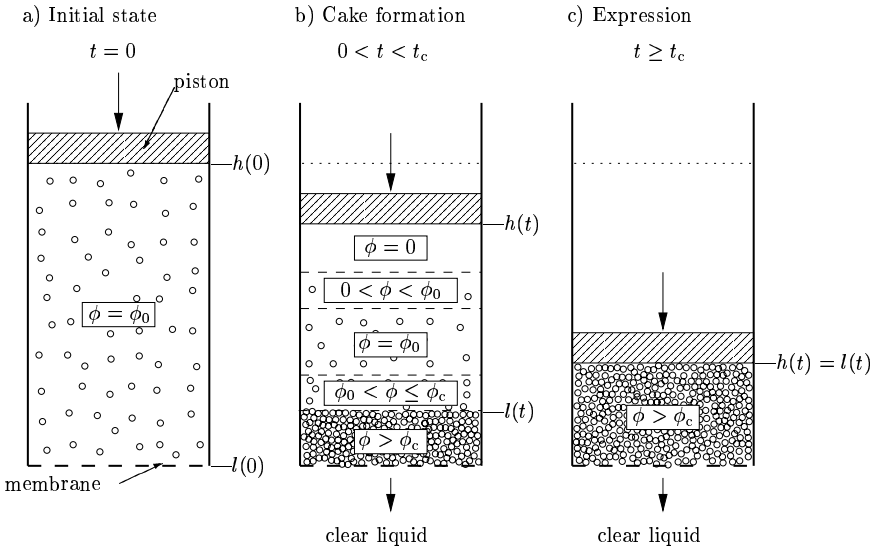


Fig. 3. Pressure filtration of a flocculated suspension [21]

initial value problem

$$S(x)\partial_t u + \partial_x g(u,x) = 0\ ,\ x \in \mathbb{R}\ ,\ t > 0\ ; \quad u(x,0) = u_0(x) \in [0,1]\ ,\ x \in \mathbb{R}\ ;$$

$$g(u,x) := \begin{cases} Q_{\mathrm{L}}(u - u_{\mathrm{F}}) & \text{für } x < -1\ , \\ Q_{\mathrm{L}}(u - u_{\mathrm{F}}) + S(x)h(u) & \text{für } -1 < x < 0\ , \\ Q_{\mathrm{R}}(u - u_{\mathrm{F}}) + S(x)h(u) & \text{für } 0 < x < 1\ , \\ Q_{\mathrm{R}}(u - u_{\mathrm{F}}) & \text{für } x > 1\ , \end{cases} \tag{18}$$

where $Q_{\mathrm{R}} \geq 0$ and $Q_{\mathrm{F}} \geq 0$ are the prescribed volume fluxes of the suspension, which leave the container at the bottom or are fed into the container, respectively, $Q_{\mathrm{L}} = Q_{\mathrm{F}} - Q_{\mathrm{R}}$ is the resulting volume flux at the overflow, and u_{F} is the solids feed concentration. For simplicity, temporally constant control functions Q_{L}, Q_{R} and u_{F} are assumed. The most reported results deal in the first instance with the case of constant cross section $S \equiv$ const. with $q_{\mathrm{L}} = Q_{\mathrm{L}}/S$ and $q_{\mathrm{R}} = Q_{\mathrm{R}}/S$ (Fig. 2 (b)).

The analysis of clarifier-thickener models, including well-posedness, convergence of numerical methods and extensions to flocculated and polydisperse suspensions, has opened a new line of research within the project that generated a series of papers [10, 22, 32, 33, 36, 37, 39, 40, 41, 43, 44, 45]. Mathematical and numerical aspects of clarifier-thickener models are further discussed in Sects. 3.4 and 5.2, respectively.

3.2 Pressure Filtration of Flocculated Suspensions

The sedimentation-consolidation theory was also extended to pressure filtration of flocculated suspensions. A filtration device is sketched in Fig. 3.

The suspension is contained in a cylindrical volume, which is bounded at the top at height $h = h(t)$ by a movable piston and at the bottom by a fixed filter (or "membrane"). This membrane is permeable for the fluid only. When the pressure $\sigma(t)$ is applied to the piston, the fluid is squeezed through the filter element. On the membrane a so-called filter cake builds up. The filter cake grows continuously and its hydraulic resistance increases accordingly, which hinders the movement of the piston.

This filtration process (sedimentation and cake formation) with the mixture flow caused by the piston movement can also be described by the sedimentation-consolidation model in one space dimension. The essential idea is that a complete mathematical model can be obtained and numerically simulated by the choice of appropriate boundary conditions and the coupling of applied pressure $\sigma(t)$ and piston height $h(t)$. Here the sedimentation of the suspension inside the container is also included.

The resulting pressure filtration model can be stated as follows. The field equation for the concentration ϕ as function of time t and height z is

$$\partial_t \phi + \partial_z \big(h'(t)\phi + f_{\mathrm{bk}}(\phi) \big) = \partial_z^2 A(\phi) \ , \quad 0 < z < h(t) \ , \ t > 0 \ . \tag{19}$$

The pore pressure p, which is also sought, can be calculated from $\phi(z,t)$ by

$$\partial_z p = -\varrho(\phi) g - \partial_z \sigma_{\mathrm{e}}(\phi) \ , \quad 0 < z < h(t) \ , \ t > 0 \ , \tag{20}$$

where $\varrho(\phi) = \phi \varrho_{\mathrm{s}} + (1-\phi)\varrho_{\mathrm{f}}$ is the local density of the mixture, if ϱ_{s} and ϱ_{f} are the densities of the solid and the liquid, respectively.

Equations (19) and (20) are supplied with the initial conditions

$$h(0) = h_0 \ ; \quad \phi(0,z) = \phi_0(z) \ , \quad 0 \le z \le h_0 \ , \tag{21}$$

(h_0 is the initial height and ϕ_0 is the initial concentration) and the kinematic boundary conditions

$$\big(f_{\mathrm{bk}}(\phi) - \partial_z A(\phi) \big)(0,t) = -h'(t)\phi(0,t) \ , \quad t > 0 \ , \tag{22}$$

$$\big(f_{\mathrm{bk}}(\phi) - \partial_z A(\phi) \big)\big(h(t),t\big) = 0 \ , \quad t > 0 \ . \tag{23}$$

Conditions (22) and (23) state that the solid phase is held up at $z = 0$ and transported at $z = h(t)$ with the piston velocity $h'(t)$. The coupling between $\sigma(t)$ and $h(t)$ is described by the dynamic boundary condition

$$\sigma(t) = \sigma_{\mathrm{e}}\big(\phi(0,t)\big) - g\big(m_0 + \varrho_{\mathrm{f}}(h(t) - h_0)\big) - \mu_{\mathrm{f}} R_{\mathrm{m}} h'(t) \ . \tag{24}$$

Here m_0 is the initial mass of the suspension, divided by the container cross section, μ_{f} is the dynamical viscosity of the pure fluid, and R_{m} is the hydraulic

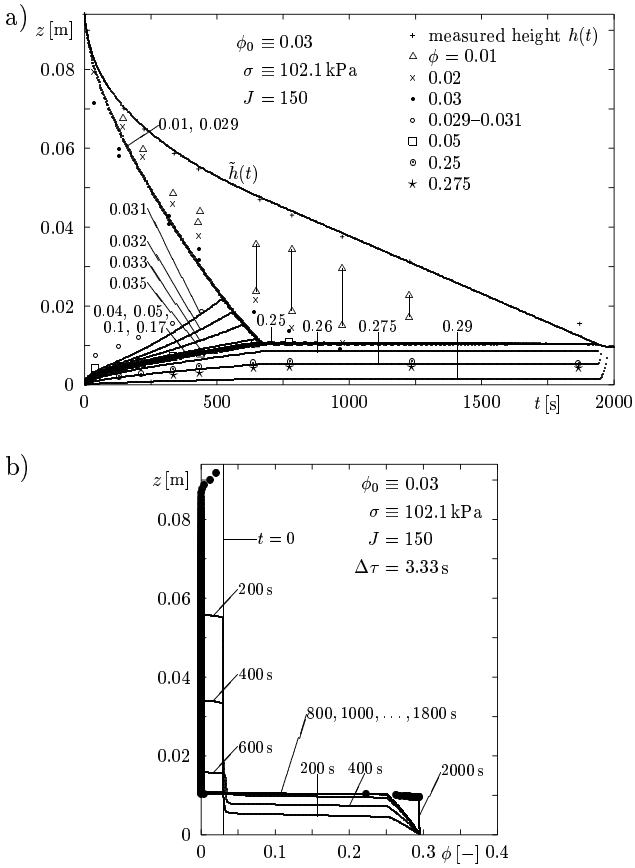


Fig. 4. Simulation of a filtration experiment [111]: **(a)** Measured (symbols) and simulated iso-concentration lines, **(b)** simulated concentration profiles and concentration (dots) in height $\tilde{h}(t)$, plotted in time intervals of length $\Delta\tau$ [21]

resistance of the filter element. The material properties of the suspension and the filter cake are thus described by the functions $f_{\mathrm{bk}}(\phi)$ and $\sigma_{\mathrm{e}}(\phi)$ as well as by the constants $\bar{\varrho}$ und μ_{f}. The filtration device is characterized by the initial height h_0 and the filter resistance R_{m}.

The externally prescribed control function is either the applied pressure $\sigma(t)$ or the piston velocity $h'(t)$, which corresponds to the filtration rate. The choice of the control function determines the resulting mathematical model. If the piston trajectory $h(t)$ (or the corresponding filtration rate) is prescribed, then (19) and (21)–(23) form an initial-boundary value problem and the applied pressure $\sigma(t)$ necessary to perform the filtration process follows from (24). The more important and also more interesting case occurs when $\sigma(t)$ is given and the piston trajectory (or the filtrate rate) $h(t)$ is sought. Since the

coupling of $\sigma(t)$ and $h(t)$ by (24) involves the evaluation of the solution at $z = 0$, which in turn depends on $h(t)$, we obtain a free boundary value problem for a strongly degenerate parabolic equation. This formulation is presented in [21], along with numerical simulations of pressure filtration processes, see Fig. 4.

For the analysis of the free boundary value problem, a new approach by the theory of divergence-measure fields is needed [31]. We elaborate on this in Sect. 3.3.

3.3 Strongly Degenerate Parabolic Equations

The analysis of scalar, strongly degenerate parabolic-hyperbolic PDEs was initially focused on generalized entropy solutions in the space $BV(Q_T)$, where Q_T is the cylindrical computational domain [27, 52, 53]. It turned out that the BV framework is an excessive limitation, even for spatially one-dimensional initial-boundary value problems, which causes severe problems for the uniqueness analysis; for example, the BV analysis in [115] is based on the assumption of additional regularity properties of the weak solutions that can hardly be verified in practice. Now these problems can be solved by the recent theory of divergence-measure fields, which will briefly be outlined.

To show that the generalized solution u of a conservation law or a strongly degenerate parabolic equation belongs to $BV(Q_T)$, one needs to derive estimates of $\|\partial_x u^\varepsilon\|_{L^1(Q_T)}$ and $\|\partial_t u^\varepsilon\|_{L^1(Q_T)}$ for solutions u^ε of the regularized problem. These estimates need to be uniform with respect to the regularizing parameter ε. In combination with an L^∞ bound on u^ε, the assumptions of Kolmogorov's compactness criterion are satisfied, which entails the existence of a limit $u \in L^\infty(Q_T) \cap BV(Q_T)$, which represents the sought generalized solution. The significance of the space $BV(Q_T)$ is the existence of traces of the limit function u at the spatial boundary of Q_T provided that $u \in BV(Q_T)$. The analysis in [27] illustrated that these traces indeed are necessary for the existence of generalized solutions.

For various reasons, the BV approach represents a strong restriction. The obvious problem lies in the difficulty, and sometimes impossibility, to obtain the required uniform estimates. Of course, this problem is even more severe when passing to multi-dimensional equations of the form

$$\partial_t u + \nabla_{\mathbf{x}} \cdot \mathbf{f}(u) = \Delta A(u) \,, \quad (\mathbf{x}, t) \in Q_T := \Omega \times (0, T) \,, \quad \Omega \subset \mathbb{R}^n \,. \tag{25}$$

Here, one needs estimates on $\|\nabla_{\mathbf{x}} u^\varepsilon\|_{L^1(Q_T)}$. Whenever it is possible to estimate the latter quantity, but not $\partial_t u^\varepsilon$, one may use Kružkov's "interpolation lemma" [91, Lemma 5] to conclude that u^ε converges for $\varepsilon \to 0$ to a limit $u \in BV_{1,1/2}(Q_T) \supset BV(Q_T)$, which means that there exists a constant K such that

$$\iint_{Q_T} \big|u(\mathbf{x}+\Delta\mathbf{x},t)-u(\mathbf{x},t)\big|\mathrm{d}\mathbf{x}\mathrm{d}t \le K|\Delta\mathbf{x}| \,,$$
$$\iint_{Q_T} \big|u(\mathbf{x},t+\Delta t)-u(\mathbf{x},t)\big|\mathrm{d}\mathbf{x}\mathrm{d}t \le K|\Delta t|^{1/2} \,.$$

The $BV_{1,1/2}$ estimates on $\{u^\varepsilon\}$ are sufficient for the application of Kolmogorov's compactness criterion, which implies the existence of a limit u. The problem consists in the boundary conditions and in the uniqueness of u, since the existence of traces of a function $u \in BV_{1,1/2}(Q_T)$ is not ensured, hence boundary conditions need to be formulated without the concept of trace, and the uniqueness of generalized solutions is not obvious then. A further limitation of the BV approach becomes apparent in [46] through the restriction to a rather narrow class of admissible initial functions, which are necessary to achieve a uniform estimate of the time derivative.

These difficulties of the BV approach motivated the search for a more general concept of generalized solutions. Here, the concept of so-called divergence-measure fields, which were introduced by Anzellotti in [1], turned out to be useful. The corresponding analysis in [31] is based on the formulation by Chen and Frid [63].

We recall that $u \in L^\infty(Q) \cap BV(Q)$, where $Q \subset \mathbb{R}^N$, if and only if

$$\|u\|_{BV(Q)} := \sup\left\{ \int_Q u\nabla\cdot\boldsymbol{\varphi}\,\mathrm{d}\mathbf{x} : \boldsymbol{\varphi} \in \big(C_0^1(Q)\big)^N,\ \|\boldsymbol{\varphi}\|_{L^\infty(Q)} \le 1 \right\}$$

is finite. Then the basic idea in [63] consists in replacing the property $u \in L^\infty(Q) \cap BV(Q)$ by the requirement that a vector field $\mathbf{F} \in L^p(Q,\mathbb{R}^N)$ associated with u satisfies the condition $|\mathrm{div}\,\mathbf{F}|(Q) < \infty$ with

$$|\mathrm{div}\,\mathbf{F}|(Q) := \sup\left\{ \int_Q \mathbf{F}\cdot\nabla\varphi\,\mathrm{d}\mathbf{x} : \varphi \in C_0^1(Q;\mathbb{R}),\ \|\varphi\|_{L^\infty(Q)} < 1 \right\} .$$

Here, we define the class of L^p divergence-measure fields by

$$\mathcal{DM}^p(Q) = \left\{ \mathbf{F} \in L^p(Q;\mathbb{R}^N) :\ |\mathrm{div}\,\mathbf{F}|(Q) < \infty \right\} .$$

For $\mathbf{F} \in \mathcal{DM}^p(Q)$ we have that $\mathrm{div}\,\mathbf{F}$ is a Radon measure on Q. If the components of $\mathbf{F}$ are Lipschitz continuous with respect to u, as in the application to conservation laws, it becomes clear that $u \in L^\infty(Q) \cap BV(Q)$ implies $\mathbf{F} \in \mathcal{DM}^\infty(Q)$. Properties of L^∞ divergence-measure fields are derived in [63]. In particular, a generalized Gauß-Green formula can be established for a class of bounded domains, which then allows the definition of traces.

For scalar conservation laws, every convex entropy pair is an L^∞ divergence-measure field on $Q = Q_T \subset \mathbb{R}^N$ if we consider $Q_T = \Omega \times (0,T)$ with a bounded spatial domain $\Omega \subset \mathbb{R}^{N-1}$. Using the Gauß-Green formula, Chen and Frid [63] find a corresponding weak solution for L^∞ (not BV) solutions to scalar conservation laws with boundary conditions. Moreover, they derive

the entropy boundary conditions with entropy boundary fluxes introduced by Otto [100, 104]. Most of the properties of L^p divergence-measure fields derived in [63] are also valid for $1 < p < \infty$. The case $p = 2$ is of particular interest for the analysis of strongly degenerate parabolic equations, since standard a priori estimates allow to show that the Kružkov entropy pair of a strongly degenerate parabolic equation represents an L^2 divergence-measure field over Q_T. This was exploited in [101] for the proof of well-posedness of the inhomogeneous Dirichlet problem of the strongly degenerate parabolic equation (25).

In [52] entropy boundary conditions are derived for strongly degenerate parabolic equations in the application to sedimentation with compression. Traces of the solution near the boundary of the computational domain can be defined only if the diffusion coefficient $a(u)$ has certain regularity properties (for example, Lipschitz continuity).

Even though the free boundary problem of pressure filtration is spatially one-dimensional, the estimate on $\partial_t u^\varepsilon$ required for the BV approach could not be obtained here yet. So, for the analysis the theory of divergence-measure fields is applied in [31].

3.4 Conservation Laws and Related Equations with Discontinuous Coefficients

To model clarifier-thickeners, we consider conservation laws of the type

$$\partial_t u + \partial_x f\big(\gamma(x), u\big) = 0 \,, \qquad x \in \mathbb{R} \,, \quad t > 0 \,, \tag{26}$$

where $u(x,t)$ is the scalar unknown and $f(\gamma, u)$ is a given flux function. The salient feature of (26) is the explicit dependence of the convection term on the position x through a possibly discontinuous parameter $\gamma(x)$. PDEs of this form appear in various applications such as flows in porous media [82], sedimentation processes [36, 68, 69] and models of traffic flow [35]. If the coefficient $\gamma(x)$ is discontinuous, the usual Kružkov theory [91] of entropy solutions breaks down. In this case, (26) is frequently written as the following 2×2 system of equations:

$$\partial_t \gamma = 0 \,, \qquad \partial_t u + \partial_x f(\gamma, u) = 0 \,. \tag{27}$$

When $u \mapsto f_u(\gamma, u)$ changes sign, this system is not strictly hyperbolic and becomes resonant. One consequence of resonance is the lack of an a priori bound on the total variation of the conserved quantity u [110]. In general, no spatial BV bound for u is available, so that a singular mapping approach is employed to prove the convergence of numerical schemes and the existence of weak solutions. This method was introduced by Temple [110] and had an enormous impact: it is used in [96, 97] to prove convergence of the 2×2 Godunov method, in [113, 114] to show convergence of the scalar Engquist-Osher and Godunov methods, in [36, 82, 88, 89, 90] to study front tracking methods based on 2×2 Riemann solvers and in [86] to analyze scalar relaxation

methods. The singular mapping approach consists of the construction of a mapping $\Psi(u)$ that is a continuous monotone function of u, and the derivation of a uniform bound of the total variation of $\Psi(u_\Delta)$, where u_Δ is the approximate solution. This bound ensures the strong convergence of $\Psi(u_\Delta)$ to a function $\bar{\Psi}$. Finally, one proves that $u := \Psi^{-1}(\bar{\Psi})$ is a weak solution. The singular mapping approach was applied for an entropy solution concept and an associated uniqueness proof in [39] for an equation with discontinuous coefficients which describes the continuous sedimentation of a monodisperse suspension.

3.5 Systems of Scalar Equations Coupled with Equations of Motion

The papers [49] and [50] are devoted to multi-dimensional systems of model equations for sedimentation-consolidation processes. In [50] the coupling between the conservation of mass equation for the solid phase with equations of motion for the mixture is considered, and energy estimates for several different regularizations of this system are derived. These energy estimates may be used for the future design of numerical schemes.

In [49] simplified models for the sedimentation of suspensions in closed, spatially two-dimensional vessels are considered. These models are based on the theory of kinematic waves. It is proved that these models, in which the motion of the mixture is coupled with the concentration fronts by boundary conditions only, are in general not well posed due to the absence of inertial or viscous terms.

4 Polydisperse Suspensions

4.1 Model Equations of Polydisperse Suspensions

We consider polydisperse suspensions of small spherical particles of a finite number N of species having the diameters $d_1, \ldots, d_N$ and the densities $\varrho_1, \ldots, \varrho_N$, where $d_i \neq d_j$ or $\varrho_i \neq \varrho_j$ for $i \neq j$. Each of these species is modeled as a separate solid phase, which leads to systems of conservation laws. In [46], the following multi-dimensional model equations for polydisperse mixtures are derived, starting from the mass and linear momentum balances, followed by constitutive assumptions and an order-of-magnitude analysis:

$$\partial_t \phi_i + \nabla \cdot \big(\phi_i \mathbf{q} + f_i(\Phi)\mathbf{k}\big) = 0, \quad i = 1, \ldots, N \,, \tag{28}$$

$$\nabla \cdot \mathbf{q} = 0 \,, \tag{29}$$

$$\nabla p = -\varrho(\Phi) g \mathbf{k} + \frac{1}{1-\phi} \nabla \cdot \mathbf{T}_{\mathrm{f}}^{\mathrm{E}}(\Phi, \nabla \mathbf{q}) \,. \tag{30}$$

Here, ϕ_i is the volumetric concentration of species i (having diameter d_i and density ϱ_i) and $\Phi := (\phi_1, \ldots, \phi_N)^{\mathrm{T}}$, t is time, $\mathbf{q}$ is the volume averaged mixture

flow velocity, $\mathbf{k}$ is the upwards-pointing unit vector, p is the pore pressure, $\varrho(\Phi) := \varrho_1\phi_1 + \cdots + \varrho_N\phi_N + (1-\phi)\varrho_\mathrm{f}$ is the local density of the mixture, $\phi := \phi_1 + \cdots + \phi_N$ is the total solids concentration, g is the acceleration of gravity, ϱ_f is the density of the fluid, and $\mathbf{T}_\mathrm{f}^\mathrm{E}$ is the viscous stress tensor of the fluid. The decisive ingredient of these equations is the solids flux density vector $\mathbf{f} := (f_1, \ldots, f_N)^\mathrm{T}$, where the components $f_1, \ldots, f_N$ are functions of Φ. Such functions have frequently been proposed in the literature as generalizations of the flux density function f_bk. One example is the so-called Masliyah-Lockett-Bassoon (MLB) model for polydisperse suspensions of spheres which might differ in size and in density [98, 102]. For $i = 1, \ldots, N$ one obtains

$$f_i^\mathrm{M}(\Phi) = \mu(1-\phi)^{n-2}\phi_i\left[\delta_i(\bar\varrho_i - \bar{\boldsymbol{\varrho}}^\mathrm{T}\Phi) - \sum_{k=1}^{N}\delta_k\phi_k(\bar\varrho_k - \bar{\boldsymbol{\varrho}}^\mathrm{T}\Phi)\right] , \qquad (31)$$

where $\mu = -gd_1^2/(18\mu_\mathrm{f})$ (μ_f is the fluid viscosity), $n > 2$, $\delta_i := d_i^2/d_1^2$, $\bar\varrho_i := \varrho_i - \varrho_\mathrm{f}$, $i = 1, \ldots, N$ and $\bar{\boldsymbol{\varrho}} := (\bar\varrho_1, \ldots, \bar\varrho_N)^\mathrm{T}$.

In one space dimension, only (28) needs to be solved, where $q \equiv 0$ in a closed column of height L, which follows from (29) and $q_z = 0$ at the bottom. This leads to a system of conservation laws of the form

$$\partial_t\phi_i + \partial_z f_i(\Phi) = 0 , \quad i = 1, \ldots, N , \qquad (32)$$

$$\Phi(z,0) = \Phi_0(z),\ 0 \le z \le L ; \quad \mathbf{f}|_{z=0} = \mathbf{f}|_{z=L} = 0 , \quad t > 0 . \qquad (33)$$

As is well known, solutions of (32) are discontinuous in general, and the propagation velocity $\sigma(\Phi^+, \Phi^-)$ of the discontinuity which separates the states Φ^+ and Φ^-, is given by the Rankine-Hugoniot condition $f_i(\Phi^+) - f_i(\Phi^-) = \sigma(\phi_i^+ - \phi_i^-)$. The system (32) is hyperbolic when all eigenvalues of the Jacobian $\mathcal{J}_\mathbf{f}(\Phi) := (\partial_{\phi_k} f_i)_{1\le i,k\le N}$ are real, and strictly hyperbolic when, in addition, these are pairwise different. A system with only pairs of complex-conjugate non-real eigenvalues is elliptic.

Depending on the choice of the flux function $\mathbf{f}(\Phi)$ and parameters, system (32) can become non-hyperbolic or, in the case $N = 2$, changes from hyperbolic to elliptic type [46] in a subregion of the phase space. There, for $\Phi \in E$, $E \subset \mathcal{D}_1 := \{\Phi \in \mathbb{R}^N : \Phi \ge 0, \phi \le 1\}$, the system is non-hyperbolic (or elliptic) with $\mathcal{D}\backslash E \ne \varnothing$. The appearance of these type changes depends on the sizes and densities of the particles involved. The ellipticity is equivalent to the criterion given in [3] for the appearance of instabilities such as blobs and finger-type structures, which have been also observed experimentally.

In [11], the polydisperse sedimentation model given by (28)–(30) was extended to compressible sediments. To this end, the model assumptions that had been used so far for monodisperse flocculated suspensions, in particular the effective solid stress σ_e, were extended to polydisperse suspensions. This leads to the following system of equations (instead of (28)), where we confine ourselves to only one space dimension:

$$\partial_t\Phi + \partial_z\big(q\Phi + \mathbf{f}^\mathrm{M}(\Phi)\big) = \partial_z\big(\mathbf{A}(\Phi)\partial_z\Phi\big) . \qquad (34)$$

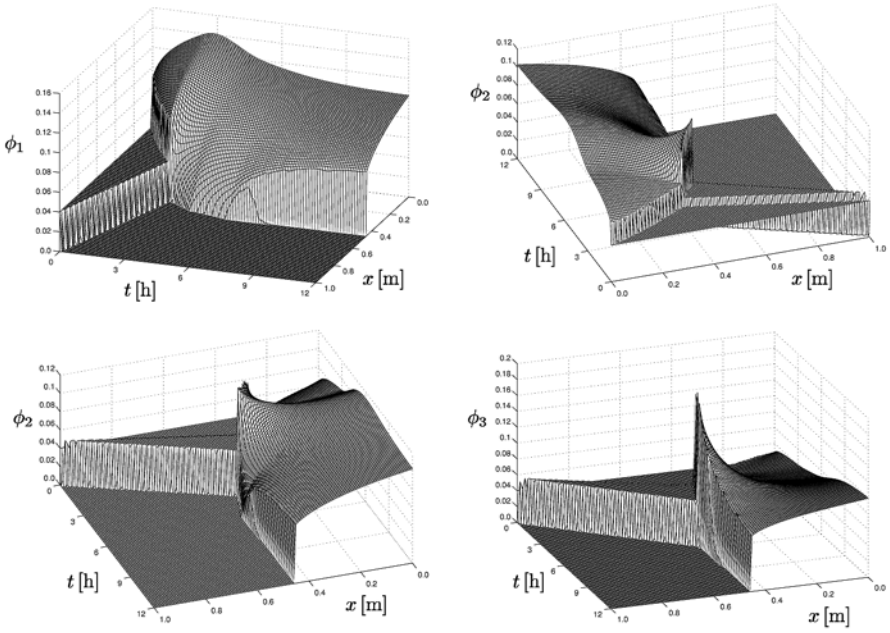


Fig. 5. Simulation of a tridisperse suspension with equal-density particles, $d_1 > d_2 > d_3$, and formation of a compressible sediment [11]

The diffusion matrix $\mathbf{A}(\Phi)$, which includes the function $\sigma_e(\phi)$ and its derivative, vanishes for $\phi \leq \phi_c$, and is usually non-sparse and non-symmetric.

An eigenvalue analysis shows that system (34) is parabolic for all non-trivial Φ satisfying $\phi_c < \phi < \phi_{max}$ and that thus the application to polydisperse suspensions produces a parabolic-hyperbolic degenerate quasilinear system of PDEs. In [11] the Kurganov-Tadmor method [92] was used for the simulation of polydisperse sedimentation with compressible sediment layers modeled by (34), see Fig. 5.

4.2 Mixed Hyperbolic-Elliptic Systems

Systems of mixed hyperbolic-elliptic type also appear in transonic flow, traffic flow, one-dimensional instationary flow of a Van-der-Waals gas, in the propagation of phase boundaries in an elastic beam, and multi-phase flows in porous media. There is a particular similarity between systems of conservation laws for three-phase flows in porous media and those of the sedimentation of a bidisperse suspension. Models of multi-phase flows are the main motivation for studying systems with type change. Surveys on the theory of mixed systems of conservation laws and their applications are given in [73, 87]. The appearance of ellipticity regions raises the question about the actual effects of complex eigenvalues. In first practical numerical computations, ellipticity

regions did not cause instabilities. Therefore, the type change did not draw particular attention [4]. Also in the particular sample calculations for polydisperse suspensions performed in [20, 74] no oscillations appeared. (In contrast to [20], both systems solved numerically in [29] are hyperbolic.) A reason for the non-appearance of oscillations is numerical diffusion, which is artificially introduced by most finite difference schemes and thus transforms a system of conservation laws into a well-posed parabolic system. The possibility of (measure-valued) oscillatory solutions for non-hyperbolic systems of first order, however, is demonstrated analytically and numerically in [75, 76].

In [99] it is shown that the structure of the diffusion matrix $\mathbf{D}$ is of importance since it determines the instability region. Since the stability or admissibility of shock may depend on the form of the (nonlinear) diffusion matrix, a mixed system cannot be comprehended offhand as the viscosity limit of a parabolic system. For the MLB model, numerical and experimental investigations of hydrodynamic diffusion of polydisperse suspensions [29, 67, 112] may provide guidance.

Summarizing, one can say that the mathematical and numerical theory and the general understanding of mixed systems has made enormous progress since those systems had first been investigated in applications [4]. The main questions are, however, still open. In particular, there is no general theory and no generally accepted shock admissibility criterion. Finally, one should emphasize that the complicated wave structures in the solutions of mixed systems are usually not observed in experiments. This observation has led to the conclusion that the main reason of the emergence of mixed systems is poor modeling since the mixed type often goes back to the introduction of closures of balance equations. This conclusion, however, is *not* valid for the models of sedimentation of polydisperse suspensions, since the instabilities predicted by the type change have been indeed observed [119].

By a perturbation approach it is shown in [46] that loss of hyperbolicity allows the appearance of instabilities also for arbitrary N. For $N = 3$, the discriminant

$$\begin{aligned} I_3(\Phi) &:= 4s^3 - s^2r^2 + 27t^2 + 4r^3t - 18rst\ ,\ r := -\operatorname{tr}\mathcal{J}_{\mathbf{f}}\ ,\ t := -\det\mathcal{J}_{\mathbf{f}}\ , \\ s &:= -\big(\partial_{\phi_3}f_1\partial_{\phi_1}f_3 + \partial_{\phi_2}f_1\partial_{\phi_1}f_2 + \partial_{\phi_3}f_2\partial_{\phi_2}f_3 \\ &\qquad - \partial_{\phi_1}f_1\partial_{\phi_2}f_2 - \partial_{\phi_1}f_1\partial_{\phi_3}f_3 - \partial_{\phi_2}f_2\partial_{\phi_3}f_3\big)\ , \end{aligned} \tag{35}$$

of the characteristic polynomial of $\mathcal{J}_{\mathbf{f}}(\Phi)$ shows that hyperbolicity is lost precisely where $I_3 > 0$. With $\phi_3 = 0$, this criterion is valid also for $N = 2$.

In [46] we numerically evaluate I_2 and I_3 and determine instability regions for three different choices of $\mathbf{f}(\Phi)$ (see Fig. 6). Moreover, it is shown that system (32) is strictly hyperbolic for all $\Phi \in \mathcal{D}_1$ with $\phi < 1$ for equal-density bidisperse suspensions whenever the flux vector (31) is used.

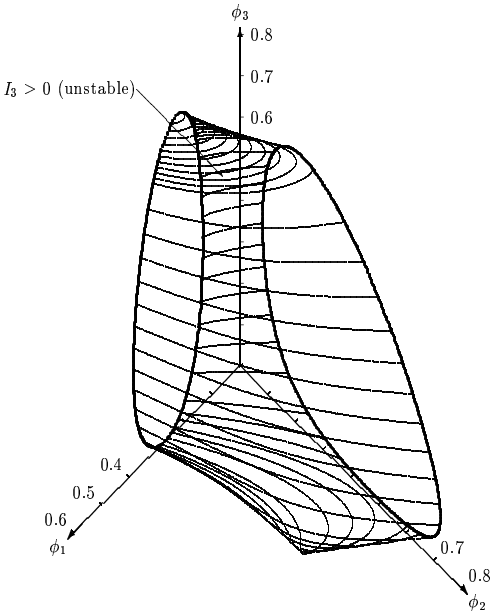


Fig. 6. Three-dimensional instability region for the MLB model for $N = 3$ with $\delta_2^2 = \delta_3^2 = 0.5$, $\bar{\varrho}_2/\bar{\varrho}_1 = 1$, $\bar{\varrho}_3/\bar{\varrho}_1 = -1/2$ and $n = 4.65$ [46]

The conjecture that the MLB model is strictly hyperbolic for particles of equal density (but different sizes) was proved in [11] by exploiting that the structure of the functional matrix $\mathcal{J}_{\mathbf{f}}(\Phi)$ admits elimination possibilities that lead to a closed formula for its characteristic polynomial.

4.3 Quasilinear Parabolic Systems

In [9], the well-posedness of a class of Neumann problems for $n \times n$ quasilinear parabolic systems modeling the sedimentation of polydisperse suspensions with compression is discussed. After the transformation of the diffusion matrix to an upper triangular form, the classical Hölder space theory [94] is applied. Since the Neumann boundary condition induces (in contrast to the Dirichlet boundary) a nonlinear coupling of the equations, a time stepping procedure is introduced, where the boundary conditions are piecewise linear. While for the standard (zero-flux) Neumann problem only existence can be shown, for regularized boundary problems the well-posedness can be proved. See [9] for details.

5 Numerical Methods

5.1 Monotone Schemes for Initial-Boundary Value Problems of Strongly Degenerate Parabolic Equations

Strongly degenerate convection-diffusion equations cannot be solved by standard methods for parabolic equations, since these in general converge to wrong solutions. Suitable methods can be constructed, e.g., by extending a scheme for conservation laws by a conservative discretization of the degenerating diffusion term. A review of suitable methods is given in [28]. In [34] an extension of the Engquist-Osher scheme to strongly degenerate convection-diffusion equations is given, which also includes a second-order method obtained by MUSCL-type extrapolation. For the first-order method introduced in [34] convergence to the entropy solution of the initial-boundary value problem is shown in [25]. A variant of this method is analyzed in [24].

5.2 Methods for Conservation Laws with Discontinuous Flux

In [58] continuous sedimentation processes are described by initial-boundary value problems of a scalar conservation law. However, an improved model can be achieved if the boundary conditions are replaced by continuous transitions between different flux functions, and the feed mechanism is described by a singular source term, which can be incorporated into the discontinuous change of flux functions. These so-called clarifier-thickener models may be described as initial-value problems without boundary conditions. Similar conservation equations with discontinuous flux appear in traffic flow models with abruptly changing road surface conditions and in multiphase flow in heterogeneous porous media. The convergence of the front tracking method for the simulation of continuous separation in such units is proved in [36]. Alternatively, a relaxation method may be employed [37]. For the case of a discontinuously varying vessel cross-sectional area, convergence of a monotone finite difference scheme is established in [39]. The results of [36] are summarized in [40]. For a vessel with constant cross-sectional area, convergence to the unique entropy solution is shown in [41]. The convergence of a monotone method and essential parts of the analysis could also be extended to a clarifier-thickener model for flocculated suspensions having an additional diffusion term [45].

Numerical methods for systems of conservation laws with discontinuous coefficients, which appear, e.g., when the polydisperse sedimentation model is combined with the clarifier-thickener setup, are compared in [10]. The numerical simulation of (in part, flocculated) suspensions in clarifier-thickeners is treated in [22, 25, 43, 103].

The finite-difference scheme which is used to solve problem (18) is a variant of the known Engquist-Osher upwind scheme [71], which now in addition considers the spatial variation of the flux in (18), which corresponds to a variable container cross section $S(x)$. Here, the flux $g(x, u)$ depends on a

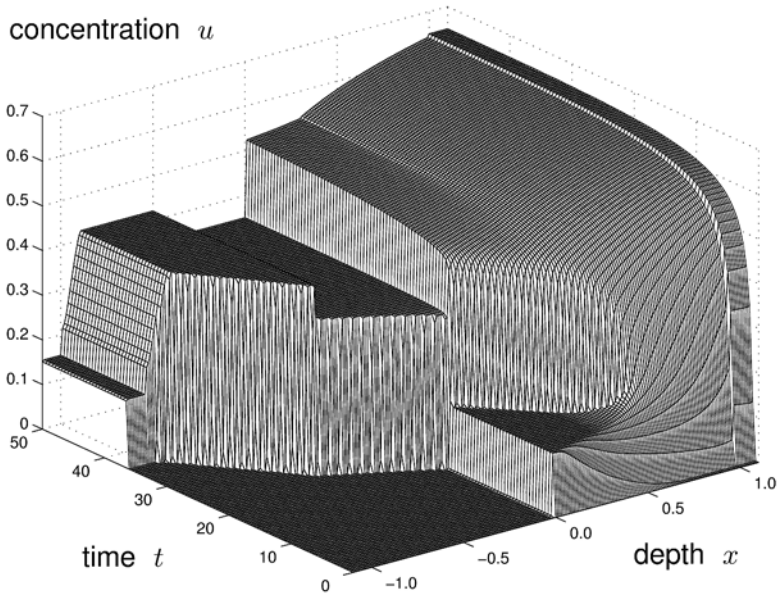


Fig. 7. Simulation of a continuous clarifier-thickener with variable cross section [39]

pair $\boldsymbol{\gamma}(x) := (\gamma^1(x), \gamma^2(x))$ of spatially varying parameters, i.e. $g(x,u) = f(\boldsymbol{\gamma}(x),u)$. The parameter vector $\boldsymbol{\gamma}$ is discretized on a grid that is staggered against that of the conserved quantity u.

This leads to the following scheme: if $\Delta x > 0$ is chosen, one sets $x_j := j\Delta x$ and discretizes the parameter vector $\boldsymbol{\gamma}$, the initial data and the cross-sectional area function by

$$\boldsymbol{\gamma}_{j+\frac{1}{2}} := \frac{1}{\Delta x}\int_{x_j}^{x_{j+1}} \boldsymbol{\gamma}(x)\,\mathrm{d}x\ , \qquad U_j^0 := \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)\,\mathrm{d}x\ ,$$

$$S_j := \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} S(x)\,\mathrm{d}x\ ,$$

respectively, then the scheme for the approximation U_j^n reads:

$$U_j^{n+1} = U_j^n - \lambda_j \Delta_- f^{\mathrm{EO}}\big(\boldsymbol{\gamma}_{j+\frac{1}{2}}, U_{j+1}^n, U_j^n\big)\ ,\ j \in \mathbb{Z}\ ,\ n = 1,2,3,\ldots$$

with $\lambda_j := \Delta t/(S_j \Delta x)$, $\Delta_- V_j := V_j - V_{j-1}$, and the Engquist-Osher flux

$$f^{\mathrm{EO}}(\boldsymbol{\gamma},v,u) := \frac{1}{2}\left[f(\boldsymbol{\gamma},u) + f(\boldsymbol{\gamma},v) - \int_u^v |f_u(\boldsymbol{\gamma},w)|\mathrm{d}w\right]. \tag{36}$$

By staggering the grid, the appearance of 2×2-Riemann problems is avoided, which otherwise would emerge at each cell boundary. By the nu-

merical flux function (36) a so-called upwind scheme is defined, i.e. the differences in the scheme are directed towards incoming information. This allows the representation of shocks of the exact solution without major smearing. Figure 7 shows a simulation for the container which is sketched in Fig. 2 (a). In this example an initially empty (only filled by water) container is filled with constant feed rate. At the end of the simulation the operation becomes stationary. The choice of the EO-flux is also motivated by its close relationship to, on the one hand, the so-called Kružkov entropy flux $F(\boldsymbol{\gamma}, u) := \operatorname{sgn}(u - c)(f(\boldsymbol{\gamma}, u) - f(\boldsymbol{\gamma}, c))$ [91], and on the other hand to the so-called Temple functional [110], which enables the convergence proof for the scheme, see Sect. 3.4.

For the case of *variable* container cross section in [39] the convergence of the difference scheme towards a weak solution is shown. Numerical computations are delivered in [38]. For the case of *constant* cross section in [41], a new entropy concept is introduced, which is based on the generalized BV space BV_t (only the weak time derivative of the generalized solution is contained in BV). For solutions in this adapted solution space, an entropy inequality with Kružkov entropy functions and entropy fluxes is employed and it is shown that this entropy solution depends continuously on the initial data. This global result is new. Moreover, it is shown that the stated numerical scheme converges to the entropy solution, if the discretization parameters $\Delta t, \Delta x$ converge to 0 and a CFL stability condition is satisfied. The front tracking scheme and the difference scheme are recapitulated in [40].

For the clarifier-thickener model with constant container cross section, additionally a relaxation scheme was used in [37]. The basic idea of a relaxation scheme [85] is that the conservation law of interest $\partial_t u + \partial_x f(\boldsymbol{\gamma}(x), u) = 0$ (here with the discontinuity parameter $\boldsymbol{\gamma}$) is approximated by the system

$$\partial_t u^\tau + \partial_x v^\tau = 0, \quad \partial_t v^\tau + a^2 \partial_x u^\tau = \tau^{-1}\left[f\left(\boldsymbol{\gamma}(x), u^\tau\right) - v^\tau\right]$$

with linear flux terms and then to consider the limit case $\tau \to 0$. In [37] the convergence of this scheme to a weak solution could be shown.

5.3 Numerical Schemes for Systems of Conservation Laws

Paper [20] is a case study in which, for the first time, settling processes of polydisperse suspensions are simulated by solving the system of conservation equations by a modern shock capturing method (the Nessyahu-Tadmor method). In [29] the Kurganov-Tadmor central scheme for systems of conservation laws is applied to the simulation of polydisperse sedimentation processes, which includes the discretization of boundary conditions. Recently, mathematical models and numerical simulations for polydisperse suspensions were applied to centrifugation [6] and fluidization [12, 14]. They were also extended to include reaction terms and applied to model sedimentation biodetectors [106]. An overview of recent developments in polydisperse sedimentation models is given in [13].

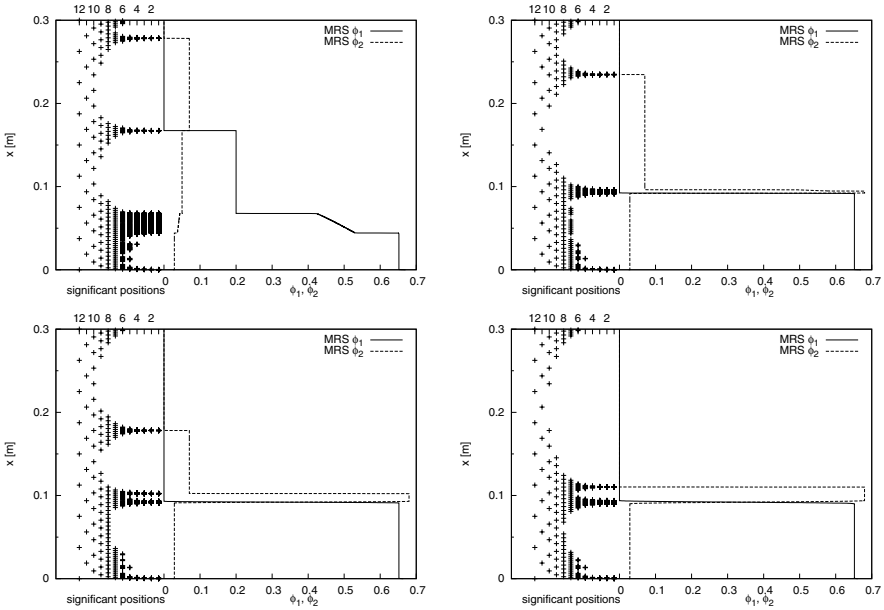


Fig. 8. Simulation of the settling of a bidisperse suspension with small (Species 1) and large (Species 2) equal-density spherical particles, showing concentration profiles and significant positions at four different times [47]

5.4 Multiresolution Methods

Multi-species kinematic flow models, such as the polydisperse sedimentation model, lead to strongly coupled, nonlinear systems of first-order, spatially one-dimensional conservation laws. The number of unknowns (the concentrations of the species) may be arbitrarily high. Models of this class also include a multi-species generalization of the Lighthill-Whitham-Richards traffic model. Their solutions typically involve kinematic shocks separating areas of constancy, and should be approximated by high resolution schemes. In [47] a fifth-order weighted essentially non-oscillatory (WENO) scheme is combined with a multiresolution wavelet technique that adaptively generates a sparse point representation (SPR) of the evolving numerical solution. Thus, computational effort is concentrated on zones of strong variation near shocks. Numerical examples of traffic and sedimentation models demonstrate the efficiency of the resulting WENO multiresolution (WENO-MRS) scheme. We show in Fig. 8 the simulation of the settling of a bidisperse suspension of equal-density particles, where ϕ_1 is the concentration of the larger and ϕ_2 that of the smaller particles ($\delta_2 = 0.0635$). The parameters in this case have been chosen in accordance with [109]. A similar multiresolution technique has

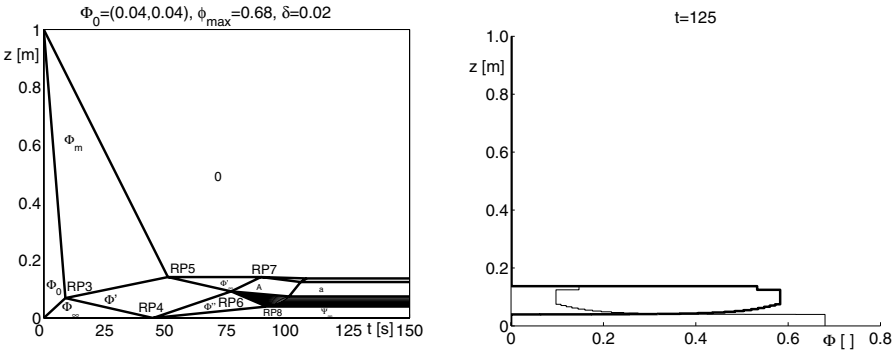


Fig. 9. Fronts (left) and stationary profile (right) for Moritomi's example, $\Phi_0 = (0.04, 0.04)$, $\phi_{\max} = 0.68$ [5]

also been applied to scalar degenerate parabolic equations modeling batch sedimentation or a diffusively corrected traffic model [48, 108].

5.5 Front Tracking for Systems

A method alternative to finite differences is the front tracking procedure [84]. The basis of the front tracking algorithm consists of the explicit determination of solutions of a family of Riemann problems, which are solved semi-analytically by the concatenation of elementary waves [5]. Since our systems are only piecewise genuinely nonlinear, one needs to employ the Liu entropy condition.

For a bidisperse suspension [109], the modes of sedimentation can be identified similar as for scalar equations. Using the output of the front tracking method, the global solution can be characterized by domains in the x-t-plane with either constant states or transitional fans. In practice, suspensions consisting of particles of different densities in addition to different sizes [14] are used to produce so-called functionally graded materials.

6 Simulation Software, Inverse Problems, Parameter Identification and Other Applications

The numerical solution of the various mathematical models can be employed for the simulation of solid-liquid separation processes and therefore for the design and control of equipment. Comparisons with experimental data performed in [17, 23, 78] confirm that batch and continuous settling processes of numerous real materials can be adequately described by a strongly degenerate convection-diffusion equation. The numerical methods have been implemented

in a user-friendly software package for the simulation of industrial thickeners [79, 80].

The sedimentation models studied so far rely on material specific flux density functions and diffusion coefficients. In practice, these functions need to be determined experimentally. This leads to the inverse problem of parameter identification for a strongly degenerate parabolic equation. This problem may be stated as an optimization problem for a suitably defined cost functional. The formal gradient of this cost functional is determined by the solution of an adjoint problem, which appears here as a backward linear parabolic equation. This approach is employed in [7, 8, 66] for the parameter identification from laboratory centrifuge data. Future applications of this methodology include the determination of the particle size distribution of polydisperse mixtures.

The new existence and uniqueness results for conservation equations with discontinuous flux and the convergence of a corresponding discretization are applied in [35] to establish well-posedness and to simulate numerically a model of traffic flow with driver reaction and abruptly changing road surface conditions. In the latter application, the governing equation assumes the form

$$\partial_t \rho + \partial_x \big(\gamma(x) f(\rho)\big) = \partial_x^2 D(\rho) \;, \quad x \in \mathbb{R} \;, \quad t > 0 \;, \tag{37}$$

where $\rho = \rho(x,t)$ is the local density of cars, measured in cars per mile, $\gamma(x)$ is a piecewise constant function describing the maximum velocity, which here depends on x, the function $f(\rho)$ is given by $f(\rho) = \rho V(\rho)$, where $V(\rho)$ is a hindrance factor with $V(0) = 1$, $V'(\rho) \le 0$ and $V(\rho_{\max}) = 0$ ($\rho_{\max}$ is a maximum car density), and the diffusion function $D(\rho)$ is given by

$$D(\rho) := \int_0^\rho d(s)\,\mathrm{d}s \;,$$

$$d(\rho) := \begin{cases} 0 & \text{if } \rho < \rho_{\mathrm{c}} \;, \\ -\rho v_{\max} V'(\rho)(L(\rho) + \tau v_{\max} \rho V'(\rho)) & \text{if } \rho > \rho_{\mathrm{c}} \;, \end{cases}$$

where $v_{\max}$ is a default maximum velocity, $L(\rho)$ is a density-dependent anticipation distance, τ is a reaction time, and ρ_{c} is a critical density beyond which the anticipation distance and the reaction time, which are both elements of driver psychology, enter into effect. (Equation (37) is the governing equation of the so-called diffusively corrected kinematic-wave traffic model (DCKWM).) If we use the Dick-Greenberg model defined by $V(\rho) = \min\{1, C \ln(\rho_{\max}/\rho)\}$ with the parameters $C = \mathrm{e}/7$ and $\rho_{\max} = 220\,\mathrm{cars/mi}$, we obtain $\rho_{\mathrm{c}} = 16.7512\,\mathrm{cars/mi}$. Figure 10 shows a numerical example from [35], in which we simulate the evolution of an initial traffic platoon given by

$$\rho(x,0) = \rho_0(x) := \begin{cases} 100\,\mathrm{cars/mi} & \text{for } x \in [-2\,\mathrm{mi}, 2\,\mathrm{mi}] \;, \\ 0 & \text{otherwise} \end{cases}$$

on an (infinite) road admitting the maximal velocity

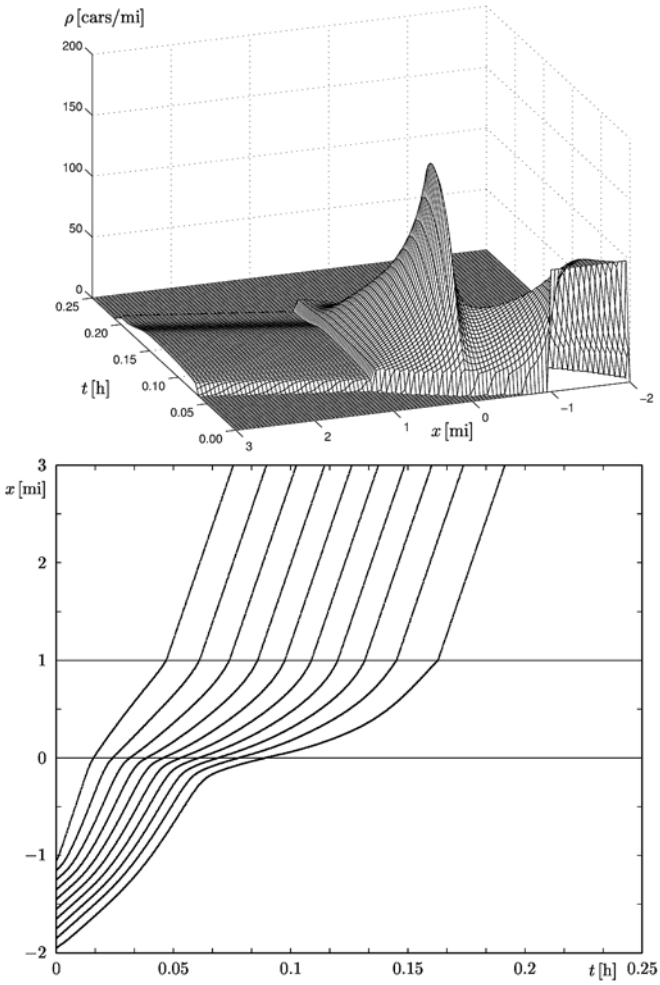


Fig. 10. Simulation of the traffic density on a road with an exceptional reduced maximum velocity on a finite interval [35]: car density (left) and car trajectories (right)

$$\gamma(x) = \begin{cases} 70\,\text{mph} & \text{for } x \le 0 \text{ or } x \ge 1\,\text{mi}\,, \\ 25\,\text{mph} & \text{for } 0 < x < 1\,\text{mi}\,. \end{cases}$$

Furthermore, a sub-case of conservation laws with discontinuous flux are transport equations with a discontinuous coefficient. Such an equation arises from a population balance model for the wear of steel balls in grinding mills used in mineral processing. This model is analyzed and simulated in [42].

7 Conclusions

The research directions we are interested to pursue in the near future include the following problems.

Up to now, the analysis has been focused on one-dimensional equations, where the priority is on the solids flow, while the equations for the fluid motion are subordinated. The one-dimensional results should be extended to the analysis of multi-dimensional coupled systems of equations of conservation and motion. The model reduction to one space dimension with discontinuous coefficients has to be taken into account. Numerical methods have to be developed and implemented for the coupling of the mass balance for the solids settling with the momentum balance for the fluid flow. Of particular interest are models for clarifier-thickeners, since they belong to the most common industrial equipment. For the analysis of entropy solutions of strongly degenerate convection-diffusion problems with initial and boundary conditions in several space dimensions (existence, uniqueness and stability) the application of the newly developed theory of divergence-measure fields is planned.

While in the past there was an emphasis on the treatment of monodisperse suspensions, there is a recently started and ongoing investigation of polydisperse suspensions. The goal is the characterization and determination of discontinuous solutions of systems of conservation laws modeling multispecies (polydisperse) sedimentation and multiphase flow in porous media via the solution of Riemann problems.

The analytical and numerical methods have already been extended to new applications as to traffic flow and grinding mills. There are further promising applications to flow in porous media and mathematical biology. The idea is to build on analogies in the model or equation structure. During the project period numerous cooperations with academics (in mathematics, science, civil, metallurgical or chemical engineering), could be established. This multidisciplinary context is going to be extended by utilizing the accumulated knowledge to industrial large scale problems.

Acknowledgments

This research has been supported by the German Research Foundation (DFG) within the Collaborative Research Center (SFB) 404 “Multifield Problems in Fluid and Solid Mechanics”, subproject A2.

Since March 2005, Raimund Bürger has been supported by Conicyt (Chile) through Fondecyt project 1050728 and Fondap in Applied Mathematics.

References

1. G. Anzellotti. Pairings between measures and functions and compensated compactness. *Ann. Mat. Pura Appl.*, 135:293–318, 1983.
2. C. Bardos, A. Y. Le Roux, and J. C. Nédélec. First order quasilinear equations with boundary conditions. *Comm. PDE*, 4:1017–1034, 1979.
3. G. K. Batchelor and R. Janse van Rensburg. Structure formation in bidisperse sedimentation. *J. Fluid Mech.*, 166:379–407, 1986.
4. J. Bell, J. Trangenstein, and G. Shubin. Conservation laws of mixed type describing three-phase flow in porous media. *SIAM J. Appl. Math.*, 46:1000–1017, 1986.
5. S. Berres. *Modeling, Analysis and Numerical Simulation of Polydisperse Suspensions.* Doctoral Thesis, University of Stuttgart, 2006.
6. S. Berres and R. Bürger. On gravity and centrifugal settling of polydisperse suspensions forming compressible sediments. *Int. J. Solids Structures*, 40:4965–4987, 2003.
7. S. Berres, R. Bürger, A. Coronel, and M. Sepúlveda. Numerical identification of parameters for a flocculated suspension from concentration measurements during batch centrifugation. *Chem. Eng. J.*, 111:91–103, 2005.
8. S. Berres, R. Bürger, A. Coronel, and M. Sepúlveda. Numerical identification of parameters for a strongly degenerate convection-diffusion problem modelling centrifugation of flocculated suspensions. *Appl. Numer. Math.*, 52:311–337, 2005.
9. S. Berres, R. Bürger, and H. Frid. Neumann problems for quasilinear parabolic systems modelling polydisperse suspensions. *SIAM J. Math. Anal.* To appear.
10. S. Berres, R. Bürger, and K. H. Karlsen. Central schemes and systems of conservation laws with discontinuous coefficients modeling gravity separation of polydisperse suspensions. *J. Comp. Appl. Math.*, 164-165:53–80, 2004.
11. S. Berres, R. Bürger, K. H. Karlsen, and E. M. Tory. Strongly degenerate parabolic-hyperbolic systems modeling polydisperse sedimentation with compression. *SIAM J. Appl. Math.*, 64:41–80, 2003.
12. S. Berres, R. Bürger, and E. M. Tory. Mathematical model and numerical simulation of the liquid fluidization of polydisperse solid particle mixtures. *Comput. Visual. Sci.*, 6:67–74, 2004.
13. S. Berres, R. Bürger, and E. M. Tory. Applications of polydisperse sedimentation models. *Chem. Eng. J.*, 111:105–117, 2005.
14. S. Berres, R. Bürger, and E. M. Tory. On mathematical models and numerical simulation of the fluidization of polydisperse suspensions. *Appl. Math. Modelling*, 29:159–193, 2005.
15. R. Bürger. *Ein Anfangs-Randwertproblem einer quasilinearen parabolischen entarteten Gleichung in der Theorie der Sedimentation mit Kompression.* Doctoral Thesis, University of Stuttgart, 1996.
16. R. Bürger. Phenomenological foundation and mathematical theory of sedimentation-consolidation processes. *Chem. Eng. J.*, 80:177–188, 2000.
17. R. Bürger, M. C. Bustos, and F. Concha. Settling velocities of particulate systems: 9. Phenomenological theory of sedimentation processes: Numerical simulation of the transient behaviour of flocculated suspensions in an ideal batch or continuous thickener. *Int. J. Mineral Process.*, 55:267–282, 1999.

18. R. Bürger and F. Concha. Mathematical model and numerical simulation of the settling of flocculated suspensions. *Int. J. Multiphase Flow*, 24:1005–1023, 1998.
19. R. Bürger and F. Concha. Settling velocities of particulate systems: 12. Batch centrifugation of flocculated suspensions. *Int. J. Mineral Process.*, 63:115–145, 2001.
20. R. Bürger, F. Concha, K.-K. Fjelde, and K. H. Karlsen. Numerical simulation of the settling of polydisperse suspensions of spheres. *Powder Technol.*, 113:30–54, 2000.
21. R. Bürger, F. Concha, and K. H. Karlsen. Phenomenological model of filtration processes: 1. Cake formation and expression. *Chem. Eng. Sci.*, 56:4537–4553, 2001.
22. R. Bürger, F. Concha, K. H. Karlsen, and A. Narváez. Numerical simulation of clarifier-thickener units treating ideal suspensions with a flux density function having two inflection points. *Math. Comp. Modelling*, 44:255–275, 2006.
23. R. Bürger, F. Concha, and F. M. Tiller. Applications of the phenomenological theory to several published experimental cases of sedimentation processes. *Chem. Eng. J.*, 80:105–117, 2000.
24. R. Bürger, A. Coronel, and M. Sepúlveda. On an upwind difference scheme for strongly degenerate parabolic equations modelling the settling of suspensions in centrifuges and non-cylindrical vessels. *Appl. Numer. Math.* To appear.
25. R. Bürger, A. Coronel, and M. Sepúlveda. A semi-implicit monotone difference scheme for an initial-boundary value problem of a strongly degenerate parabolic equation modelling sedimentation-consolidation processes. *Math. Comp.*, 75:91–112, 2006.
26. R. Bürger, J. J. R. Damasceno, and K. H. Karlsen. A mathematical model for batch and continuous thickening in vessels with varying cross section. *Int. J. Mineral Process.*, 73:183–208, 2004.
27. R. Bürger, S. Evje, and K. H. Karlsen. On strongly degenerating convection-diffusion problems modeling sedimentation-consolidation processes. *J. Math. Anal. Appl.*, 247:517–556, 2000.
28. R. Bürger, S. Evje, K. H. Karlsen, and K. A. Lie. Numerical methods for the simulation of the settling of flocculated suspensions. *Chem. Eng. J.*, 80:91–104, 2000.
29. R. Bürger, K.-K. Fjelde, K. Höfler, and K. H. Karlsen. Central difference solutions of the kinematic model of settling of polydisperse suspensions and three-dimensional particle-scale simulations. *J. Eng. Math.*, 41:167–187, 2001.
30. R. Bürger, H. Frid, and K. H. Karlsen. On the well-posedness of entropy solutions to conservation laws with a zero-flux boundary condition. *J. Math. Anal. Appl.* To appear.
31. R. Bürger, H. Frid, and K. H. Karlsen. On a free boundary problem for a strongly degenerate quasilinear parabolic equation with an application to a model of pressure filtration. *SIAM J. Math. Anal.*, 34:611–635, 2003.
32. R. Bürger, A. García, K. H. Karlsen, and J. D. Towers. A note on an extended clarifier-thickener model with singular source and sink terms. *Sci. Ser. A Math. Sci. (N.S.).* To appear.
33. R. Bürger, A. García, K. H. Karlsen, and J. D. Towers. On an extended clarifier-thickener model with singular source and sink terms. *Eur. J. Appl. Math.* To appear.

34. R. Bürger and K. H. Karlsen. On some upwind schemes for the phenomenological sedimentation-consolidation model. *J. Eng. Math.*, 41:145–166, 2001.
35. R. Bürger and K. H. Karlsen. On a diffusively corrected kinematic-wave traffic model with changing road surface conditions. *Math. Models Meth. Appl. Sci.*, 13:1767–1799, 2003.
36. R. Bürger, K. H. Karlsen, C. Klingenberg, and N. H. Risebro. A front tracking approach to a model of continuous sedimentation in ideal clarifier-thickener units. *Nonlin. Anal. Real World Appl.*, 4:457–481, 2003.
37. R. Bürger, K. H. Karlsen, and N. H. Risebro. A relaxation scheme for continuous sedimentation in ideal clarifier-thickener units. *Comput. Math. Applic.*, 50:993–1009, 2005.
38. R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers. On a model for continuous sedimentation in vessels with discontinuous cross-sectional area. In T. Y. Hou and E. Tadmor, editors, *Hyperbolic Problems: Theory, Numerics, Applications. Proceedings of the Ninth International Conference on Hyperbolic Problems (Pasadena, 2002)*, pages 397–406. Springer-Verlag, Berlin, 2003.
39. R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers. Monotone difference approximations for the simulation of clarifier-thickener units. *Comput. Visual. Sci.*, 6:83–91, 2004.
40. R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers. Numerical methods for the simulation of continuous sedimentation in ideal clarifier-thickener units. *Int. J. Mineral Process.*, 73:209–228, 2004.
41. R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers. Well-posedness in BV_t and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units. *Numer. Math.*, 97:25–65, 2004.
42. R. Bürger, K. H. Karlsen, and J. D. Towers. Closed-form and finite difference solutions to a population balance model of grinding mills. *J. Eng. Math.*, 51:165–195, 2005.
43. R. Bürger, K. H. Karlsen, and J. D. Towers. Mathematical model and numerical simulation of the dynamics of flocculated suspensions in clarifier-thickeners. *Chem. Eng. J.*, 111:119–134, 2005.
44. R. Bürger, K. H. Karlsen, and J. D. Towers. A mathematical model of clarifier-thickener units. *PAMM Proc. Appl. Math. Mech.*, 5:589–590, 2005.
45. R. Bürger, K. H. Karlsen, and J. D. Towers. A model of continuous sedimentation of flocculated suspensions in clarifier-thickener units. *SIAM J. Appl. Math.*, 65:882–940, 2005.
46. R. Bürger, K. H. Karlsen, W. L. Wendland, and E. M. Tory. Model equations and instability regions for the sedimentation of polydisperse suspensions of spheres. *ZAMM Z. Angew. Math. Mech.*, 82:699–722, 2002.
47. R. Bürger and A. Kozakevicius. Adaptive multiresolution WENO schemes for multi-species kinematic flow models. *J. Comp. Phys.* Submitted.
48. R. Bürger, A. Kozakevicius, and M. Sepúlveda. Multiresolution schemes for strongly degenerate parabolic equations. *Numer. Meth. Partial Diff. Eqns.* Submitted.
49. R. Bürger and M. Kunik. A critical look at the kinematic-wave theory for sedimentation-consolidation processes in closed vessels. *Math. Meth. Appl. Sci.*, 24:1257–1273, 2001.
50. R. Bürger, C. Liu, and W. L. Wendland. Existence and stability for mathematical models of sedimentation-consolidation processes in several space dimensions. *J. Math. Anal. Appl.*, 264:288–310, 2001.

51. R. Bürger and E. M. Tory. On upper rarefaction waves in batch settling. *Powder Technol.*, 108:74–87, 2000.
52. R. Bürger and W. L. Wendland. Entropy boundary and jump conditions in the theory of sedimentation with compression. *Math. Meth. Appl. Sci.*, 21:865–882, 1998.
53. R. Bürger and W. L. Wendland. Existence, uniqueness and stability of generalized solutions of an initial-boundary value problem for a degenerating parabolic equation. *J. Math. Anal. Appl.*, 218:207–239, 1998.
54. R. Bürger and W. L. Wendland. Sedimentation and suspension flows: Historical perspective and some recent developments. *J. Eng. Math.*, 41:101–116, 2001.
55. R. Bürger, W. L. Wendland, and F. Concha. A mathematical model for sedimentation-consolidation processes. *ZAMM Z. Angew. Math. Mech.*, 80:S177–S178, 2000.
56. R. Bürger, W. L. Wendland, and F. Concha. Model equations for gravitational sedimentation-consolidation processes. *ZAMM Z. Angew. Math. Mech.*, 80:79–92, 2000.
57. M. C. Bustos. *On the Existence and Determination of Discontinuous Solutions to Hyperbolic Conservation Laws in the Theory of Sedimentation.* Doctoral Thesis, TH Darmstadt, 1984.
58. M. C. Bustos, R. Bürger, F. Concha, and E. M. Tory. *Sedimentation and Thickening.* Kluwer Academic Publishers, Dordrecht, 1999.
59. M. C. Bustos, F. Concha, and W. L. Wendland. Global weak solutions to the problem of continuous sedimentation of an ideal suspension. *Math. Meth. Appl. Sci.*, 13:1–22, 1990.
60. M. C. Bustos, F. Paiva, and W. L. Wendland. Control of continuous sedimentation of ideal suspensions as an initial and boundary value problem. *Math. Meth. Appl. Sci.*, 12:533–548, 1990.
61. M. C. Bustos, F. Paiva, and W. L. Wendland. Entropy boundary conditions in the theory of sedimentation of ideal suspensions. *Math. Meth. Appl. Sci.*, 19:679–697, 1996.
62. J. Chancelier, M. Cohen de Lara, and F. Pacard. Analysis of a conservation PDE with discontinuous flux: a model of settler. *SIAM J. Appl. Math.*, 54:954–995, 1994.
63. G.-Q. Chen and H. Frid. Divergence-measure fields and hyperbolic conservation laws. *Arch. Rat. Mech. Anal.*, 147:89–118, 1999.
64. F. Concha and R. Bürger. A century of research in sedimentation and thickening. *KONA Powder and Particle*, 20:38–70, 2002.
65. F. Concha and R. Bürger. Thickening in the 20th century: A historical perspective. *Minerals & Metallurgical Process.*, 20:57–67, 2003.
66. A. Coronel. *Estudio de un Problema Inverso para una Ecuación Parabólica Degenerada con Aplicaciones a la Teoría de la Sedimentación.* PhD thesis, Universidad de Concepción, Chile, 2004.
67. R. H. Davis. Hydrodynamic diffusion of suspended particles: A symposium. *J. Fluid Mech.*, 310:325–335, 1996.
68. S. Diehl. A conservation law with point source and discontinuous flux function modelling continuous sedimentation. *SIAM J. Appl. Math.*, 56:388–419, 1996.
69. S. Diehl. On scalar conservation laws with point source and discontinuous flux function. *SIAM J. Math. Anal.*, 26:1425–1451, 1996.
70. F. Dubois and P. Le Floch. Boundary conditions for nonlinear hyperbolic systems of conservation laws. *J. Diff. Eqns.*, 71:93–122, 1988.

71. B. Engquist and S. Osher. Stable and entropy satisfying approximations for transonic flow calculations. *Math. Comp.*, 34:45–75, 1980.
72. S. Evje and K. H. Karlsen. Monotone difference approximation of BV solutions to degenerate convection-diffusion equations. *SIAM J. Numer. Anal.*, 37:1838–1860, 2000.
73. A. Fitt. Mixed systems of conservation laws in industrial mathematical modelling. *Surv. Math. Indust.*, 6:21–53, 1996.
74. K.-K. Fjelde. *Numerical Schemes for Complex Nonlinear Hyperbolic Systems of Equations.* PhD thesis, University of Bergen, Norway, 2000.
75. H. Frid. Existence and asymptotic behavior of measure-valued solutions for three-phase flow in porous media. *J. Math. Anal. Appl.*, 196:614–627, 1995.
76. H. Frid and I.-S. Liu. Oscillation waves in Riemann problems inside elliptic regions for conservation laws of mixed type. *Z. Angew. Math. Phys.*, 46:913–931, 1995.
77. P. Garrido. *Validación, Simulación y Determinación de Parámetros en un Proceso de Espesamiento.* PhD thesis, Universidad de Concepción, Chile, 2005.
78. P. Garrido, R. Bürger, and F. Concha. Settling velocities of particulate systems: 11. Comparison of the phenomenological sedimentation-consolidation model with published experimental results. *Int. J. Mineral Process.*, 60:213–227, 2000.
79. P. Garrido, R. Burgos, R. Bürger, and F. Concha. Software for the design and simulation of gravity thickeners. *Minerals Eng.*, 16:85–92, 2003.
80. P. Garrido, R. Burgos, R. Bürger, and F. Concha. Settling velocities of particulate systems: 13. Software for the batch and continuous sedimentation of flocculated suspensions. *Int. J. Mineral Process.*, 73:131–144, 2004.
81. P. Garrido, F. Concha, and R. Bürger. Settling velocities of particulate systems: 14. Unified model of sedimentation, centrifugation and filtration of flocculated suspensions. *Int. J. Mineral Process.*, 72:57–74, 2003.
82. T. Gimse and N. Risebro. Solution of the Cauchy problem for a conservation law with a discontinuous flux function. *SIAM J. Math. Anal.*, 23:635–648, 1992.
83. P. Grassmann and R. Straumann. Entstehen und Wandern von Unstetigkeiten der Feststoffkonzentration in Suspensionen. *Chem.-Ing.-Techn.*, 35:477–482, 1963.
84. H. Holden and N. Risebro. *Front Tracking for Conservation Laws.* Springer-Verlag, Berlin, 2002.
85. S. Jin and Z. Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Comm. Pure Appl. Math.*, 48:235–276, 1995.
86. K. H. Karlsen, C. Klingenberg, and N. H. Risebro. A relaxation scheme for conservation laws with a discontinuous coefficient. *Math. Comp.*, 73:1235–1259, 2004.
87. B. Keyfitz. A geometric theory of conservation laws which change type. *ZAMM Z. Angew. Math. Mech.*, 75:571–581, 1995.
88. R. Klausen and N. Risebro. Stability of conservation laws with discontinuous coefficients. *J. Diff. Eqns.*, 157:41–60, 1999.
89. C. Klingenberg and N. Risebro. Convex conservation laws with discontinuous coefficients. *Comm. PDE*, 20:1959–1990, 1995.
90. C. Klingenberg and N. Risebro. Stability of a resonant system of conservation laws modeling polymer flow with gravitation. *J. Diff. Eqns.*, 170:344–380, 2001.

91. S. Kružkov. First order quasilinear equations in several independent variables. *Math. USSR Sb.*, 10:217–243, 1970.
92. A. Kurganov and E. Tadmor. New high resolution central schemes for nonlinear conservation laws and convection-diffusion equations. *J. Comput. Phys.*, 160:241–282, 2000.
93. G. Kynch. A theory of sedimentation. *Trans. Faraday Soc.*, 48:166–176, 1952.
94. O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva. *Linear and Quasilinear Equations of Parabolic Type*. American Math. Soc., Providence, R.I., 1968.
95. O. Lev, E. Rubin, and M. Sheintuch. Steady state analysis of a continuous clarifier-thickener system. *AIChE J.*, 32:1516–1525, 1986.
96. L. Lin, B. Temple, and H. Wang. A comparison of convergence rates for Godunov's method and Glimm's method in resonant nonlinear systems of conservation laws. *SIAM J. Numer. Anal*, 32:824–840, 1995.
97. L. Lin, B. Temple, and H. Wang. Suppression of oscillations in Godunov's method for a resonant non-strictly hyperbolic system. *SIAM J. Numer. Anal.*, 32:841–864, 1995.
98. M. Lockett and K. Bassoon. Sedimentation of binary particle mixtures. *Powder Technol.*, 4:1–7, 1979.
99. A. Majda and R. Pego. Stable viscosity matrices for system of conservation laws. *J. Diff. Eqns.*, 56:229–262, 1985.
100. J. Málek, J. Nečas, M. Rokyta, and M. Ružička. *Weak and Measure-Valued Solutions to Evolutionary PDEs*. Chapman & Hall, London, UK, 1996.
101. C. Mascia, A. Porretta, and A. Terracina. Non-homogeneous Dirichlet problems for degenerate parabolic-hyperbolic equations. *Arch. Rat. Mech. Anal.*, 163:87–124, 2002.
102. J. Masliyah. Hindered settling in a multiple-species particle system. *Chem. Eng. Sci.*, 34:1166–1168, 1979.
103. A. Narváez. Simulación numérica de la sedimentación de suspensiones ideales en un espesador-clarificador. Master's thesis, Universidad de Concepción, Chile, 2004.
104. F. Otto. *Ein Randwertproblem für Erhaltungssätze*. Doctoral Thesis, Universität Bonn, 1993.
105. C. A. Petty. Continuous sedimentation of a suspension with a nonconvex flux law. *Chem. Eng. Sci.*, 30:1451–1458, 1975.
106. S. Qian, R. Bürger, and H. H. Bau. Analysis of sedimentation biodetectors. *Chem. Eng. Sci.*, 60:2585–2598, 2005.
107. J. F. Richardson and W. N. Zaki. Sedimentation and fluidization: Part I. *Trans. Instn. Chem. Engrs. (London)*, 32:35–53, 1954.
108. R. Ruiz. Métodos de multiresolución y su aplicación a un modelo de ingeniería. Master's thesis, Universidad de Concepción, Chile, 2005.
109. W. Schneider, G. Anestis, and U. Schaflinger. Sediment composition due to settling of particles of different sizes. *Int. J. Multiphase Flow*, 11:419–423, 1985.
110. B. Temple. Global solution of the Cauchy problem for a 2×2 non-strictly hyperbolic system of conservation laws. *Adv. Appl. Math.*, 3:335–375, 1982.
111. F. Tiller, N. Hsyung, and D. Cong. Role of porosity in filtration: XII. Filtration with sedimentation. *AIChE J.*, 41:1153–1164, 1995.
112. E. M. Tory. Stochastic sedimentation and hydrodynamic diffusion. *Chem. Eng. J.*, 80:81–89, 2000.

113. J. Towers. Convergence of a difference scheme for conservation laws with a discontinuous flux. *SIAM J. Numer. Anal.*, 38:681–698, 2000.
114. J. Towers. A difference scheme for conservation laws with a discontinuous flux—the nonconvex case. *SIAM J. Numer. Anal.*, 39:1197–1218, 2001.
115. A. Volpert. Generalized solutions of degenerate second-order quasilinear parabolic and elliptic equations. *Adv. Diff. Eqns*, 5:1453–1518, 2000.
116. A. I. Volpert. The spaces BV and quasilinear equations. *Math. USSR Sb.*, 2:225–267, 1967.
117. A. I. Volpert and S. I. Hudjaev. Cauchy's problem for degenerate second order quasilinear parabolic equations. *Math. USSR Sb.*, 7:365–387, 1969.
118. G. Wallis. A simplified one-dimensional representation off two-component vertical flow and its application to batch sedimentation. In *Proceedings of the Symposium on the Interaction Between Fluids and Particles, London, June 20–22, 1962*, pages 9–16. Instn. Chem. Engrs., London, 1962.
119. R. H. Weiland, Y. P. Fessas, and B. Ramarao. On instabilities arising during sedimentation of two-component mixtures of solids. *J. Fluid Mech.*, 142:383–389, 1984.
120. Z. Wu. A note on the first boundary value problem for quasilinear degenerate parabolic equations. *Acta Math. Sci*, 4:361–373, 1982.
121. Z. Wu. A boundary value problem for quasilinear degenerate parabolic equation. MRC Technical Summary Report 2484, University of Wisconsin, USA, 1983.
122. Z. Wu and J. Wang. Some results on quasilinear degenerate parabolic equations of second order. In *Proceedings of the 1980 Beijing Symposium on Differential Geometry and Differential Equations, Vol. 3*, pages 1593–1609. Science Press, Beijing, Gordon & Breach, Science Publishers Inc., New York, 1982.
123. Z. Wu and J. Yin. Some properties of functions in BV_x and their applications to the uniqueness of solutions for degenerate quasilinear parabolic equations. *Northeastern Math. J.*, 5:395–422, 1989.

Multiphase Processes in Porous Media*

Holger Class, Rainer Helmig, Jennifer Niessner, and Ulrich Ölmann

Institute of Hydraulic Engineering, Chair of Hydromechanics and Modeling of Hydrosystems, University of Stuttgart, Pfaffenwaldring 61, 70569 Stuttgart, Germany
holger.class@iws.uni-stuttgart.de
rainer.helmig@iws.uni-stuttgart.de
jennifer.niessner@iws.uni-stuttgart.de
ulrich.oelmann@iws.uni-stuttgart.de

Summary. Models for multiphase flow in porous media are widespread today and can be found in many places in science and engineering. More complex multiphase-multicomponent models that even allow phase changes to occur need sophisticated numerical algorithms. Research in this area has been very successful with a versatile result.

Another challenge are simulations on the field scale. Here the idea of upscaling is a very promising concept. In these models the necessary amount of details is limited while they still preserve the ability to forecast the interesting information. New effects arise like the direction-dependence of permeabilities.

For the latter a new mathematical description and new numerical fluxes with new properties have been constructed. The fluxes are based upon two- and multi-point flux approximations. To perform tests a heterogeneous Buckley-Leverett-problem has been set up and solved quasi-analytically using the method of characteristics. The utilizability of the numerical fluxes is then demonstrated by application to a test-problem.

Unfortunately the concept of upscaling cannot be applied in general. There are problem classes where processes from the finer scale have to be integrated in more detail by solving local subproblems on that scale. This multi-scale approach is the future orientation of the project.

Keywords: Multiphase flow in porous media, upscaling, MPFA, anisotropic relative permeabilities, multi-scale modeling

1 Introduction

The flow of multiple fluid phases inside a porous medium is a complex process with special challenges for scientists and practical engineers. In the following,

* Research Project A3 "Multiphase Processes in Porous Media"

a rough overview of the topic itself as well as a short introduction to the special domain of interest of this research project are given.

Although the discipline of multiphase flow processes in porous media is comparatively young, it has a wide application area. The motivation for pressing on with the theoretical description originally came from the simulation of groundwater flow in order to find answers to questions connected with the recovery of drinking water. A similar task which is also classically of great importance is the extraction of oil and natural gas from fossil deposits.

These traditional fields come back into play with current questions from environmental engineering. One example is the problem of groundwater contamination by man-made pollutants. Here, there is a strong demand for models capable of forecasting where and how these substances will spread within aquifers. This is strongly connected to the simulation of cost intensive remediation techniques aiming at their optimization. Another area of current research is CO_2 sequestration which originates in considerations in the field of climate protection: instead of releasing carbon dioxide into the atmosphere where it contributes to the greenhouse effect, the concept is to store it within exploited deposits in the subsurface.

There is also a large number of challenging industrial applications for this kind of model, e. g. the processes that occur inside the cathode of a fuel cell or fluids flowing through a fill of a catalyzer in chemical engineering. These applications lead to multiphase or multiphase-multicomponent processes which might be non-isothermal in porous media. Porous media represent in nature a complex structure which shows multi-scale properties, see e.g. [26].

When upscaling methods are applied to porous media with distinctive anisotropically structured block heterogeneities, a pronounced direction-dependence of the relative permeability on the macroscale results. This was noticed as a new effect of multiphase flow in porous media on a higher scale, but there was no general precise theoretical characterization of this phenomenon leading to an extended mathematical model. Only if the main axes of this upscaled relative permeability coincide with the directions of a Cartesian computational grid parallel to the x- and y-direction both – the extension of the model and of the numerical algorithm – are quite obvious as this constraint leads to a diagonal tensor. It is therefore only for these special cases that numerical simulations could be realized.

At first, the challenge thus lies in the development of a concise and sound theoretical description of this new effect. Based upon that a new numerical technique has to be constructed that is on the one hand able to cope with almost arbitrary anisotropies of the relative permeabilities on any unstructured grid. On the other hand, its concept should simultaneously integrate the numerical idea of upwinding, which is known to be of great importance for the simulation of convection-dominated problems.

At the outset of the work on this topic, no such numerical tool existed. The method of multi-point flux approximation (MPFA) developed in the Nineties

at the same time, yet independently, by Ivar Aavatsmark [2] and Michael G. Edwards [17] appeared most promising as a good starting point.

2 Mathematical Model for Two-Phase Flow and Its Numerical Solution

To give an overview of the mathematics applied in this field, we take a look at the simple case of two immiscible fluids that flow in a porous medium. For a more detailed deduction, consult the publications [24, 18, 12, 14, 8] of Rainer Helmig and his workgroup or the technical literature [6, 32]. As one would expect the model that is used basically consists of two scalar balance equations for the mass of the fluids, two vector-valued balance equations for the momentum of both phases and, if exchange of energy is of concern, another scalar equation that expresses the conservation of energy. The latter is not of special interest here and is therefore disregarded for the time being. It will be essential, however, in the context of Sect. 3.

The path from physics to these equations is arduous, as it comprises volume averaging of multiphase Navier-Stokes equations on the pore scale, see [41]. In the case of the momentum balance these equations can be simplified under reasonable assumptions to Darcy's law, which is known from empirical examinations:

$$\mathbf{v}_\alpha = -\frac{k_{\mathrm{r}\alpha}}{\mu_\alpha}\mathsf{K}\left(\nabla p_\alpha - \rho_\alpha \mathbf{g}\right) \ , \quad \alpha \in \{\mathrm{w}, \mathrm{n}\} \ . \tag{1}$$

Here, the index α denotes the wetting or the non-wetting phase, $\mathbf{v}_\alpha$ is the Darcy velocity, $k_{\mathrm{r}\alpha}$ the relative permeability, μ_α the viscosity, K the symmetric and positive definite absolute permeability, ρ_α the density and $\mathbf{g}$ denotes the gravitational acceleration.

Taking note of the porosity Φ and the phase saturation S_α which arise by the process of volume averaging, the mass balance equations for the phases appear in differential form as

$$\frac{\partial}{\partial t}\Phi\rho_\alpha S_\alpha = -\nabla\cdot\left(\rho_\alpha \mathbf{v}_\alpha\right) + \rho_\alpha q_\alpha \ , \quad \alpha \in \{\mathrm{w}, \mathrm{n}\} \ , \tag{2}$$

where q_α names the volume source for phase α. Equations (1) and (2) as a whole are still under-determined, further statements are needed. This additional information completely consists exclusively of algebraic equations: the pores are entirely filled with both phases

$$S_\mathrm{w} + S_\mathrm{n} = 1 \ , \tag{3}$$

phase pressures are not independent but related by the capillary pressure

$$p_\mathrm{c} = p_\mathrm{n} - p_\mathrm{w} \tag{4}$$

and the capillary pressure as well as the relative permeabilities are modeled as functions of the saturation, which are given in terms of nonlinear constitutive relations

$$p_{\mathrm{c}} := p_{\mathrm{c}}\left(S_{\mathrm{w}}\right) \ , \tag{5}$$

$$k_{\mathrm{r}\alpha} := k_{\mathrm{r}\alpha}\left(S_{\mathrm{w}}\right) \ , \quad \alpha \in \{\mathrm{w}, \mathrm{n}\} \ . \tag{6}$$

Equations (1)–(6) are two vector valued and seven scalar equations for the two unknown vector fields $\mathbf{v}_\alpha$ and the seven unknown scalar fields p_{c}, $k_{\mathrm{r}\alpha}$, p_α and S_α, $\alpha \in \{\mathrm{w}, \mathrm{n}\}$. Therefore, the system is now closed and, because of the two partial differential equations (2), it has only to be completed by two boundary conditions for every boundary point to make the solution unique. These boundary conditions can in general either be Dirichlet-type conditions for the scalar unknowns or Neumann conditions in the form of prescribed Darcy fluxes.

The most trivial solution attempt would be to solve the nonlinear system (1)–(6) as a whole. But a more clever approach consists of decoupling the algebraic equations from the differential equations. This is done by choosing two of the scalar unknowns and rearranging (3)–(6) so that the other scalar unknowns are explicitly given in terms of the chosen two. By putting these expressions into (1) and (2) in combination with inserting (1) directly into (2), one ends up with a strongly coupled nonlinear system of two partial differential equations for two scalar unknowns. These unknowns are termed *primary variables* as they can be obtained in a primary step by solving this reduced system of partial differential equations. Utilizing this solution in a second step in the prepared expressions for the remaining unknowns, the *secondary variables* can be computed. There is more than one choice of primary variables and therefore the resulting decoupled system, called a *formulation*, is only one of a few possibilities. Of course, all formulations are mathematically equivalent. But they tend to show different behavior in terms of robustness and convergence when solved numerically.

Discretization of the system of partial differential equations in space is carried out using the finite volume (FV) method. To construct a FV-mesh at

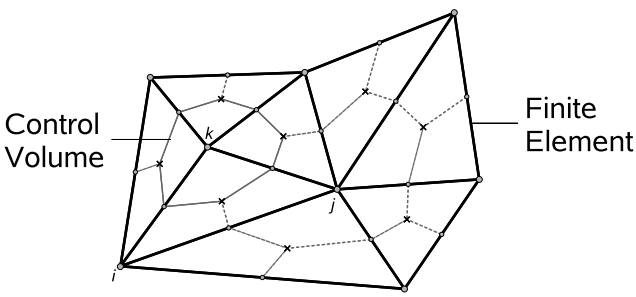


Fig. 1. Finite element mesh and dual mesh constructed to be used as control volumes

first the computational domain is discretized with a finite element (FE) grid consisting of arbitrary miscellaneous element types. On this, a dual mesh of control volumes (or cells) is constructed by connecting element barycenters with edge barycenters, see Fig. 1. The unknown scalar primary variables are spatially discretized on the corners of the finite element mesh. Inside an element, ordinary finite element interpolations are used for approximating the exact solution. All volume and surface integrals that occur using the finite volume method are calculated numerically by a midpoint rule. For the approximation of the fluxes between the cells (cf. (1)), the gradient of the FE interpolations of phase pressure in combination with harmonic averaging of the absolute permeability is used. To stabilize the resulting numerical scheme in the case of convection-dominated problems, an upwinding concept is introduced by taking the relative permeability from the upstream node.

Time discretization of the resulting system of ordinary differential equations can be done in various ways, whereas the strong nonlinear character of the system demands an implicit time-stepping scheme. Often, an implicit Euler scheme is used, but other time discretizations with different properties like Crank-Nicolson or a diagonally implicit Runge-Kutta method are used as well.

The outcome of this procedure is a coupled, strongly nonlinear algebraic system of equations that has to be solved for every time step. To handle this, Newton's method with a line search strategy is utilized. In close combination with that, control of the step length for the time discretization takes place: if the defect does not get better with regard to a specifically tailored criterion from one line search to the next, the actual step length is discarded and the time step is halved instead. On the other hand, if convergence inside Newton's method gets better again, the step size grows.

Last but not least, the linear system that arises in every Newton step has to be addressed. For this, a stabilized preconditioned biconjugate gradient method is applied. It turned out that the solution is gained very effectively if preconditioning is performed via a multigrid method based on a V-cycle and an incomplete LU-decomposition for pre- and postsmoothing (cf. [9]).

The realization of the specified algorithm in cooperation with the former SFB 404 project C4 leads to the simulator MUFTE-UG. It is constructed on top of the powerful framework UG (see [7]) which offers the necessary infrastructure for the adoption of parallel adaptive multigrid methods on unstructured grids.

The previous approach only permitted Dirichlet-type boundary conditions for the primary variables. As a consequence, the spectrum of treatable problems clearly suffered. One possible way out lies in dynamically choosing the primary variables and with them the whole formulation in different parts of the computational domain. This formulation adaptivity was supposed to address the problem of varying numerical robustness of the different formulations as well.

An alternative way out, which became apparent while working on the formulation adaptivity, was to face the lack of robustness by well considered regularizations of the problematic constitutive relations. In combination with a new concept for the treatment of Dirichlet-type boundary conditions that resembles the idea of getting to a formulation, this would be a remedy as well.

As the latter alternative is clearly less complex and hence a lot less error-prone, the idea of formulation adaptivity has been rejected in favor of this new extended boundary condition concept. This concept has been successfully worked out and implemented in MUFTE-UG within the final project period. It is examined and evaluated in [23].

3 Non-isothermal Multiphase-Multicomponent Systems

Modeling nonisothermal compositional multiphase flow and transport processes in porous media requires the consideration of the transfer of mass and energy between the phases in addition to the flow processes such as advection and diffusion. Therefore, it is a significant and important extension of the model of Sect. 2. Such processes occur in several fields of environmental engineering and reservoir engineering in the subsurface. While first applications for multicomponent models in porous media appeared in the oil industry in the 1970s [15], these concepts have subsequently been transfered, adapted and extended to issues of groundwater protection and remediation [19, 14, 13, 20] as well as to the storage of CO_2 in deep geological formations [10, 33, 34]. Recently, multiphase multicomponent models have also been developed for technical applications on a much smaller scale for the simulation of the water management in the cathode diffusion layer of fuel cells [4].

Characteristic of compositional multiphase models is that the phases are not only matter of a single chemical substance. Instead, their composition in general includes several species. In the following, we give some basic definitions and assumptions that are required for the formulation of the model concept below. As an example, we take a three-phase three-component system water-NAPL-gas [14]. The modification for other multicomponent systems is straightforward and can be found, e. g., in [10, 27, 4]. Discretization of the resulting partial differential equations is done analogously to the procedure of Sect. 2.

3.1 Basic Definitions and Assumptions for the Compositional Model Concept

Components: The term *component* stands for constituents of the phases which can be associated with a unique chemical species. In this work, we simplify air to a single pseudo-component and neglect its composition of N_2, O_2, and CO_2. We assume that our water-gas-NAPL systems are composed of the phases water (subscript w), gas (g), and NAPL (n). These phases are

composed of the components water (superscript w), air (a), and the organic contaminant (c) (see Fig. 2).

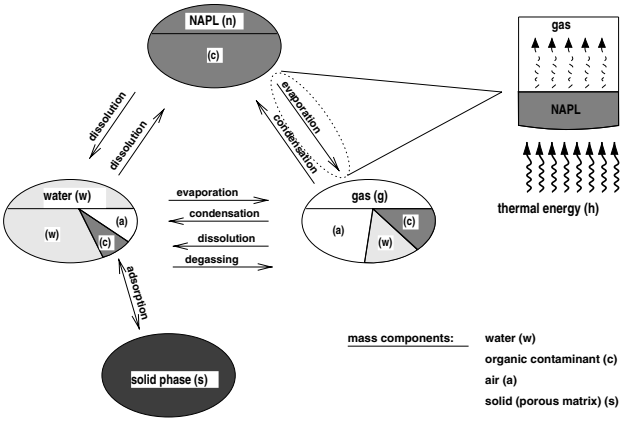


Fig. 2. Mass and energy transfer between the phases

Thermal, chemical, and mechanical equilibrium: For the nonisothermal multiphase processes in porous media under consideration, we state that the assumption of local thermal equilibrium is valid since flow velocities are small. We neglect chemical reactions and biological decomposition and assume chemical equilibrium. Mechanical equilibrium is not valid in a porous medium since discontinuities in pressure can occur across a fluid-fluid interface due to capillary effects.

3.2 Additional Model Parameters

The mathematical description of compositional effects in multiphase flow requires the quantification of the amounts of specific components in the locally existing fluid phases. Typically, this is achieved by introducing *mole fractions* x_α^κ or *mass fractions* X_α^κ of component κ in phase α. Their calculation is basically different for low-pressure gaseous and liquid phases or gas phases at near-critical pressures, as in the case of CO_2 sequestration, cf. [10]. The characteristic gas phase in the shallow unsaturated zone is at a pressure condition in the order of magnitude of atmospheric pressure and thus far below critical pressure for all of the components in the gas phase. Thus, the assumption of ideal gases is reasonable. This allows us to apply the ideal gas law for the computation of the molar gas phase density by

$$\varrho_{\mathrm{mol,g}} = \frac{p_\mathrm{g}}{\mathcal{R}T} \tag{7}$$

with $\mathcal{R} = 8.314\,\mathrm{J/(mol\,K)}$ representing the universal gas constant and furthermore the use of Dalton's law for expressing the pressure of the gas phase

as the sum of the partial pressure of the components

$$p_{\mathrm{g}} = \sum_{\kappa} p_{\mathrm{g}}^{\kappa} \,. \tag{8}$$

The mole fractions in the gaseous phase can then be obtained by

$$x_{\mathrm{g}}^{\kappa} = \frac{p_{\mathrm{g}}^{\kappa}}{p_{\mathrm{g}}} \,, \qquad \kappa \in \{\mathrm{w,c}\} \,. \tag{9}$$

The computation of the partial pressures of the gaseous components depends on the co-existence of the corresponding liquid phase. For example, the partial gas phase pressure of water vapor is equal to the saturation vapor pressure if the liquid water is present at the same time. Otherwise, it is lower than the saturation vapor and it is necessary to use $x_{\mathrm{g}}^{\mathrm{w}}$ as a primary variable (see below). For the computation of mole fractions in the liquid phases, it is necessary to have equations of state dependent on temperature and pressures. In cases of low soluble components, it is in general appropriate to apply Henry's law

$$x_{\mathrm{w}}^{\kappa} = \frac{p_{\mathrm{g}}^{\kappa}}{H_{\mathrm{w}}^{\kappa}} \,, \qquad \kappa \in \{\mathrm{a,c}\} \tag{10}$$

with H_{w}^{κ} representing the Henry coefficient.

Since it is convenient to formulate the continuity equation on the basis of moles instead of kilograms (see (15)), one has to distinguish between the *molar density* ϱ_{mol} and the *mass density* ϱ_{mass} of a fluid.

As it is characteristic of mass transfer like evaporation or condensation that heat exchange is a relevant phenomenon, one has to include the caloric state variables *specific internal energy* u and *specific enthalpy* h in the set of model parameters. Both are related via

$$h_{\alpha} = u_{\alpha} + p_{\alpha} v_{\alpha} = u_{\alpha} + \frac{p_{\alpha}}{\varrho_{\mathrm{mass},\alpha}} \,, \tag{11}$$

the volume-changing work making the difference between h and u. While this can be neglected in many cases for low-compressible liquid phases, it is in general significant for gases. Values for h and u can be found as functions of temperature and pressure in the literature, cf. [28] for water or [38] for a large number of organic and other substances.

Another thermal phenomenon is the conductive transport of heat. The heat flow due to conduction across the boundary Γ of a given control volume can be expressed by

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \int_{\Gamma} \lambda \nabla T \, \mathrm{d}\Gamma \,. \tag{12}$$

Since we postulated above that thermal equilibrium should hold, we use here an averaged *heat conductivity* λ_{pm} for the fluid-filled porous medium.

Furthermore, we need to quantify the amount of stored thermal energy in the porous matrix which can be achieved by considering the *specific heat capacity* c_s of the soil grains.

Apart from additional model parameters that we have to consider in the case of a multicomponent system, there is also a process that only occurs in such systems, namely *diffusion*. In general, there are complex mechanisms interacting that cause a diffusive/dispersive spreading of components. The influence of molecular diffusion in comparison to dispersive mixing decreases with increasing phase velocities. Mechanical dispersion is a complex phenomenon and depends not only on the velocity but also strongly on the degree of heterogeneity. In the following, for the sake of simplicity, we use only scalar diffusion coefficients to quantify diffusivity. Thus, a diffusive flux of component κ in phase α is given by

$$J_\alpha^\kappa = -\varrho_{\mathrm{mol},\alpha}\, D_{\mathrm{pm}}^\kappa\, \nabla x_\alpha^\kappa \;, \tag{13}$$

where

$$D_{\mathrm{pm}}^\kappa = \tau\phi S_\alpha D_\alpha^\kappa \;. \tag{14}$$

τ stands for the tortuosity of the porous medium. The diffusivity D_α^κ is specific for each component.

3.3 Balance Equations

Analogously to the derivation of (2), we formulate the balance equations for multicomponent systems. In many cases, it is convenient to use a molar formulation of the continuity equation. Considering the mass conservation for each component allows us to drop source/sink terms for describing the mass transfer between phases since it is included inherently in the formulation. Then, the mass balance (molar) can be written as:

$$\begin{aligned}
&\phi\frac{\partial(\sum_\alpha \varrho_{\mathrm{mol},\alpha} x_\alpha^\kappa S_\alpha)}{\partial t} \\
&-\sum_\alpha \nabla\cdot\left\{\frac{k_{\mathrm{r}\alpha}}{\mu_\alpha}\varrho_{\mathrm{mol},\alpha} x_\alpha^\kappa \mathbf{K}(\nabla p_\alpha - \varrho_{\mathrm{mass},\alpha}\mathbf{g})\right\} \\
&-\nabla\cdot\left\{D_{\mathrm{pm}}^\kappa \varrho_{\mathrm{mol},\alpha}\nabla x_\alpha^\kappa\right\} \\
&-q^\kappa = 0 \qquad \kappa\in\{\mathrm{w,a,c}\},\ \alpha\in\{\mathrm{w,n,g}\}\;.
\end{aligned} \tag{15}$$

In the case of non-isothermal systems, we further have to balance the thermal energy. This commonly occurs on the basis of some simplifying assumptions. The most important one is that of local thermal equilibrium, which allows us to formulate a single heat balance for the fluid-filled porous medium. Other simplifications concern the neglection of dissipative effects, the heat

transport due to molecular diffusion, or the assumption of fully reversible processes, which violates the fact that entropy has to increase. In this case, entropy is not needed as a model parameter. The heat balance can then be formulated as:

$$\phi \frac{\partial \left(\sum_\alpha \varrho_{\mathrm{mass},\alpha} u_\alpha S_\alpha \right)}{\partial t} + (1-\phi) \frac{\partial \varrho_\mathrm{s} c_\mathrm{s} T}{\partial t}$$

$$-\nabla \cdot (\lambda_\mathrm{pm} \nabla T)$$

$$-\sum_\alpha \nabla \cdot \left\{ \frac{k_{\mathrm{r}\alpha}}{\mu_\alpha} \varrho_{\mathrm{mass},\alpha} h_\alpha \mathbf{K} \left(\nabla p_\alpha - \varrho_{\mathrm{mass},\alpha} \mathbf{g} \right) \right\}$$

$$-q^h \;=\; 0 \qquad \kappa \in \{\mathrm{w,a,c}\},\ \alpha \in \{\mathrm{w,n,g}\}\ . \tag{16}$$

Supplementary constraints for capillary pressure, saturations and mole fractions are valid analogously to (3) and (4).

3.4 Adaptive Choice of Primary Variables

According to the Gibbsian phase rule, the number of degrees of freedom in a non-isothermal multiphase multicomponent system is equal to the number of components plus one. This means we need as many independent unknowns in the mathematical system description (balance equations plus supplementary constraints plus constitutive relationships). The available primary variables are, e. g., saturations, mole/mass fractions, temperature, pressures, etc.

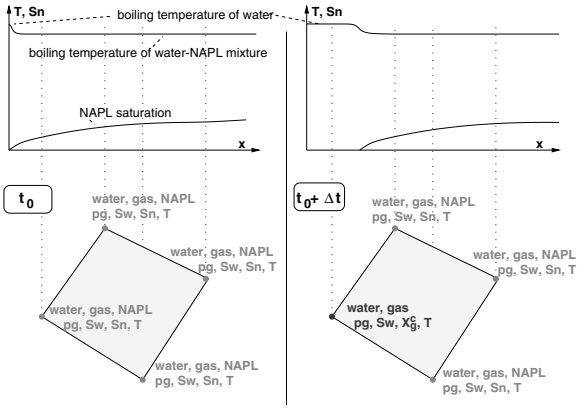


Fig. 3. Process-adaptive substitution of primary variables after a change of the phase state due to complete local evaporation of the liquid NAPL phase

Table 1. Phase states, corresponding primary variables, and criteria for the substitution in the case of phase appearance

phase state	present phases	primary variables	appearance of phase: water	appearance of phase: NAPL	appearance of phase: Gas
1	w,n,g	S_W, S_n, p_g, T	-	-	-
2	w	x^c_w, x^a_w, p_g, T	-	$x^\mathrm{c}_\mathrm{w} > \chi^\mathrm{c}_\mathrm{w}$	$p^\mathrm{w}_\mathrm{sat} + H^\mathrm{a}_\mathrm{w} x^\mathrm{a}_\mathrm{w} > p_\mathrm{g}$
3	n,g	S_n, x^w_g, p_g, T	$x^\mathrm{w}_\mathrm{g} p_\mathrm{g} > p^\mathrm{w}_\mathrm{sat}$	-	-
4	w,n	S_n, x^a_w, p_g, T	-	-	$H^\mathrm{a}_\mathrm{w} x^\mathrm{a}_\mathrm{w} + \sum_K p^K_\mathrm{sat} > p_\mathrm{g}$
5	g	x^c_g, x^w_g, p_g, T	$x^\mathrm{w}_\mathrm{g} p_\mathrm{g} > p^\mathrm{w}_\mathrm{sat}$	$x^\mathrm{c}_\mathrm{g} p_\mathrm{g} > p^\mathrm{c}_\mathrm{sat}$	-
6	w,g	x^c_g, S_W, p_g, T	-	$x^\mathrm{c}_\mathrm{g} p_\mathrm{g} > p^\mathrm{c}_\mathrm{sat}$	-

For the following explanations, we choose a non-isothermal water-NAPL-gas system to simulate steam-injection and remediation of a NAPL-contaminated unsaturated zone. If all three phases exist locally, it is possible to choose two of the phase saturations (the third is then trivial) together with one of the pressures (the others are coupled via capillary pressure-saturation functions) and the temperature. However, this holds only until one of the phases disappears. This is illustrated in Fig. 3, where the NAPL phase is fully evaporated at one of the nodes. Thus, the saturation of NAPL is also trivial and a new parameter is required as the primary unknown for balancing the remaining NAPL content as vapor in the gas phase. For this reason, we choose the mole fraction of NAPL vapor in the gas phase.

Table 1 summarizes the possible phase states, the corresponding sets of primary variables, and criteria for the indication of the (re-)appearance of fluid phases. The disappearance of phases is simply recognized by negative values of the respective saturations.

The concept of adaptive phase states and primary variables is embedded in the solution process, as the algorithm in Fig. 4 illustrates.

4 Upscaling

Effective parameters for the macroscale are required for these simulations in order to avoid a detailed discretization of the heterogeneous structure. Starting from the observed influence of heterogeneities on two-phase flow processes on the macroscale (see Sect. 2), we show an upscaling procedure from the local to the macroscale for the derivation of constitutive relationships for multiphase flow processes. Here, the localscale corresponds to an averaged scale, where porescale properties are averaged over a so-called representative elemenary volume, while the macroscale is a scale larger then the localscale scale and represents the scale, one would like to model on. The approach is based on the assumption of an equilibrium of (capillary) forces, which allows the application of a percolation model. This results in saturation distributions for different capillary pressures. Averaging these distributions gives rise to a macroscopic

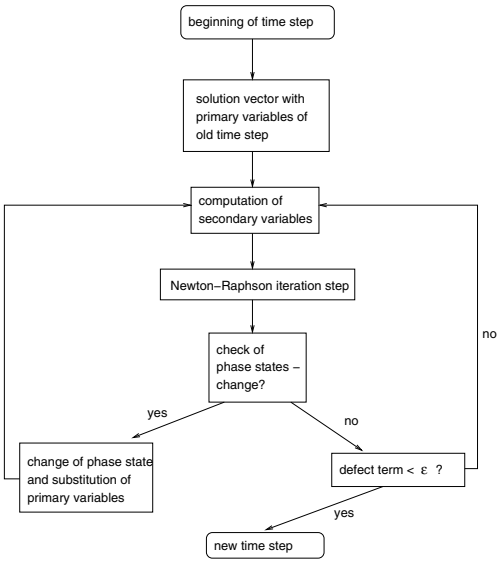


Fig. 4. Algorithm for the substitution of the primary variables during the computation of a time step

capillary pressure-saturation-relationship. For the saturation distribution relative permeabilities and effective conductivities are computed depending on the structure and the flow direction. These are averaged with the help of the renormalization method. The evolving relative permeability-saturation-relationship for the macroscale shows a saturation-dependent anisotropy and pronounced residual saturations of the non-wetting phase (which were not assumed for the local scale). The anisotropy reflects the underlying structure of the considered system which needs not to be known in detail [11].

4.1 Description of the Upscaling Approach

In this section, we describe the different methods that comprise the proposed upscaling approach used to derive effective parameters on the macroscale for the simulation of two-phase flow processes.

Starting from the local scale we have to determine the distribution of all material parameters and constitutive relationships (see Fig. 5). Applying the capillary equilibrium assumption to a distribution of local P_c-S_w relationships we can determine the saturation distribution for a given capillary pressure. Doing so, we compute the saturation distribution with a quasi-static displacement site-percolation model, cf. [39]. The weighted arithmetic mean of the saturation distribution gives one point on the macroscopic capillary pressure-saturation-relationship. We neglect hysteresis effects for the simple geometries, only the drainage curve is computed in these cases.

In all the above-quoted examples gravity effects are neglected for the averaging procedure of the saturation distribution.

For both the wetting and the non-wetting phase the conductivity-distribution can be computed for the aforementioned saturation distribution and thus relative permeability-saturation distribution. Averaging this distribution leads to a point of the macroscopic relative permeability-saturation relationship. If we assume capillary equilibrium we can average the multiphase conductivities with the same techniques known from single-phase flow systems. In the context of this paper, we use the renormalization method. Renormalization methods are a compromise between numerical efficiency and physically interpretable results. They treat easy-to-handle subdomains of the domain [29, 42]. Of course also other methods can be used, as, e. g., described by Wen and Gómez-Hernández [40].

Cycling through this procedure with different capillary pressures as boundary conditions one can determine macroscopic capillary pressure-saturation and relative permeability-saturation relationships.

4.2 Derivation of Constitutive Relationships for Two-Phase Flow in a Perfectly Layered Medium Applying the Upscaling Toolbox

In the following, the upscaling approach is applied to a perfectly layered medium. This can be conceptualized as a simple description of a heterogeneous aquifer. The utilization of such a medium has two advantages. The derived parameters are valid infinitely parallel to the layers and the main axis of the permeability tensor coincides with the main axis of the system. For other media, the coordinate system might have to be transformed.

In a perfectly layered medium the layers and their properties such as porosity or the constitutive relationships alternate periodically with a length of

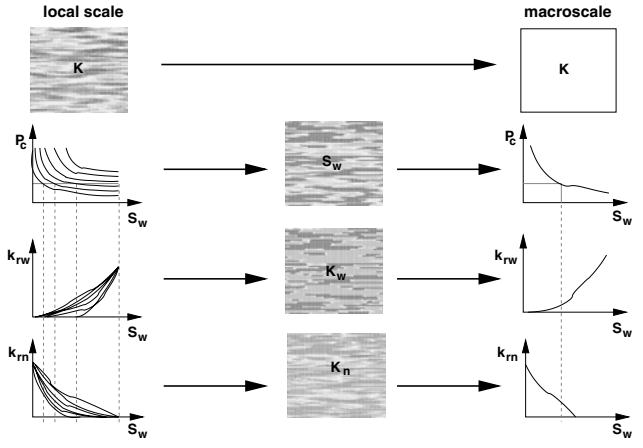


Fig. 5. Averaging constitutive relations from the local to the macroscale

period a (see Fig. 6). The single layers are assumed to be isotropic and homogeneous.

In our case, one period of the horizontally layered medium contains a 14 cm thick layer of coarse sand (index c) and a 2 cm thick layer of fine sand (index f). The material properties of the two sands can be taken from Table 2. For the description of the capillary pressure-saturation-relationship, the parametrization after Brooks & Corey is applied. Because $\lambda, S_{\mathrm{wr}}, S_{\mathrm{nr}}$ do not differ, the relative permeability-saturation-relationship is the same for both sands. This system is subsequently named the reference case.

Table 2. Material properties for the coarse and the fine sand

	k_0 [m^2]	ϕ [-]	P_{D} [Pa]	λ [-]	S_{wr} [-]	S_{nr} [-]
coarse sand	10^{-10}	0.4	1000	2	0.0	0.0
fine sand	10^{-11}	0.4	$\sqrt{10} \times 1000$	2	0.0	0.0

4.3 Upscaling of the Capillary Pressure-Saturation-Relationship

Again, we use the simple case of a layered medium to demonstrate the averaging process for the transition from the local to the macroscale for the capillary pressure-saturation-relationship.

With the assumption of a capillary equilibrium and neglecting gravity effects we compute the macroscopic capillary pressure-saturation-relationship as described at the beginning of this section. A striking result of a comparison of the capillary pressure-saturation-relationships of the two sands with the averaged, macroscopic capillary pressure-saturation-relationship (see Fig. 7) is that the macroscopic P_{c}-S_{w}-relation equals the P_{c}-S_{w}-relation of the coarse sand especially for high saturations. Moreover, in the case of low saturations, the curves do not differ much either. This is due to the higher volume fraction of the coarse sand.

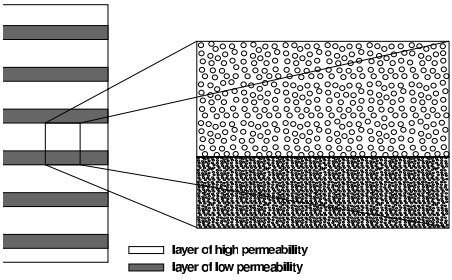


Fig. 6. Perfectly layered medium

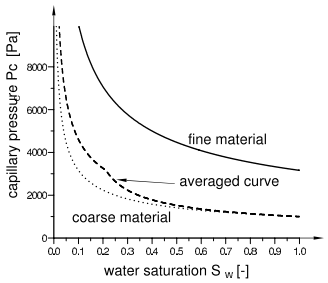


Fig. 7. Averaged capillary pressure-saturation-relationship

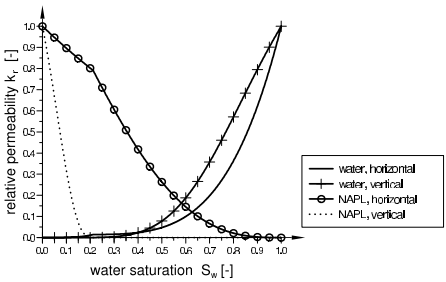


Fig. 8. Macroscopic relative permeability-saturation-relationship for layered medium

If the averaged macroscopic capillary pressure-saturation-relationship were to be used for a numerical simulation of the infiltration of a DNAPL into a water-saturated domain, the horizontal mobilization of the DNAPL would not be represented correctly. In such a case, the entry pressure of finer materials delays the vertical spreading of the DNAPL, yet the entry pressure of the fine material is not represented in the macroscopic capillary pressure-saturation-relationship. Consequently, for the upscaling procedure presented here, another averaged parameter must be found which represents the interaction of the layered medium with the two fluids more appropriately. It is worth to point out that even if we figured out capillary forces to be one of the dominant processes in the upscaling procedure, averaged capillary pressure-saturation-relationships do not explain the macroscopic behavior of multiphase flow in heterogeneous systems.

4.4 Upscaling of the Relative Permeabilities

Based on the saturation distributions of the wetting and the non-wetting phase and on the Brooks-Corey functions for the relative permeabilities, a distribution of the relative permeabilities can be calculated for the periodically layered medium as depicted in Fig. 6. For a perfectly layered medium, the averages can be obtained analytically. Equation (17) defines the relative permeability of the wetting phase in the horizontal direction which is parallel to the layers:

$$\begin{aligned} \overline{k_{\mathrm{rwh}}}(\overline{S_{\mathrm{w}}}) &= \frac{k_{\mathrm{h}}(\overline{S_{\mathrm{w}}})}{k_{\mathrm{h}}(\overline{S_{\mathrm{w}}}=1)} \\ &= \frac{b \times k_{\mathrm{rwf}} \times k_{\mathrm{f}} + (1-b) \times k_{\mathrm{rwc}} \times k_{\mathrm{c}}}{b \times k_{\mathrm{f}} + (1-b) \times k_{\mathrm{c}}} \,. \end{aligned} \tag{17}$$

Equation (18) gives the relative permeability for the wetting phase in the vertical direction perpendicular to layers:

$$\overline{k_{\mathrm{rwv}}}(\overline{S_{\mathrm{w}}}) = \frac{k_{\mathrm{v}}(\overline{S_{\mathrm{w}}})}{k_{\mathrm{v}}(\overline{S_{\mathrm{w}}} = 1)} = \frac{k_{\mathrm{rwf}} \times k_{\mathrm{rwc}} \times ((1-b) \times k_{\mathrm{f}} + b \times k_{\mathrm{c}})}{(1-b) \times k_{\mathrm{rwf}} \times k_{\mathrm{f}} + b \times k_{\mathrm{rwc}} \times k_{\mathrm{c}}}\,, \tag{18}$$

where the index w denotes the wetting phase. The indices f,c identify the fine and the coarse sand respectively. The equations for the relative permeability for the non-wetting phase can be set up accordingly. The results of the averaging procedure are shown in Fig. 8.

The averaged relative permeabilities show courses of the curves for the wetting and the non-wetting phase parallel to the layers (horizontal flow) different to those for the phases perpendicular to the layers (vertical flow). Consequently, a *macroscopic saturation-dependent anisotropy* evolves, depending solely on the heterogeneities on the macroscale. On the local scale, the layers are assumed to be isotropic. In order to illustrate the saturation dependent macroscopic anisotropy further, the relation of the horizontal conductivity (k_{h}) and the vertical conductivity (k_{v}) for water and DNAPL $a = k_{\mathrm{v}}/k_{\mathrm{h}}$ is plotted against the water saturation (see Fig. 9).

For a water saturation of $S_{\mathrm{w}} = 1.0$, the anisotropy of the conductivity of the water and of the single phase flow case (horizontal line in Fig. 9) coincide (marking 3 in Fig. 9). When the water saturation is reduced, the water still occupies the fine sand alone, but retreats from the coarse sand where the DNAPL starts to infiltrate. Thus, the relative permeability for the water in the coarse sand decreases. At a saturation of $S_{\mathrm{w}} = 0.617$, the conductivity for water in the coarse sand is reduced to 10% of the saturated conductivity. Consequently, the two layers possess the same conductivity for the water phase ($k_{\mathrm{v}} = k_{\mathrm{h}}$) and the relation of the anisotropy becomes $a = 1.0$ (marking 2 in Fig. 9). When the water saturation is reduced further the system becomes less and less permeable for water. Now the fine sand is more permeable than the coarse sand because of the fact that at low saturations the water retreats

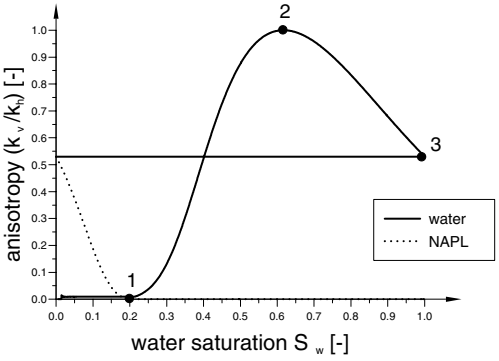


Fig. 9. Macroscopic anisotropy of the effective permeabilities in a layered medium, horizontal line: anisotropy for single-phase flow

to the finer material. Thus especially in the vertical direction the effective conductivity for water is reduced.

For $S_w < 0.17$, the relative permeability and thus the conductivity approach zero (marking 1 in Fig. 9).

The anisotropy of the DNAPL conductivities does not exceed $a = 0.53$ for a saturation of DNAPL of $S_n = 1.0$ (see Fig. 9). In the case of high DNAPL and low water saturations, the water infiltrates into the fine sand first, while the DNAPL still occupies the coarse sand with its higher intrinsic permeability. The relative permeability for the DNAPL phase in the fine sand decreases first, and the relative permeability for the non-wetting phase in the coarse sand decreases as well. As both values decrease the anisotropy remains at values near zero. When the water saturation has reached $S_w = 0.2125$, the water as the wetting phase completely occupies the fine sand. Consequently, the relative permeability for the non-wetting phase in these layers decreases to zero and the whole medium becomes impermeable for the non-wetting phase in the flow direction perpendicular to the layers. As the conductivity k_v thus becomes zero, the anisotropy also becomes zero.

On the macroscale, the vertical conductivities of the DNAPL can only exceed zero when the entry pressure of the fine sand has been overcome. This happens at a macroscopic DNAPL saturation of $S_n = 0.7875$. This saturation is therefore called the *macroscopic residual saturation of the non-wetting phase*. The corresponding water saturation is called the *macroscopic residual saturation of the wetting phase*. Again, these residual saturations can only be identified on the macroscale whereas there are no residual saturations defined at the same time on the local scale. The size of macroscopic residual saturations depends on the entry pressure of the fine material and the thickness ratio of both layers [11].

The above described upscaling procedure was validated with a numerical experiment using a periodically layered 2D system where DNAPL infiltrated from the top. The results were published in [11].

4.5 Conclusion

The application of our upscaling procedure proved that the structures of a porous medium on the local scale, such as layers or lenses, have an important influence on the effective parameters on the macroscale. The incorporation of the geometry of these structures in the upscaling process enhances the quality of the effective parameters on the macroscale and we are thus able, e. g., to reproduce the infiltration behavior of a DNAPL. The theoretical background as well as the numerical tools that are necessary to do so are presented in the following section.

5 Anisotropic Relative Permeabilities and Their Treatment with Multi-point Flux Approximations

From Sect. 4 it is known that a new phenomenon may arise in the context of upscaling: relative permeability can become direction-dependent on the target scale. In general, none of the standard tools is able to handle this and hence one can only gain solutions of such models on that scale for special cases. Consequently, a numerical flux capable of managing a tensorial relative permeability has to be developed. Because of the special challenges that a convection-dominated problem raises this flux should by its design incorporate an upwinding concept from an early stage.

To keep the considerations as simple as possible, again only two-phase flow is considered.

5.1 Extension of the Model

Jumping from one scale to another, one cannot expect the model not to change (for example from the pore scale described by multiphase Navier-Stokes equations to the macroscale where the model equations from Sect. 2 hold). Therefore, model equations or properties of involved quantities have to be adjusted adequately.

The direction-dependent relative permeability that is presented in the preceding section is an upscaling result that is already refined. The original outcome consists of $\mathsf{K}_{\mathrm{tot},\alpha}(S_\alpha)$, the so-called *total permeability* of phase α which is a saturation-dependent, generally matrix-valued quantity. It appears in a flux law that is a preliminary stage of the well-known extended version of Darcy's law (1):

$$\mathbf{v}_\alpha = -\frac{1}{\mu_\alpha}\mathsf{K}_{\mathrm{tot},\alpha}(S_\alpha)\left(\nabla p_\alpha - \rho_\alpha \mathbf{g}\right) , \quad \alpha \in \{\mathrm{w},\mathrm{n}\} . \tag{19}$$

Comparing (1) with (19), it becomes obvious that $\mathsf{K}_{\mathrm{tot},\alpha}(S_\alpha)$ and $k_{\mathrm{r}\alpha}(S_\alpha)\times\mathsf{K}$ should fulfill identical requirements: they should be symmetric so they have real eigenvalues (which are the permeabilities acting in the direction of the associated eigenvectors) and they sould be positive definite (which means that the eigenvalues are positive so that flow always takes place downhill). These properties are essential and the upscaling method is required to guarantee them for every saturation $S_\alpha \in \left(0,1\right]$!

In the fully saturated case, the total permeability is called *absolute permeability*

$$\mathsf{K} := \mathsf{K}_{\mathrm{tot,w}}(1) = \mathsf{K}_{\mathrm{tot,n}}(1) \tag{20}$$

and the upscaling method again has to guarantee that the second equality holds. The unsaturated medium is impermeable, $\mathsf{K}_{\mathrm{tot},\alpha}(0) = \mathsf{0}$, and then the total permeability is, of course, still symmetric but no longer positive definite anymore.

With the help of the inverse K^{-1}, which always exists under the given assumptions, the *relative permeability* can now generally be defined as

$$\mathsf{k}_{\mathrm{r}\alpha}\left(S_\alpha\right) := \mathsf{K}_{\mathrm{tot},\alpha}\left(S_\alpha\right) \times \mathsf{K}^{-1}\,, \quad \alpha \in \{\mathrm{w}, \mathrm{n}\} \tag{21}$$

and it is obvious that it does not need to be a scalar-valued function of the saturation any more (as has been shown in Sect. 4). In general, it is not even necessarily positive definite even for a positive saturation. But there are cases where it is possible to make a statement. To be able to do so, Theorem 2.42. and Theorem 8.61. of [30] are summarized into the following

Theorem 1. *Let* A *and* $\mathsf{B} \in \mathbb{C}^{n\times n}$ *each be similar to a diagonal matrix. Then one has*

$$\mathsf{AB} = \mathsf{BA}$$
$$\iff \quad \mathsf{A} \text{ and } \mathsf{B} \text{ have } n \text{ linearly independent eigenvectors in common.}$$

Further on, there is the trivial

Lemma 1. *Let* A *and* $\mathsf{B} \in \mathbb{R}^{n\times n}$ *be symmetric. Then it follows*

$$\textit{(i)} \quad \mathsf{AB} = \mathsf{BA} \quad \Longleftrightarrow \quad \mathsf{AB} \text{ is symmetric.}$$

If in addition to that A *and* B *are positive definite, then one can state that*

$$\textit{(ii)} \quad \mathsf{AB} = \mathsf{BA} \quad \Longrightarrow \quad \mathsf{AB} \text{ is positive definite.}$$

And from that follows the remarkable

Lemma 2. $\mathsf{k}_{r\alpha}\left(S_\alpha\right)$ *is symmetric* $\Longrightarrow$ $\mathsf{k}_{r\alpha}\left(S_\alpha\right)$ *is positive definite.*

These tools are necessary to prove the following Theorem 2, which yields more insight into the inner structures of the connection between total and relative permeabilities. It allows us to characterize when the relative permeability is symmetric (and then, because of Lemma 2, automatically positive definite as well):

Theorem 2. *The statements given below are equivalent:*

(i) $\mathsf{k}_{r\alpha}\left(S_\alpha\right)$ *is symmetric for every* $S_\alpha \in [0,1]$.

(ii) *The* n *linearly independent eigenvectors of* $\mathsf{K}_{tot,\alpha}\left(S_\alpha\right)$ *are the same for every* $S_\alpha \in [0,1]$.

(iii) $\mathsf{k}_{r\alpha}\left(S_\alpha\right)$ *and* $\mathsf{K}_{tot,\alpha}\left(S_\alpha\right)$ *have* n *linearly independent eigenvectors in common for every* $S_\alpha \in [0,1]$.

Apparently, the relative permeability is a quantity whose interpretation is possible in a descriptive way if it is symmetric. If so, it can be seen as a bundle of positive saturation-dependent factors, each acting in the direction of the associated eigenvector of the absolute permeability. In that case, the total permeability is also of notable shape and has main axes independent of the saturation. This has convenient consequences in practice as it reduces the amount of information that is necessary for the description of $\mathsf{K}_{\mathrm{tot},\alpha}(S_\alpha)$ considerably: only the fixed main axes and the saturation-dependent eigenvalues have to be known in contrast to a complete saturation-dependent triangular matrix. As a last comment, it should be noted that, in the case of fixed eigenvectors, the dependence on saturation only remains in the anisotropy ratio.

Because of these interesting properties, it is always assumed in the following that the permeability has fixed main axes. Thus, with a symmetric and positive definite tensorial relative permeability, one has now, in comparison to (1), an extended form of Darcy's law:

$$\mathbf{v}_\alpha = -\frac{\mathsf{k}_{\mathrm{r}\alpha}}{\mu_\alpha}\mathsf{K}\left(\nabla p_\alpha - \rho_\alpha \mathbf{g}\right) \ , \quad \alpha \in \{\mathrm{w}, \mathrm{n}\} \ . \tag{22}$$

To conclude these considerations, the model on the target scale of the upscaling procedure which is the focus of interest here is identical to the model of Sect. 2 with (22) replacing (1) and the associated definitions (20) and (21) substituting (6) and the classical definition of the absolute permeability.

5.2 Survey of Current Numerical Flux Functions

In the majority of cases, general purpose finite volume codes for groundwater simulation or reservoir engineering traditionally utilize two-point flux approximations (TPFA) and descendants as numerical flux functions. However, within the last decade, the technique of multi-point flux approximations (MPFA) has been developed and, because of its superior properties, it is slowly replacing the former in modern simulators.

Both classes of numerical flux functions are related to each other insofar as the MPFA can be interpreted from one point of view as a conceptual upgrade of the TPFA. They were originally designed for flux laws like

$$\mathbf{v} = -\mathsf{K}\nabla p \tag{23}$$

which basically describe one-phase flow (for the sake of simplicity, gravity and viscosity are neglected here). The extension to multiphase flow with the possible inclusion of an upwinding strategy is straightforward. It is a second step that is considered later.

TPFA are motivated by classical finite difference approaches and, therefore, so can be looked on as a comparatively low-end numerical tool. In the simplest approaches an arithmetic average of the involved permeabilities is

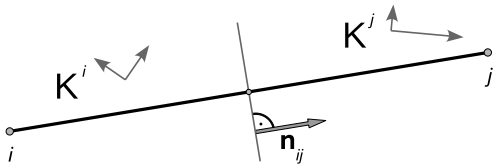


Fig. 10. Quasi one-dimensional reference problem for motivation of harmonic average

conducted. More elaborate attempts that can be deduced from a quasi-one-dimensional reference problem include a harmonic average instead. This leads to a better reproduction of impermeable layers in the discrete model. The aforementioned reference problem is sketched in Fig. 10 and shows one half of each control volume for discretization points i and j with the respective permeability tensors and the control volume interface in-between. The procedure that leads to TPFA with a harmonic average is as follows: approximate the pressure in each cell i and j linearly. At first, each linear function should interpolate the nodal value of the pressure at node i and j. Secondly, there should be a continuous transition from the cell i approximation to the cell j approximation at the control volume interface. As the gradient of a linear function is constant, the fluxes that come out of the linear approximations of pressure using (23) are constant as well. So, as a third postulation, the scheme is explicitly designed to be locally conservative by the condition that the constant flux that leaves cell i to the right has the same value as the flux that enters cell j from the left.

This setup leads to a small linear system of equations that can be solved explicitly. As a result, one gets the following approximation for the sought flux crossing the cell interface

$$v_{ij} := \mathbf{v} \cdot \mathbf{n}_{ij} \approx -2 \left(\frac{1}{\mathbf{n}_{ij}^{\mathrm{T}} \cdot \mathsf{K}^i \cdot \mathbf{n}_{ij}} + \frac{1}{\mathbf{n}_{ij}^{\mathrm{T}} \cdot \mathsf{K}^j \cdot \mathbf{n}_{ij}} \right)^{-1} \times \frac{p_j - p_i}{\|\mathbf{x}_j - \mathbf{x}_i\|_2} \tag{24}$$

which harmonically averages the effective components of the permeability tensors. A more detailed derivation and discussion can be found in [36].

Of course, the control volume interface being exactly perpendicular to the edge from node i to j is a very rough approximation of real grids (compare Fig. 1). Another point is that the TPFA (24) is known to be a consistent approximation only for the cases where the anisotropies of grid and permeability are somehow specially related (keyword: K-orthogonality, see [3, 1]).

A way out of this deficiency is presented by the MPFA. The fundamental concept is simple: the flux going from cell i to cell j is split up into one fraction that flows above the edge from i to j through the control volume interface and a second one that flows below. Thus, this split is made by means of the involved elements. Figure 11 exemplarily shows the element that belongs to the upper part of the flux from i to j (cp. Fig. 1). In analogy to the TPFA

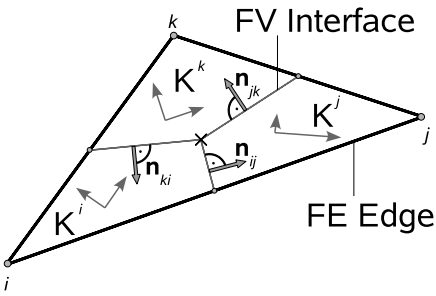


Fig. 11. Example element for motivation of MPFA

procedure, the idea consists of using a linear approximation for the pressure in every piece of the control volumes found in this element. In this way, one has three degrees of freedom for every sub-control volume, and a total of nine in a triangular element. Like the TPFA, these approximations are postulated to interpolate the nodal value of pressure at the vertices. Furthermore, the approximation of pressure at an edge midpoint has to be the same from both sides. And it is postulated that the flux leaving one control volume is equal to that entering the next one (local conservativity). As pressure has been approximated linearly, this flux is again constant. This results in nine statements leading to nine equations. By clever reformulation and insertion, the problem can be reduced to the solution of a linear system of three equations with three unknowns. As an example, the equation describing conservativity for the flux alongside the edge from i to j is

$$\begin{aligned} v_{ij} &\approx \mathbf{n}_{ij}^{\mathrm{T}} \cdot \mathsf{K}^i \times \frac{1}{2F_i} \left[(p_{ij} - p_i)\, \boldsymbol{\nu}_{ij}^{(i)} + (p_{ik} - p_i)\, \boldsymbol{\nu}_{ik}^{(i)} \right] \\ &= \mathbf{n}_{ij}^{\mathrm{T}} \cdot \mathsf{K}^j \times \frac{1}{2F_j} \left[(p_{ij} - p_j)\, \boldsymbol{\nu}_{ij}^{(j)} + (p_{jk} - p_j)\, \boldsymbol{\nu}_{jk}^{(j)} \right] . \end{aligned} \tag{25}$$

The definition of the new quantities (F_i, p_{ij}, $\boldsymbol{\nu}_{ij}^{(i)}$ etc.) is not of interest here and can be found in [36, 1]. Coefficients t_{ij}^i, t_{ij}^j, t_{ij}^k, the so-called *transmissibilities*, can be obtained by solving this system and refining its solution. They deliver the sought flux simply by linear combination:

$$v_{ij} \approx t_{ij}^i \times p_i + t_{ij}^j \times p_j + t_{ij}^k \times p_k \, . \tag{26}$$

Transmissibilities for the fluxes v_{jk} and v_{ik} are gained in the same manner. The whole procedure can be accomplished in a single preprocessing step as the transmissibilities do not change during the simulation in this case. In contrast to the TPFA, all edge points of the element participate in the approximation. This is why they are called multi-point flux approximations. For K-orthogonal grids the described MPFA method reduces to TPFA with harmonic averages. As the former can be applied to any full-tensor permeability fields on arbitrary grids, they are a clear step forward and should be the tool of choice.

The transfer of the presented concepts from one-phase flow (23) to multiphase flow (1) is managed in the following way: at first using the described TPFA (24) or MPFA method (26) for (23) with the phase pressure p_α instead of the pressure p, a flux $u_{\alpha,ij}$ that is similar to one-phase flow is calculated. A second step determines whether this flux is going from i to j or vice versa on the basis of its sign. If it flows from i to j, then node i is the upstream node and the mobility $k_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right)/\mu_\alpha^{\mathrm{up}}$ is evaluated with up $= i$. If u_{ij} is directed the other way round, the mobility is evaluated for node j. This upstream mobility is then multiplied by $u_{\alpha,ij}$ to finally get the approximation for $v_{\alpha,ij}$.

5.3 Enhancement of Two- and Multi-point Flux Approximations with Regard to Anisotropic Relative Permeabilities

Upscaling the model for two-phase flow of Sect. 2 to the extended model of Sect. 5.1, it is easily seen that the TPFA or MPFA flux is no longer applicable in the classical way illustrated in the preceding Sect. 5.2: the outcome of formulas (24) or (26) is a scalar that would be multiplied by the matrix-valued upstream mobility in a second step. The course of action clearly has to be adapted somehow to this new situation! This adaptation is motivated exemplarily using TPFA as it is easier to comprehend, as already here in this special case of MPFA the difficulties become evident and as the transfer to MPFA is obvious. In the next steps of this development, isotropic relative permeabilities are still used.

If the two steps of computing the flux $u_{\alpha,ij}$ similar to the one-phase-flow and multiplying it by the upwinding mobility are formally combined, this leads to

$$v_{\alpha,ij} \approx -\frac{k_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right)}{\mu_\alpha^{\mathrm{up}}} \times 2\left(\frac{1}{\mathbf{n}_{ij}^{\mathrm{T}}\cdot\mathsf{K}^i\cdot\mathbf{n}_{ij}}+\frac{1}{\mathbf{n}_{ij}^{\mathrm{T}}\cdot\mathsf{K}^j\cdot\mathbf{n}_{ij}}\right)^{-1}\times\frac{p_{\alpha j}-p_{\alpha i}}{\left\|\mathbf{x}_j-\mathbf{x}_i\right\|_2} \tag{27}$$

and further reshaping yields

$$v_{\alpha,ij} \approx -2\left[\frac{\mu_\alpha^{\mathrm{up}}}{k_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right)}\left(\frac{1}{\mathbf{n}_{ij}^{\mathrm{T}}\cdot\mathsf{K}^i\cdot\mathbf{n}_{ij}}+\frac{1}{\mathbf{n}_{ij}^{\mathrm{T}}\cdot\mathsf{K}^j\cdot\mathbf{n}_{ij}}\right)\right]^{-1}\times\frac{p_{\alpha j}-p_{\alpha i}}{\left\|\mathbf{x}_j-\mathbf{x}_i\right\|_2} \tag{28}$$

$$=-2\left(\frac{\mu_\alpha^{\mathrm{up}}}{\mathbf{n}_{ij}^{\mathrm{T}}\cdot k_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right)\mathsf{K}^i\cdot\mathbf{n}_{ij}}+\frac{\mu_\alpha^{\mathrm{up}}}{\mathbf{n}_{ij}^{\mathrm{T}}\cdot k_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right)\mathsf{K}^j\cdot\mathbf{n}_{ij}}\right)^{-1}\times\frac{p_{\alpha j}-p_{\alpha i}}{\left\|\mathbf{x}_j-\mathbf{x}_i\right\|_2}\,. \tag{29}$$

The right hand term of (29) is the sought-after expression, it is ready to be equipped with a tensorial relative permeability:

$$v_{\alpha,ij}=-2\left(\frac{\mu_\alpha^{\mathrm{up}}}{\mathbf{n}_{ij}^{\mathrm{T}}\cdot\mathsf{k}_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right)\mathsf{K}^i\cdot\mathbf{n}_{ij}}+\frac{\mu_\alpha^{\mathrm{up}}}{\mathbf{n}_{ij}^{\mathrm{T}}\cdot\mathsf{k}_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right)\mathsf{K}^j\cdot\mathbf{n}_{ij}}\right)^{-1}\times\frac{p_{\alpha j}-p_{\alpha i}}{\left\|\mathbf{x}_j-\mathbf{x}_i\right\|_2}\,. \tag{30}$$

As the previous mathematical modifications have been equivalence transformations, a true generalization of the traditional TPFA has been achieved. It is denoted *conventional upwinding approach* in the future.

However, a problem with this approach soon becomes apparent: up $= i$ or up $= j$; therefore, one of the products of relative and absolute permeability in the denominators is a product where both permeabilities come from the same node whereas the other product is a mixed product. Why is this bad? Think of an upscaled problem where upscaling has been performed individually for the left and for the right half so that on the higher scale the problem is not homogeneous but heterogeneous. And now think of cell i belonging to the left and cell j belonging to the right half. The absolute permeability is symmetric and positive definite. Even if the relative permeability has these properties as well, in the mixed product both are not guaranteed to have the same principal axes as they stem from different subdomains. Now Theorem 1 of Sect. 5.1 does not hold any more and nothing can be said about the definiteness of this mixed product. In fact, it is possible that the fraction with this mixed product becomes negative. Then the whole term could become negative and, as a result, water could flow uphill. It could even happen that this negative fraction has nearly the same absolute value as the other fraction. In that case, the whole term would explode. To sum up: the method potentially delivers nonsense if applied to problems with heterogeneous anisotropic relative permeabilities.

The following idea presents a possible loophole: do not use the whole mobility for upwinding purposes, but only use the quantity that has the hyperbolic character which requires upwinding. Give this role exclusively to the saturation! Then (30) has to be modified to result in

$$v_{\alpha,ij} = -2\left(\frac{\mu_\alpha^i}{\mathbf{n}_{ij}^{\mathrm{T}} \cdot \underbrace{\mathsf{k}_{\mathrm{r}\alpha}^i\left(S_\alpha^{\mathrm{up}}\right)\mathsf{K}^i}_{=\mathsf{K}_{\mathrm{tot},\alpha}^i\left(S_\alpha^{\mathrm{up}}\right)} \cdot \mathbf{n}_{ij}} + \frac{\mu_\alpha^j}{\mathbf{n}_{ij}^{\mathrm{T}} \cdot \underbrace{\mathsf{k}_{\mathrm{r}\alpha}^j\left(S_\alpha^{\mathrm{up}}\right)\mathsf{K}^j}_{=\mathsf{K}_{\mathrm{tot},\alpha}^j\left(S_\alpha^{\mathrm{up}}\right)} \cdot \mathbf{n}_{ij}}\right)^{-1} \times \frac{p_{\alpha j} - p_{\alpha i}}{\|\mathbf{x}_j - \mathbf{x}_i\|_2}\,, \tag{31}$$

known as *modified upwinding approach* from now on.

Some things concerning these two upwinding approaches are striking:

- With (31) an explicit splitting of the total permeability is no longer necessary!
- If used inside a homogeneous subdomain where $\mathsf{k}_{\mathrm{r}\alpha}^i$ and $\mathsf{k}_{\mathrm{r}\alpha}^j$ do not differ, (30) and (31) are identical and hence provide the same result.
- If there are various subdomains with different relative permeabilities which have varying residual saturations the following could happen when (31) is applied: if the upwinding saturation is smaller than the residual saturation of the relative permeability at the downstream node, the latter will effectively be evaluated to be $\mathsf{k}_{\mathrm{r}\alpha}^{\mathrm{down}}\left(S_\alpha^{\mathrm{up}}\right) = 0$. As a consequence, the harmonic average and the whole flux will vanish. It is rather what one expects from physical reality compared to (30) which gives a non-zero flux. This behavior of (31) is similar to the effect of entry pressure in the PPSIC

discretization (cf. [16, 25]) although it is caused by a completely different effect. One important fact should be noted here: this diverse behavior of the conventional and the modified upwinding approach exists even in the case of isotropic relative permeability. Thus, the modified flux could be of interest in traditional two-phase flow simulators as well!

The transfer of the concept to MPFA is obvious: instead of using the absolute permeability in the equation on its own, include the relative permeability as well! Pursuing this idea, one gets from (25) the conventional

$$\begin{aligned} v_{ij} &\approx \mathbf{n}_{ij}^{\mathrm{T}} \cdot \frac{1}{\mu_\alpha^{\mathrm{up}}} \mathsf{k}_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right) \mathsf{K}^i \times \frac{1}{2F_i} \left[\left(p_{ij} - p_i\right) \boldsymbol{\nu}_{ij}^{(i)} + \left(p_{ik} - p_i\right) \boldsymbol{\nu}_{ik}^{(i)} \right] \\ &= \mathbf{n}_{ij}^{\mathrm{T}} \cdot \frac{1}{\mu_\alpha^{\mathrm{up}}} \mathsf{k}_{\mathrm{r}\alpha}^{\mathrm{up}}\left(S_\alpha^{\mathrm{up}}\right) \mathsf{K}^j \times \frac{1}{2F_j} \left[\left(p_{ij} - p_j\right) \boldsymbol{\nu}_{ij}^{(j)} + \left(p_{jk} - p_j\right) \boldsymbol{\nu}_{jk}^{(j)} \right] \end{aligned} \tag{32}$$

and the modified

$$\begin{aligned} v_{ij} &\approx \mathbf{n}_{ij}^{\mathrm{T}} \cdot \frac{1}{\mu_\alpha^{i}} \underbrace{\mathsf{k}_{\mathrm{r}\alpha}^{i}\left(S_\alpha^{\mathrm{up}}\right) \mathsf{K}^i}_{=\mathsf{K}_{\mathrm{tot},\alpha}^{i}\left(S_\alpha^{\mathrm{up}}\right)} \times \frac{1}{2F_i} \left[\left(p_{ij} - p_i\right) \boldsymbol{\nu}_{ij}^{(i)} + \left(p_{ik} - p_i\right) \boldsymbol{\nu}_{ik}^{(i)} \right] \\ &= \mathbf{n}_{ij}^{\mathrm{T}} \cdot \frac{1}{\mu_\alpha^{j}} \underbrace{\mathsf{k}_{\mathrm{r}\alpha}^{j}\left(S_\alpha^{\mathrm{up}}\right) \mathsf{K}^j}_{=\mathsf{K}_{\mathrm{tot},\alpha}^{j}\left(S_\alpha^{\mathrm{up}}\right)} \times \frac{1}{2F_j} \left[\left(p_{ij} - p_j\right) \boldsymbol{\nu}_{ij}^{(j)} + \left(p_{jk} - p_j\right) \boldsymbol{\nu}_{jk}^{(j)} \right] \end{aligned} \tag{33}$$

upwinding extension of MPFA for anisotropic relative permeabilities. MPFA are a generalization of TPFA; hence, the same remarks are valid here.

This extension of TPFA and MPFA with the goal of coping with tensorial relative permeabilities is won at the price of the resulting transmissibilities depending on saturation. Therefore, they cannot be computed in a preprocessing step once and for all but have to be recalculated every time the involved saturations change.

Another big issue that has been ignored up to now is the question of how to find the upwinding node. With this extension, it is no longer possible to gain a priori knowledge of that information. Instead, the only solution known so far is brute force: start a fresh computation for all possible values for this information and check afterwards whether the result is consistent with the chosen assumption. This topic is dealt with in more detail in Sect. 5.5.

5.4 Comparison of Conventional and Modified Upwinding Strategy

Up to now both strategies for upwinding have only been derived and solely some analytically accessible distinctions within their properties have been discussed. Not so obvious differences should be discoverable in practical usage by performing real simulations for special problems.

Construction of Reference Problem: Heterogeneous Buckley-Leverett

In a situation like this where alternatives have to be evaluated one has to find some kind of measure for that purpose. A high-class way to do that in numerics is the quantitative comparison of simulation results with an analytical solution. Thus, analytically addressable problems are of great interest.

What should a problem for especially this situation look like? The numerical flux-functions (32) and (33) do not differ in a homogeneous case. Therefore, the problem to be constructed should comprise many heterogeneities. For the time being it does not need to have a tensorial relative permeability because (32) and (33) already differ in the scalar case. And last but not least as the concept we are looking at is an upwinding strategy the problem has to be dominated by advective flux.

In the case of the simple model of two-phase flow a well-known reference problem is that of Buckley and Leverett. Starting with the model of Sect. 2 and performing some simplifications (cp. [36]) one ends up with the hyperbolic PDE

$$\frac{\partial S_{\mathrm{w}}}{\partial t} + \frac{\partial}{\partial x} f\left(S_{\mathrm{w}}\right) = 0 \tag{34}$$

where the flux function f only depends on saturation S_{w} and has the typical S-shape (see Fig. 12). The actual Riemann-problem then consists of (34) and initial conditions with $1 - S_{\mathrm{nr}}$ as the left and S_{wr} as the right state, cf. Fig. 12. Applying the method of characteristics to this hyperbolic problem at least a quasi-analytical solution of high precision can be easily constructed (see [24, 31]).

What is needed here is a heterogeneous version of this problem where the relative permeability is changing with space. To become a little bit more concrete, flux functions which basically follow a design pattern like

$$f\left(S_{\mathrm{w}}, x\right) = \begin{cases} f_{\lambda_1}\left(S_{\mathrm{w}}\right) & \text{for } x \leq x_{\mathrm{het}} \\ f_{\lambda_2}\left(S_{\mathrm{w}}\right) & \text{for } x > x_{\mathrm{het}} \end{cases} \tag{35}$$

are in the focus of interest. Here λ_1 and λ_2 denote the λ-parameter of the widespread Brooks-Corey parametrization for the relative permeability (cp. [24]). The flux functions that result for two choices of λ can bee seen in Fig. 12. Of course, in general one is interested in more than one heterogeneity in f — the extension should be clear.

It is shown in [21, 22] that a solution to this kind of problem exists and that this solution is unique. But the proof is not constructive. Therefore using the method of characteristics a solution at least for a subset of this problem-class has been constructed and extensively discussed in [36, 37]. Here only a rough outline of the structures and phenomena that are encountered shall be presented:

- The solution is constructed recursively in the form of a new level of recursion for every new (in itself homogeneous) subdomain.

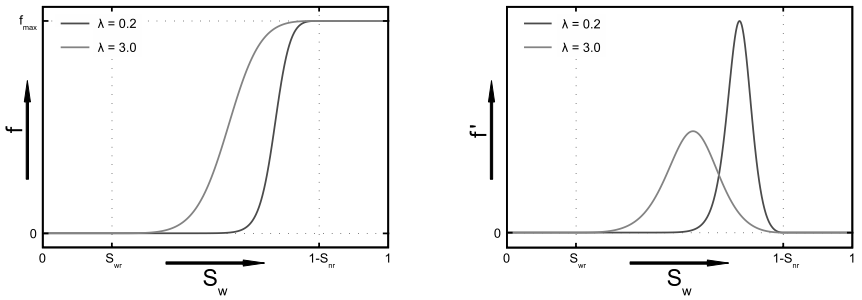


Fig. 12. Flux function of Buckley-Leverett problem and its derivative for different choice of shape-parameter λ

- Every level can be mapped into one of two subclasses:
  - One subclass where additional to the rarefaction wave that is known from the solution of the homogeneous Buckley-Leverett problem a new rarefaction wave appears at the heterogeneity which follows a shock traveling at constant speed, see Fig. 13 ($\lambda_1 = 0.2$, $\lambda_2 = 3.0$).
  - A second subclass where characteristics run into the shock which as a result raises its propagation speed and asymptotically approximates one of the characteristics, see Fig. 14 ($\lambda_1 = 3.0$, $\lambda_2 = 0.2$).
- The solution exhibits stationary discontinuities at the points where the homogeneous subdomains meet.
- At these places the characteristics bend and change their propagation speed from high to low and vice versa.

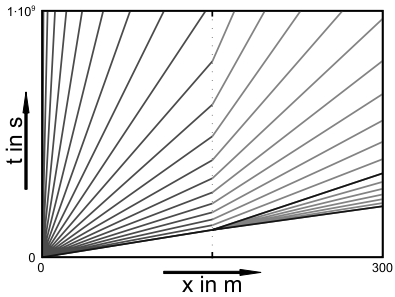


Fig. 13. Diagramm with course of characteristics for the first subclass of the recursion levels

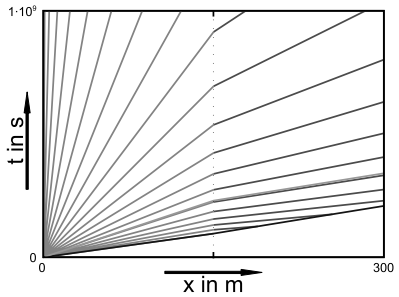


Fig. 14. Diagramm with course of characteristics for the second subclass of the recursion levels

Table 3. Rate of convergence r for the introduced test-cases

	rate of convergence r	
test-case	conv. flux	mod. flux
problem A	0.331703	0.312071
problem B	0.292567	0.297602
problem C	0.316063	0.292819
problem D	0.303027	0.304333

Numerical Evaluation of Alternative Strategies

Taking the heterogeneous Buckley-Leverett problem as a reference problem for the purpose of validating the two competing upwind fluxes (32) and (33), examples have been set up with a great number of heterogeneities to give potential differences the chance to result in noticeable numerical effects.

Four test-problems have been designed, each consisting of twelve homogeneous subdomains so that eleven heterogeneity-frontiers arise. The first problem (problem A) is based upon the parameter λ decreasing from the left to the right. In the second case (problem B) λ increases from the left to the right and an alternating parameter λ makes up test-problem C. Last but not least the fourth test-case, problem D, consists of a random choice of λ.

For every test-problem the quasi-analytical solution has been calculated by means of the developed recursion. Then the discretized model of Sect. 2 using the discussed numerical fluxes has been solved. Grid convergence of the numerical solutions has been tested and measured against the reference solution. For that purpose, the numerical simulations of the test-problems have been performed on a sequence of grids with 80, 160, 320, ..., 5120 elements.

The comparison of the quasi-analytical solution $S_{\mathrm{w}}(t,x)$ and the numerical results $\hat{S}_{\mathrm{w}}(t,x)$ has been conducted using the L_2-norm. At a fixed $\tilde{t} \in \mathbb{R}$ that is great enough so that the shock front already passed all heterogeneities the simulations were stopped. Figures 15–18 show the reference solution and the numerical results calculated on the coarsest mesh (80 elements) applying both fluxes. On the finer grids the difference between the numerical and the quasi-analytical solution very fast becomes hard to detect optically. Thus only coarse grid solutions are pictured here.

To get a feeling of the rate of convergence r with respect to the space discretization h in the estimation

$$\left\| S_{\mathrm{w}}\left(\tilde{t},x\right) - \hat{S}_{\mathrm{w}}\left(\tilde{t},x\right) \right\|_{\mathrm{L}_2} \leq C \times h^r \tag{36}$$

least-squares-fits of the computed L_2-norms on the left of (36) against the right hand side have been realized. The results of this fits can be seen in Table 3. From that the following conclusions can be drawn:

- Within all tests A–D both numerical flux functions produce a converging solution.

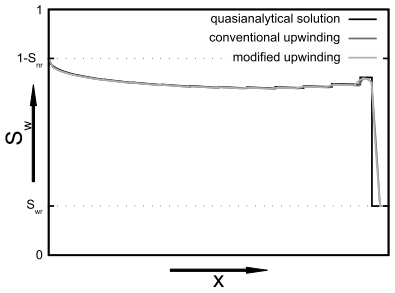


Fig. 15. Solutions of problem A

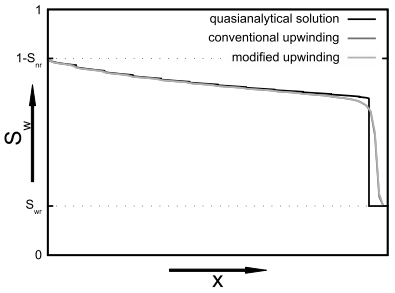


Fig. 16. Solutions of problem B

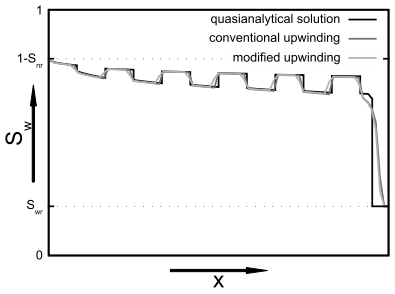


Fig. 17. Solutions of problem C

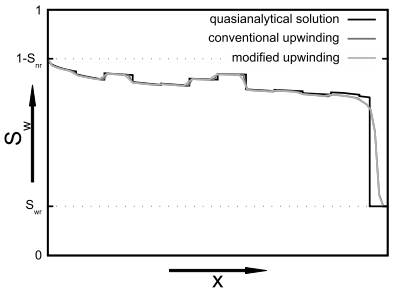


Fig. 18. Solutions of problem D

- The rate of convergence r for the conventional upwinding flux compared to that of the modified upwinding flux is similarly small. But that is an expected property of the simulations as the convergence for problems with discontinuities is known to be bad, cf. [31].

One difference is very apparent looking at the simulation results. It can be recognized best in the outcome of problem C (see Fig. 19): the result that was computed using the conventional flux function needs one element (two gridpoints) for jumps, whereas the modified flux function needs two elements (three gridpoints). The additional second element is always located in front of

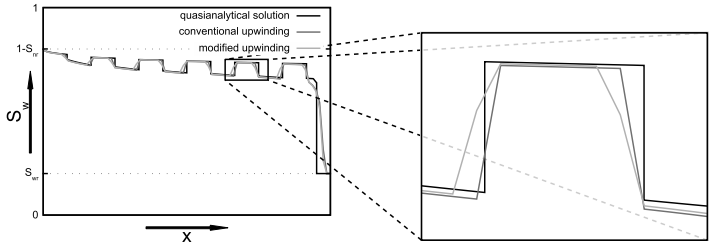


Fig. 19. Zoom into two discontinuities of problem C

the heterogeneity. In ostensible contradiction to the direction of information propagation using an upwinding scheme, the solution knows of the heterogeneity before effectively reaching it. This can be explained with the pressure being a quantity with elliptic character. As pressure and saturation are computed using a fully coupled system of equations such a behavior is possible. Besides those elements close-by the discontinuities the difference in the simulation results is marginal. But even on the finest grid the two numerical fluxes show the aforementioned qualitative discrepancy around discontinuities.

To conclude this section it can definitely be stated that with the present experiences the new modified upwinding flux seems to be ready for operation.

5.5 Algorithmic Challenges

Two challenges are encountered in the implementation of the ideas described before. For both a solution strategy is presented which cannot be justified with hard mathematical facts, but is more of heuristic or intuitive character.

Finding the Upstream Node

Within current numerical codes for the solution of two- or multiphase flow in porous media where anisotropic relative permeabilities do not play a role the computation of an upwinding flux using TPFA or MPFA can be divided into two steps, see Sect. 5.2: the first one provides the information which one of the involved nodes has to be taken as the upstream node. And in the second step the resulting upwinding flux is calculated. If both steps are integrated into one, because direction-dependent relative permeabilities are considered or because one of the modified upwinding approaches (31) or (33) has to be used, a priori knowledge of the direction of flux is necessary for computation of the flux itself. But this information is not available prior to the calculation. What to do? The only solution up to now is brute force. This approach of trial and error is illuminated for MPFA and a triangular element (cf. Fig. 11) in the following.

The local linear system of equations for the calculation of the transmissibilities consists of equations of the type (33) which express the continuity of the associated flux. For each flux and therewith for each local equation one of the nodes has to be specified as the upstream node. This is done independently from the specification for the other equations. Therefore, with two choices for every flux and three fluxes in a triangular element one has a total of $2^3 = 8$ combinations for this element.

For any such combination the calculation of the transmissibilities has to be performed which essentially consists of the solution of the local linear system of equations and of two matrix-vector products. As a result one gets three fluxes which are assigned to that combination. And now from these eight combinations one candidate has to be chosen appropriately as the one that is finally used for the numerical calculation. For this purpose suitable selection

criteria had been constructed which are of heuristic nature and which are simply presented here without any motivation or discussion (more details are given in [36]).

At first, for every combination the orientation of each flux and therefore the resulting upstream node have to be compared with the original assumption for the upwinding direction in this combination. If both match then the flux is called consistent, otherwise inconsistent. Every combination which contains at least one inconsistent flux is also marked with the attribute of being inconsistent, whereas the remaining combinations are called consistent.

If for an element there are only inconsistent combinations, the one is chosen which has the smallest sum of absolute values of the inconsistent fluxes. This case is not of academic nature but really occurs in practice. Presumably it originates from problems with monotonicity (cp. [35]). It will be looked at in more detail in future work.

In the case that consistent combinations exist they are sorted for the number of non-vanishing fluxes. If there is more than one combination with maximal number of non-vanishing fluxes take the one with the greatest sum of absolute values of fluxes.

The described procedure leads to good results, see Sect. 5.6. So far so good. But if for example a three-dimensional domain is meshed with hexahedrons the following practical problem arises: every hexahedron has twelve fluxes. Therefore, $2^{12} = 4096$ combinations have to be evaluated per element. For every combination a local linear system of twelve coupled equations has to be solved. Thus, even for a coarse mesh this brute force approach now results in a tremendous effort!

Hence, other ways have to be found. One attempt could be to start with the last successful combination for this element. Then evaluate only those combinations which differ maximally by one switched upwinding direction compared to this last successful combination. That heuristic would reduce the search-space enormously! If this does not succeed allow two upwinding directions to switch and so on. Up to now this procedure is only a promising idea. Future research work has to examine it further and probably additional alternative heuristics can be found.

Degeneracy of MPFA Equations

If in (32) or (33) the upstream saturation drops below the residual saturation of the relative permeability the latter becomes zero. This is no problem if it only happens on one side of the equation. However, if the relative permeability vanishes on the left *and* on the right side the whole equation degenerates to the unconditional true statement $0 = 0$. At first sight this does not look like a problem as the flux, the continuity of which is described by this equation, is a priori known to be zero. But for the computation of the transmissibilities a coupled system of these equations has to be solved. Now one statement is

missing in this system. Therefore, it is underdetermined and unfortunately, there is no other obvious statement that could replace this information!

The current idea for regularization of the system is the following: the equation degenerated because saturation became too small. Instead of that now take a look at the limiting statement which should reflect the searched information best. This is done numerically in the most simple way by replacing a saturation below the residual saturation with the residual saturation plus a very small $\varepsilon > 0$. Now solve the resulting MPFA system of equations. And then for the rest of the MPFA procedure again use the degenerated equations which ensures that the affected flux becomes zero exactly and not only approximately. This concept successfully proved its worth in practice up to now. And furthermore it guarantees a smooth transition from the area below the residual saturation to the area above and vice versa.

5.6 First Results

An experimental setup of J. Allan et al. (cf. [5]) serves as the basis for the test-case. It has been simulated utilizing the newly developed technique for the first time. The test-case consists of a 1.2 m wide and 0.5 m high cuvette (see Fig. 20) filled with a medium sand as background material and loose coarse sand lenses (dark-gray) and fine sand lenses (light-gray). The experiment starts fully water-saturated. 29 ml/min dense non-aqueous phase liquid (DNAPL) is infiltrated at the middle of the top boundary. After being injected it spreads due to gravity.

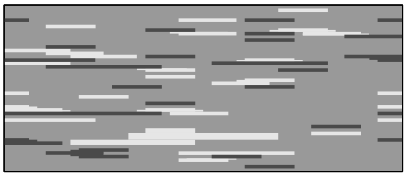


Fig. 20. Distribution of sands in experimental setup

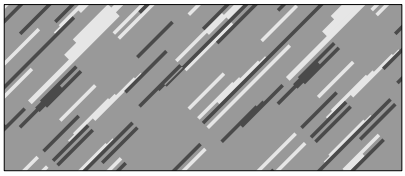


Fig. 21. Rotation of the main axes of the system by 45 degrees

Because of entry pressure effects the DNAPL pools up directly above fine sand lenses. As a result the geometrical structure of the lenses now leads to pronounced lateral spreading of the DNAPL. A similar effect is induced by the coarse sand lenses which at first fill up with DNAPL to a certain extend before the DNAPL enters the medium sand again.

After applying the upscaling procedure of Sect. 4 this effect is reflected in the permeabilities on the macroscale. Both the relative and absolute permeability become anisotropic: the horizontal components are greater compared with the vertical ones. In the relative permeability this behavior is pronounced stronger than in the absolute permeability. The outcome of the upscaling

method has been refined in a special way (see [36]) to smooth out oscillations in the saturation-dependent anisotropy ratio. Figure 22 shows the result of this refinement. The main axes of the upscaled permeabilities are identical

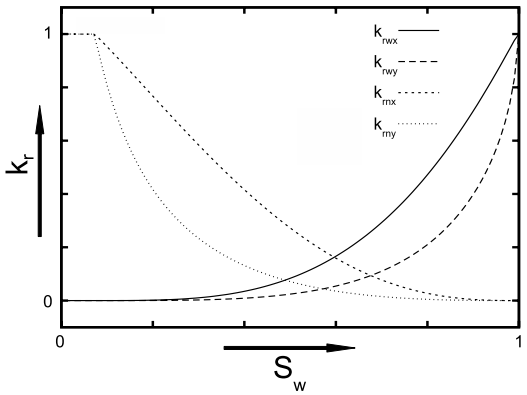


Fig. 22. Refined upscaling results for the relative permeabilities

and are oriented in horizontal and vertical direction. Discretizing the domain of interest with a simple Cartesian grid the test-case leads to a numerical problem with a diagonal tensor and, therefore, can be solved easily with slightly modified default fluxes.

Things get interesting in the case when the grid axes and the main axes of the permeabilities are not aligned anymore: now the new techniques of Sect. 5 have to be applied. The most challenging case as a starting point is constructed by rotating the tensors of the permeabilities by 45 degrees. This matches the setup of Fig. 21.

A simulation run produces results as expected. They are shown in form of snapshots at different points in time in Figs. 23–25. One clearly sees that spreading of DNAPL is significantly affected by the orientation of the permeabilities' main axes and that it obviously shows the expected anisotropy.

Things get even more interesting if on the macroscale the anisotropic permeabilities are not homogeneous anymore. Then the conventional and the modified upwinding approach actually differ. As a consequence now the modified MPFA method (33) can demonstrate its advantages compared with the conventional approach (32) which might fail in this case. Results of such simulation runs are expected in the near future and will be published in [36].

6 Outlook

This work has shown that physical processes occurring on different scales need special treatment. Depending on the considered process either a higher

Fig. 23. Distribution of DNAPL-saturation after 500 s

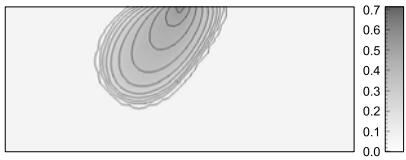


Fig. 24. Distribution of DNAPL-saturation after 2500 s

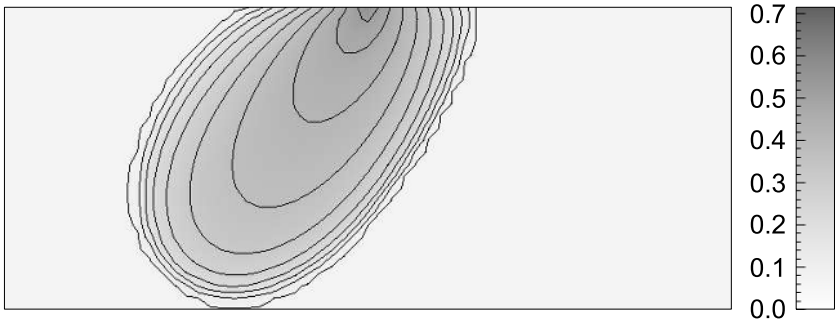


Fig. 25. Distribution of DNAPL-saturation after 4500 s

or a smaller fine-scale dependence can be detected. Mass transfer processes, e. g., are highly dependent on the interfacial areas between different phases and thus, show a high fine-scale dependence. Similarly, all processes that are dependent on the heterogeneities in the soil structure require a fine resolution of the interfacial area between materials with different properties and show thus a high fine-scale dependence.

In general, one has three different possibilities how to handle multi-scale problems:

1. The most accurate solution is a very well-resolved fine-scale solution. Sometimes, however, the physical laws describing the fine-scale processes are not known in detail. Furthermore, fine-scale modeling requires a large amount of data and takes long computational times.
2. To reduce the amount of data to be collected and to speed up the computation, upscaling approaches have been developed for a variety of flow and transport equations. Upscaled models represent a mathematically justified description of fine-scale processes on a larger scale. Nevertheless, depending on the fine-scale equations and the considered scales, such an upscaled model can not always be designed.
3. A further possibility to deal with multi-scale problems are multi-scale approaches which model different processes on different scales depending on their impact on each scale.

Within the frame of this work, the issue of multi-scale processes has been tackled using the technique of upscaling.

The more complex the physical and mathematical problem gets, the more data is needed for the first possibility, i.e. the fine-scale simulation, and the more expensive it becomes. Furthermore, considering the second variant, the derivation of an upscaled model becomes more and more difficult with increasing model complexity if at all it is possible to develop a mathematically justified coarse-scale description. Therefore, in order to be able to tackle highly complex systems, a multi-scale approach is appropriate which allows the adequate treatment of processes taking place on different scales. Our future research points into the direction of multi-scale modeling where processes are modeled on a scale which is as coarse as possible and as fine as is required. Furthermore, we plan to further increase efficiency by modeling processes only in the spatial domain where they are actually relevant. And finally, we plan to profit of existing upscaled models and model those processes for which mathematically and physically justified upscaling approaches exist on the upscaled (coarse) scale.

7 Conclusions

Models for multiphase flow in porous media nowadays can be found in many areas of science and engineering. Within the current project work has been done to develop sophisticated multiphase-multicomponent models which can be applied to many fields of interest from techniques used for groundwater remediation to CO_2-sequestration in the subsurface. But besides the new application areas these models also have new requirements for the numerical algorithm: there is the adaptive switching of phase state and of the associated model equations. This task could under the additional challenge of utilizing multigrid methods be solved satisfactorily in the current project. In the meantime, the resulting numerical code is used successfully in many application-oriented research projects.

A more fundamental field of research in this project are the activities in the area of upscaling methods. In addition to preceding theoretical investigations a toolbox for upscaling purposes has been provided. In the context of its application the upscaled relative permeabilities turned out to be direction-dependent. For this new effect a well-founded theoretical description has been worked out. On the numerical side of this extended macroscale model two new numerical flux functions have been constructed. Motivating the ideas using TPFA and finally transferring them to MPFA in order to cope with this anisotropy they differ only in their upwinding approach. The successful applicability of both has been shown using a specially tailored heterogeneous variety of the Buckley-Leverett-problem. Existence and uniqueness of its solution was known before but the proof was not constructive. Therefore, using the

method of characteristics a quasi-analytical solution was constructed. It then served as a reference for the numerical validation of the two flux functions.

Using an example problem motivated from experimental practice it was demonstrated that the basic upwinding idea for inclusion of anisotropic relative permeabilities into the numerical flux on the macroscale is working and gives the expected results.

The future orientation of this project exceeds present upscaling approaches: in a multiscale model effects of the finer scale are reflected differently on the macroscale. This is done by solving local subproblems on the finer scale and integrating their solution into the model on the macroscale.

References

1. I. Aavatsmark. An introduction to multipoint flux approximations for quadrilateral grids. *Computational Geosciences*, 6:405–432, 2002.
2. I. Aavatsmark, T. Barkve, Ø. Bøe, and T. Mannseth. Discretization on non-orthogonal, curvilinear grids for multiphase flow. In *Proc. of the 4th European Conf. on the Mathematics of Oil Recovery*, Norway, 1994.
3. I. Aavatsmark, T. Barkve, Ø. Bøe, and T. Mannseth. Discretization on non-orthogonal, quadrilateral grids for inhomogeneous anisotropic media. *Journal of Computational Physics*, 127:2–14, 1996.
4. M. Acosta, C. Merten, G. Eigenberger, H. Class, R. Helmig, B. Thoben, and H. Müller-Steinhagen. Modeling non-isothermal two-phase multicomponent flow in the cathode of pem fuel cells. *Journal of Power Sources*, page in print, 2006.
5. J. Allan, J. Ewing, R. Helmig, and J. Braun. Scale effects in multiphase flow modeling. In G. Wickramanayake and R. Hinchee, editors, *1. International conference on remediation of chlorinated and recalcitrant compounds*, Monterey, California, USA, 18th–21st of may 1998. Battelle Press, Columbus, OH, USA.
6. K. Aziz and A. Settari. *Petroleum Reservoir Simulation*. Applied Science Publishers, London, 1979.
7. P. Bastian, K. Birken, S. Lang, K. Johannsen, N. Neuss, H. Rentz-Reichert, and C. Wieners. *UG: A flexible software toolbox for solving partial differential equations.*, volume 1 of *Computing and Visualization in Science*, pages 27–40. Springer Verlag, 1997.
8. P. Bastian, Z. Chen, R. E. Ewing, R. Helmig, J. H., and R. V. *Numerical Simulation of Multiphase Flow in Fractured Porous Media.*, pages 52–71. Lecture Notes in Physics, Chen, Ewing and Shi (eds.). Springer Verlag, 2000.
9. P. Bastian and R. Helmig. Efficient fully-coupled solution techniques for two phase flow in porous media. parallel multigrid solution and large scale computations. *Advances in Water Resources*, 1999.
10. A. Bielinski. *Numerical Simulation of CO_2 Sequestration in Geological Formations*. PhD thesis, Institut für Wasserbau, Universität Stuttgart, 2006.
11. C. Braun, R. Helmig, and S. Manthey. Determination of constitutive relationships for two-phase flow processes in heterogeneous porous media with emphasis on the relative permeability-saturation-relationship. *Journal of Contaminant Hydrology*, 2005.

12. H. Class. *Theorie und numerische Modellierung nichtisothermer Mehrphasenprozesse in NAPL-kontaminierten porösen Medien*, volume 105 of *Mitteilungsheft*. Institut für Wasserbau, Universität Stuttgart, 2001.
13. H. Class and R. Helmig. Numerical simulation of nonisothermal multiphase multicomponent processes in porous media – 2. applications for the injection of steam and air. *Advances in Water Resources*, 25:551–564, 2002.
14. H. Class, R. Helmig, and P. Bastian. Numerical simulation of nonisothermal multiphase multicomponent processes in porous media – 1. an efficient solution technique. *Advances in Water Resources*, 25:533–550, 2002.
15. K. Coats, W. Chieh Chu, and B. Marcum. Three-dimensional simulation of steamflooding. *Society of Petroleum Engineers Journal*, December 1974.
16. M. De Neef and J. Molenaar. Analysis of dnapl infiltration in a medium with a low-permeable lens. *Computational Geosciences*, 1:191–214, 1997.
17. M. G. Edwards and C. F. Rogers. A flux continuous scheme for the full tensor pressure equation. In *Proc. of the 4th European Conf. on the Mathematics of Oil Recovery*, Norway, 1994.
18. M. Emmert. *Numerische Simulation von isothermen/nichtisothermen Mehrphasenprozessen unter Berücksichtigung der Veränderung der Fluideigenschaften*. PhD thesis, Institut für Wasserbau, Universität Stuttgart, 1997.
19. R. Falta, K. Pruess, I. Javandel, and P. Witherspoon. Numerical modeling of steam injection for the removal of nonaqueous phase liquids from the subsurface. 1. numerical formulation. *Water Resoures Research*, 28,2:433–449, 1992.
20. P. Forsyth. Three dimensional modeling of steam flush for dnapl site remediation. Technical report, Dep. of Computer Science, University of Waterloo, 1993. CS-93-56.
21. T. Gimse and N. H. Risebro. Riemann problems with a discontinuous flux function. In *Proc. 3rd Internat. Conf. Hyperbolic Problems*, pages 488–502, Uppsala, 1991.
22. T. Gimse and N. H. Risebro. Solution of the cauchy problem for a conservation law with a discontinuous flux function. *SIAM J. Math. Anal.*, 23(3):635–648, 1992.
23. C. Grass. Untersuchung von randbedingungen bei der numerischen simulation von zweiphasenströmungen in porösen medien. Master's thesis, Institut für Wasserbau, Universität Stuttgart, January 2005.
24. R. Helmig. *Multiphase Flow and Transport Processes in the Subsurface — A Contribution to the Modeling of Hydrosystems*. Springer Verlag, 1997.
25. R. Helmig, H. Class, H. Jakobs, A. Bierlinski, and U. Ölmann. *Arbeits- und Ergebnisbericht 2003*, chapter A3, pages 69–98. SFB 404, May 2003.
26. R. Helmig, C. T. Miller, H. Jakobs, H. Class, M. Hilpert, C. E. Kees, and J. Niessner. Multiphase Flow and Transport Modeling in Heterogeneous Porous Media. In A. Di Bucchianico, R. M. M. Mattheij, and M. A. Peletier, editors, *Progress in Industrial Mathematics at ECMI 2004*, pages 449–488, Eindhoven University of Technology, 6 2006. Springer-Verlag. 3-540-28072-3.
27. S. Hölzemann, H. Class, and R. Helmig. A new concept for the numerical simulation and parameter identification of multiphase flow and transport processes in cohesive soils. In T. Schanz, editor, *Unsaturated Soils: Numerical and Theoretical Approaches – Proceedings of the International Conference "From Experimental Evidence towards Numerical Modelling of Unsaturated Soils" (18. - 19. September 2003, Bauhaus-Universität Weimar)*. Springer-Verlag, 2004. ISBN: 3-540-21122-5.

28. IAPWS (The International Association for the Properties of Water and Steam). Revised release on the iaps formulation 1985 for the viscosity of ordinary water substance. http://www.iapws.org/, 2003.
29. P. R. King. Upscaling permeability: Error analysis for renormalisation. *Transport in Porous Media*, 23:337–354, 1996.
30. P. Lancaster. *Theory of Matrices*. Academic Press, Inc. (London) Ltd., 1969.
31. R. J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser Verlag, Basel Boston, Berlin, 1998.
32. B. B. Looney and R. W. Falta. *Vadose Zone*. Batelle Press, Columbus OH, 2000.
33. J. Nordbotten, M. Celia, and S. Bachu. Injection and storage of CO_2 in deep saline aquifers: Analytical solution for CO_2 plume evolution during injection. *Transport in Porous Media*, 58(3):339–360, 2005.
34. J. Nordbotten, M. Celia, S. Bachu, and H. Dahle. Semi-analytical solution for CO_2 leakage through an abandoned well. *Environmental Science and Technology*, 39(2):602–611, 2005.
35. J. M. Nordbotten, I. Aavatsmark, and G. T. Eigestad. Monotonicity of control volume methods. submitted to Numerische Mathematik, 2005.
36. U. Ölmann. *Behandlung anisotroper Mobilitäten als Resultat von Upscalingverfahren mittels Mehrpunktflußapproximationen*. PhD thesis, Institut für Wasserbau, Universität Stuttgart, to be published 2006.
37. U. Ölmann, I. Aavatsmark, and R. Helmig. Buckley-Leverett heterogen — Konstruktion der Lösung mit der Charakteristikenmethode. Preprint-Reihe des SFB404, March 2006. 2006/05.
38. R. Reid, J. Prausnitz, and B. Poling. *The Properties of Gases and Liquids*. McGraw-Hill Inc., 1987.
39. F. Stauffer and A. Aharnony. *Introduction to Percolation Theory*. Taylor & Francis, 1994.
40. X. H. Wen and J. J. Gómez-Hernández. Upscaling hydraulic conductivities in heterogenous media: An overview. *Journal of Hydrology*, 183:ix–xxxii, 1996.
41. S. Whitaker. *The Method of Volume Averaging*, volume 13 of *Theory and Applications of Transport in Porous Media*. Kluwer Academic Publishers, Dordrecht, 1999.
42. J. K. Williams. Simple renormalisation schemes for calculating effective properties of heterogeneous reservoirs. *1st European Conference on the Mathematics of Oil Recovery, Cambridge, UK, July 1989*, 1989.

Localization Analysis of Porous Granular Materials*

Wolfgang Ehlers and Bernd Scholz

Institute of Applied Mechanics (CE), University of Stuttgart, Pfaffenwaldring 7, 70569 Stuttgart, Germany
ehlers@mechbau.uni-stuttgart.de
scholz@mechbau.uni-stuttgart.de

Summary. In the present article the work of the project A6 of the SFB 404 is presented, which is concerned with the description of localization phenomena in granular material. The entire work in this project can be divided in two parts. In the first part, a thermodynamic consistent formulation of the micropolar theory in the framework of the Theory of Porous Media is developed. This part includes also the numerical implementation in a finite element code and the computation of some exemplary boundary-value problems. With regard to the application of the model, i. e., the computation of realistic problems, the work in the second part is devoted to the identification of the material parameters included in the model. The main challenge thereby is the determination of the additional parameters associated with the micropolar theory, which requires a fully inverse computation in the sense of a backward analysis of the underlying boundary-value problem. In the present article, this procedure is carried out by use of a semi-discrete sensitivity analysis. Finally, the whole model is applied to the experimental data of Hostun sand obtained at the Universities of Grenoble and Stuttgart.

Keywords: Granular material, Cosserat theory, Theory of Porous Media, parameter identification, sensitivity analysis

1 Introduction

In the project A6 of the SFB 404, the localization phenomena of granular material is analyzed. On the micro level, such materials exhibit a complex stress-strain behavior governed by the local grain deformations as well as by the translational and rotational degrees of freedom of the grain motion associated with contact forces and contact moments. Proceeding from homogenization techniques applied to representative elementary volumes (REV) of

* Research Project A6 "Material Behavior and Localization Phenomena in the Theory of Porous Media (TPM)"

particle ensembles, the average of the local forces and moments of the microstructure results in macroscopic stresses and couple stresses, thus basically defining granular materials as a micropolar continuum in the sense of the Cosserat brothers [4]. Concerning the homogenization procedure of particle ensembles towards micropolar continua, the interested reader is referred to the work by Ehlers et al. [15] and the quotations therein. While the complex stress-strain behavior of granular materials like sand has been observed and described within the range of cohesive-frictional, elasto-plastic materials, e. g., by Lade and Duncan [23], by Yamada and Ishihara [39], Desrues et al. [6], or by Lade and Kim [21, 24, 25], there has been no hint to micropolar rotations within the usual range of applied deformations. This result is not surprising, since standard experiments, e. g., on cylindrical samples tested in triaxial devices are carried out as homogeneous as possible in order to interpret the corresponding stress-strain curves as the governing material behavior. On the other hand, real samples as well as real boundary conditions in real experiments are not ideally homogeneous. As a result, one observes localization, bifurcation and instability phenomena after having passed a maximally applicable external load. A careful investigation of these phenomena both on the microscopic and on the macroscopic level reveals that inhomogeneous deformations result in local concentrations of plastic strains in narrow zones forming shear bands in 2-dimensional applications like the biaxial test and shear domains in general 3-dimensional problems. Furthermore it has been observed that with the onset of shear zones, there is a local switch of the material behavior from non-polar to micropolar. This behavior is due to the fact that the grains of granular materials start to roll upon each other, thus initiating rotational degrees of freedom on the micro level corresponding to micropolar rotations on the macro level. In the literature, these effects are described on the experimental side by, e. g., Viggiani et al. [37] and on the numerical side by, e. g., Ehlers and Volk [16]. From these articles and the related work on micropolar material, cf., e. g., the works by Kafadar and Eringen [20], by Nowacki [33] or by Mühlhaus and Vardoulakis [31], it is clearly seen that granular materials behave, in a macroscopic setting, as a standard non-polar continuum as far as there are no localizing effects and as a micropolar continuum, whenever localization occurs. Since, in general, the location of upcoming localizations is a priori completely unknown and must be found, e. g., during numerical computations by the applied algorithm, the overall model of granular materials must basically be understood as a micropolar continuum with elasto-plastic material properties. As a result of this procedure, there is no pathological mesh dependency in numerical computations proceeding from the Finite Element Method (FEM), which is well-known to occur within the framework of the standard non-polar continuum, where the shear band width is shrinking with the mesh size up to a singularity. In contrast to this unphysical behavior, the real shear band width of granular materials range about 10 to 30 average particle diameters. Following this, the shear band width can be used for the determination of the material parameters of the micropolar part

of the overall constitutive model. Finally, granular materials contain voids, filled with a pore fluid (usually air or water), and must thus be described within the well-founded framework of the Theory of Porous Media (TPM). Concerning a general overview of the different approaches to describe porous materials, the reader is referred to the works by Biot [2], Bowen [3], de Boer [5] and Ehlers [8, 10].

Describing cohesive-frictional, elasto-plastic material in the framework of a micropolar continuum requires the identification of a variety of material parameters. Given a natural, non-preloaded state, granular materials exhibit elasto-plastic behavior starting at the onset of deformation. On the other hand, unloading and reloading circles reveal purely elastic behavior, however, in a physically non-linear manner, even in the geometrically linear domain [14, 32]. In the framework of plasticity, granular materials fall in the category of non-associate plasticity models [24, 25, 8]. Following this, one has not only to define an appropriate yield function, which, in the present article, is taken as the single-surface yield criterion by Ehlers [9], but one additionally needs a plastic potential function to define the direction of the plastic evolution. Furthermore, the parameters of the plastic range are subjected to hardening properties, thus leading to additional material parameters that have to be determined. Finally, the standard elasto-plasticity model of porous materials has to be extended towards micropolar material properties yielding further independent parameters. In conclusion, the overall model discussed in this article is governed by a total set of 23 material parameters. The determination of such an amount of parameters generally causes considerable problems in the course of the identification procedure. Based on measured experimental data, it is the goal of any identification process to find an optimal set of material parameters in the sense of the best approximation. Following this leads to an inverse problem within the wide range of Nonlinear Optimization, cf., e. g., the works by Luenberger [26] or by Heinz and Spellucci [19]. The methods described therein have been applied to the parameter identification of inelastic problems, e. g., by Mahnken and Stein [27], Mahnken and Steinmann [29] or by Ehlers and Müllerschön [14]. A general overview of the different approaches to identification methods can be found, e. g., in the thesis by Thielecke [35].

Proceeding from the considerable number of material parameters of the model under discussion, it is necessary to come along with a partitioned identification procedure governed by two basic steps. These steps fall into the determination of the parameters corresponding to the non-polar part, on the one hand, and to the micropolar part of the material behavior, on the other hand. While the first set of material parameters can be obtained from homogeneous tests like the triaxial experiment, the second set must be determined from inhomogeneous tests like the biaxial experiment, where the onset and the development of shear bands must be included. Especially the last step yields considerable numerical costs. Although one is generally interested in the determination of the global minimum, stochastic methods like the group of genetic algorithms must be excluded as well as the method of neural net-

works because of their enormous computational costs. Following this, a deterministic, gradient-based method is applied, although this method, depending on the initial trial set of parameters, generally runs into local minima rather than into the global minimum. Proceeding from a gradient-based method, the computational costs can furthermore be reduced by the application of the semi-analytical sensitivity analysis, where use is made of the exact derivative of the set of discrete system equations. For further details, the interested reader is referred to the works by Adelmann and Hafka [1] or by Kleiber [22].

In the present contribution, we start with the basic equations describing granular, micropolar, cohesive-frictional porous materials in the framework of the geometrically linear elasto-plasticity (Sect. 2). This includes the kinematics of micropolar materials, the balance equations of mass, linear and angular momentum as well as the constitutive equations of porous media elasto-plasticity. In a second step (Sect. 3), the discretization of the model is discussed in detail prior to the description of the parameter identification procedure applied to the material under study (Sect. 4). Furthermore, the method of the sensitivity analysis technique is explained in (Sect. 5) followed by the numerical examples (Sect. 6). Therein on the one hand the effects of the micropolar theory are shown on two exemplary boundary-value problems and one the other hand the identification of the parameters of micropolar material behavior is illustrated.

2 Basic Equations

In this section, the basic model for the description of granular materials is presented in the frame of continuum mechanics. As a result of the particle rotations of granular materials, the model represents a micropolar solid, where in addition to the translational degrees of freedom of standard non-polar continua, a supplementary and independent rotational field is considered. Furthermore, since the homogenization of a granular microstructure towards a macroscopic continuum results in a porous solid material, the basic setting of a granular continuum is embedded in the well-founded Theory of Porous Media (TPM).

2.1 Kinematics of Micropolar Porous Materials

In the framework of the TPM, a porous solid skeleton filled by a pore fluid (liquid or gas) is described by application of the concept of superimposed continua, where every single constituent φ^α follows its own motion function

$$\mathbf{x} = \boldsymbol{\chi}_\alpha(\mathbf{X}_\alpha, t)\,. \tag{1}$$

Thus, each constituent with the reference position $\mathbf{X}_\alpha$ at time t_0 occupies the same actual position $\mathbf{x}$. Furthermore, in the framework of the TPM the

concept of volume fractions is introduced, whereby the single volume fractions n^α are defined as the local ratios of the volume element $\mathrm{d}v^\alpha$ of the constituent φ^α with respect to the volume element of the mixture $\mathrm{d}v$:

$$n^\alpha = \frac{\mathrm{d}v^\alpha}{\mathrm{d}v}\,. \tag{2}$$

Since no vacant space is allowed in the overall medium, the summation over all volume fractions yields the saturation condition:

$$\sum_{\alpha=1}^{k} n^\alpha = 1\,. \tag{3}$$

Furthermore, with the definition of the volume fractions (2) two different density functions can be introduced, the material density $\rho^{\alpha R}$ as the realistic density of the constituent φ^α and the partial density ρ^α as the density of φ^α with respect to the bulk volume of the mixture $\mathrm{d}v$. Both density functions are coupled via the volume fractions n^α:

$$\rho^\alpha = \frac{\mathrm{d}m^\alpha}{\mathrm{d}v} = n^\alpha \frac{\mathrm{d}m^\alpha}{\mathrm{d}v^\alpha} = n^\alpha \rho^{\alpha R}\,. \tag{4}$$

Therewith, it is obvious that the incompressibility of the single constituents is not equivalent with the incompressibility of the mixture, as the partial densities can still change through changes in the volume fractions n^α.

Following (1), each constituent has its own velocity and acceleration field, which can be expressed in a Langrangean description

$$\overset{\prime}{\mathbf{x}}_\alpha = \frac{\partial \boldsymbol{\chi}_\alpha(\mathbf{X}_\alpha, t)}{\partial t} \quad \text{and} \quad \overset{\prime\prime}{\mathbf{x}}_\alpha = \frac{\partial^2 \boldsymbol{\chi}_\alpha(\mathbf{X}_\alpha, t)}{\partial t^2} \tag{5}$$

or, by use of the inverse motion function $\mathbf{X} = \boldsymbol{\chi}_\alpha^{-1}(\mathbf{x}, t)$ in the Eulerian notation

$$\mathbf{v}_\alpha(\mathbf{x},\, t) = \overset{\prime}{\mathbf{x}}_\alpha\,(\boldsymbol{\chi}_\alpha^{-1}(\mathbf{x},\, t),\, t) \quad \text{and} \quad (\mathbf{v}_\alpha)_\alpha(\mathbf{x},\, t) = \overset{\prime\prime}{\mathbf{x}}_\alpha\,(\boldsymbol{\chi}_\alpha^{-1}(\mathbf{x},\, t),\, t)\,. \tag{6}$$

Therein, the symbol $(\cdot)'_\alpha$ denotes the material time derivative following the motion of the constituent φ^α.

Describing coupled solid-fluid problems, the motion of the solid skeleton is commonly described by use of a Langrangean setting with the displacement vector $\mathbf{u}_S$, whereas the fluid is described by use of a modified *Euler*ian method with the seepage velocity $\mathbf{w}_F$:

$$\mathbf{u}_S = \mathbf{x} - \mathbf{X}_S\,, \quad \mathbf{w}_F = \overset{\prime}{\mathbf{x}}_F - \overset{\prime}{\mathbf{x}}_S\,. \tag{7}$$

In the framework of the micropolar theory, which is applied for the description of the granular solid skeleton, a material point is seen as a rigid micro

body. Thus, additionally to the motion function (1), a second kinematic mechanism, the so-called micromotion

$$\boldsymbol{\xi}_S \;=\; \bar{\mathbf{R}}_S\,(\mathbf{X}_S,\,t)\,\boldsymbol{\Xi}_S\,, \tag{8}$$

has to be considered in the kinematic description, cf. Fig. 1. Therein, $\mathbf{x}$ and $\mathbf{X}_S$ represent the locations of the material point $\mathcal{P}_S$ in the actual configuration at time t and in the reference configuration at time t_0. Furthermore, $\boldsymbol{\xi}_S$ and $\boldsymbol{\Xi}_S$ can be understood as rigid directors fixed to the material points in order to represent the local orientation of $\mathcal{P}_S$. While the motion is given by the vector function $\boldsymbol{\chi}_S\,(\mathbf{X}_S,\,t)$, the micromotion is governed by the proper orthogonal tensor function $\bar{\mathbf{R}}_S\,(\mathbf{X}_S,\,t)$. Following this, the micromotion describes the homogenized image of the change of the particle orientation in the different configurations around the angle $\bar{\varphi}_S$.

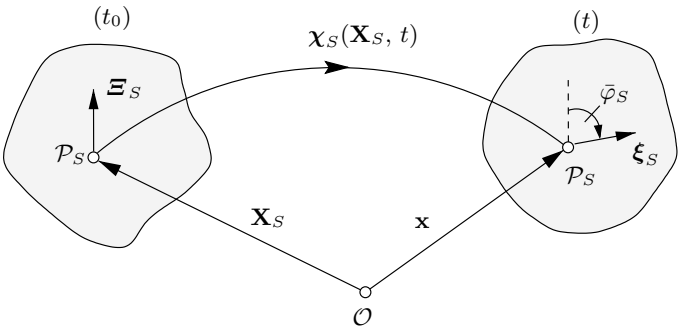


Fig. 1. Micromotion of a micropolar body $\mathcal{B}$

Given both the motion and the micromotion functions, the deformation gradient can be expressed by

$$\mathbf{F}_S \;=\; \mathrm{Grad}_S\,\boldsymbol{\chi}(\mathbf{X}_S,\,t) \;=\; \mathbf{R}_S\,\mathbf{U}_S \;=\; \bar{\mathbf{R}}_S\,\bar{\mathbf{U}}_S\,. \tag{9}$$

Therein, the operator "$\mathrm{Grad}_S\,(\,\cdot\,)$" denotes the partial derivative of $(\,\cdot\,)$ with respect to the reference position $\mathbf{X}_S$ of φ^S. Furthermore, the above representations of $\mathbf{F}_S$ exhibit the polar decompositions of the deformation gradient in case of nonpolar and of micropolar materials. In addition to the micromotion $\bar{\mathbf{R}}_S$, $\mathbf{R}_S$ is the proper orthogonal continuum rotation, while $\mathbf{U}_S$ and $\bar{\mathbf{U}}_S$ are the right stretch tensors of non-polar and micropolar materials. Note in passing that $\mathbf{U}_S$ is symmetric while $\bar{\mathbf{U}}_S$ is necessarily not. Proceeding from an Euler-Rodrigues representation of the spatial rotation, $\mathbf{R}_S$ and $\bar{\mathbf{R}}_S$ can be given by

$$\begin{aligned}
\mathbf{R}_S &= \mathbf{e}_S \otimes \mathbf{e}_S + (\,\mathbf{I} - \mathbf{e}_S \otimes \mathbf{e}_S)\cos\varphi_S + (\mathbf{e}_S \times \mathbf{I}\,)\sin\varphi_S\,,\\
\bar{\mathbf{R}}_S &= \bar{\mathbf{e}}_S \otimes \bar{\mathbf{e}}_S + (\,\mathbf{I} - \bar{\mathbf{e}}_S \otimes \bar{\mathbf{e}}_S)\cos\bar{\varphi}_S + (\bar{\mathbf{e}}_S \times \mathbf{I}\,)\sin\bar{\varphi}_S\,,
\end{aligned} \tag{10}$$

where $\mathbf{e}_S \times \mathbf{I} = -\overset{3}{\mathbf{E}}\,\mathbf{e}_S$ as well as $\bar{\mathbf{e}}_S \times \mathbf{I}$ are skew-symmetric. Therein, $\overset{3}{\mathbf{E}}$ characterizes the Ricci permutation tensor and $\mathbf{I}$ the second-order identity. Furthermore, $\mathbf{e}_S$ and $\bar{\mathbf{e}}_S$ represent the axes of rotation such that the continuum and the micropolar rotations can be expressed by the vectors

$$\boldsymbol{\varphi}_S = \varphi_S\,\mathbf{e}_S \quad \text{and} \quad \bar{\boldsymbol{\varphi}}_S = \bar{\varphi}_S\,\bar{\mathbf{e}}_S\,. \tag{11}$$

Proceeding from micropolar materials, $\bar{\mathbf{R}}_S$ describes the rotation of the directors or the microparticles as a joint rotation of the line elements and the independent particle rotations. Since the line elements rotate through $\mathbf{R}_S$, the additional micropolar rotation can be given by

$$\overset{*}{\mathbf{R}}_S = \overset{*}{\mathbf{e}}_S \otimes \overset{*}{\mathbf{e}}_S + (\mathbf{I} - \overset{*}{\mathbf{e}}_S \otimes \overset{*}{\mathbf{e}}_S)\cos\overset{*}{\varphi}_S + (\overset{*}{\mathbf{e}}_S \times \mathbf{I})\sin\overset{*}{\varphi}_S\,. \tag{12}$$

Assuming $\bar{\mathbf{R}}_S$ to be a result of $\mathbf{R}_S$ and $\overset{*}{\mathbf{R}}_S$ through

$$\bar{\mathbf{R}}_S = \mathbf{R}_S\,\overset{*}{\mathbf{R}}_S\,, \tag{13}$$

$\bar{\mathbf{U}}_S$ can be expressed as a function of $\overset{*}{\mathbf{R}}_S$ and $\mathbf{U}_S$:

$$\bar{\mathbf{U}}_S = \overset{*}{\mathbf{R}}{}^T_S\mathbf{U}_S\,. \tag{14}$$

In the frame of micropolar materials, the Green-Lagrangean strain of the standard continuum formulation is substituted by the so-called first Cosserat strain together with the curvature tensor of third order:

$$\bar{\mathbf{E}}_S = \bar{\mathbf{U}}_S - \mathbf{I}\,, \qquad \overset{3}{\bar{\mathbf{C}}}_S = (\bar{\mathbf{R}}^T_S\,\mathrm{Grad}_S\,\bar{\mathbf{R}}_S)\underline{^3}\,. \tag{15}$$

Note in passing, that in the present considerations, $\bar{\mathbf{E}}_S$ and $\overset{3}{\bar{\mathbf{C}}}_S$ are given with respect to the referential frame. Furthermore, the symbol $(\,\cdot\,)\underline{^i}$ defines a contraction of the arguments in parentheses towards a tensor of i-th order. Concerning the introduction of further curvature measures, the reader is referred to the work by Ehlers [10], Steinmann [34] and Volk [38]. Given $(15)_2$, it is concluded that the curvature tensor is skew-symmetric with respect to the first two basis systems. Following this results in

$$\overset{3}{\bar{\mathbf{C}}}_S = -\,\overset{3}{\bar{\mathbf{C}}}{}^{12}_S{}^T\,. \tag{16}$$

Based on the skew-symmetry condition, the axial tensor corresponding to $\overset{3}{\bar{\mathbf{C}}}_S$ reads

$$\mathrm{axl}\,\overset{3}{\bar{\mathbf{C}}}_S =: \overset{A}{\bar{\mathbf{C}}}_S = \tfrac{1}{2}(\overset{3}{\mathbf{E}}\,\overset{3}{\bar{\mathbf{C}}}{}^{12}_S{}^T)\underline{^2}\,, \tag{17}$$

thus defining the curvature tensor of second order.

Restricting the following considerations to geometrically linear applications, a formal linearization of $\bar{\mathbf{R}}_S$ and $\bar{\mathbf{E}}_S$ around the natural undeformed and unrotated state results in

$$\bar{\mathbf{R}}_{S\,\mathrm{lin.}} = \mathbf{I} - \overset{3}{\mathbf{E}}_S\,\bar{\boldsymbol{\varphi}}_S \quad \text{and} \quad \bar{\mathbf{E}}_{S\,\mathrm{lin.}} = \mathbf{H}_S + \overset{3}{\mathbf{E}}\,\bar{\boldsymbol{\varphi}}_S\,, \tag{18}$$

where $\mathbf{H}_S = \mathrm{Grad}_S\,\mathbf{u}_S$ is the displacement gradient with the displacement vector $\mathbf{u}_S = \mathbf{x} - \mathbf{X}_S$. In addition to $(18)_1$, a linearization of the continuum rotation and the additional micropolar rotation yields

$$\mathbf{R}_{S\,\mathrm{lin.}} = \mathbf{I} - \overset{3}{\mathbf{E}}\,\boldsymbol{\varphi}_S \quad \text{and} \quad \overset{*}{\mathbf{R}}_{S\,\mathrm{lin.}} = \mathbf{I} - \overset{3}{\mathbf{E}}\,\overset{*}{\boldsymbol{\varphi}}_S\,. \tag{19}$$

Finally, one obtains the formal linearization of the curvature tensor as

$$\bar{\boldsymbol{\kappa}}_S := \overset{A}{\bar{\mathbf{C}}}_{S\,\mathrm{lin.}} = \tfrac{1}{2}(\overset{3}{\mathbf{E}}\,\overset{3}{\bar{\mathbf{C}}}_{S\,\mathrm{lin.}}) = \mathrm{Grad}_S\,\bar{\boldsymbol{\varphi}}_S\,, \tag{20}$$

thus recovering the classical result $\bar{\boldsymbol{\kappa}}_S = \mathrm{Grad}_S\bar{\boldsymbol{\varphi}}_S$. Concerning the formal linearization process to obtain (18)–(20), the interested reader is referred to Ehlers [10]. In addition to $(18)_2$, the symmetric and skew-symmetric parts of the Cosserat strain read

$$\bar{\boldsymbol{\varepsilon}}_S := \bar{\mathbf{E}}_{S\,\mathrm{lin.}}\,, \quad \text{where} \quad \begin{cases} \bar{\boldsymbol{\varepsilon}}_{S\,\mathrm{sym}} = \frac{1}{2}\,(\mathbf{H}_S + \mathbf{H}_S^T)\,, \\ \bar{\boldsymbol{\varepsilon}}_{S\,\mathrm{skw}} = \frac{1}{2}\,(\mathbf{H}_S - \mathbf{H}_S^T) + \overset{3}{\mathbf{E}}\,\bar{\boldsymbol{\varphi}}_S\,. \end{cases} \tag{21}$$

In this representation, $\bar{\boldsymbol{\varepsilon}}_{S\,\mathrm{sym}}$ equals the linearized Lagrangean strain $\boldsymbol{\varepsilon}_S$ of non-polar materials, whereas

$$\bar{\boldsymbol{\varepsilon}}_{S\,\mathrm{skw}} = \overset{3}{\mathbf{E}}\,(\bar{\boldsymbol{\varphi}}_S - \boldsymbol{\varphi}_S) = \overset{3}{\mathbf{E}}\,\overset{*}{\boldsymbol{\varphi}}_S \tag{22}$$

exhibits a tensorial measure for the additional micropolar rotation $\overset{*}{\boldsymbol{\varphi}}_S$. Note in passing that, in contrast to the geometrically non-linear approach, where the rotation vectors are non-additive pseudo-vectors, an addition of the geometrically linear rotation measures yields

$$\bar{\boldsymbol{\varphi}}_S = \boldsymbol{\varphi}_S + \overset{*}{\boldsymbol{\varphi}}_S\,. \tag{23}$$

This result is easily concluded from (13) together with the linearizations of $\bar{\mathbf{R}}_S$, $\mathbf{R}_S$ and $\overset{*}{\mathbf{R}}_S$, cf. $(18)_1$ and (19).

2.2 Balance Equations

The balance equations of a multiphasic material differ from the balance equations of a single phase material by the appearance of interaction terms, which occur as production terms in the balance equations of the individual phases

φ^α, cf. e. g. [3]. Furthermore, in consequence of the use of the micropolar theory, the balance equations have to be extended, particularly, the balance of moment of momentum (m. o. m.), cf. [34] or [10]. Restricting the following considerations to isothermal problems, one is mainly interested in the mechanical balances of mass, momentum and angular momentum. These balances read:

$$\begin{aligned}
&\bullet \text{ mass:} && (\rho^\alpha)'_\alpha + \rho^\alpha \operatorname{div} \overset{\prime}{\mathbf{x}}_\alpha = \hat{\rho}^\alpha \,, \\
&\bullet \text{ momentum:} && \rho^\alpha \overset{\prime\prime}{\mathbf{x}}_\alpha = \operatorname{div} \mathbf{T}^\alpha + \rho^\alpha \mathbf{b}^\alpha + \hat{\mathbf{p}}^\alpha \,, \\
&\bullet \text{ m. o. m.:} && \rho^\alpha (\bar{\boldsymbol{\Theta}}_\alpha \bar{\boldsymbol{\omega}}_\alpha)'_\alpha = \mathbf{I} \times \mathbf{T}^\alpha + \operatorname{div} \mathbf{M}^\alpha + \rho^\alpha \mathbf{c}^\alpha + \hat{\mathbf{m}}^\alpha \,.
\end{aligned} \tag{24}$$

In the above equations, $\hat{\rho}^\alpha$ and $\hat{\mathbf{p}}^\alpha$ are the production terms for the mass and the momentum, respectively. Furthermore, $\mathbf{T}^\alpha$ is the Cauchy stress, and $\mathbf{M}^\alpha$ is the couple stress, while $\mathbf{b}^\alpha$ and $\mathbf{c}^\alpha$ generally denote the body forces and the body couples per unit mass, $\bar{\boldsymbol{\Theta}}^\alpha$ is the tensor of microinertia per unit mass, and $\bar{\boldsymbol{\omega}}_\alpha = (\bar{\boldsymbol{\varphi}}_\alpha)'_\alpha$ is the total rotational velocity. Finally, $\operatorname{div}(\,\cdot\,)$ is the divergence operator corresponding to the spatial gradient $\operatorname{grad}(\,\cdot\,) = \operatorname{Grad}_\alpha(\,\cdot\,)\,\mathbf{F}_\alpha^{-1}$. As a result of the existence of couple stresses and the effect of microinertia, it is directly concluded from $(24)_3$ that $\mathbf{T}^\alpha$ is generally non-symmetric.

Assuming quasi static conditions, where $\overset{\prime\prime}{\mathbf{x}}_\alpha$ and $\bar{\boldsymbol{\omega}}_\alpha$ vanish together with its time derivative, and the considered special case of a micropolar solid skeleton combined with a non-polar pore fluid without any mass exchange ($\hat{\rho}^\alpha \equiv 0$) the above set of equations reduces to

$$\begin{aligned}
&\bullet \text{ mass:} && (\rho^\alpha)'_\alpha + \rho^\alpha \operatorname{div}(\overset{\prime}{\mathbf{x}}_\alpha) = 0 \\
&\bullet \text{ momentum:} && \mathbf{0} = \operatorname{div} \mathbf{T}^\alpha + \rho^\alpha \mathbf{g} + \hat{\mathbf{p}}^\alpha, \\
&\bullet \text{ m. o. m. (solid skeleton):} && \mathbf{0} = \mathbf{I} \times \mathbf{T}^S + \operatorname{div} \mathbf{M}^{\mathrm{S}} \;\longrightarrow\; \mathbf{T}^{\mathrm{S}} \neq (\mathbf{T}^{\mathrm{S}})^{\mathrm{T}} \,.
\end{aligned} \tag{25}$$

To obtain (25) from (24), it has been assumed that the body forces $\mathbf{b}^\alpha$ of all constituents are equal to the gravitation $\mathbf{g}$ and $\mathbf{c}^S$ is set to zero.

By considering materially incompressible constituents, one obtains from the sum of the mass balances over all constituents φ^α, by use of the saturation condition (3) and the concept of the volume fractions (2), an overall volume balance relation, cf. [8]. Thus, finally, restricting the considerations to the geometrically linear approach yields

$$\begin{aligned}
&\bullet \text{ overall volume:} && \operatorname{div}[n^F \mathbf{w}_F + (\mathbf{u}_S)'_S] = 0 \,, \\
&\bullet \text{ momentum:} && \mathbf{0} = \operatorname{div} \boldsymbol{\sigma}^\alpha + \rho^\alpha \mathbf{g} + \hat{\mathbf{p}}^\alpha \,, \\
&\bullet \text{ m. o. m. (solid skeleton):} && \mathbf{0} = \mathbf{I} \times \boldsymbol{\sigma}^S + \operatorname{div} \boldsymbol{\mu}^{\mathrm{S}} \;\longrightarrow\; \boldsymbol{\sigma}^{\mathrm{S}} \neq (\boldsymbol{\sigma}^{\mathrm{S}})^{\mathrm{T}} \,,
\end{aligned} \tag{26}$$

where $\boldsymbol{\sigma}^\alpha$ and $\boldsymbol{\mu}^S$ are the geometrically linear versions of $\mathbf{T}^\alpha$ and $\mathbf{M}^S$.

2.3 Constitutive Equations

For the complete description of the considered two phase material, additionally to the above kinematic and balance equations constitutive equations are required, which give a relation between the geometrically linear strain and curvature measurements of the solid skeleton ($\bar{\boldsymbol{\varepsilon}}_S$ and $\bar{\boldsymbol{\kappa}}_S$) and the corresponding stress and couple stresses ($\boldsymbol{\sigma}^S$ and $\boldsymbol{\mu}^S$). Furthermore, constitutive equations for the stress of the fluid $\boldsymbol{\sigma}^F$ and the momentum production term $\hat{\mathbf{p}}^F$ are needed.

With the assumption of material incompressibility for both constituents, the following relations are valid:

$$\begin{aligned} \boldsymbol{\sigma}^\alpha &= -n^\alpha\, p\, \mathbf{I} + \boldsymbol{\sigma}^\alpha_E\,, \\ \hat{\mathbf{p}}^F &= p\, \operatorname{grad} n^F + \hat{\mathbf{p}}^F_E = -\hat{\mathbf{p}}^S\,. \end{aligned} \tag{27}$$

Therein, p represents the pore pressure and the quantities indicated by $(\cdot)_E$ are the so-called extra quantities, which have to be specified by constitutive equations. Firstly, by a dimensional analysis it can be shown that the extra stress of the fluid $\boldsymbol{\sigma}^F_E$, which results from the fluid viscosity, can be neglected in relation to the momentum production $\hat{\mathbf{p}}^F_E$, cf. [11]. Thus, the fluid viscosity is only considered in the extra term of the momentum production representing the local interaction force between the solid skeleton and the pore fluid,

$$\hat{\mathbf{p}}^F_E = -(n^F)^2\, \frac{\gamma^{FR}}{k^F}\, \mathbf{w}_F\,. \tag{28}$$

Therein, $\gamma^{FR} = \rho^{FR}\,|\mathbf{g}|$ is the real specific weight of the pore fluid and k^F represents the Darcy permeability coefficient.

The constitutive equations of the solid skeleton are formulated in the framework of the geometrically linear theory of elasto-plasticity. This approach requires the split of the deformation in elastic and plastic parts. In the framework of the general non-linear approach, this is usually done by a multiplicative split of the deformation gradient connected with the image of a plastic intermediate configuration, compare, e. g., Ehlers [7] or Haupt [18] and the quotations therein. Based on this multiplicative split, an additive decomposition of strain tensors can be obtained, which in the geometrically linear frame results in the usual additive decomposition of the linear Lagrangean strain tensor. With respect to micropolar materials, this decomposition reads

$$\bar{\boldsymbol{\varepsilon}}_S = \bar{\boldsymbol{\varepsilon}}_{Se} + \bar{\boldsymbol{\varepsilon}}_{Sp} \quad \text{and} \quad \bar{\boldsymbol{\kappa}}_S = \bar{\boldsymbol{\kappa}}_{Se} + \bar{\boldsymbol{\kappa}}_{Sp}\,, \tag{29}$$

wherein $\bar{\boldsymbol{\varepsilon}}_{Se}$ and $\bar{\boldsymbol{\kappa}}_{Se}$ represent the elastic parts and $\bar{\boldsymbol{\varepsilon}}_{Sp}$ and $\bar{\boldsymbol{\kappa}}_{Sp}$ are the plastic parts. In the framework of the geometrically linear theory, the solid volume fraction can be expressed dependently on the volumetric strain ε^v_S:

$$n^S = n^S_0\,(1 - \varepsilon^v_S)\,, \quad \text{where} \quad \varepsilon^v_S = \bar{\boldsymbol{\varepsilon}}_S \cdot \mathbf{I} = \operatorname{div} \mathbf{u}_S\,. \tag{30}$$

Proceeding from (30) and the image of a plastic intermediate configuration, the solid volume fraction can be computed via

$$\left.\begin{aligned} n^S &= n_p^S\,(1-\varepsilon_{Se}^v) \\ n_p^S &= n_0^S\,(1-\varepsilon_{Sp}^v) \end{aligned}\right\}, \quad \text{where} \quad \left\{\begin{aligned} \varepsilon_{Se}^v &= \bar{\boldsymbol{\varepsilon}}_{Se}\cdot\mathbf{I}\,, \\ \varepsilon_{Sp}^v &= \bar{\boldsymbol{\varepsilon}}_{Sp}\cdot\mathbf{I}\,. \end{aligned}\right. \tag{31}$$

Therein, n_p^S can be understood as the solid volume fraction in the plastic intermediate configuration. Note in passing that the above result is in full agreement with results of the geometrically non-linear approach, where the plastic deformation relates the line elements of the reference configuration to the line elements of the plastic intermediate configuration. This configuration can then be assumed to be a local reference configuration of the elastic part of the deformation relating the line elements of the plastic intermediate configuration to the line elements of the actual configuration. Furthermore, it should be noted that, to recover (30) from (31), the additive split of $\bar{\boldsymbol{\varepsilon}}_S$ together with a neglect of higher order terms has been used:

$$n^S \;=\; n_0^S\,(1-\varepsilon_{Sp}^v)\,(1-\varepsilon_{Se}^v)\,. \tag{32}$$

Elastic Material Properties

Proceeding from the image of a plastic intermediate configuration, the elastic deformation takes place between this intermediate configuration with the volume fraction n_p^S and the actual configuration with the volume fraction n^S. Furthermore, since the individual grains of granular media are assumed to be incompressible ($\rho^{SR} = \text{const.}$) and, in addition, uncrushable, the minimum of the pore space or the maximum of the solid volume fraction is given by the constant

$$n_{\max}^S \;<\; 1\,. \tag{33}$$

Following this implies a critical volumetric strain given with the aid of (31) as the maximal volumetric compression

$$\varepsilon_{S\,\mathrm{crit}}^v = 1 - \frac{n_{\max}^S}{n_0^S}\,, \quad \text{where} \quad \varepsilon_{Se\,\mathrm{crit}}^v = 1 - \frac{n_{\max}^S}{n_p^S} \tag{34}$$

is the corresponding elastic part. Note in passing that, following (34), $\varepsilon_{S\,\mathrm{crit}}^v$ is a constant, while $\varepsilon_{Se\,\mathrm{crit}}^v$ is a constant only with respect to the purely elastic process, which is parameterized through the plastic volume fraction n_p^S.

Proceeding from an additive split of the free energy function ψ^S of the solid skeleton, the purely elastic part ψ^{Se} of ψ^S or the stored elastic energy of the solid skeleton per reference volume is assumed to be given as

$$\begin{aligned} \rho_0^S\,\psi^{Se} = {}& \mu\;\bar{\boldsymbol{\varepsilon}}_{Se\,\mathrm{sym}}\cdot\bar{\boldsymbol{\varepsilon}}_{Se\,\mathrm{sym}} \;+\; \mu_c\;\bar{\boldsymbol{\varepsilon}}_{Se\,\mathrm{skw}}\cdot\bar{\boldsymbol{\varepsilon}}_{Se\,\mathrm{skw}} \;+\; \mu_c\,(l_c)^2\;\bar{\boldsymbol{\kappa}}_{Se}\cdot\bar{\boldsymbol{\kappa}}_{Se} + \\ & + \lambda\;\varepsilon_{Se\,\mathrm{crit}}^v\,\{\,\varepsilon_{Se\,\mathrm{crit}}^v\,[\,\ln(\varepsilon_{Se\,\mathrm{crit}}^v) - \ln(\varepsilon_{Se\,\mathrm{crit}}^v - \varepsilon_{Se}^v)\,]\,\} - \varepsilon_{Se}^v\,\}\,. \end{aligned} \tag{35}$$

Therein μ and λ are the two Lamé constants, where λ is explicitly coupled with the volumetric strain ε^v_{Se}. In consequence of the use of the micropolar theory, the parameter μ_c is used for the inclusion of the skew symmetric part of the strain, whereas l_c can be interpreted as an internal length.

Taking the derivatives of $\rho^S_0\,\psi^{Se}$ with respect to $\bar{\boldsymbol{\varepsilon}}_{Se}$ and to $\bar{\boldsymbol{\kappa}}_{Se}$ yields the stress and the couple stress tensors

$$\begin{aligned}
\boldsymbol{\sigma}^S_E &= \rho^S_0\,\frac{\partial\,\psi^{Se}}{\partial\,\bar{\boldsymbol{\varepsilon}}_{Se}} \\
&= 2\,\mu\;\bar{\boldsymbol{\varepsilon}}_{Se\,\mathrm{sym}} \;+\; 2\,\mu_c\;\bar{\boldsymbol{\varepsilon}}_{Se\,\mathrm{skw}} \;+\; \lambda\;\frac{\varepsilon^v_{Se\,\mathrm{crit}}}{\varepsilon^v_{Se\,\mathrm{crit}} - \varepsilon^v_{Se}} \;-\; \varepsilon^v_{Se}\;\mathbf{I}\,, \qquad (36)\\
\boldsymbol{\mu}^S &= \rho^S_0\,\frac{\partial\,\psi^{Se}}{\partial\,\bar{\boldsymbol{\kappa}}_{Se}} \;=\; 2\,\mu_c\,(l_c)^2\,\bar{\boldsymbol{\kappa}}_{Se}\,.
\end{aligned}$$

Plastic Material Properties

The specification of the plastic material properties requires a yield condition for the limitation of the elastic range as well as an evolution equation which defines the evolution of the plastic strains. Additionally, the variation of the yield condition can be described by the usage of a hardening law or softening law, respectively.

Due to the fact, that granular materials generally can be described as frictional materials, the yield condition in Ehlers [8] is applied, which is especially developed for such materials. In consequence of the usage of polar materials, this condition is extended by the skew-symmetric stress and the curvature tensors:

$$\begin{aligned}
F &= \sqrt{\Gamma\,\mathrm{II}^D_{\mathrm{sym}} + k_\sigma \mathrm{II}_{\mathrm{skw}} + k_\mu\,(\boldsymbol{\mu}^S\cdot\boldsymbol{\mu}^S)^{1/2} + \tfrac{1}{2}\,\alpha\,\mathrm{I}^2 + \delta^2\,\mathrm{I}^4} + \beta\,\mathrm{I} + \epsilon\,\mathrm{I}^2 - \kappa = 0\,,\\
\Gamma &= \left(1 + \gamma\,\frac{\mathrm{III}^D_{\mathrm{sym}}}{\left(\mathrm{II}^D_{\mathrm{sym}}\right)^{3/2}}\right)^m\,.
\end{aligned} \qquad (37)$$

Therein, I is the first invariant, $\mathrm{II}^D_{\mathrm{sym}}$ and $\mathrm{III}^D_{\mathrm{sym}}$ are the negative second and third invariants of the symmetric part of the deviator of the stress tensor $\boldsymbol{\sigma}^S_E$, whereas $\mathrm{II}_{\mathrm{skw}}$ denotes the second negative invariant of the skew-symmetric part of $\boldsymbol{\sigma}^S_E$. The above form of the yield condition contains nine material parameters, (α, β, δ, ϵ, κ, γ, m, k_σ, k_μ). By neglecting the very small cohesion of dry sand the parameters κ and α are assumed to be zero. Furthermore, the parameters γ and m can be merged using the convexity condition of the yield criterion, cf. Ehlers [8]:

$$m = \frac{2}{9} + \frac{\sqrt{3}}{3\,\gamma}\,. \qquad (38)$$

Thus, six parameters (β, δ, ϵ, γ, k_σ, k_μ) remain in the yield condition to be identified.

Due to the fact that frictional materials generally show a non-associated yielding, the introduction of a plastic potential

$$G = \sqrt{\psi_1 \, \mathrm{II}^D_{\mathrm{sym}} + k_\sigma \mathrm{II}_{\mathrm{skw}} + k_\mu \, (\boldsymbol{\mu}^S \cdot \boldsymbol{\mu}^S)^{1/2} + \frac{1}{2}\,\alpha\,\mathrm{I}^2 + \delta^2\,\mathrm{I}} + \psi_2\,\beta\,\mathrm{I} + \epsilon\,\mathrm{I}^2 \quad (39)$$

with the additional parameters ψ_1 and ψ_2 is necessary. Thus, the evolution of the plastic strain and curvature follow from,

$$(\bar{\boldsymbol{\varepsilon}}_{Sp})'_S = \Lambda \, \frac{\partial G}{\partial \boldsymbol{\sigma}^S_E} \,, \quad (\bar{\boldsymbol{\kappa}}_{Sp})'_S = \Lambda \, \frac{\partial G}{\partial \boldsymbol{\mu}^S} \,, \quad (40)$$

with the plastic multiplier Λ.

In granular materials, plastic deformations occur with the onset of the loading. Consequently, the yield surface has to be assumed as a nearly singular point at the beginning of the loading and increases with the plastic deformations. Furthermore, in triaxial compression tests on dense sand, one can observe the increase of the applied load in addition to a decrease of the load after exceeding the maximum load. This observation can be interpreted as a result of the increasing pore volume due to the dilatancy. Both effects, the increasing and the decreasing of the yield condition have to be taken into account for the formulation of the evolution of the yield condition. Therefore, the evolution of the parameters of the yield condition p_i, summarized in the vector $\boldsymbol{p}$, are split into a volumetric and a deviatoric part p_i^v and p_i^D:

$$\begin{aligned} &\dot{p}_i = \dot{p}_i^v + \dot{p}_i^D = (\overset{\star}{p}_i - p_i) \, (C^V_{pi} \, (\varepsilon^v_{S\mathrm{p}})'_S + C^D_{pi} \, \|(\bar{\boldsymbol{\varepsilon}}_{Sp})'_S\|) \\ &\text{with} \quad p_i(t=0) = p_{i\,0} \quad \text{and} \quad \boldsymbol{p} = (\beta,\, 1/\delta,\, 1/\epsilon,\, \gamma)^T \,. \end{aligned} \quad (41)$$

Therein, $\overset{\star}{p}_i$ is the maximum value of p_i and C^V_{pi} and C^D_{pi} represent the volumetric and deviatoric evolution constants. Note in consequence that the yield condition increases by the reciprocal values of δ and ϵ, which are considered in $\boldsymbol{p}$. The parameters k_σ and k_μ, which are associated with the Cosserat theory, are excluded from the evolution, because the identification of evolution constants would be impossible.

3 Discretization in Space and Time

For the solution of initial boundary-value problems numerical methods in the framework of the Finite Element Method (FEM) are applied. For this purpose, the weak formulations of the balance equations have to be discretized in space and time.

3.1 Weak Formulations

The presented model contains three independent fields of primary variables: The pore pressure p, the translation field $\mathbf{u}_S$ and the field of the total rotations

$\bar{\boldsymbol{\varphi}}_S$. Consequently, the overall volume balance and the balance of momentum of the mixture as well as the balance of moment of momentum of the solid skeleton are needed as weak formulations. These weak formulations are obtained from the strong formulations (26) by multiplication with the corresponding test functions δp, $\delta \mathbf{u}_S$ and $\delta \bar{\boldsymbol{\varphi}}_S$ and integration over the spatial domain Ω:

- Volume balance of the mixture

$$\begin{aligned} \mathcal{G}_V^M = & \int_\Omega \operatorname{div}(\mathbf{u}_S)'_S \, \delta p \, \mathrm{d}v + \int_\Omega \frac{k^F}{\gamma^{FR}} \left(\operatorname{grad} p - \rho^{FR} \, \mathbf{g}\right) \cdot \operatorname{grad} \delta p \, \mathrm{d}v + \\ & + \int_{\Gamma_p} n^F \mathbf{w}_F \cdot \mathbf{n} \, \delta p \, \mathrm{d}a = 0 \,, \end{aligned} \tag{42}$$

- Balance of momentum of momentum of the mixture

$$\begin{aligned} \mathcal{G}_M^M = & \int_\Omega \left(\boldsymbol{\sigma}_E^S - p \, \mathbf{I}\right) \cdot \operatorname{grad}(\delta \mathbf{u}_S) \, \mathrm{d}v - \int_\Omega \left(n^F \, \rho^{FR} + n^S \, \rho^{SR}\right) \mathbf{g} \cdot \delta \mathbf{u}_S \, \mathrm{d}v - \\ & - \int_{\Gamma_\sigma} \boldsymbol{\sigma}_E^S \, \mathbf{n} \cdot \delta \mathbf{u}_S \, \mathrm{d}a = 0 \,, \end{aligned} \tag{43}$$

- Balance of moment of the solid skeleton

$$\mathcal{G}_{MoM}^S = \int_\Omega \boldsymbol{\mu}^S \cdot \operatorname{grad}(\delta \bar{\boldsymbol{\varphi}}_S) \, \mathrm{d}v - \int_\Omega \left(\mathbf{I} \times \boldsymbol{\sigma}_E^S\right) \cdot \delta \bar{\boldsymbol{\varphi}}_S \, \mathrm{d}v + \int_{\Gamma_\mu} \boldsymbol{\mu}^S \, \mathbf{n} \cdot \delta \bar{\boldsymbol{\varphi}}_S \, \mathrm{d}a = 0 \,. \tag{44}$$

In the above equations $\mathbf{n}$ denotes the outward oriented unit surface normal vector of the Neumann boundaries Γ_p, Γ_σ and Γ_μ.

3.2 Discretization in Space

The spatial discretization of the field of primary variables

$$\boldsymbol{y} = (p, \, \mathbf{u}_S, \, \bar{\boldsymbol{\varphi}}_S)^T \tag{45}$$

is carried out by use of quadratic shape functions $\boldsymbol{\phi}_i$ for the approximated displacement $\mathbf{u}_S^h$ and linear shape functions $\boldsymbol{\eta}_i$ and $\boldsymbol{\psi}_i$ for the approximated pore pressure p^h and the total rotations $\bar{\boldsymbol{\varphi}}_S^h$:

$$\begin{aligned} p \approx p^h = \sum_{i=1}^{M} \boldsymbol{\eta}_i \, p_i \,, \qquad & \mathbf{u}_S \approx \mathbf{u}_S^h = \sum_{i=1}^{N} \boldsymbol{\phi}_i \, \mathbf{u}_{Si} \,, \qquad & \bar{\boldsymbol{\varphi}}_S \approx \bar{\boldsymbol{\varphi}}_S^h = \sum_{i=1}^{M} \boldsymbol{\psi}_i \, \bar{\boldsymbol{\varphi}}_{Si} \,, \\ \delta p^h = \sum_{i=1}^{M} \boldsymbol{\eta}_i \, \delta p_i \,, \qquad & \delta \mathbf{u}_S^h = \sum_{i=1}^{N} \boldsymbol{\phi}_i \, \delta \mathbf{u}_{Si} \,, \qquad & \delta \bar{\boldsymbol{\varphi}}_S^h = \sum_{i=1}^{M} \boldsymbol{\psi}_i \, \delta \bar{\boldsymbol{\varphi}}_{Si} \,. \end{aligned} \tag{46}$$

Thereby, N represents the number of translational and M the number of rotational nodes. Moreover, in application of the Bubnov-Galerkin method,

the test functions δp_i, $\delta \mathbf{u}_i$ and $\delta \bar{\boldsymbol{\varphi}}_{Sj}$ are identical with the shape functions $\boldsymbol{\eta}_i$, $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_i$. The stress tensor $\boldsymbol{\sigma}_E^S$, the couple stress tensor $\boldsymbol{\mu}^S$, the hardening variables $\boldsymbol{p}$ and the internal variables

$$\boldsymbol{q} = (\varepsilon_p^{11}, \ldots, \varepsilon_p^{33}, \kappa_p^{11}, \ldots, \kappa_p^{33}, \Lambda)^T \tag{47}$$

are computed at the integration points of the numerical quadrature.

3.3 Discretization in Time

The spatially discretized weak formulations (43) and (44) are summarized to the following system of equations:

$$\mathcal{G}(t, \boldsymbol{y}, \boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{p}, \dot{\boldsymbol{p}}) = \mathcal{G}_V^M + \mathcal{G}_M^M + \mathcal{G}_{MoM}^S = 0\,. \tag{48}$$

Furthermore, for the viscoplastic case, the plastic evolution (40) can be expressed as,

$$\boldsymbol{l} = \boldsymbol{A} \begin{pmatrix} \dot{\boldsymbol{q}} \\ \dot{\boldsymbol{p}} \end{pmatrix} - \boldsymbol{r}(\boldsymbol{y}, \boldsymbol{q}, \boldsymbol{p}) \equiv \begin{pmatrix} (\bar{\boldsymbol{\varepsilon}}_{Sp})'_S \\ (\bar{\boldsymbol{\kappa}}_{Sp})'_S \\ 0 \\ \dot{\boldsymbol{p}} \end{pmatrix} - \begin{pmatrix} \Lambda \dfrac{\partial G}{\partial \boldsymbol{\sigma}_E^S} \\ \Lambda \dfrac{\partial G}{\partial \boldsymbol{\mu}^S} \\ \Lambda - \dfrac{1}{\eta} \left\langle \dfrac{F}{\sigma_0} \right\rangle^r \\ \boldsymbol{f}((\varepsilon_{\mathrm{Sp}}^v)'_S, \|(\bar{\boldsymbol{\varepsilon}}_{Sp})'_S\|) \end{pmatrix} = \boldsymbol{0}\,, \tag{49}$$

cf. Ehlers and Ellsiepen [12], whereby η is the relaxation time, σ_0 is the reference stress and r is the viscoplastic exponent. Furthermore, $\langle\cdot\rangle$ are the Macaulay brackets, which are defined as: $\langle\cdot\rangle := \frac{1}{2}\left[(\cdot) + |(\cdot)|\right]$. Accordingly, the model equations can be written as a system of Differential-Algebraic Equations (DAE),

$$\boldsymbol{F}(t, \boldsymbol{y}) = \begin{bmatrix} \mathcal{G}(t, \boldsymbol{y}, \boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{p}, \dot{\boldsymbol{p}}) \\ \boldsymbol{l}(t, \boldsymbol{y}, \boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{p}, \dot{\boldsymbol{p}}) \end{bmatrix} = \boldsymbol{0}\,, \tag{50}$$

which can efficiently be solved by use of a Singly Diagonal Implicit Runge-Kutta method (SDIRK) with an embedded time step control, see Ehlers et al. [13] and Ellsiepen [17] for details.

The final set of seven equations to be solved on the Gauß point level at the time step n is summarized in Table 1. In particular, the first two, $\mathbf{G}_n^1$ and $\mathbf{G}_n^2$, are tensorial equations representing the elastic law (36), the third equation g_n^3 governs the viscoplasticity and finally, the hardening law (41) is represented by the equations g_n^4 until g_n^7, where in g_n^5 and g_n^6 the inverses of the parameters δ and ϵ are used for the evolution.

Table 1. System equations to be solved at the integration points at time step n

Elasticity:	$\mathbf{G}^1_n = \boldsymbol{\sigma}^S_{E\,n} - 2\,\mu\,\bar{\boldsymbol{\varepsilon}}_{Sen} + \bar{\lambda}\,\varepsilon^v_{Sen}\,\mathbf{I} + 2\,\mu_c\,\bar{\boldsymbol{\varepsilon}}_{Sen\,\mathrm{skw}} = \mathbf{0}$
	$\mathbf{G}^2_n = \boldsymbol{\mu}^S_n - 2\,\mu_c\,(l_c)^2\,\bar{\boldsymbol{\kappa}}_{Sen} = \mathbf{0}$
	with $\begin{cases} \bar{\lambda} = \lambda\,\dfrac{\varepsilon^v_{Se\,\mathrm{crit}}}{\varepsilon^v_{Se\,\mathrm{crit}} - \varepsilon^v_{Sen}} \\ \bar{\boldsymbol{\varepsilon}}_{Se} = \bar{\boldsymbol{\varepsilon}}_S - \bar{\boldsymbol{\varepsilon}}_{Sp}\,, \quad \bar{\boldsymbol{\kappa}}_{Se} = \bar{\boldsymbol{\kappa}}_S - \bar{\boldsymbol{\kappa}}_{Sp} \end{cases}$
Plastic yielding:	$(\bar{\boldsymbol{\varepsilon}}_{Spn})'_S = \Lambda\,\dfrac{\partial G(\boldsymbol{\sigma}^S_{E\,n},\,\boldsymbol{\mu}^S_n)}{\partial \boldsymbol{\sigma}^S_{E\,n}}$
	$(\bar{\boldsymbol{\kappa}}_{Spn})'_S = \Lambda\,\dfrac{\partial G(\boldsymbol{\sigma}^S_{E\,n},\,\boldsymbol{\mu}^S_n)}{\partial \boldsymbol{\mu}^S_n}$
	with the plastic potential G from (39)
Viscoplasticity:	$g^3_n = \Lambda_n - \dfrac{1}{\eta}\left\langle \dfrac{F(\boldsymbol{\sigma}^S_{E\,n},\,\boldsymbol{\mu}^S_n)}{\sigma_0} \right\rangle^r = 0$
	with the yield condition F from (37)
Plastic hardening:	$g^4_n = \dot{\beta}_n - (\overset{\star}{\beta}_n - \beta_n)\,(C^V_\beta\,(\varepsilon^v_{Sen})'_S + C^D_\beta\,\|(\bar{\boldsymbol{\varepsilon}}_{Spn})'_S\|) = 0$
	$g^5_n = \dot{\delta}_n - (\delta_n - \dfrac{\delta^2_n}{\overset{\star}{\delta}})\,(C^V_\delta\,(\varepsilon^v_{Sen})'_S + C^D_\delta\,\|(\bar{\boldsymbol{\varepsilon}}_{Spn})'_S\|) = 0$
	$g^6_n = \dot{\epsilon}_n - (\epsilon_n - \dfrac{\epsilon^2_n}{\overset{\star}{\epsilon}})\,(C^V_\epsilon\,(\varepsilon^v_{Sen})'_S + C^D_\epsilon\,\|(\bar{\boldsymbol{\varepsilon}}_{Spn})'_S\|) = 0$
	$g^7_n = \dot{\gamma}_n - (\overset{\star}{\gamma}_n - \gamma_n)\,(C^V_\gamma\,(\varepsilon^v_{Sen})'_S + C^D_\gamma\,\|(\bar{\boldsymbol{\varepsilon}}_{Spn})'_S\|) = 0$
Time discretization:	$\dot{k}_n = \dfrac{k_n - k_{n-1}}{t_n - t_{n-1}}$
	with $k = \bar{\boldsymbol{\varepsilon}}_{Sp},\,\bar{\boldsymbol{\kappa}}_{Sp},\,\beta,\,\delta,\,\epsilon,\,\gamma$

(51)

4 Parameter Identification

In the above sections, a model for the simulation of the stress-strain behavior of sand as well as its numerical implementation was illustrated. For the application of the model to real boundary-value problems with a specified material, all material constants must be known. On this account, the methods of parameter identification are required.

In this section, firstly, the applied method for the identification will be introduced. Afterwards, the stepwise approach of the determination of the various material parameters is illustrated. Finally, the computation of the thickness and the direction of the shear band is explained, as this information is required for the identification of the material parameters of the Cosserat theory.

4.1 Identification Method

The determination of the material constants is carried out on basis of measured data $\tilde{\phi}$ of homogeneous triaxial tests and inhomogeneous biaxial tests. These tests are simulated as boundary-value problems with an assumed set of material parameters $\boldsymbol{s}$. Naturally, the obtained simulated results $\phi(\boldsymbol{s})$ are different to the measured data. This difference is expressed by use of the least squares method, which yields the objective function

$$f(\boldsymbol{s}) = \sum_{n=1}^{N} w_n \left(\phi_n(\boldsymbol{s}) - \tilde{\phi}_n\right)^2 . \tag{52}$$

Herein, w_n is a weighting factor and N is the number of data points, which are considered in the identification process. The task of the identification is to find the parameter set $\bar{\boldsymbol{s}}$, which yields the global minimum of the objective function $f(\boldsymbol{s})$. Due to the fact that with the measured data $\tilde{\phi}$ the result of the boundary-value problem is given, while the material parameters are unknown, this problem is called an inverse or indirect problem. In contrast, the solution of the boundary-value problem itself is denoted as the direct problem.

For the complete formulation of the inverse problem restrictions for the values of the parameters to physically reasonable ranges by N_{iq} inequality constraints are required. Furthermore, in some cases, the formulation of N_{eq} equality constraints is necessary:

$$\begin{aligned} \text{inequality constraints: } g_j(\boldsymbol{s}) \geq 0 \quad &\text{with} \quad j = 1, \ldots, N_{iq} , \\ \text{equality constraints: } h_i(\boldsymbol{s}) = 0 \quad &\text{with} \quad i = 1, \ldots, N_{eq} . \end{aligned} \tag{53}$$

Thus, with the objective function (52) and the restrictions (53), the complete inverse problem is defined.

In this contribution, the solution of the inverse problem is carried out with the gradient based Sequential Quadratic Programming (SQP) by use of the

BFGS method, which is well described in the textbook of Töring & Spellucci [36]. The major advantage of this method is the low computational costs, in contrast to gradient-free methods. Consequently, an effective computation of the gradient of the objective function $f(\boldsymbol{s})$ with respect to the material parameters $\boldsymbol{s}$ by use of the semi-analytical sensitivity analysis (cf. Sect. 5) is essential.

4.2 Stepwise Identification

The presented model contains a lot of different material parameters, which have to be identified. Hence, the overall identification process is separated into three steps:

1. The identification of the parameters of the homogeneous elastic behavior (μ, λ , $n^S_{\max}$), cf. (34) and (36), on the basis of unloading cycles in triaxial tests.
2. Identification of the parameters of the homogeneous plastic behavior (C^V_{pi}, C^D_{pi} and $p_{i\,0}$ with $p_i = \beta$, δ, ϵ, γ), cf. (41), by use of data of homogeneous triaxial tests.
3. Identification of the additional parameters associated with the Cosserat theory (μ_c, l_c, k_σ, k_μ), cf. (36) and (37), on the basis of inhomogeneous biaxial tests.

Thereby, the parameters, which are determined in the first step are assumed to be given in the second step and similarly, the parameters, identified at the steps one and two, are assumed as given for the last step.

4.3 Shear Band Thickness and Direction

Whereas, for the first two steps standard stress-strain data of homogeneous triaxial tests is sufficient, for the third step geometrical strain data of the incoming shear band are additionally required. One method for the inclusion of such data in the identification process is the direct comparison of the strain data in discrete spatial points. This technique is used in most papers, which deal with parameter identification by use of inhomogeneous tests. In this contribution such an approach will not be convenient, because the angle and the thickness of the shear band, as most important values, can thereby not separately taken into account. Thus, a new method for the determination of these values on the basis of strain data has been developed, see Fig. 2. In particular, in Fig. 2 (a), the setup of the biaxial test is shown: A plane sand specimen is stabilized by a side load, then a load platen at the top is slowly driven down (for details see [30]). After overstepping a critical load level, a shear band occurs as the result of strain localization, see Fig. 2 (b), wherein the norm of the strain $\|\bar{\boldsymbol{\varepsilon}}_S\|$ is pictured. Now an orthographic cut through this data, Fig. 2 (c), looks like a bell-shaped curve. This observation motivates the approximation

of the norm of strain $\|\bar{\boldsymbol{\varepsilon}}_S\|$ by a bell-shaped curve moving along the axis of the shear band,

$$\epsilon(x_1,\, x_2,\, \boldsymbol{r}) = \epsilon_0 + A\, \exp\left[\frac{-1}{2\,H^2}\, \Big((x_1 - x_{10})\, \cos(\alpha_S) - x_2\, \sin(\alpha_S)\Big)^{\,2}\right],$$
$$\text{with} \quad \boldsymbol{r} = (\epsilon_0,\, A,\, H,\, x_{10},\, \alpha_S)^T, \tag{54}$$

cf. Fig. 2 (d). Therein, the function ϵ depends on the one hand on the spatial coordinates x_1 and x_2, cf. Fig. 2 (d) and on the other hand on the parameter vector $\boldsymbol{r}$. Thus, the in $\boldsymbol{r}$ included parameters can be used for the fit of the function $\epsilon(\boldsymbol{r})$ to the calculated or measured norm of strains $\|\bar{\boldsymbol{\varepsilon}}(\boldsymbol{s})\|$. By application of the least-squares method, one obtains the objective function

$$f(\boldsymbol{r}) = \sum_{i=0}^{M} \Big(\epsilon(x_{1\,i},\, x_{2\,i},\, \boldsymbol{r}) - \|\bar{\boldsymbol{\varepsilon}}_{Si}(\boldsymbol{s})\|\Big)^{\,2} \tag{55}$$

with M measurement or integration points, which has to be minimized. The resulting unconstrained optimization problem yields the nonlinear system of equations

$$\boldsymbol{h}(\boldsymbol{r}) = \frac{\partial f}{\partial \boldsymbol{r}} = 2 \sum_{i=1}^{N} \Big(\epsilon_i - \|\bar{\boldsymbol{\varepsilon}}_{Si}(\boldsymbol{s})\|\Big)\, \frac{\partial \epsilon_i}{\partial \boldsymbol{r}} = \boldsymbol{0}\,, \tag{56}$$

wherein ϵ_i stands for $\epsilon(x_{1\,i},\, x_{2\,i},\, \boldsymbol{r})$. The solution of this system of equations for $\boldsymbol{r}$ is carried out by the Newton's method with the Jacobi matrix

$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{r}} = 2 \sum_{i=1}^{N} \left[\frac{\partial \epsilon_i}{\partial \boldsymbol{r}} \left(\frac{\partial \epsilon_i}{\partial \boldsymbol{r}}\right)^{T} + (\epsilon_i(\boldsymbol{r}) - \|\bar{\boldsymbol{\varepsilon}}_{Si}(\boldsymbol{s})\|)\, \frac{\partial^2 \epsilon_i}{\partial \boldsymbol{r}\, \partial \boldsymbol{r}^T}\right], \tag{57}$$

which can be calculated analytically.

5 Sensitivity Analysis

The effective solution of the minimization problem of the parameter identification requires a numerically cheap computation of the parameter sensitivities. Additionally, the sensitivities are needed for the evaluation of the correlations between the obtained material parameters.

The task of the sensitivity analysis is the computation of the gradient of the objective function (52) with respect to the material parameters $\boldsymbol{s}$,

$$\frac{\mathrm{d} f(\boldsymbol{s})}{\mathrm{d} \boldsymbol{s}} = 2 \sum_{n=1}^{N} w_n\, \frac{\mathrm{d} \phi_n(\boldsymbol{s})}{\mathrm{d} \boldsymbol{s}}\, (\phi_n(\boldsymbol{s}) - \tilde{\phi}_n)\,, \tag{58}$$

which requires the gradient of the computed data $\phi(\boldsymbol{s})$ with respect to the material parameters.

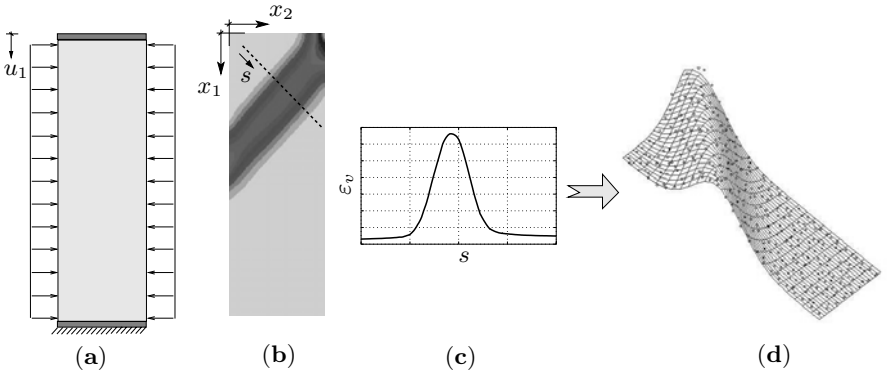


Fig. 2. Determination of the thickness and direction of the shear band: (**a**) setup of the biaxial test, (**b**) computed norm of the total strain, (**c**) cut of this data, (**d**) approximation with a bell-shaped curve

The simplest way for the computation of this gradient is the numerical evaluation of the differential quotient for every material parameter s_i:

$$\frac{\mathrm{d}\phi(s_i)}{\mathrm{d}s_i} \approx \frac{\phi(s_i + \Delta s_i) - \phi(s_i)}{\Delta s_i} \,. \tag{59}$$

Whereas, the implementation of this method is very simple, the main disadvantage is the enormous numerical cost, which annihilates the advantage of the gradient based optimization methods.

For this reason, the semi-analytical sensitivity analysis is applied, which is based on the analytical determination of the gradient of the numerical result. Such methods were developed in the framework of structure optimization in the 80's, cf. Adelmann and Haftka [1]. A further development for applications of the parameter identification in the context of viscoplasticity is shown in Mahnken and Stein [28] and a good overview of the several methods is given in the book of Kleiber [22].

Firstly, in this section the determination of the sensitivities of the primary variables is explained. Therewith, the required sensitivities of strain, stress as well as the hardening and internal variables can be computed. As a consequence of using the geometrical data of the shear band in the identification process, the sensitivities of this data are also required. The computation of these sensitivities is described in the last part of this section. Due to the fact that the identification of the parameters is only applied on the material parameters of the solid skeleton, the sensitivity analysis has also been applied on an empty porous solid skeleton.

5.1 Sensitivities of the Primary Variables

The first step in computing the sensitivities is the determination of the sensitivities of the primary variables, which requires the differentiation of the weak formulations (48) with respect to the material parameters $\boldsymbol{s}$. In consequence of the usage of an empty porous solid skeleton, the overall volume balance as well as the production terms in the balance of momentum can be neglected, and furthermore the pore pressure can be set to zero. Therewith, the derivative of the reduced weak forms reads:

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{G}}{\mathrm{d}\boldsymbol{s}} &= \int_{\Omega} \frac{\partial \boldsymbol{\sigma}_E^S}{\partial \boldsymbol{s}} \cdot \operatorname{grad} \delta \mathbf{u}_S \, \mathrm{d}v + \int_{\Omega} \left[\frac{\partial \boldsymbol{\sigma}_E^S}{\partial \bar{\boldsymbol{\varepsilon}}_S} \frac{\mathrm{d}\bar{\boldsymbol{\varepsilon}}_S}{\mathrm{d}\boldsymbol{s}} + \frac{\partial \boldsymbol{\sigma}_E^S}{\partial \bar{\boldsymbol{\kappa}}_S} \frac{\mathrm{d}\bar{\boldsymbol{\kappa}}_S}{\mathrm{d}\boldsymbol{s}} \right] \cdot \operatorname{grad} \delta \mathbf{u}_S \, \mathrm{d}v + \\
&+ \int_{\Omega} \frac{\partial \boldsymbol{\mu}^S}{\partial \boldsymbol{s}} \cdot \operatorname{grad} \delta \bar{\boldsymbol{\varphi}}_S \, \mathrm{d}v + \int_{\Omega} \left[\frac{\partial \boldsymbol{\mu}^S}{\partial \bar{\boldsymbol{\varepsilon}}_S} \frac{\mathrm{d}\bar{\boldsymbol{\varepsilon}}_S}{\mathrm{d}\boldsymbol{s}} + \frac{\partial \boldsymbol{\mu}^S}{\partial \bar{\boldsymbol{\kappa}}_S} \frac{\mathrm{d}\bar{\boldsymbol{\kappa}}_S}{\mathrm{d}\boldsymbol{s}} \right] \cdot \operatorname{grad} \delta \bar{\boldsymbol{\varphi}}_S \, \mathrm{d}v - \\
&- \int_{\Omega} (\mathbf{I} \times \frac{\partial \boldsymbol{\sigma}_E^S}{\partial \boldsymbol{s}}) \cdot \delta \bar{\boldsymbol{\varphi}}_S \, \mathrm{d}v - \int_{\Omega} \left[\mathbf{I} \times (\frac{\partial \boldsymbol{\sigma}_E^S}{\partial \bar{\boldsymbol{\varepsilon}}_S} \frac{\mathrm{d}\bar{\boldsymbol{\varepsilon}}_S}{\mathrm{d}\boldsymbol{s}} + \frac{\partial \boldsymbol{\sigma}_E^S}{\partial \bar{\boldsymbol{\kappa}}_S} \frac{\mathrm{d}\bar{\boldsymbol{\kappa}}_S}{\mathrm{d}\boldsymbol{s}}) \right] \cdot \delta \bar{\boldsymbol{\varphi}}_S \, \mathrm{d}v \stackrel{!}{=} 0 \,.
\end{aligned} \tag{60}$$

Using the shape functions (46), one obtains the discrete sensitivities of the strain and curvature tensors:

$$\begin{aligned}
\frac{\mathrm{d}\bar{\boldsymbol{\varepsilon}}_S}{\mathrm{d}\boldsymbol{s}} &= \sum_{i=1}^{N} \operatorname{grad} \boldsymbol{\phi}^i(\mathbf{x}) \frac{\mathrm{d}\mathbf{u}_{Si}}{\mathrm{d}\boldsymbol{s}} + \overset{3}{\mathbf{E}} \sum_{i=1}^{M} \boldsymbol{\psi}^i(\mathbf{x}) \frac{\mathrm{d}\bar{\boldsymbol{\varphi}}_{Si}}{\mathrm{d}\boldsymbol{s}} \,, \\
\frac{\mathrm{d}\bar{\boldsymbol{\kappa}}_S}{\mathrm{d}\boldsymbol{s}} &= \sum_{i=1}^{M} \operatorname{grad} \boldsymbol{\psi}^i(\mathbf{x}) \frac{\mathrm{d}\bar{\boldsymbol{\varphi}}_{Si}}{\mathrm{d}\boldsymbol{s}} \,.
\end{aligned} \tag{61}$$

In combination with (60), this yields a linear system of equations for the sensitivities of the primary variables at time step n:

$$\boldsymbol{K}_n \frac{\mathrm{d}\boldsymbol{y}_n}{\mathrm{d}\boldsymbol{s}} - \boldsymbol{f}_{sn} = \mathbf{0} \,. \tag{62}$$

Therein, the matrix $\boldsymbol{K}_n$ is the stiffness or Jacobian matrix at the time step n, and

$$\boldsymbol{f}_{sn} = \int_{\Omega} \frac{\partial \boldsymbol{\sigma}_E^S}{\partial \boldsymbol{s}} \cdot \operatorname{grad} \delta \mathbf{u}_s \, \mathrm{d}v + \int_{\Omega} \frac{\partial \boldsymbol{\mu}^S}{\partial \boldsymbol{s}} \cdot \operatorname{grad} \delta \bar{\boldsymbol{\varphi}}_S \, \mathrm{d}v - \int_{\Omega} (\mathbf{I} \times \frac{\partial \boldsymbol{\sigma}_E^S}{\partial \boldsymbol{s}}) \cdot \delta \bar{\boldsymbol{\varphi}}_S \, \mathrm{d}v \tag{63}$$

represents the so-called Sensitivity Load Term. Due to the fact that the same Jacobi matrix is used for the solution of the direct problem, the inverse of this matrix has already been computed. Thus, the numerical effort for the solution of this system of equations simply consists in a matrix multiplication.

5.2 Partial Derivative of the Stress (Sensitivity Load Term)

For the computation of the Sensitivity Load Term $\boldsymbol{f}_{sn}$, the model equations (51) in Table 1, which can be summarized to

$$\boldsymbol{G} = (\mathbf{G}_1,\, \mathbf{G}_2,\, g_3,\, g_4,\, g_5,\, g_6,\, g_7)^T \tag{64}$$

at a frozen deformation state ($\boldsymbol{y}$ = const.), are required. With the vector $\boldsymbol{p} = (\epsilon,\, \delta,\, \gamma,\, \beta)^T$ for the hardening variables and $\boldsymbol{q} = (\bar{\varepsilon}_{p\,ij},\, \bar{\kappa}_{p\,ij})^T$ for the internal variables and furthermore the generalized stresses and strains

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma}_E^S \\ \boldsymbol{\mu}^S \end{pmatrix} \quad \text{and} \quad \mathcal{E} = \begin{pmatrix} \bar{\boldsymbol{\varepsilon}}_S \\ \bar{\boldsymbol{\kappa}}_S \end{pmatrix}, \tag{65}$$

the following dependences are obtained:

$$\boldsymbol{G} = \boldsymbol{G}[\boldsymbol{s},\, \boldsymbol{\Sigma}_n(\boldsymbol{s}),\, \boldsymbol{p}_n(\boldsymbol{s}),\, \boldsymbol{p}_{n-1}(\boldsymbol{s}),\, \Lambda_n,\, \boldsymbol{q}_{n-1}(\boldsymbol{s})] \equiv \mathbf{0}\,. \tag{66}$$

Analogously to the global weak form (60), the system of equations $\boldsymbol{G}$ are differentiated with respect to the material parameters $\boldsymbol{s}$,

$$\begin{aligned} \left.\frac{\mathrm{d}\boldsymbol{G}}{\mathrm{d}\boldsymbol{s}}\right|_{\mathcal{E}_n} = \frac{\partial \boldsymbol{G}}{\partial \boldsymbol{s}} + \frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\Sigma}_n} \left.\frac{\mathrm{d}\boldsymbol{\Sigma}_n}{\mathrm{d}\boldsymbol{s}}\right|_{\mathcal{E}_n} + \frac{\partial \boldsymbol{G}}{\partial \boldsymbol{p}_n} \left.\frac{\mathrm{d}\boldsymbol{p}_n}{\mathrm{d}\boldsymbol{s}}\right|_{\mathcal{E}_n} + \frac{\partial \boldsymbol{G}}{\partial \boldsymbol{p}_{n-1}} \frac{\mathrm{d}\boldsymbol{p}_{n-1}}{\mathrm{d}\boldsymbol{s}} + \\ + \frac{\partial \boldsymbol{G}}{\partial \Lambda_n} \left.\frac{\mathrm{d}\Lambda_n}{\mathrm{d}\boldsymbol{s}}\right|_{\mathcal{E}_n} + \frac{\partial \boldsymbol{G}}{\partial \boldsymbol{q}_{n-1}} \frac{\mathrm{d}\boldsymbol{q}_{n-1}}{\mathrm{d}\boldsymbol{s}} \overset{!}{=} \mathbf{0}\,, \end{aligned} \tag{67}$$

at the generalized strain state $\mathcal{E}_n$. This procedure yields a linear system of equations for the derivatives of the generalized stress $\boldsymbol{\Sigma}$ and of the hardening variables $\boldsymbol{p}$ with respect to the material parameters $\boldsymbol{s}$. The load terms in this equation, i. e., $\partial \boldsymbol{G}/\partial \boldsymbol{s}$, can be easily determined by a direct differentiation of the space and time discrete system equations (51). Furthermore, the total derivatives of the history and hardening variables at the time step n–1 appear in (67). Therewith, the computation of the sensitivities becomes an iterative procedure.

5.3 Sensitivities of the Stress, Hardening and Internal Variables

After the solution of (67), the sensitivity load term $\boldsymbol{f}_{sn}$ is known and the system of equations (62) can be solved for the sensitivities of the primary variables $\mathrm{d}\boldsymbol{y}/\mathrm{d}\boldsymbol{s}$. Therewith, the sensitivities of the strain can be computed by use of the shape functions (61). Hence, the sensitivities of the stress as well as the hardening variables result from

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{\Sigma}_n}{\mathrm{d}\boldsymbol{s}} &= \left.\frac{\mathrm{d}\boldsymbol{\Sigma}_n}{\mathrm{d}\boldsymbol{s}}\right|_{\mathcal{E}_n} + \frac{\partial \boldsymbol{\Sigma}_n}{\partial \mathcal{E}_n} \frac{\mathrm{d}\mathcal{E}_n}{\mathrm{d}\boldsymbol{s}}\,, \\ \frac{\mathrm{d}\boldsymbol{p}_n}{\mathrm{d}\boldsymbol{s}} &= \left.\frac{\mathrm{d}\boldsymbol{p}_n}{\mathrm{d}\boldsymbol{s}}\right|_{\mathcal{E}_n} + \frac{\partial \boldsymbol{p}_n}{\partial \mathcal{E}_n} \frac{\mathrm{d}\mathcal{E}_n}{\mathrm{d}\boldsymbol{s}}\,. \end{aligned} \tag{68}$$

Next, the sensitivities of the history variables, which are needed in (67), have to be determined by an iterative procedure:

$$\begin{aligned}
&\frac{\mathrm{d}\boldsymbol{q}_n}{\mathrm{d}\boldsymbol{s}} = \frac{\mathrm{d}\boldsymbol{q}_{n-1}}{\mathrm{d}\boldsymbol{s}} + \frac{\mathrm{d}\Delta\boldsymbol{q}_n}{\mathrm{d}\boldsymbol{s}} \\
&\text{with} \quad \frac{\mathrm{d}\Delta\boldsymbol{q}_n}{\mathrm{d}\boldsymbol{s}} = \frac{\partial\Delta\boldsymbol{q}_n}{\partial\boldsymbol{s}} + \frac{\partial\Delta\boldsymbol{q}_n}{\partial\mathcal{E}_n}\frac{\mathrm{d}\mathcal{E}_n}{\mathrm{d}\boldsymbol{s}} + \\
&\qquad + \frac{\partial\Delta\boldsymbol{q}_n}{\partial\boldsymbol{\Sigma}_n}\frac{\mathrm{d}\boldsymbol{\Sigma}_n}{\mathrm{d}\boldsymbol{s}} + \frac{\partial\Delta\boldsymbol{q}_n}{\partial\boldsymbol{p}_n}\frac{\mathrm{d}\boldsymbol{p}_n}{\mathrm{d}\boldsymbol{s}} + \frac{\partial\Delta\boldsymbol{q}_n}{\partial\Lambda_n}\frac{\mathrm{d}\Lambda_n}{\mathrm{d}\boldsymbol{s}}\,.
\end{aligned} \tag{69}$$

5.4 Sensitivity of the Shear Band Data

The approach for the computation of the sensitivities of the shear band data $\mathrm{d}\boldsymbol{r}/\mathrm{d}\boldsymbol{s}$ is similar to the computation of the sensitivities of the primary variables. The equation for the determination of the shear band data (56) is implicitly differentiated with respect to the material parameters, which yields a system of equations for the determination of the corresponding sensitivities:

$$\boldsymbol{h}(\boldsymbol{r}(\boldsymbol{s}),\,\boldsymbol{s}) \equiv \boldsymbol{0} \quad \Longrightarrow \quad \frac{\mathrm{d}\boldsymbol{h}}{\mathrm{d}\boldsymbol{s}} = \frac{\partial\boldsymbol{h}}{\partial\boldsymbol{s}} + \frac{\partial\boldsymbol{h}}{\partial\boldsymbol{r}}\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}\boldsymbol{s}} = \boldsymbol{0}\,. \tag{70}$$

In the above equation, the partial derivative of $\boldsymbol{h}$ with respect to the parameter vector $\boldsymbol{r}$ (54) is already computed in (57). Furthermore, the load term results in

$$\frac{\partial\boldsymbol{h}}{\partial\boldsymbol{s}} = -2\sum_{i=1}^{N}\frac{\partial\|\bar{\boldsymbol{\varepsilon}}_i\|}{\partial\boldsymbol{s}}\left(\frac{\partial\epsilon_i}{\partial\boldsymbol{r}}\right)^T. \tag{71}$$

Herein, the first term represents the sensitivities of the norm of the strain tensor and can be computed by use of the sensitivities of the primary variables. The second term, the derivative of ϵ with respect to $\boldsymbol{r}$, can be analytically determined using (54).

6 Numerical Examples

The numerical examples are presented in two parts. In the first part two different boundary-value problems are computed with the above presented theory, where the basic effects of the micropolar theory are shown. In the second part the identification of the additional material parameters of the micropolar theory is illustrated on the basis of measured data of Hostun sand.

In the first example a standard biaxial test is computed, which is used for material tests, cf. [30]. In particular, a plane specimen is stabilized by the side pressure p, then a load platen is driven down slowly, cf. Fig. 3. In consequence of the material softening due to the dilatancy, the plastic strain $\bar{\boldsymbol{\varepsilon}}_{Sp}$ is localized in a shear band. The regularization effect of the micropolar

theory can be observed due to the fact that the localization zone includes a few elements, whereas in the case of an ill-posed problem the localization zone has a thickness of about one element. In Fig. 3 it can also be seen that the total rotation $\bar{\boldsymbol{\varphi}}_S$ is also concentrated along this shear band. Hence, the gradient of the total rotation has a maximum left and right of the shear band, which yields the typical allocation of the couple stress $\boldsymbol{\mu}^S$ left and right along the failure line. This observation shows that micropolar effects are only relevant in the location zone (shear band).

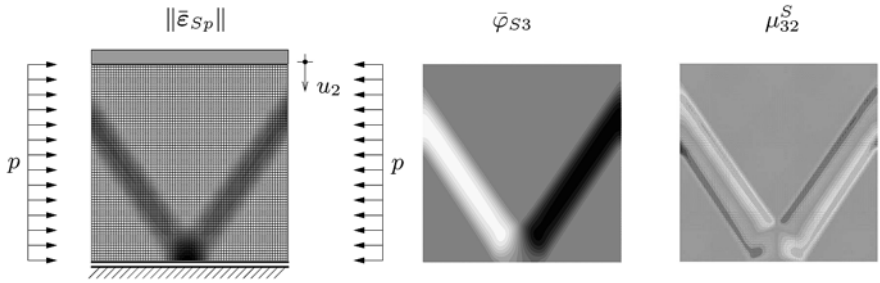


Fig. 3. Norm of the plastic strains $\bar{\boldsymbol{\varepsilon}}_{Sp}$, total rotations $\bar{\varphi}_{S3}$ and the couple stress component μ^S_{32}

The second example exhibits the in the geomechanics field well-known base failure problem, cf. Fig. 4. Thereby, an external load q of a rigid strip footing is applied onto the surface of a liquid-saturated half-space. The numerically obtained results for the plastic strains $\bar{\boldsymbol{\varepsilon}}_{Sp}$ are shown in Fig. 4. In contrast to the above example of the biaxial test, the incoming of different temporary shear bands can be observed until one shear band, which represents the final failure line, reaches the surface of the half space. This problem is computed by

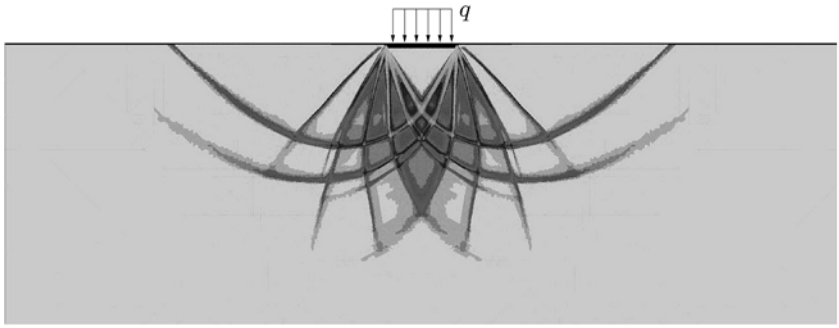


Fig. 4. Norm of the plastic strains $\bar{\boldsymbol{\varepsilon}}_{Sp}$ for the base failure problem

use of an adaptive algorithm, which yields 15 254 elements with 68 981 degrees of freedom in the final state. For a detailed description of this example, cf. [10].

After the usability of the model to geotechnical problems has been shown, the second part of the numerical examples is concerned with the identification of the additional material parameters due to the application of the micropolar theory, namely μ_c, l_c, k_σ and k_μ. This task requires a full computation of the boundary-value problem of the biaxial test with an adequate discretization. Thus, the numerical effort is very high and therewith the effectivity of the identification method is an important factor. For that reason, firstly, the semi-analytical sensitivities are verified by use of numerically computed values and secondly, the whole inhomogeneous identification process is demonstrated.

The verification of the semi-analytical computed sensitivities is carried out on the basis of the vertical stress as well as the thickness of the incoming shear band of biaxial tests for the parameters μ_c, l_c, k_σ and k_M, cf. Fig. 5 and Fig. 6.

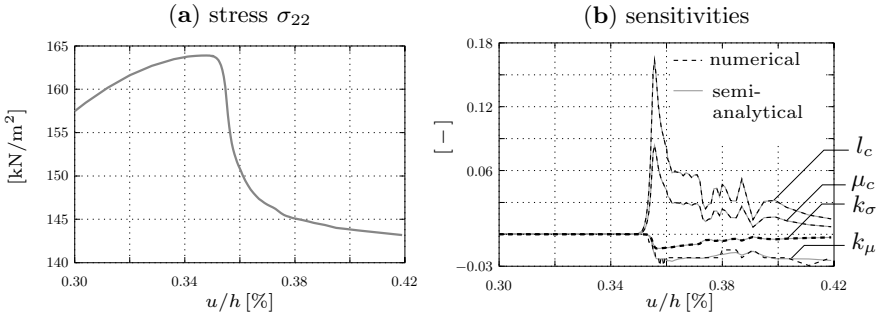


Fig. 5. Comparison between numerically and semi-analytically calculated sensitivities of the vertical stress σ^S_{E22}

In Fig. 5 (a), the stress versus the vertical strain is shown, where the decrease of the stress coincides with the incoming of the shear band. The comparison of the semi-analytical and the numerical sensitivities is visualized in Fig. 5 (b) showing a very good correlation, i. e., the lines for l_c^S, μ_c^S, k_σ and k_μ are congruent. Furthermore, it can be seen that an influence of the Cosserat parameters on the first occurs after the formation of the shear band. In Fig. 6 (a) the computed thickness of the shear band is illustrated and in Fig. 6 (b) the corresponding sensitivities are shown. Again a good correlation between numerical and the semi-analytical computed values can be observed.

In the last example the complete procedure of the parameter identification is demonstrated. For the realization of the parameter identification, measured stress-strain data as well as strain measurements on the surface of the specimen by use of the stereophotogrammetry are required. Therewith, the

thickness of the incoming shear band can be considered in the identification, cf. Sect. 4.3. This procedure yields the objective function

$$f(\boldsymbol{s}) = \frac{1}{N_\sigma\, \sigma_0^2} \sum_{i=1}^{N_\sigma} [\,\sigma_{11}(\boldsymbol{s}) - \tilde{\sigma}_{11}\,]^2 + \frac{1}{N_H\, H_0^2} \sum_{i=1}^{N_H} [\,H(\boldsymbol{s}) - \tilde{H}\,]^2\,, \tag{72}$$

wherein σ_{11} represents the vertical stress under the load platen, which is considered in N_σ load steps, H is the thickness of the shear band, which is taken in account in N_H steps. Furthermore, σ_0 and H_0 are normalizing factors for the stress and the shear band thickness, respectively. In principle, all parameters of the Cosserat theory have to be identified in this way. But in consequence of former results of a parameter re-identification, where a correlation between the parameters μ_c and l_c was observed, only the parameters l_c and k_σ are taken into account. The results of the identification are shown in Fig. 7, where the measured stress versus the computed values with the identified parameters

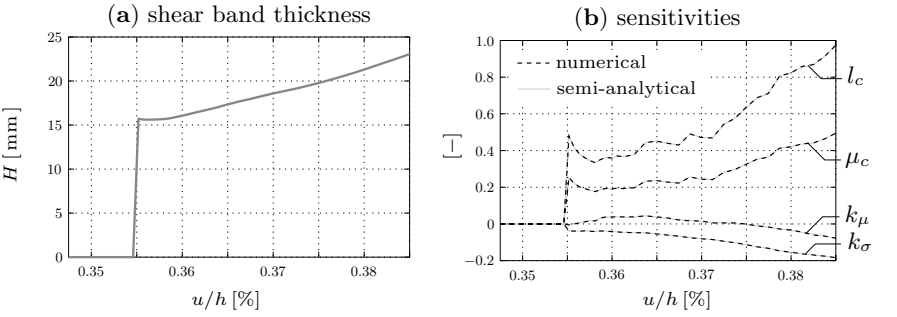


Fig. 6. Comparison between numerically and semi-analytically calculated sensitivities of the shear band thickness

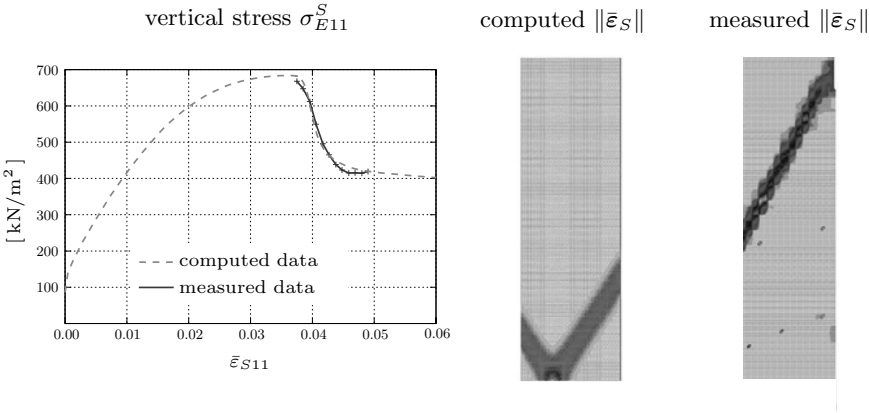


Fig. 7. Identification of the parameters of the micropolar theory by use of stress strain data in combination with the shear band thickness

as well as the measured and the computed distribution of the strain are depicted. Due to the fact that the incoming of the shear band varies during the identification, only the decrease of the stress is used for the identification. By comparing the strain distribution, it can be seen that the angle of the shear band fits also well, although the shear band angle has not been considered in the identification.

7 Conclusions

The intention of the project A6 of the SFB 404 was the investigation of localization problems in granular materials in the framework of the Theory of Porous Media. In consequence of the ill-posedness of the localization problem, a regularization method is required. Therefore, under consideration of the material properties on the microscale, the Cosserat theory as a physical correct way was applied. In this article, the embedding of this theory into the Theory of Porous Media was shown and furthermore, the numerical implementation in the Finite Element program PANDAS was illustrated with two boundary-value problems, the standard biaxial test and the base failure problem. Thereby, for the computation of the biaxial test a fixed mesh was sufficient, whereas for the computation of the larger base failure problem space adaptive strategies, including appropriate error indicators, are required.

In view of the computation of realistic problems, the second step in this project consists in the determination of the respective material parameters. This task can be split in a part concerning the homogeneous material behavior and another part concerning the inhomogeneous properties, in which the influence of the micropolar theory came into play. Due to the fact that the second part requires a complete computation of the inhomogeneous boundary-value problem of the biaxial test including the incoming of a shear band, this part requires an efficient identification method. Thus, the gradient-based SQP-method with the semi analytically computation of the sensitivities is used. The determination of the sensitivities was demonstrated by the comparison with numerically obtained values, where the difference was in the range of one percent. Furthermore, for the identification of the additional micropolar material parameters, the consideration of the thickness of the incoming shear band is necessary, whereas the angle of the shear band is primarily governed by the parameters of the plastic potential. In this project, a method, which enables a separate consideration of the thickness and the angle of the shear band was developed and used for the identification of the additionally micropolar material parameters on the basis of test data from Hostun sand.

For further works, the consideration of the effects of the liquid phase on the material behavior could be interesting, as the parameter identification in this project was only applied to an empty porous solid skeleton. Thus, it should be verified whether the results are affected by a pore fluid. This could

be realized by the validation of the model by use of undrained fluid saturated biaxial tests.

Furthermore, future work on this topic could be in the improvement of the numerical methods, in particular with regard to an extension to 3-dimensional problems, which yields a rapid increase of the number of degrees of freedom. Due to the fact, that the additional rotational degrees of freedom are only active in the localization zones, a reduction strategy of the set of equations in the nearly homogeneous areas can be developed.

References

1. H. Adelmann and R. Haftka. Sensitivity analysis of discrete structural systems. *AIAA Journal*, 24:823–832, 1986.
2. M. A. Biot. Theory of propagation of elastic waves in a fluid-saturated porous solid, I. Low frequency range. *Journal of the Acoustical Society of America*, 28:168–178, 1956.
3. R. M. Bowen. Incompressible porous media models by use of the theory of mixtures. *International Journal of Engineering Sciences*, 18:1129–1148, 1980.
4. E. Cosserat and F. Cosserat. *Théorie des Corps Déformable*. A. Hermann et Fils, Paris, 1909.
5. R. de Boer. *Theory of Porous Media*. Springer-Verlag, Berlin, 2000.
6. J. Desrues, J. Lanier, and P. Stutz. Localization of the deformation in tests on sand sample. *Engineering Fracture Mechanics*, 21(4):909–921, 1985.
7. W. Ehlers. Toward finite theories of liquid-saturated elasto-plastic porous media. *International Journal of Plasticity*, 7:443–475, 1991.
8. W. Ehlers. Constitutive equations for granular materials in geomechanical context. In K. Hutter, editor, *Continuum Mechanics in Environmental Sciences and Geophysics*, pages 313–402. Springer-Verlag, Wien, 1993.
9. W. Ehlers. A single surface yield function for geomaterials. *Archive of Applied Mechanics*, 65:246–259, 1995.
10. W. Ehlers. Foundations of multiphasic and porous materials. In W. Ehlers and J. Bluhm, editors, *Porous Media: Theory, Experiments and Numerical Applications*, pages 3–86. Springer-Verlag, Berlin, 2002.
11. W. Ehlers, S. Diebels, D. Mahnkopf, and P. Ellsiepen. *Theoretische und numerische Studien zur Lösung von Rand- und Anfangswertproblemen in der Theorie poröser Medien*. Zwischenbericht zum DFG-Forschungsvorhaben Eh 107/6-2. Bericht aus dem Institut für Mechanik (Bauwesen), Nr. 96-II-1. Universität Stuttgart, 1996.
12. W. Ehlers and P. Ellsiepen. Theoretical and numerical methods in environmental continuum mechanics based on the theory of porous media. In B. A. Schrefler, editor, *CISM Courses and Lectures No. 417*, pages 1–81. Springer-Verlag, Wien, 2001.
13. W. Ehlers, P. Ellsiepen, and M. Ammann. Time- and space-adaptive methods applied to localization phenomena in empty and saturated micropolar and standard porous materials. *International Journal of Numerical Methods in Engineering*, 52:503–526, 2001.

14. W. Ehlers and H. Müllerschön. Parameter identification of a macroscopic granular soil model applied to dense Berlin sand. *Granular Matter*, 2:105–112, 2000.
15. W. Ehlers, E. Ramm, S. Diebels, and J. D'Addetta. From particle ensembles to cosserat continua: Homogenization of contact forces towards stress and couple stresses. *International Journal of Solids and Structures*, 40:6681–6702, 2003.
16. W. Ehlers and W. Volk. On theoretical and numerical methods in the theory of porous media based on polar and non-polar elasto-plastic porous solid materials. *International Journal of Solids and Structures*, 35:4597–4617, 1998.
17. P. Ellsiepen. *Zeit- und ortsadaptive Verfahren angewandt auf Mehrphasenprobleme poröser Medien*. Bericht aus dem Institut für Mechanik im Bauwesen, Nr. II-3. Universität Stuttgart, 1999.
18. P. Haupt. *Continuum Mechanics and Theory of Materials*. Springer-Verlag, Berlin, 2000.
19. J. Heinz and P. Spellucci. A succesful implementation of the pantoja-mayne sqp-method. *Optimization Methods and Software*, 4:1–28, 1994.
20. C. B. Kafadar and A. C. Eringen. Micropolar media I the classical theory. *International Journal of Engineering Sciences*, 9:271–305, 1971.
21. M. K. Kim and P. V. Lade. Single Hardening Constitutive Model for Frictional Materials I. Plastic Potential Function. *Computers and Geotechnics*, 5:307–324, 1988.
22. M. Kleiber. *Parameter Sensitivity in Nonlinear Mechanics*. John Wiley & Sons, Chichester, 1997.
23. P. V. Lade and J. M. Duncan. Cubical triaxial tests on cohesionless soil. *Journal of the Soil Mechanics and Foundations Division*, 99:793–812, 1973.
24. P. V. Lade and M. K. Kim. Single Hardening Constitutive Model for Frictional Materials II. Yield Criterion and Plastic Work Contours. *Computers and Geotechnics*, 6:13–29, 1988.
25. P. V. Lade and M. K. Kim. Single hardening constitutive model for frictional materials III. Comparisons with experimental data. *Computers and Geotechnics*, 6:31–47, 1988.
26. D. V. Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, Massachusetts, 1973.
27. R. Mahnken and E. Stein. Parameter identification for viscoplastic models based on analytical derivatives of a least-squares functional and stability investigations. *International Journal of Plasticity*, 12:451–479, 1995.
28. R. Mahnken and E. Stein. A unified approach for parameter identification of inelastic material models in the frame of the finite element method. *Computer Methods in Applied Mechanics and Engineering*, 136:225–258, 1996.
29. R. Mahnken and E. Steinmann. A finite element algorithm for parameter identification of material models for fluid saturated porous media. *International Journal for Numerical and Analytical Methods in Geomechanics*, 25:415–434, 2001.
30. M. Mokni and J. Desrues. Strain localization measurements in undrained plane-strain biaxial tests on hostun RF sand. *Mechanics of cohesive-frictional materials*, 4:419–441, 1998.
31. H. B. Mühlhaus and I. Vardoulakis. The thickness of shear bands in granular materials. *Géotechnique*, 37(3):271–283, 1987.
32. H. Müllerschön. *Spannungs-Verformungsverhalten granularer Materialien am Beispiel von Berliner Sand*. Bericht aus dem Institut für Mechanik im Bauwesen, Nr. II-6. Universität Stuttgart, 2000.

33. W. Nowacki. *Theory of asymmetric elasticity*. Pergamon Press, Oxford, 1986.
34. P. Steinmann. A micropolar theory of finite deformation and finite rotation multiplicative elastoplasticity. *International Journal of Solids and Structures*, 31(8):1063–1084, 1994.
35. F. Thielecke. *Parameteridentifizierung von Simulationsmodellen für das viskoplastische Verhalten von Metallen – Theorie, Numerik, Anwendung –*. Braunschweiger Schriften zur Mechanik Nr. 34-1998. Universität Braunschweig, 1998.
36. W. Törnig and P. Spellucci. *Numerische Mathematik für Ingenieure und Physiker – Numerische Methoden der Analysis*, volume 1, 2. Springer-Verlag, Berlin, 1990.
37. G. Viggiani, M. Küntz, and J. Desrues. An experimental investigation of the relationships between grain size distribution and shear banding in sand. In P. A. Vermeer, S. Diebels, W. Ehlers, H. J. Herrmann, S. Luding, and E. Ramm, editors, *Continuous and Discontinuous Modelling of Cohesive-Frictional Materials*, pages 301–350. Springer-Verlag, Berlin, 2001.
38. W. Volk. *Untersuchung des Lokalisierungsverhaltens mikropolarer poröser Medien mit Hilfe der Cosserat-Theorie*. Bericht aus dem Institut für Mechanik im Bauwesen, Nr. II-2. Universität Stuttgart, 1999.
39. Y. Yamada and K. Ishihara. Anisotropic deformation characteristic of sand under three dimensional stress conditions. *Soils and Foundation*, 19:79–94, 1979.

Computer Simulation of Particle Suspensions*

Jens Harting[1], Martin Hecht[1], Hans J. Herrmann[2], and Sean McNamara[1]

[1] Institute for Computational Physics, University of Stuttgart, Pfaffenwaldring 27, 70569 Stuttgart, Germany
jens@icp.uni-stuttgart.de
Martin.Hecht@icp.uni-stuttgart.de
S.McNamara@icp.uni-stuttgart.de
[2] Institute for Building Materials, ETH Hönggerberg, HIF E 12, 8093 Zürich, Switzerland
hans@icp.uni-stuttgart.de

Summary. Particle suspensions are ubiquitous in our daily life, but are not well understood due to their complexity. During the last twenty years, various simulation methods have been developed in order to model these systems. Due to varying properties of the solved particles and the solvents, one has to choose the simulation method properly in order to use the available compute resources most effectively with resolving the system as well as needed. Various techniques for the simulation of particle suspensions have been implemented at the Institute for Computational Physics allowing us to study the properties of clay-like systems, where Brownian motion is important, more macroscopic particles like glass spheres or fibers solved in liquids, or even the pneumatic transport of powders in pipes. In this paper we will present the various methods we applied and developed and discuss their individual advantages.

Keywords: Particle suspensions, molecular dynamics, stochastic rotation dynamics, lattice Boltzmann method

1 Introduction

Adding a fluid to a dry granulate causes the behavior of the mixture to change dramatically and a host of unexpected phenomena arises. A good example can be studied by anyone on the beach: whereas it is impossible to build a sand castle from dry sand, once just a little bit of water has been stirred into the sand, one can shape the resulting material almost arbitrarily into surprisingly complex arrangements. Adding even more fluid might result in the material

* Research Project A7 "Microscopic Simulations of Suspensions – Analysis of Local Behavior of Polydisperse Systems and Nonspherical Particles"

loosing this stability. If we stir such a mixture, it behaves like a liquid of increased density. Other very common particle-fluid mixtures are ubiquitous in our daily life and include the cacao drink which keeps separating into its constituents, tooth paste and wall paint which are mixtures of finely ground solid ingredients in fluids or blood which is made up of red and white blood cells suspended in a solvent. An extreme example is the sand on the beach which can be blown away by the wind. It is important for industrial applications to obtain a detailed knowledge of those systems in order to optimize production processes or to prevent accidents.

Long-range fluid-mediated hydrodynamic interactions often dictate the behavior of particle-fluid mixtures. The majority of analytical results for the particle scale behavior of suspensions has been obtained in the regime of vanishing Reynolds numbers (viscous flow). For large systems, scientists aim at an average, continuum description of the large-scale behavior. However, this requires time-consuming and sometimes very difficult experimental measurements of phenomenological quantities such as the mean settling speed of a suspension, the stress contributions in the system of the individual components (solid and fluid) as functions of, e.g., the solid volume fraction of the constituents.

Computer simulation methods are indispensable for many-particle systems, for the inclusion of inertia effects (Reynolds numbers > 1) and Brownian motion (Peclet number of order 1). These systems often contain a large number of important time scales which differ by many orders of magnitude, but nevertheless have to be resolved by the simulation, leading to a large numerical effort. However, simulations have the potential to increase our knowledge of elementary processes and to enable us to find the aforementioned relations from simulations instead of experiments.

Various simulation methods have been developed to simulate particle-fluid mixtures. All of them have their inherent strengths but also some disadvantages. For example, simplified Brownian Dynamics does not contain long-ranged hydrodynamic interactions among particles at all [28]. Brownian Dynamics with full hydrodynamic interactions utilizes a mobility matrix which is based on tensor approximations which are exact in the limit of zero Reynolds number and zero particle volume fraction [55, 2]. However, the computational effort scales with the cube of the particle number due to the inversion of matrices. Pair-Drag simulations have been proposed by Silbert et al. [67], and include hydrodynamic interactions in an approximative way. They have focused on suspensions with high densities (up to 50 %) of uncharged spherical colloidal particles. Stokesian Dynamics has been presented by Bossis and Brady in the 80s and applied by many authors [7, 66, 56, 8]. For example, Melrose and Ball have performed detailed studies of shear thickening colloids using Stokesian Dynamics simulations [50, 49]. However, this method is limited to Reynolds numbers close to zero and the computational effort is very high for dynamical simulations. Even with today's powerful computers it is not possible to study the dynamics of more than a few hundred particles.

The method is still widely used due to its physical motivation and its robustness, but is complicated to code. Boundary-element methods are more flexible than Stokesian dynamics and can also be used to simulate non-spherical or deformable particles, but the computational effort is even higher [44, 40].

All these methods assume that hydrodynamic interactions are fully developed and that the dynamics of the fluid and of the solved particles can be treated as fully separated. In reality, this is not the case. Hydrodynamic interactions are time dependent due to local stresses at the fluid-particle interfaces. A number of more recent methods attempt to describe the time dependent long-range hydrodynamics properly with the computational effort scaling linearly with the number of particles. These include recent mesoscopic methods like dissipative particle dynamics [15, 14, 6], the lattice-Boltzmann method [11, 40, 38, 39, 53, 35], or stochastic rotation dynamics [46, 47, 24, 23]. However, for small Reynolds numbers, the computational gain of these methods is lost due to the additional effort needed to describe the motion of the fluid. Finite element or finite difference methods need a proper meshing of the computational domain which is not trivial for complicated boundary conditions as in the case of dense suspensions. Therefore, many authors only simulated a limited number of static configurations rather than the full dynamics of the system. Advances in remeshing techniques as well as more powerful computers have allowed to overcome these problems. Also, in order to avoid remeshing at all, uniform grids can be used [17, 63, 25]. These methods are flexible and robust. They can properly treat non-Newtonian effects and incorporate inertia, but are complicated to code.

For a more detailed description of the simulation methods, experiments or theoretical approaches not addressed in this paper, the reader is referred to one of the various books on colloid science [45, 41, 64, 52, 62, 26].

The remainder of this paper focuses on three different simulation techniques which have recently been applied to particle-laden flows in our group. First, we introduce a method developed by Malevanets and Kapral to model a solvent with thermal fluctuations. This approach is used to study the properties of claylike colloids [24, 23]. For larger particles, thermal fluctuations are undesirable. Here, the lattice Boltzmann method and its extension to particle suspension is a very good candidate to study the dynamics of glass spheres in a sugar solution [35, 22]. The method is easy to code and has been applied to suspensions of spherical and non-spherical particles by various authors. If the particles are very massive and the density of the fluid is very low, a full hydrodynamic treatment of the solvent is not needed anymore. In the last chapter we describe an algorithm based on a coarse-grained description of the fluid, so that it is resolved on a length scale larger than the particles. Much larger systems can be treated this way, but the coarse-graining is justified only in certain situations. As an example we model the pneumatic transport of a powder in a pipe which is a common process in many industrial applications [48, 68, 69].

A more computational demanding and not as easy to code method is a Navier Stokes solver for the fluid which is coupled to the particles. The method has been successfully applied to the simulation of sedimentation processes of spherical or non-spherical particles and profits from its well established physical background and long standing experience with similar fluid solvers in engineering disciplines [72, 63, 25, 37, 18, 19]. These methods have a long standing history in our group, but have been described in detail elsewhere and will not be covered in this paper.

2 Simulation of Claylike Colloids: Stochastic Rotation Dynamics

Dense suspensions of small strongly interacting particles are complex systems, which are rarely understood on the microscopic level. We investigate properties of dense suspensions and sediments of small spherical Al_2O_3 particles by means of a combined Molecular Dynamics (MD) and Stochastic Rotation Dynamics (SRD) simulation. Stochastic Rotation Dynamics is a simulation method developed by Malevanets and Kapral [46, 47] for a fluctuating fluid. The work this chapter is dealing with is presented in more detail in references [24, 23].

We simulate claylike colloids, for which in many cases the attractive Vander-Waals forces are relevant. They are often called "peloids" (Greek: claylike). The colloidal particles have diameters in the range of some nm up to some µm. The term "*peloid*" originally comes from soil mechanics, but particles of this size are also important in many engineering processes. Our model systems of Al_2O_3-particles of about half a µm in diameter suspended in water are often used ceramics and play an important role in technical processes. In soil mechanics [59] and ceramics science [54], questions on the shear viscosity and compressibility as well as on porosity of the microscopic structure which is formed by the particles, arise [73, 43]. In both areas, usually high volume fractions ($\Phi > 20\%$) are of interest. The mechanical properties of these suspensions are difficult to understand. Apart from the attractive forces, electrostatic repulsion strongly determines the properties of the suspension. Depending on the surface potential, one can either observe formation of clusters or the particles are stabilized in suspension and do sediment only very slowly. The surface potential can be adjusted by the pH-value of the solvent. Within Debye-Hückel theory one can derive a so-called $2pK$ charge regulation model which relates the simulation parameters with the pH-value and ionic strength I adjusted in the experiment. In addition to the static interactions hydrodynamic effects are also important for a complete description of the suspension. Since typical Peclet numbers are of order one in our system, Brownian motion cannot be neglected.

2.1 Molecular Dynamics (MD): Simulation of the Suspended Particles

We study colloidal particles, composing the solid fraction, suspended in a fluid solvent. The colloidal particles are simulated with molecular dynamics (MD), whereas the solvent is modeled with stochastic rotation dynamics (SRD) as described in Sect. 2.2.

In the MD part of our simulation we include effective electrostatic interactions and van der Waals attraction, a lubrication force and Hertzian contact forces. In order to correctly model the statics and dynamics when approaching stationary states, realistic potentials are needed. The interaction between the particles is described by DLVO theory [28, 61, 43]. If the colloidal particles are suspended in a solvent, typically water, ions move into solution, whereas their counter ions remain in the particle due to a different resolvability. Thus, the colloidal particle carries a charge. The ions in solution are attracted by the charge on the particles and form the electric double layer. It has been shown (see [61]), that the resulting electrostatic interaction between two of these particles can be described by an exponentially screened Coulomb potential

$$V_{\mathrm{Coul}} = \pi\varepsilon_r\varepsilon_0 \left[\frac{2+\kappa d}{1+\kappa d}\cdot\frac{4k_{\mathrm{B}}T}{ze}\tanh\left(\frac{ze\zeta}{4k_{\mathrm{B}}T}\right)\right]^2 \times \frac{d^2}{r}\exp(-\kappa[r-d])\ , \quad (1)$$

where d denotes the particle diameter and r is the distance between the particle centers. e is the elementary charge, T the temperature, k_{B} the Boltzmann constant, and z is the valency of the ions of added salt. Within DLVO theory one assumes linear screening, mainly by one species of ions with valency z (e.g. $z = +1$ for NH_4^+). The first fraction in (1) is a correction to the original DLVO potential, which takes the surface curvature into account and is valid for spherical particles [5].

The effective surface potential ζ is the electrostatic potential at the border between the diffuse layer and the compact layer, it may therefore be identified with *the* ζ-potential. It includes the effect of the bare charge of the colloidal particle itself, as well as the charge of the ions in the Stern layer, where the ions are bound permanently to the colloidal particle. In other words, DLVO theory uses a renormalized surface charge. This charge can be related to the *p*H value of the solvent within Debye-Hückel theory [23].

ε_0 is the permittivity of the vacuum, ε_r the relative dielectric constant of the solvent. κ is the inverse Debye length defined by $\kappa^2 = 8\pi\ell_B I$, with the ionic strength I and the Bjerrum length ℓ_B. We use 81 for water, which implies $\ell_B = 7\,\text{Å}$.

The Coulomb term of the DLVO potential competes with the attractive van der Waals term

$$V_{\mathrm{VdW}} = -\frac{A_{\mathrm{H}}}{12}\left[\frac{d^2}{r^2-d^2}+\frac{d^2}{r^2}\ +2\ln\left(\frac{r^2-d^2}{r^2}\right)\right]\ . \quad (2)$$

$A_{\mathrm{H}} = 4.76 \cdot 10^{-20}\,\mathrm{J}$ is the Hamaker constant [27] which involves the polarizability of the particles. It is kept constant in our simulations since it only depends on the material of the particles and on the solvent. The DLVO potentials are plotted in Fig. 1 for six typical examples with different depth of the secondary minimum. The primary minimum has to be modeled separately, as discussed below. Long range hydrodynamic interactions are taken

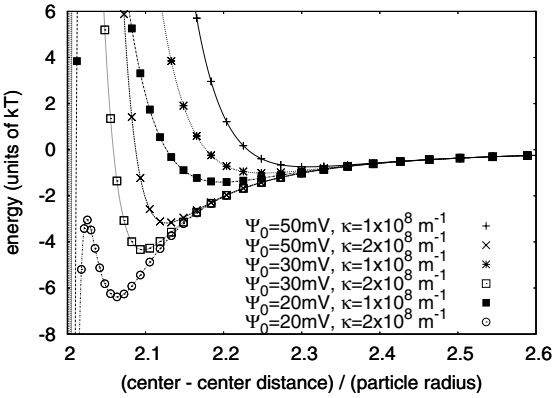


Fig. 1. DLVO potentials for Al_2O_3 spheres of $R = 0.5\,\mu\mathrm{m}$ diameter suspended in water. These are typical potentials used for our simulations as described below. The primary minimum at $d/R = 2.0$ is not reproduced correctly by the DLVO theory. It has to be modeled separately. In most of our cases the existence of the secondary minimum determines the properties of the simulated system

into account in the simulation for the fluid as described below. This can only reproduce interactions correctly down to a certain length scale. On shorter distances, a lubrication force has to be introduced explicitly in the molecular dynamics simulation. The most dominant mode, the so-called squeezing mode, is an additional force

$$\mathbf{F}_{\mathrm{lub}} = -(\mathbf{v}_{\mathrm{rel}}, \hat{\mathbf{r}})\hat{\mathbf{r}} \frac{6\pi\eta r_{\mathrm{red}}^2}{r - r_1 - r_2}\,, \tag{3}$$

$$\text{with} \quad r_{\mathrm{red}} = \frac{r_1 r_2}{r_1 + r_2} \tag{4}$$

between two spheres with radii r_1, r_2 and the relative velocity $\mathbf{v}_{\mathrm{rel}}$. η is the dynamic viscosity of the fluid. In contrast to the DLVO potentials the lubrication force is a dissipative force. When two particles approach each other very closely, this force becomes very large. To ensure numerical stability of the simulation, one has to limit $\mathbf{F}_{\mathrm{lub}}$. We do this by introducing a small cutoff radius r_{sc}. Instead of calculating $\mathbf{F}_{\mathrm{lub}}(r)$ we take the value for $\mathbf{F}_{\mathrm{lub}}(r + r_{\mathrm{sc}})$. In addition, since the force decays for large particle distances, we can introduce

a large cutoff radius r_{lc} for which we assume $\mathbf{F}_{\mathrm{lub}}(r) \equiv 0$ if $r > r_{\mathrm{lc}}$. As the intention of $\mathbf{F}_{\mathrm{lub}}$ is to correct the finite resolution of the fluid simulation, r_{sc} and r_{lc} have to be adjusted in a way that the dynamic properties, i.e., the viscosity of a simulated particle suspension with weak DLVO interactions fit the measurements. It turns out that $r_{\mathrm{sc}} = 1.05(r_1 + r_2)$ and $r_{\mathrm{lc}} = 2.5(r_1 + r_2)$ work best.

To avoid that the particles penetrate each other, one needs a repulsive force depending on their overlap. We are using a Hertz force described by the potential

$$V_{\mathrm{Hertz}} = K(d - r)^{5/2} \quad \text{if} \quad r < d \,, \tag{5}$$

where K could be expressed by the elastic modulus of Al_2O_3. This would determine the simulation time step, but to keep the computational effort relatively small, we determine the time step using the DLVO-potentials as described later on and then choose a value for K. Two aspects have to be considered: K has to be big enough so that the particles do not penetrate each other by more than approximately 10% and it may not be too big, so that numerical errors are kept small, which is the case when the collision time is resolved with about 20 time steps. Otherwise total energy and momentum are not conserved very well in the collision.

The Hertz force also contains a damping term in normal direction,

$$\mathbf{F}_{\mathrm{Damp}} = -(\mathbf{v}_{\mathrm{rel}}, \hat{\mathbf{r}})\hat{\mathbf{r}}\beta\sqrt{r - r_1 - r_2} \,, \tag{6}$$

with a damping constant β and for the transverse direction a viscous friction proportional to the relative velocity of the particle surfaces is applied.

Since DLVO theory contains the assumption of linear polarizability, it holds only for large distances, i.e., the singularity when the two spheres touch does not exist in reality. Nevertheless, there *is* an energy minimum about $30\,k_{\mathrm{B}}T$ deep, so that particles which come that close would very rarely become free again. To obtain numerical stability of our simulation, we model this minimum by a parabolic potential, some $k_{\mathrm{B}}T$ deep (e.g. $6\,k_{\mathrm{B}}T$). The depth of the minimum in our model is much less than in reality, but the probability for particles to be trapped in the minimum has to be kept low enough so that only few of them might escape during simulation time.

For the integration of the translational motion we utilize a velocity Verlet algorithm [3] to update the velocity and position of particle i according to the equations

$$\mathbf{x}_i(t + \delta t) = \mathbf{x}_i(t) + \delta t\, \mathbf{v}_i(t) + \delta t^2\, \frac{F_i(t)}{m} \,, \tag{7}$$

$$\mathbf{v}_i(t + \delta t) = \mathbf{v}_i(t) + \delta t\, \frac{F_i(t) + F_i(t + \delta t)}{2m} \,. \tag{8}$$

For the rotation, a simple Euler algorithm is applied:

$$\omega_i(t + \delta t) = \omega_i(t) + \delta t\, \mathbf{T}_i \,, \tag{9}$$

$$\vartheta_i(t + \delta t) = \vartheta_i(t) + F(\vartheta_i, \omega_i, \delta t) \,, \tag{10}$$

where $\omega_i(t)$ is the angular velocity of particle i at time t, $\mathbf{T}_i$ is the torque exerted by non central forces on the particle i, $\vartheta_i(t)$ is the orientation of particle i at time t, expressed by a quaternion, and $F(\vartheta_i, \omega_i, \delta t)$ gives the evolution of ϑ_i of particle i rotating with the angular velocity $\omega_i(t)$ at time t. The concept of quaternions [3] is often used to calculate rotational motions in simulations, because the Euler angles and rotation matrices can easily be derived from quaternions. Using Euler angles to describe the orientation would give rise to singularities for the two orientations with $\vartheta = \pm 90°$. The numerical problems related to this fact and the relatively high computational effort of a matrix inversion can be avoided using quaternions.

2.2 Stochastic Rotation Dynamics (SRD): Simulation of the Fluid

The Stochastic Rotation Dynamics method (SRD) introduced by Malevanets and Kapral [46, 47] is a promising tool for a coarse-grained description of a fluctuating solvent, in particular for colloidal and polymer suspensions. The method is also known as "Real-coded Lattice Gas" [33] or as "multi-particle-collision dynamics" (MPCD) [60]. It can be seen as a "hydrodynamic heat bath", whose details are not fully resolved but which provides the correct hydrodynamic interaction among embedded particles [42]. SRD is especially well suited for flow problems with Peclet numbers of order one and Reynolds numbers on the particle scale between 0.05 and 20 for ensembles of many particles. The method is based on so-called fluid particles with continuous positions and velocities. Each time step is composed of two simple steps: One streaming step and one interaction step. In the streaming step the positions of the fluid particles are updated as in the Euler integration scheme known from Molecular Dynamics simulations:

$$\mathbf{r}_i(t+\tau) = \mathbf{r}_i(t) + \tau\ \mathbf{v}_i(t) i\ , \tag{11}$$

where $\mathbf{r}_i(t)$ denotes the position of the particle i at time t, $\mathbf{v}_i(t)$ its velocity at time t and τ is the time step used for the SRD simulation. After updating the positions of all fluid particles they interact collectively in an interaction step which is constructed to preserve momentum, energy and particle number. The fluid particles are sorted into cubic cells of a regular lattice and only the particles within the same cell are involved in the interaction step. First, their mean velocity $\mathbf{u}_j(t') = \frac{1}{N_j(t')} \sum_{i=1}^{N_j(t')} \mathbf{v}_i(t)$ is calculated, where $\mathbf{u}_j(t')$ denotes the mean velocity of cell j containing $N_j(t')$ fluid particles at time $t' = t + \tau$. Then, the velocities of each fluid particle in cell j are updated as:

$$\mathbf{v}_i(t+\tau) = \mathbf{u}_j(t') + \mathbf{\Omega}_j(t') \cdot [\mathbf{v}_i(t) - \mathbf{u}_j(t')]\ . \tag{12}$$

$\mathbf{\Omega}_j(t')$ is a rotation matrix, which is independently chosen randomly for each time step and each cell. We use rotations about one of the coordinate axes by an angle $\pm\alpha$, with α fixed. This has been suggested by M. Strauß in [70].

The coordinate axis as well as the sign of the rotation are chosen by random, resulting in six possible rotation matrices. The mean velocity $\mathbf{u}_j(t)$ in the cell j can be seen as streaming velocity of the fluid at the position of the cell j at the time t, whereas the difference $[\mathbf{v}_i(t) - \mathbf{u}_j(t')]$ entering the interaction step can be interpreted as a contribution to the thermal fluctuations.

In order to remove low temperature anomalies and to achieve exact Galilean-invariance, we use a modification of the original algorithm [29]: all particles are shifted by the *same* random vector with components in the interval $[-a/2, a/2]$ before the collision step. Particles are then shifted back by the same amount after the collision. The random vectors of consecutive iterations are uncorrelated. Ihle and Kroll have discussed in [30, 31] why this simple procedure works and shown that it leads to transport coefficients independent of an imposed homogeneous flow field. In [32] and [34] analytical calculations of the transport coefficient of this method are presented.

Two different methods to couple the SRD and the MD simulation have been introduced in the literature. Inoue et al. proposed a way to implement no slip boundary conditions on the particle surface [33], whereas Falck et al. [16] have developed a "more coarse grained" method we describe shortly in the following section.

2.3 Coupling of the MD and the SRD Simulation Part

To couple the two parts of the simulation, MD on the one hand and SRD on the other one, the colloidal particles are sorted into the SRD boxes and their velocities are included in the rotation step. This technique has been used to model protein chains suspended in a liquid [16, 74]. Since the mass of the fluid particles is much smaller than the mass of the colloidal particles, one has to use the mass of each particle—colloidal or fluid particle—as a weight factor when calculating the mean velocity

$$\mathbf{u}_j(t') = \frac{1}{M_j(t')} \sum_{i=1}^{N_j(t')} \mathbf{v}_i(t) m_i \ , \tag{13}$$

$$\text{with} \qquad M_j(t') = \sum_{i=1}^{N_j(t')} m_i \ , \tag{14}$$

where we sum over all colloidal and fluid particles in the cell, so that $N_j(t')$ is the total number of both particles together. m_k is the mass of the particle with index i and therefore $M_j(t')$ gives the total mass contained in cell j at the time $t' = t + \tau$.

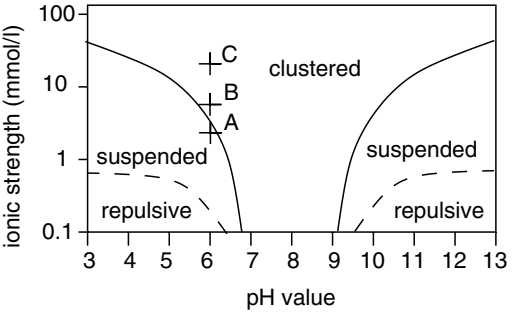


Fig. 2. Schematic phase diagram for volume fraction $\Phi = 35\%$ in terms of pH-value and ionic strength involving three different phases: a clustering regime due to van der Waals attraction, stable suspensions where the charge of the colloidal particles prevents clustering, and a repulsive structure for further increased electrostatic repulsion. This work concentrates on state A ($p\mathrm{H} = 6$, $I = 3\,\mathrm{mmol/l}$) in the suspended phase, state B ($p\mathrm{H} = 6$, $I = 7\,\mathrm{mmol/l}$) close to the phase border but already in the clustered phase, and state C ($p\mathrm{H} = 6$, $I = 25\,\mathrm{mmol/l}$) in the clustered phase [23]

2.4 Results

Phase Diagram

Depending on the experimental conditions, one can obtain three different phases: A clustered region, a suspended phase, and a repulsive structure. These phases can be reproduced in the simulations and we can quantitatively relate interaction potentials to certain experimental conditions. A schematic picture of the phase diagram is shown in Fig. 2. Close to the isoelectric point ($p\mathrm{H} = 8.7$), the particles form clusters for all ionic strengths since they are not charged. At lower or higher pH values one can prepare a stable suspension for low ionic strengths because of the charge, which is carried by the colloidal particles. At even more extreme pH values, one can obtain a repulsive structure due to very strong electrostatic potentials (up to $\zeta = 170\,\mathrm{mV}$ for $p\mathrm{H} = 4$ and $I = 1\,\mathrm{mmol/l}$, according to our model). The repulsive structure is characterized by an increased shear viscosity. In the following we focus on three states: State A ($p\mathrm{H} = 6$, $I = 3\,\mathrm{mmol/l}$) is in the suspended phase, state B ($p\mathrm{H} = 6$, $I = 7\,\mathrm{mmol/l}$) is a point already in the clustered phase but still close to the phase border, and state C ($p\mathrm{H} = 6$, $I = 25\,\mathrm{mmol/l}$) is located well in the clustered phase.

Some typical examples for the different phases are shown in Figs. 3a)–d). These examples are meant to be only illustrative and do not correspond exactly to the cases A–C in Fig. 2 denoted by uppercase letters. In the suspended case (a), the particles are mainly coupled by hydrodynamic interactions. One can find a linear velocity profile and a slight shear thinning. If one increases

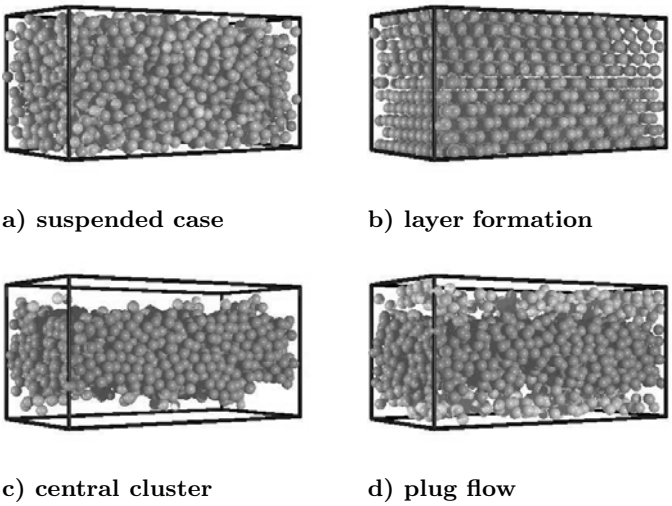


Fig. 3. Images of four different cases. For better visibility we have chosen smaller systems than we usually use for the calculation of the viscosity. The colors denote velocities: Dark particles are slow, bright ones move fast. The potentials do not correspond exactly to the cases *A*–*C* in Fig. 2, but they show qualitatively the differences between the different states: **a)** Suspension like in state *A*, at low shear rates. **b)** Layer formation, which occurs in the repulsive regime, but also in the suspension (state *A*) at high shear rates. **c)** Strong clustering, like in state *C*, so that the single cluster in the simulation is deformed. **d)** Weak clustering close to the phase border like in state *B*, where the cluster can be broken into pieces, which follow the flow of the fluid (plug flow)

the shear rate $\dot{\gamma} > 500/\mathrm{s}$, the particles arrange in layers. The same can be observed if the Debye-screening length of the electrostatic potential is increased (b), which means that the solvent contains less ions ($I < 0.3\,\mathrm{mmol/l}$) to screen the particle charges. On the other hand, if one increases the salt concentration, electrostatic repulsion is screened even more and attractive van der Waals interaction becomes dominant ($I > 4\,\mathrm{mmol/l}$). Then the particles form clusters (c), and viscosity rises. A special case, called "plug flow", can be observed for high shear rates, where it is possible to tear the clusters apart and smaller parts of them follow with the flow of the solvent (d). This happens in our simulations for $I = 25\,\mathrm{mmol/l}$ (state *C*) at a shear rate of $\dot{\gamma} > 500/\mathrm{s}$. However, as long as there are only one or two big clusters in the system, it is too small to expect quantitative agreement with experiments. In these cases we have to focus on state *B* ($I = 7\,\mathrm{mmol/l}$) close to the phase border.

We restrict ourselves to the region around $p\mathrm{H} = 6$ where we find the phase border between the suspended region and the clustered regime at about $I = 4\,\mathrm{mmol/l}$ in the simulations as well as in the experiments. Also the shear

rate dependence of the viscosity is comparable in simulations and experiments as discussed below.

Shear Profile and Shear Viscosity

In each of the three phases a typical velocity profile of the shear flow occurs. In the suspended phase one finds a linear velocity profile (Fig. 4a)) with nearly Newtonian flow. The particles are distributed homogeneously, thus the density profile is structureless (Fig. 5a)). The motion of the particles is only weakly coupled by the hydrodynamic forces. At high enough shear rates ($\dot{\gamma} > 500$) the particles arrange in layers parallel to the shear plane, as can be seen in the density profile Fig. 5b), too. This arrangement minimizes collisions between the particles. As a result, the shear viscosity descents as shown in Fig. 6, which we discuss in more detail below. Shear induced layer formation has been reported in literature for different systems. Vermant and Solomon have reviewed this topic recently [71].

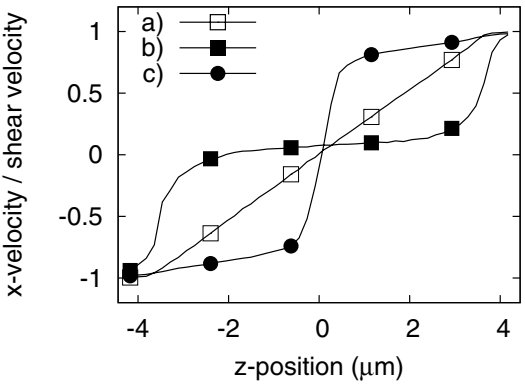


Fig. 4. Profiles of tangential velocity component in normal direction: a) Linear profile in the suspended regime, state A of Fig. 2 ($I = 3\,\mathrm{mmol/l}$) at $\dot{\gamma} = 500/\mathrm{s}$)
b) Cluster formation in state C ($I = 25\,\mathrm{mmol/l}$) at $\dot{\gamma} = 100/\mathrm{s}$. In principle one could determine the viscosity of one single cluster from the central plateau, but this is not the viscosity found in experiments. There, one measures the viscosity of a paste consisting of many of these clusters
c) Same as case b) but with higher shear rate (500/s). Hydrodynamic forces are large enough to break the cluster into two pieces. The velocity axis is scaled with the shear velocity v_S for better comparability

In the clustered phase, the clusters move in the fluid as a whole. They are deformed, but since the inter-particle forces are stronger than the hydrodynamic forces, the cluster moves more like a solid body than like a fluid.

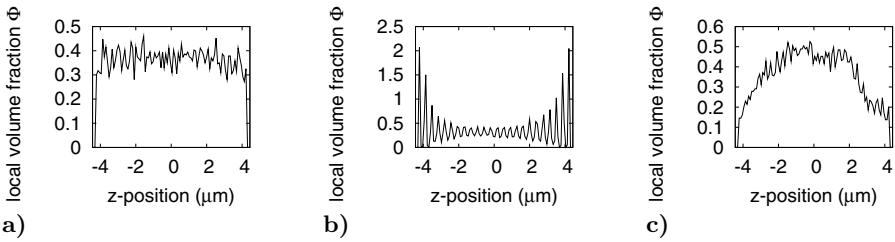


Fig. 5. Density profiles: **a)** Suspended case: State A in Fig. 2 ($I = 3\,\mathrm{mmol/l}$), at low shear rates ($\dot{\gamma} = 50/\mathrm{s}$). The density distribution is homogeneous
b) Shear induced layer formation: This is state A as in graph a) of this figure, but for a high shear rate ($\dot{\gamma} = 1000/\mathrm{s}$)
c) Strong attractive forces in state C ($I = 25\,\mathrm{mmol/l}$): For low shear rates ($\dot{\gamma} = 50/\mathrm{s}$) only one central cluster is formed, which is deformed slowly

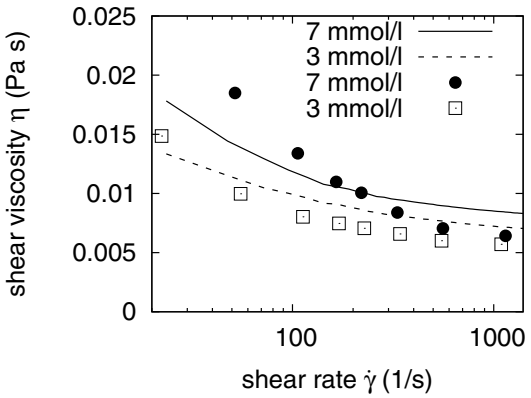


Fig. 6. Comparison between simulation and experiment: viscosity in dependence of the shear rate for the states A ($I = 3\,\mathrm{mmol/l}$) and B ($I = 7\,\mathrm{mmol/l}$) of Fig. 2. Note: shear thinning is more pronounced for the slightly attractive interactions in state B than for the suspended state A. Lines denote experimental data [58], points are results from our simulations

Often there is one big cluster that spans the whole system. The density profile (Fig. 5c)) increases in the central region and decays at the regions close to the border, since particles from there join the central cluster. When averaging the velocity profile in the shear flow, one finds a very small velocity gradient in the center of the shear cell and fast moving particles close to the wall, where the shear is imposed (Fig. 4b)). The velocity profile is non-linear on the length scale of the simulations. In the experiment the physical dimensions are much larger and therefore the velocity profile can become approximately linear again if the system consists of many large clusters. However, due to

the computational effort in simulations it is today impossible to measure the shear viscosity for these strongly inhomogeneous systems.

Closer to the phase border clusters can then be broken up into small pieces by the hydrodynamic forces at least for high shear rates. In state C of Fig. 2 this happens for the first time at $\dot{\gamma} = 500/\mathrm{s}$, so that one can find two clusters in the system moving in opposite directions. The velocity profile of this case is shown in Fig. 4c). For even higher shear rates or closer to the phase border (e.g. state B), the clusters are broken into smaller pieces. Then, they move in the shear flow with an approximately linear velocity profile. Due to van der Waals attraction the system resists with stronger shear forces and the viscosity is higher than in the suspended case (Fig. 6).

In Fig. 6 the simulation results are shown together with the experimental results, both for the two cases of a slightly clustered system in state B ($I = 7\,\mathrm{mmol/l}$) and a suspension (state A, $I = 3\,\mathrm{mmol/l}$). For the suspension (state A) the viscosity decreases with the shear rate ("shear thinning"). The experimental data and the simulation are consistent within the accuracy of our model. There are several reasons for which our model does not fit exactly the measurements: Even though we use a charge regulation model to determine the input parameters for the DLVO potentials, microscopic properties like the surface density of sites, where ions can be adsorbed on the surface of the colloidal particle, have to be determined indirectly. Measurements of the ζ-potential in certain conditions provide data to fit the unknown microscopic parameters. Furthermore, we have monodisperse spheres, which is another simplification in our model.

For the slightly clustered case (state B) an increase of the shear viscosity, compared to the suspended case, can be observed in the experiment as well as in the simulations. Shear thinning becomes more pronounced, because clusters are broken up, as mentioned above. However, the shear rate dependence is stronger in the simulations than in the experiment. This can be the first indication of finite size effects. We have studied the dependence of the simulated shear viscosity in dependence of the system size. The effect is most important for low shear rates.

3 Transport Phenomena and Structuring in Suspensions: Lattice-Boltzmann Simulations

For industrial applications, systems with rigid boundaries, e.g. a pipe wall, are of particular interest since structuring effects might occur in the solid fraction of the suspension. Such effects are known from dry granular media resting on a plane surface or gliding down an inclined chute [51, 57]. In addition, the wall causes a demixing of the solid and fluid components which might have an unwanted influence on the properties of the suspension. Near the wall one finds a thin lubrication layer which contains almost no particles and causes a

so-called "pseudo wall slip". Due to this slip the suspension can be transported substantially faster and less energy is dissipated.

We expect structuring close to a rigid wall at much smaller concentrations than in granular media because of long-range hydrodynamic interactions. In [35], we study these effects by the means of particle volume concentrations versus distance to the wall.

3.1 The Lattice-Boltzmann Method

The lattice-Boltzmann method is a simple scheme for simulating the dynamics of fluids. By incorporating solid particles into the model fluid and imposing the correct boundary condition at the solid/fluid interface, colloidal suspensions can be studied. Pioneering work on the development of this method has been done by Ladd et al. [38, 39, 40] and we use their approach to model sheared suspensions near solid walls.

The lattice-Boltzmann (hereafter LB) simulation technique which is based on the well-established connection between the dynamics of a dilute gas and the Navier-Stokes equations [10]. We consider the time evolution of the one-particle velocity distribution function $n(\mathbf{r}, \mathbf{v}, t)$, which defines the density of particles with velocity $\mathbf{v}$ around the space-time point $(\mathbf{r}, t)$. By introducing the assumption of molecular chaos, i.e. that successive binary collisions in a dilute gas are uncorrelated, Boltzmann was able to derive the integro-differential equation for n named after him [10]

$$\partial_t n + \mathbf{v} \cdot \nabla n = \left(\frac{dn}{dt} \right)_{coll} , \tag{15}$$

where the left hand side describes the change in n due to collisions.

The LB technique arose from the realization that only a small set of discrete velocities is necessary to simulate the Navier-Stokes equations [21]. Much of the kinetic theory of dilute gases can be rewritten in a discretized version. The time evolution of the distribution functions n is described by a discrete analogue of the Boltzmann equation [40]:

$$n_i(\mathbf{r} + \mathbf{c}_i \Delta t, t + \Delta t) = n_i(\mathbf{r}, t) + \Delta_i(\mathbf{r}, t) , \tag{16}$$

where Δ_i is a multi-particle collision term. Here, $n_i(\mathbf{r}, t)$ gives the density of particles with velocity $\mathbf{c}_i$ at $(\mathbf{r}, t)$. In our simulations, we use 19 different discrete velocities $\mathbf{c}_i$. The hydrodynamic fields, mass density ϱ, momentum density $\mathbf{j} = \varrho \mathbf{u}$, and momentum flux Π, are moments of this velocity distribution:

$$\varrho = \sum_i n_i , \qquad \mathbf{j} = \varrho \mathbf{u} = \sum_i n_i \mathbf{c}_i , \qquad \Pi = \sum_i n_i \mathbf{c}_i \mathbf{c}_i . \tag{17}$$

We use a linear collision operator,

$$\Delta_i(r,t) = M_{ij}(n_j - n_j^{eq}) \,, \tag{18}$$

where $M_{ij} \equiv \frac{\partial \Delta_i(n^{eq})}{\partial n_j}$ is the collision matrix and n_i^{eq} the equilibrium distribution [11], which determines the scattering rate between directions i and j. For mass and momentum conservation, M_{ij} satisfies the constraints

$$\sum_{i=1}^{M} M_{ij} = 0 \,, \qquad \sum_{i=1}^{M} \mathbf{e}_i M_{ij} = 0 \,. \tag{19}$$

We further assume that the local particle distribution relaxes to an equilibrium state at a single rate τ and obtain the lattice BGK collision term [4]

$$\Delta_i = -\frac{1}{\tau}(n_i - n_i^{eq}) \,. \tag{20}$$

By employing the Chapman-Enskog expansion [10, 20] it can be shown that the equilibrium distribution

$$n_i^{eq} = \varrho \omega^{c_i} \left[1 + 3\mathbf{c}_i \cdot \mathbf{u} + \frac{9}{2}(\mathbf{c}_i \cdot \mathbf{u})^2 - \frac{3}{2}u^2 \right] \,, \tag{21}$$

with the coefficients of the three velocities

$$\omega^0 = \frac{1}{3} \,, \qquad \omega^1 = \frac{1}{18} \,, \qquad \omega^{\sqrt{2}} = \frac{1}{36} \,, \tag{22}$$

and the kinematic viscosity [40]

$$\nu = \frac{\eta}{\varrho_f} = \frac{2\tau - 1}{9} \,, \tag{23}$$

properly recovers the Navier-Stokes equations

$$\frac{\partial u}{\partial t} + (u\nabla)u = -\frac{1}{\varrho}\nabla p + \frac{\eta}{\varrho}\Delta u \,, \qquad \nabla u = 0 \,. \tag{24}$$

3.2 Fluid-Particle Interactions

To simulate the hydrodynamic interactions between solid particles in suspensions, the lattice-Boltzmann model has to be modified to incorporate the boundary conditions imposed on the fluid by the solid particles. Stationary solid objects are introduced into the model by replacing the usual collision rules (Equation (20)) at a specified set of boundary nodes by the "link-bounce-back" collision rule [53]. When placed on the lattice, the boundary surface cuts some of the links between lattice nodes. The fluid particles moving along these links interact with the solid surface at boundary nodes placed halfway along the links. Thus, a discrete representation of the surface is obtained, which becomes more and more precise as the surface curvature gets smaller and which is exact for surfaces parallel to lattice planes.

Since the velocities in the lattice-Boltzmann model are discrete, boundary conditions for moving suspended particles cannot be implemented directly. Instead, we can modify the density of returning particles in a way that the momentum transferred to the solid is the same as in the continuous velocity case. This is implemented by introducing an additional term Δ_b in (16) [38]:

$$\Delta_{b,i} = \frac{2\omega^{c_i}\varrho_i \mathbf{u}_i \cdot \mathbf{c}_i}{c_s^2} , \tag{25}$$

with c_s being the velocity of sound and coefficients ω^{c_i} from (22).

To avoid redistributing fluid mass from lattice nodes being covered or uncovered by solids, we allow interior fluid within closed surfaces. Its movement relaxes to the movement of the solid body on much shorter time scales than the characteristic hydrodynamic interaction [38].

If two particle surfaces approach each other within one lattice spacing, no fluid nodes are available between the solid surfaces. In this case, mass is not conserved anymore since boundary updates at each link produce a mass transfer $\Delta_b a^3$ ($a \equiv$cell size) across the solid-fluid interface [38]. The total mass transfer for any closed surface is zero, but if some links are cut by two surfaces, no solid-fluid interface is available anymore. Instead, the surface of each particle is not closed at the solid-solid contacts anymore and mass can be transferred in-between suspended particles. Since fluid is constantly added or removed from the individual particles, they never reach a steady state. In such cases, the usual boundary-node update procedure is not sufficient and a symmetrical procedure which takes account of both particles simultaneously has to be used [39]. Thus, the boundary-node velocity is taken to be the average of that computed from the velocities of each particle. Using this velocity, the fluid populations are updated (Equation (25)), and the force is computed; this force is then divided equally between the two particles.

If two particles are in near contact, the fluid flow in the gap cannot be resolved by LB. For particle sizes used in our simulations ($R < 5a$), the lubrication breakdown in the calculation of the hydrodynamic interaction occurs at gaps less than $0.1R$ [53]. This effect "pushes" particles into each other.

To avoid this force, which should only occur on intermolecular distances, we use a lubrication correction method described in [53]. For each pair of particles a force

$$\mathbf{F}_{\text{lub}} = -6\pi\eta \frac{R_1 R_2}{(R_1+R_2)^2} \left(\frac{1}{h} - \frac{1}{h_N}\right) \mathbf{u}_{12} \cdot \frac{\mathbf{r}_{12}}{|\mathbf{r}_{12}|} , \qquad h < h_N \tag{26}$$

is calculated, where $\mathbf{u}_{12} = \mathbf{u}_1 - \mathbf{u}_2, h = |\mathbf{r}_{12}| - R_1 - R_2$ is the gap between the two surfaces and a cut off distance $h_N = \frac{2}{3}a$ [40]. For particle-wall contacts we apply the same formula with $R_2 \to \infty$ and $h = |\mathbf{r}_{12}| - R_1$. The tangential lubrication can also be taken into account, but since it has a weaker logarithmic divergence and its breakdown does not lead to serious problems, we do not include it in our simulations.

3.3 Particle Motion

The particle position and velocity are calculated using Newton's equations in a similar manner as in section on SRD simulations. To avoid repetition, the reader is referred to Sect. 2.1. However, particles do not feel electrostatic interactions, but behave like hard spheres in the case presented in this section.

3.4 Simulations

The purpose of our simulations is the reproduction of rheological experiments on computers. We simulate a representative volume element of the experimental setup and compare our calculations with experimentally accessible data, i.e. density profiles, time dependence of shear stress and shear rate. We also get experimentally inaccessible data from our simulations like translational and rotational velocity distributions, particle-particle and particle-wall interaction frequencies. The experimental setup consists of a rheoscope with two spherical plates, which distance can be varied. The upper plate can be rotated either by exertion of a constant force or with a constant velocity, while the complementary value is measured simultaneously. The material between the rheoscope plates consist of glass spheres suspended in a sugar-water solution. The radius of the spheres varies between 75 and 150 μm. For our simulations we assume an average particle radius of 112.5 μm. The density and viscosity of the sugar solution can also be changed. We simulate only the behavior of a representative volume element which has the experimental separation between walls, but a much lower extension in the other two dimensions than the experiment. In these directions we employ periodic boundary conditions for particles and for the fluid.

Shearing is implemented using the "link-bounce-back" rule with an additional term $\Delta_{b,i}$ at the wall in the same way as already described for particles (Equation (25) with $\mathbf{u}_i$ now being the velocity of the wall).

To compare the numerical and experimental results, we need to find characteristic dimensionless quantities of the experiment which then determine the simulation parameters. For this purpose we use the ratio of the rheoscope height and the particle size λ, the particle Reynolds number $\Re$ and the volume fraction of the particles ϕ. The simulation results are provided with units by calculating the length of the lattice constant a and the duration of one time step as described in [35].

3.5 Results

Figure 7 shows a snapshot of a suspension with 50 spheres after 5772500 time steps which are equivalent to 729 s. The vector $\mathbf{g}$ represents the direction of gravity and $\mathbf{v}_S$ depicts the velocity of the sheared wall.

The particles feel a gravitational acceleration $g = 0.8\ \mathrm{m/s^2}$, have a mass $m = 7.7 \cdot 10^{-8}$ kg, a Reynolds number $\Re = 4.066875 \cdot 10^{-4}$, and a radius

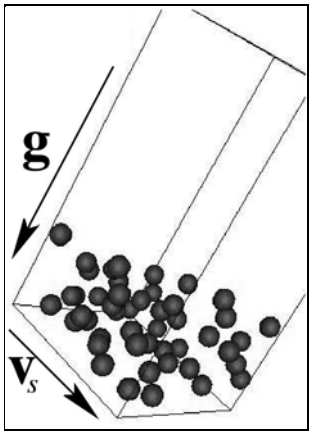


Fig. 7. A snapshot of a suspension with 50 spheres (radius $R = 1.125 \cdot 10^{-4}$ m, mass $m = 7.7 \cdot 10^{-8}$ kg) at time $t = 729$ s. The volume of the simulated system is $1.83 \cdot 10^{-3} \times 1.83 \cdot 10^{-3} \times 3.375 \cdot 10^{-3}$ m $= 11.3025 \cdot 10^{-9}$ m^3, acceleration of gravity $g = 0.80$ m/s^2, and shear velocity $v_s = 3.375 \cdot 10^{-2}$ m/s. The fluid has a viscosity $\eta = 450$ mPa $\cdot$ s and density $\varrho_f = 1446 \frac{\text{kg}}{\text{m}^3}$. This visualization is a typical example for a system that has reached a steady state: All particles have fallen to the ground due to the exerted gravitational force and most of the system has no particles [35]

$R = 1.125 \cdot 10^{-4}$ m. The system size is $1.83 \cdot 10^{-3} \times 1.83 \cdot 10^{-3} \times 3.375 \cdot 10^{-3}$ m which corresponds to a lattice size of $32 \times 32 \times 59$. The density of the fluid is set to $\varrho_f = 1446 \frac{\text{kg}}{\text{m}^3}$ and its viscosity is $\eta = 450$ mPa $\cdot$ s. The walls at the top and the bottom are sheared with a relative velocity $v_s = 3.375 \cdot 10^{-2}$ m/s. Figure 7 is a representative visualization of our simulation data and demonstrates that after the system has reached its steady state, all particles have fallen to the ground due to the influence of the gravitational force. Most of the simulation volume is free of particles.

In order to quantitatively characterize structuring effects, we calculate the particle density profile of the system by dividing the whole system into layers parallel to the walls and calculating a partial volume V_{ij} for each particle i crossing such a layer j. The scalar V_{ij} is given by the volume fraction of particle i that is part of layer j:

$$V_{ij} = \pi \left(R^2 \left(R_{ij}^{\max} - R_{ij}^{\min} \right) - \frac{1}{3} \left(R_{ij}^{\max} - R_{ij}^{\min} \right) \right) \tag{27}$$

If the component $r_{i,z}$ perpendicular to the wall of the radius vector $\mathbf{r}_i$ of the center of sphere i lies between $r_j^{\min}$ and $r_j^{\max}$, we have

$$r_j^{\min} = \left(j - \frac{1}{2} \right) \Delta L_z - R \,,$$
$$r_j^{\max} = \left(j + \frac{1}{2} \right) \Delta L_z + R \,,$$

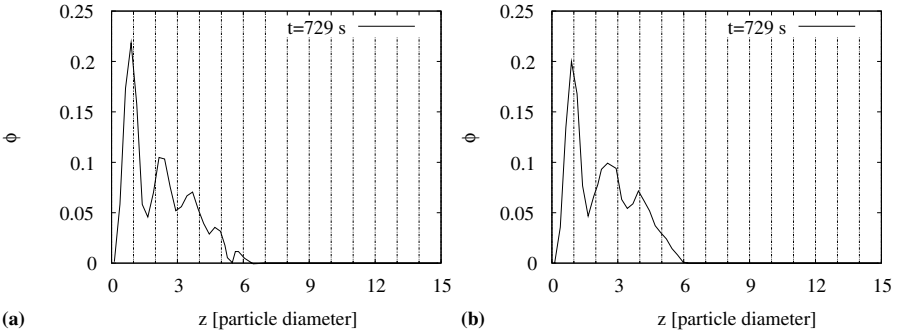


Fig. 8. Density profiles from simulations with two different shear rates $\gamma = 10\ \mathrm{s}^{-1}$ (**a**) and $\gamma = 1\ \mathrm{s}^{-1}$ (**b**). Other parameters are equal to those given in Fig. 7. (**a**) shows five peaks with separations about one particle diameter, which reveal the forming of particle layers. The number of particles per layer is decreasing with increasing distance to the wall, and the change in particle numbers is caused by gravity which is directed perpendicular to the wall at $z = 0$. Although we used the same gravity and particle numbers, there are only three peaks in (**b**) and their width is higher than in (**a**), demonstrating that the structuring effects strongly relate to the shear rate

and

$$R_{ij}^{\max} = \begin{cases} R & \text{if } r_{i,z} + R < r_j^{\max} \\ r_j^{\max} - r_{i,z} & \text{else} \end{cases} ,$$

$$R_{ij}^{\min} = \begin{cases} -R & \text{if } r_{i,z} - R > r_j^{\min} \\ r_j^{\min} - r_{i,z} & \text{else} \end{cases} .$$

Finally, the sum of all weights associated with a layer is divided by the volume of the layer

$$\phi_j = \frac{1}{L_x \cdot L_y \cdot \Delta L_z} \sum_{i=1}^{N} v_{ij}, \qquad \Delta L_z = \frac{L_z}{M} , \tag{28}$$

with L_x, L_y being the system dimensions between periodic boundaries, L_z the distance between walls, M the number of layers, and ΔL_z the width of a single layer.

Density profiles calculated by this means for systems with two different shear rates $\gamma = 10\ \mathrm{s}^{-1}$ and $\gamma = 1\ \mathrm{s}^{-1}$ are presented in Fig. 8. All other parameters are equal to the set given in the last paragraph. The peaks in Fig. 8 demonstrate that at certain distances from the wall the number of particles is substantially higher than at other positions. The first peak in both figures is slightly below one particle diameter, which can be explained by a lubricating fluid film between the first layer and the wall which is slightly thinner than one particle radius. Due to the small amount of particles, time dependent fluctuations of the width of the lubricating layer cannot be neglected and

a calculation of the exact value is not possible. The five peaks in Fig. 8**a** have similar distances which are equal to one particle diameter. These peaks can be explained by closely packed parallel layers of particles. Due to the linear velocity profile in z-direction of the fluid flow, every layer adopts the local velocity of the fluid resulting in a relative velocity difference between two layers of about $2R\gamma$. These layers stay stable in time with only a small number of particles being able to be exchanged between them.

Figure 8**b** only shows three peaks with larger distances than in Fig. 8**a**. However, the average slope of the profile is identical for both shear rates. For smaller shear rates, velocity differences between individual layers are smaller, too. As a result, particles feel less resistance while moving from one layer to another. Every inter-layer transition distorts the well defined peak structure of the density distribution resulting in only three clearly visible peaks in Fig. 8**b**.

With changing time, the first peak stays constant for both shear rates. The shape, number and position of all other peaks is slightly changing in time.

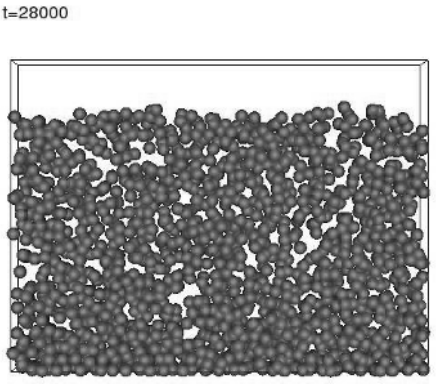


Fig. 9. A snapshot of a suspension with 1536 spheres after 28000 timesteps used to gain statistics of particle velocity distributions

We are currently investigating the occurrence of non-Gaussian velocity distributions of particles for higher particle densities and higher shear rates. For this, improvements of the method are mandatory in order to prevent instabilities of the simulation. By utilizing an implicit scheme for the update of the particle velocities [40, 53] we are able to overcome artefacts caused by numerical inaccuracies at high volume fractions or shear rates. Figure 9 shows a snapshot of a system containing 1536 particles after 28000 timesteps.

The lattice Boltzmann has been extended in order to include thermal fluctuations [40, 1]. With these modifications the method is another candidate to simulate suspensions where Brownian motion cannot be neglected.

4 Plug Conveying in Vertical or Horizontal Tubes: a Coarse Grained Model for the Fluid Flow

4.1 Model Description

Another approach to modeling two phase flow is to course-grain the fluid, so that it is resolved on a length scale larger than the grains. The advantage is that much larger systems can be treated, but the disadvantage is that this coarse-graining is justified only in certain situations. One of those situations is when the density of the fluid is small compared to that of the grains, and the Reynolds number of the grains is small. It is then possible to neglect the inertia of the fluid, which means that all momentum is contained in the grains. The fluid transfers momentum between grains, but stores no momentum itself. And when the Reynolds number of the grains is small, one can treat the granulate phase as a moving porous medium. In the following, we present the model in more detail.

Gas Model

The model for the gas simulation was first introduced by McNamara and Flekkøy [48] and has been implemented for the two-dimensional case to simulate the rising of bubbles within a fluidized bed. We developed a three-dimensional version of this algorithm.

The algorithm is based on the mass conservation of the gas and the granular medium. Conservation of grains implies that the density ϱ_p of the granular medium obeys

$$\frac{\partial \varrho_p}{\partial t} + \boldsymbol{\nabla} \cdot (\mathbf{u}\varrho_p) = 0 \,, \qquad \varrho_p = \varrho_s(1-\phi) \,, \tag{29}$$

where ϱ_p is the mass density of the material making up the particles, the porosity of the medium is ϕ (i.e. the fraction of the space available to the gas), and the velocity of the granulate is $\mathbf{u}$.

The mass conservation equation for the gas is

$$\frac{\partial \varrho_g}{\partial t} + \boldsymbol{\nabla} \cdot (\mathbf{v}_g \varrho_g) = 0 \,, \qquad \varrho_g \propto \phi P \,, \tag{30}$$

where ϱ_g is the mass density of the gas averaged over the total volume of the granular medium and $\mathbf{v}_g$ its velocity. This equation can be transformed into a differential equation for the gas pressure P using the ideal gas equation, together with the assumption of uniform temperature.

The velocity $\mathbf{v}_g$ of the gas is related to the granulate velocity $\mathbf{u}$ through the d'Arcy relation:

$$-\boldsymbol{\nabla} P = \frac{\eta}{\kappa(\phi)} \phi(\mathbf{v}_g - \mathbf{u}) \,, \tag{31}$$

where η is the dynamic viscosity of the air and κ is the permeability of the granular medium. This relation was first given by d'Arcy in 1856 [13]. The d'Arcy relation is preferred here over the Ergun equation, because it is linear in the velocity. This makes the simplification steps done later possible. For the permeability κ the Carman-Kozeny relation [9] was chosen, which provides a relation between the porosity ϕ, the particle diameter d and the permeability of a granular medium of monodisperse spheres,

$$\kappa(\phi) = \frac{d^2 \phi^3}{180(1-\phi)^2} \,. \tag{32}$$

Combining (29), (30) and (31) results in a nonlinear differential equation for the gas pressure:

$$\phi(\frac{\partial P}{\partial t} + \mathbf{u}\boldsymbol{\nabla} P) = \boldsymbol{\nabla}(P\frac{\kappa(\phi)}{\eta}\boldsymbol{\nabla} P) - P\boldsymbol{\nabla}\mathbf{u} \,. \tag{33}$$

After linearizing around the normal atmospheric pressure P_0 the resulting differential equation only depends on the relative pressure P' ($P = P_0 + P'$), the porosity ϕ and the granular velocity $\mathbf{u}$, which can be derived from the particle simulation, and three constants: the viscosity η, the particle diameter d and the pressure P_0:

$$\frac{\partial P'}{\partial t} = \frac{P_0}{\eta\phi}\boldsymbol{\nabla}(\kappa(\phi)\boldsymbol{\nabla} P') - \frac{P_0}{\phi}\boldsymbol{\nabla}\mathbf{u} \,. \tag{34}$$

This differential equation can be interpreted as a diffusion equation with a diffusion constant $D = \phi\kappa(\phi)/\eta$. The equation is solved numerically, using a Crank-Nickelson approach for the discretization. Each dimension is integrated separately.

Granulate Algorithm

The model for the granular medium simulates each grain individually using a discrete element simulation (DES). For the implementation of the discrete element simulation we used a version of the molecular dynamics method described by Cundall [12]. The particles are approximated as monodisperse spheres, rotations in three dimensions are taken into account.

The equation of motion for an individual particle is

$$m\ddot{\mathbf{x}} = m\mathbf{g} + \mathbf{F}_c - \frac{m\boldsymbol{\nabla} P}{\varrho_s(1-\phi)} \,, \tag{35}$$

where m is the mass of a particle, $\mathbf{g}$ the gravitation constant and $\mathbf{F}_c$ the sum over all contact forces. The last term, the drag force, is assumed to be a

volume force given by the pressure drop ∇P and the local mass density of the granular medium $\varrho_s(1-\phi)$, which is valid for monodisperse granular media.

The interaction between two particles in contact is given by two force components: a normal and a tangential component with respect to the particle surface. The normal force is the sum of a repulsive elastic force (Hooke's law) and a viscous damping. The tangential force opposes the relative tangential motion and is proportional to the normal force (sliding Coulomb friction) or proportional to the relative tangential velocity (viscous damping). Viscous damping is used only for small relative tangential velocities.

Gas-Grain Interaction

The simulation method uses both a continuum and a discrete element approach. While the gas algorithm uses fields, which are discretized on a cubic grid, the granulate algorithm describes particles in a continuum. A mapping is needed for the algorithms to interact. For the mapping a tent function $F(\mathbf{r})$ is used:

$$F(\mathbf{r}) = f(x)f(y)f(z), \qquad f(x) = \begin{cases} 1-|x/l|, & |x/l| \leq 1\ , \\ 0\ , & 1 < |x/l|\ , \end{cases} \tag{36}$$

where l is the grid constant used for the discretization of the gas simulation.

For the gas algorithm the porosity ϕ_j and the granular velocity $\mathbf{u}_j$ must be derived from the particle positions $\mathbf{r}_i$ and velocities $\mathbf{v}_i$, where i is the index of particle and j is the index for the grid node. The tent function distributes the particle properties around the particle position smoothly on the grid:

$$\phi_j = 1 - \sum_i F(\mathbf{r}_i - \mathbf{r}_j)\ , \qquad \mathbf{u}_j = \frac{1}{1-\phi_j} \sum_i \mathbf{v}_i F(\mathbf{r}_i - \mathbf{r}_j)\ , \tag{37}$$

where r_j is the position of the grid point and the sum is taken over all particles.

For the computation of the drag force on a particle the pressure drop ∇P_i and the porosity ϕ_i at the position of the particle are needed. These can be obtained by a linear interpolation of the fields ∇P_j and ϕ_j from the gas algorithm:

$$\phi_i = \sum_j \phi_j F(\mathbf{r}_j - \mathbf{r}_i)\ , \qquad \nabla P_i = \sum_j \nabla P_j F(\mathbf{r}_j - \mathbf{r}_i)\ , \tag{38}$$

where the sum is taken over all grid points. Note that ∇P_i is a continuous function of the particle position $\mathbf{r}_i$. There are no discontinuities at all boundaries.

4.2 Application to Plug Conveying

This method was applied to study plug conveying in both vertical [68] and horizontal [69] tubes. Plug conveying is a special case of pneumatic conveying,

where grains are driven through pipes by air flow. Plug conveying occurs when the flux of grains through the pipe is relatively high. Currently plug conveying is gaining importance in industry, because it causes a lower product degradation and pipeline erosion than dilute phase conveying.

Unfortunately, current models [36, 65] of plug conveying disagree even on the prediction of such basic quantities as the pressure drop and the total mass flow, and these quantities have a great impact in industrial applications. One of the reasons for the lack of valid models is that it is difficult to study plugs experimentally in a detailed way. Usually experimental setups are limited to the measurement of the local pressure drop, the total mass flux and the velocity of plugs. Simulational studies are handicapped by the high computational costs for solving the gas flow and the particle-particle interaction, and are therefore mostly limited to two dimensions.

Using the above-described method, we were able to provide a detailed view of plugs. This approach provides access to important parameters like the porosity and velocity of the granulate and the shear stress on the wall at relatively low computational costs. Contrary to the experiments, it is possible to access these parameters at high spatial resolution and without influencing the process of transportation at all. Additional to plug profiles, characteristic curves of the pressure drop and the influence of simulation parameters can be measured.

4.3 Results

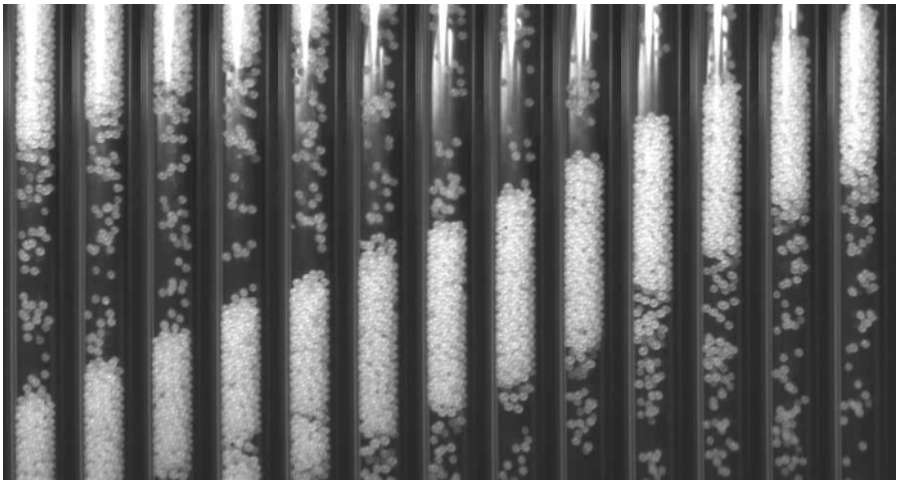


Fig. 10. A series of photos showing a plug moving upwards. The height of the shown tube is 9 cm, the frame rate is 30 Hz

In Fig. 10, we show a series of photos of plug conveying taken by Karl Sommer and Gerhard Niederreiter of TU München. The particles are wax beads of diameter $d = 1.41\,\mathrm{mm}$, density $\varrho_s = 937\,\mathrm{kg/m^3}$ and a Coulomb coefficient of 0.21. The experimental transport channel is a vertical tube (PMMA) of length $l = 1.01\,\mathrm{m}$ and of internal diameter $D_t = 7\,\mathrm{mm}$. The air is injected at a constant flow rate of $2.2\,\ell/\mathrm{min}$ at the bottom of the tube. As one can easily see, grains travel in clusters up the tube.

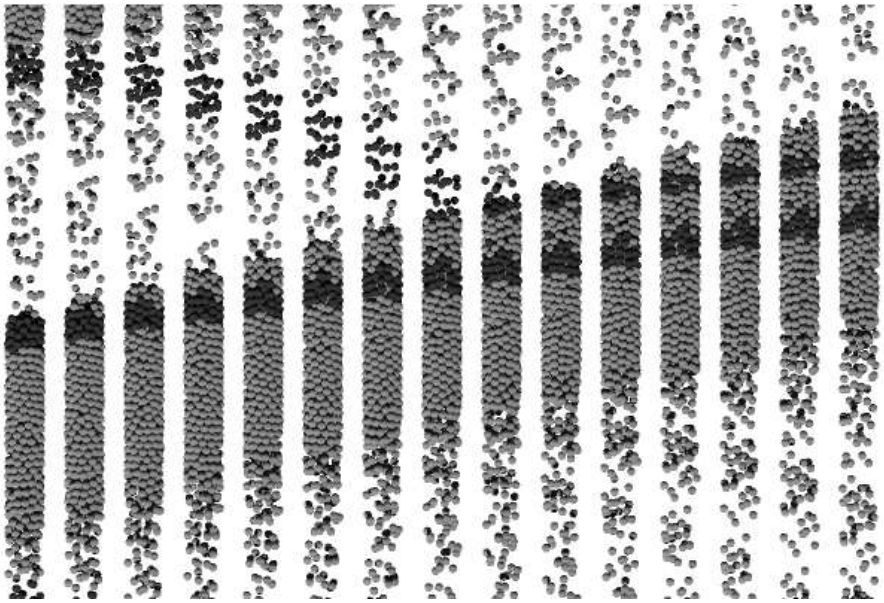


Fig. 11. A series of simulation snapshots showing a plug moving upwards. The height of the shown tube is 12 cm, the frame rate is 100 Hz

The simulations were carried out in a system that matched as closely as possible the experimental one. The same mode of transport was observed, as shown in Fig. 11. Not only is there a qualitative resemblance between Figs. 10 and 11, but the simulations give the same value for the pressure drop as the experiments. The success of the model permitted a thorough study of the plugs to be carried out. For example, so-called "characteristic curves", where pressure drop is displayed as a function of gas velocity, could be calculated. The simulations also allow the study of the effects of parameters not easily controlled experimentally, such as the air viscosity and particle friction. The speed, density, size, and number of plugs were analyzed. In addition, the detailed structure of plugs could be studied. For example, the variation of density, velocity, and different components of the stress tensor were evaluated inside the plugs. All this information should help researchers to develop better models of plug conveying.

5 Conclusion

In this paper we have discussed the properties of various simulation techniques for particles in fluids and demonstrated that there is no perfect candidate that is able to simulate all systems of interest and to utilize the available resources as efficient as possible. For each individual problem, one has to choose the method of choice carefully: while stochastic rotation dynamics is well suited to simulate systems like clay-like colloids where Brownian motion is important, the lattice Boltzmann method is not able to resolve the stochastic motion of the particles without modifications of the method. However, in cases where thermodynamic fluctuations are neglectably small, this approach is much more efficient than stochastic rotation dynamics. Like conventional Navier-Stokes solvers, the fluid flow can be resolved in great detail, but the lattice Boltzmann method is much easier to implement and to parallelize. It is of particular advantage if complicated boundary conditions like non-spherical particles or complex channel geometries come into play. The implementation of Navier-Stokes solvers on the other hand can be based on a long-standing and widespread experience with these techniques allowing to create very efficient solvers. In macroscopic systems like the movement of granular particles in air, the exact properties of the flow field are not necessary to understand experimentally observable parameters. Therefore, computationally much less demanding techniques like a coarse-grained description of the fluid should be applied.

Acknowledgements

We would like to thank all former members of the group who contributed to the projects related to the simulation of particles in fluids.

References

1. R. Adhikari, M. E. Cates, K. Stratford, and A. Wagner. Fluctuating lattice boltzmann. *condmat/0402598*, 2005.
2. P. Ahlrichs, R. Everaers, and B. Dünweg. Screening of hydrodynamic interactions in semidilute polymer solutions: A computer simulation study. *Phys. Rev. E*, 64(4):040501, 2001.
3. M. P. Allen and D. J. Tildesley. *Computer simulation of liquids.* Oxford Science Publications. Clarendon Press, 1987.
4. P. L. Bhatnagar, E. P. Gross, and M. Krook. Model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems. *Phys. Rev.*, 94(3):511–525, 1954.
5. L. Bocquet, E. Trizac, and M. Aubouy. Effective charge saturation in colloidal suspensions. *J. Chem. Phys.*, 117:8138, 2002.

6. E. S. Boek, P. V. Coveney, H. N. W. Lekkerkerker, and P. van der Schoot. Simulating the rheology of dense colloidal suspensions using dissipative particle dynamics. *Phys. Rev. E*, 55(3):3124–3133, 1997.
7. G. Bossis and J. Brady. Dynamic simulation of sheared suspensions. i. general method. *J. Chem. Phys.*, 80(10):5141–5154, 1984.
8. J. Brady and G. Bossis. Stokesian dynamics. *Ann. Rev. Fluid Mech.*, 20:111–157, 1988.
9. P. Carman. Fluid flow through granular beds. *Trans. Inst. Chem. Engng.*, 26:150–166, 1937.
10. S. Chapman and T. G. Cowling. *The Mathematical Theory of Non-uniform Gases*. Cambridge University Press, second edition, 1952.
11. S. Chen and G. Doolen. Lattice-boltzmann method for fluid flows. *Ann. Rev. Fluid Mech.*, 30:329–364, 1998.
12. P. Cundall and D. Strack. Discrete numerical-model for granular assemblies. *Geotechnique*, 29:47–65, 1979.
13. H. d'Arcy. *Les fontaines publiques de la ville de Dijon*. Victor Dalmont, 1856.
14. P. Español. A fluid particle model. *Phys. Rev. E*, 57(3):2390–2948, 1998.
15. P. Español and P. Warren. Statistical mechanics of dissipative particle dynamics. *Europhys. Lett.*, 30(4):191–196, 1995.
16. E. Falck, J. M. Lahtinen, I. Vattulainen, and T. Ala-Nissila. Influence of hydrodynamics on many-particle diffusion in 2d colloidal suspensions. *Eur. Phys. J. E*, 13:267–275, 2004.
17. A. Fogelson and C. Peskin. A fast numerical method for solving the three-dimensional stokes equations in the presence of suspended particles. *J. Comput. Phys.*, 79:50, 1988.
18. F. Fonseca and H. J. Herrmann. Sedimentation of oblate ellipsoids at low and moderate reynolds numbers. *Physica A*, 342:447–461, 2004.
19. F. Fonseca and H. J. Herrmann. Simulation of the sedimentation of a falling oblate ellipsoid. *Physica A*, 345:341–355, 2005.
20. U. Frisch, D. d'Humières, B. Hasslacher, P. Lallemand, Y. Pomeau, and J.-P. Rivet. Lattice gas hydrodynamics in two and three dimensions. *Complex Systems*, 1:649–707, 1987.
21. U. Frisch, B. Hasslacher, and Y. Pomeau. Lattice-gas automata for the Navier-Stokes equation. *Phys. Rev. Lett.*, 56(14):1505–1508, 1986.
22. J. Harting, M. Harvey, J. Chin, M. Venturoli, and P. V. Coveney. Large-scale lattice boltzmann simulations of complex fluids: advances through the advent of computational grids. *Phil. Trans. R. Soc. A*, 363:1895–1915, 2005.
23. M. Hecht, J. Harting, M. Bier, J. Reinshagen, and H. J. Herrmann. Shear viscosity of clay-like colloids: Computer simulations and experimental verification. *submitted to Phys. Rev. E*, 2006. cond-mat/0601413.
24. M. Hecht, J. Harting, T. Ihle, and H. J. Herrmann. Simulation of claylike colloids. *Physical Review E*, 72:011408, 2005.
25. K. Höfler and S. Schwarzer. Navier-stokes simulation with constraint forces: Finite-difference method for particle-laden flows and complex geometries. *Phys. Rev. E*, 61:7146, 2000.
26. R. J. Hunter. *Foundations of colloid science*. Oxford University Press, 2001.
27. M. Hütter. *Brownian Dynamics Simulation of Stable and of Coagulating Colloids in Aqueous Suspension*. PhD thesis, Swiss Federal Institute of Technology Zurich, 1999.

28. M. Hütter. Local structure evolution in particle network formation studied by brownian dynamics simulation. *Journal of Colloid and Interface Science*, 231:337–350, 2000.
29. T. Ihle and D. M. Kroll. Stochastic rotation dynamics: A galilean-invariant mesoscopis model for fluid flow. *Phys. Rev. E*, 63:020201(R), 2001.
30. T. Ihle and D. M. Kroll. Stochastic rotation dynamics i: Formalism, galilean invariance, green-kubo relations. *Phys. Rev. E*, 67:066705, 2003.
31. T. Ihle and D. M. Kroll. Stochastic rotation dynamics ii: Transport coefficients, numerics, long time tails. *Phys. Rev. E*, 67:066706, 2003.
32. T. Ihle, E. Tuzel, and D. M. Kroll. Resummed green-kubo relations for a fluctuating fluid-particle model. *Phys. Rev. E*, 70:035701(R), 2004.
33. Y. Inoue, Y. Chen, and H. Ohashi. Development of a simulation model for solid objects suspended in a fluctuating fluid. *J. Stat. Phys.*, 107(1):85–100, 2002.
34. N. Kikuchi, C. M. Pooley, J. F. Ryder, and J. M. Yeomans. Transport coefficients of a mesoscopic fluid dynamics model. *J. Chem. Phys.*, 119(12):6388–95, 2003.
35. A. Komnik, J. Harting, and H. J. Herrmann. Transport phenomena and structuring in shear flow of suspensions near solid walls. *J. Stat. Mech: Theor. Exp.*, P12003, 2004.
36. K. Konrad and T. Totah. Vertical pneumatic conveying or particle plug. *Canadian Journal of Chemical Engineering*, 67:245–252, 1989.
37. E. Kuusela, K. Höfler, and S. Schwarzer. Computation of settling speed and orientation distribution in suspensions of prolate spheroids. *J. Eng. Math.*, 41:221, 2001.
38. A. J. C. Ladd. Numerical simulations of particulate suspensions via a discretized boltzmann equation. part 1. theoretical foundation. *J. Fluid Mech.*, 271:285–309, 1994.
39. A. J. C. Ladd. Numerical simulations of particulate suspensions via a discretized boltzmann equation. part 2. numerical results. *J. Fluid Mech.*, 271:311–339, 1994.
40. A. J. C. Ladd and R. Verberg. Lattice-boltzmann simulations of particle-fluid suspensions. *J. Stat. Phys.*, 104(5):1191, 2001.
41. G. Lagaly, O. Schulz, and R. Zimehl. *Dispersionen und Emulsionen*. Dr. Dietrich Steinkopff Verlag, Darmstadt, Germany, 1997.
42. A. Lamura, G. Gompper, T. Ihle, and D. M. Kroll. Multi-particle-collision dynamics: Flow around a circular and a square cylinder. *Eur. Phys. Lett*, 56:319, 2001.
43. J. A. Lewis. Colloidal processing of ceramics. *J. Am. Ceram. Soc.*, 83:2341–59, 2000.
44. M. Loewenberg and E. Hinch. Numerical simulation of a concentrated emulsion in shear flow. *J. Fluid. Mech.*, 321:395–419, 1996.
45. J. Mahanty and B. W. Ninham. *Dispersion Forces*. Academic Press, London, 1996.
46. A. Malevanets and R. Kapral. Mesoscopic model for solvent dynamics. *J. Chem. Phys.*, 110:8605, 1999.
47. A. Malevanets and R. Kapral. Solute dynamics in mesoscale solvent. *J. Chem. Phys.*, 112:7260, 2000.
48. S. McNamara, E. Flekkøy, and K. Måløy. Grains and gas flow: Molecular dynamics with hydrodynamic interaction. *Phys. Rev. E*, 61:658–670, 2000.
49. J. Melrose and R. Ball. "contact networks" in continuously shear thickening colloids. *J. Rheo.*, 48(5):961–978, 2004.

50. J. Melrose and R. Ball. Continuous shear thickening transitions in model concentrated colloids – the role of inter-particle forces. *Journal of Rheology*, 48(5):937–960, 2004.
51. P. Mijatović. Bewegung asymmetrischer Teilchen unter stochastischen Kräften. Master-thesis, Universität Stuttgart, 2002.
52. I. D. Morrison and S. Ross. *Colloidal Dispersions: Suspensions, Emulsions and Foams*. John Wiley and Sons, New York, 2002.
53. N. Q. Nguyen and A. J. C. Ladd. Lubrication corrections for lattice-boltzmann simulations of particle suspensions. *Phys. Rev. E*, 66(4):046708, 2002.
54. R. Oberacker, J. Reinshagen, H. von Both, and M. J. Hoffmann. Ceramic slurries with bimodal particle size distributions: Rheology, suspension structure and behaviour during pressure filtration. *Ceramic Transactions*, 112:179–184, 2001.
55. D. Petera and M. Muthukumar. Brownian dynamics simulation of bead–rod chains under shear with hydrodynamic interaction. *J. Chem. Phys*, 111(16):7614–7623, 1999.
56. T. Phung, J. Brady, and G. Bossis. Stokesian dynamics simulation of brownian suspensions. *J. Fluid Mech.*, 313:181–207, 1996.
57. T. Pöschel. Granular material flowing down an inclined chute: a molecular dynamics simulation. *J. Phys. II*, 3(1):27–40, 1993.
58. J. Reinshagen, R. C. D. Cruz, R. Oberacker, and J. Hoffmann. Electrostatically stabilized alumina suspensions with defined interparticle potentials: I. influence of salt concentration on suspension conductivity and rheology. *submitted*, 2005.
59. S. Richter and G. Huber. Resonant column experiments with fine-grained model material - evidence of particle surface forces. *Granular Matter*, 5:121–128, 2003.
60. M. Ripoll, K. Mussawisade, R. G. Winkler, and G. Gompper. Low-reynolds-number hydrodynamics of complex fluids by multi-particle-collision dynamics. *Europhys. Lett.*, 68:106–12, 2004.
61. W. B. Russel, D. A. Saville, and W. Schowalter. *Colloidal Dispersions*. Cambridge Univ. Press., Cambridge, 1995.
62. K. S. Schmitz. *Macroions in Solution and Colloidal Suspension*. John Wiley and Sons, New York, 1993.
63. S. Schwarzer, K. Höfler, and B. Wachmann. Simulation of hindered settling in bidisperse suspensions of rigid spheres. *Comp. Phys. Comm.*, 268:121–122, 1999.
64. D. J. Shaw. *Introduction to Colloid and Surface Chemistry*. Butterworth-Heinemann Ltd, Oxford, 1992.
65. W. Siegel. *Pneumatische Förderung*. Vogel, 1991.
66. A. Sierou and J. Brady. Accelerated stokesian dynamics simulations. *J. Fluid Mech.*, 448:115–146, 2001.
67. L. Silbert, J. Melrose, and R. Ball. Colloidal microdynamics: Pair-drag simulations of model-concentrated aggregated systems. *Phys. Rev. E*, 56(6):7067–7077, 1997.
68. M. Strauß, H. Herrmann, S. McNamara, G. Niederreiter, and K. Sommer. Plug conveying in a vertical tube. *Particle Technology*, submitted, 2005.
69. M. Strauß, S. McNamara, and H. Herrmann. Plug conveying in a horizontal tube. *Granular Matter*, accepted, 2006.
70. E. Tuzel, M. Strauss, T. Ihle, and D. M. Kroll. Transport coefficients in three dimensional stochastic rotation dynamics. *Phys. Rev. E*, 68:036701, 2003.
71. J. Vermant and M. J. Solomon. Flow-induced structure in colloidal suspensions. *J.Phys.: Condens. Matter*, 17:R187–R216, 2005.

72. B. Wachmann and S. Schwarzer. Three dimensional massively parallel computing of suspensions. *Int. J. of Modern Physics C*, 9:759–776, 1998.
73. G. Wang, P. Sarkar, and P. S. Nicholson. Surface chemistry and rheology of electrostatically (ionically) stabillized allumina suspensions in polar media. *J. Am. Ceram. Soc.*, 82(4):849–56, 1999.
74. R. G. Winkler, K. Mussawisade, M. Ripoll, and G. Gompper. Rod-like colloids and polymers in shear flow: a multi-particle-collision dynamics study. *J. of Physics-Condensed Matter*, 16(38):S3941–54, 2004.

Multiscale Modeling of Anisotropies in Single Crystals and Polycrystals at Finite Strains*

Christian Miehe and Martin Becker

Institute of Applied Mechanics, Faculty of Civil and Environmental Engineering, University of Stuttgart, Pfaffenwaldring 7, 70569 Stuttgart, Germany
cm@mechbau.uni-stuttgart.de
becker@mechbau.uni-stuttgart.de

Summary. The paper provides an overview about recent developments in the multiscale analysis of the anisotropic material behavior of single crystals and polycrystals. We outline a distinct incremental variational formulation for the local constitutive response of standard dissipative materials where an incremental stress potential is obtained from a local minimization problem with respect to the internal variables. This potential allows for the formulation of IBVPs for standard dissipative solids as a sequence of incremental minimization problems. Particular emphasis is put on crystal plasticity where we develop alternative stress update algorithms for Schmid-type single crystal plasticity with potential hardening. Furthermore the variational setting provides for the formulation of a canonical framework of nonlinear homogenization of standard dissipative microstructures where a quasi-hyperelastic incremental macro-stress potential is obtained from a global minimization problem with respect to the fine-scale displacement fluctuation field. Finally we extend the framework of standard local crystal plasticity to the setting of dislocation density based gradient plasticity for inhomogeneously deforming crystals. Here we equip the formulation with a sound micromechanical basis where an incompatibility induced storage of geometrically necessary dislocations results in a scale dependent material behavior. The performance of the formulations is demonstrated for benchmarks of single crystal and polycrystal plasticity as well as thin films.

Keywords: Energy minimization, homogenization, microstructures, crystal plasticity, size effects

1 Introduction

The incremental variational formulation for standard dissipative materials developed in [34] and [41] provides a canonical framework for the treatment of a

* Research Project A8 "Numerical Simulation of Thermo-Mechanical Deformation Processes of Heterogeneous Materials with the Concept of Micro-Macro-Transitions"

broad class of model problems in finite elastoplasticity. Within this paper we summarize recent developments in the formulation and numerical implementation of incremental minimization principles for inelastic solids with a focus on metallic single crystals and polycrystals. The overall goal is the quantitative description of evolving anisotropies due to texture developments in crystals. This necessitates multiscale analyses for which we develop a canonical incremental variational structure. Furthermore we report on our developments in strain gradient crystal plasticity which allow for a description of size effects in nonhomogeneously deforming metal crystals.

Key aspect is the outline of *minimization principles for standard dissipative materials* which allow to recast incremental initial boundary-value problems for inelastic solids into minimization problems. The minimization structure provides a fundamental approach to inelastic solid mechanics under quasistatic conditions, with the following important consequences for the modeling and numerical implementation: (i) Stress update algorithms in plasticity, viscoplasticity and damage mechanics appear in a natural format in the form of energy minimizers. (ii) Micro-to-macro transitions for the modeling of the overall response of a priori given heterogeneous microstructures can be defined in terms of a principle of minimum averaged incremental energy. (iii) The material stability of incremental plastic deformations can be defined in terms of weak-convexity conditions of incremental stress potentials in analogy to elasticity. Starting point is a general internal variable formulation for the constitutive response of standard dissipative materials in terms of an energy storage and a dissipation function. Nonlinear standard materials cover a broad spectrum of constitutive models in finite elasticity, viscoelasticity, plasticity or damage mechanics, see for example [60] and [17]. Consistent with this type of finite inelasticity we outline a distinct *incremental variational formulation of the local constitutive response* as proposed in [34] and [41], where a quasi-hyperelastic stress potential is obtained from a local minimization problem with respect to the internal variables. It is shown that this local minimization problem determines the internal state of the material for finite increments of time. The approach extends earlier works on incremental variational formulations and variational updates in plasticity by [46, 47].

In the following we specify the incremental variational framework for the setting of local single crystal plasticity. The continuum slip theory of crystals provides a canonical framework of finite plasticity with physically well-defined roots in the dislocation mechanics of metals. Reliable algorithmic settings of crystal plasticity are not only needed for structural analyses of single crystals but also provide a cornerstone for multiscale computations of evolving anisotropic microstructures in polycrystals. The physically based phenomenological description of macroscopic plastic strains in metallic single crystals has been guided by the pioneering works [56, 45] and [25]. The mechanism of inelastic distortion in ductile single crystals was found to be governed by plastic slip on a certain set of crystallographic systems where the shear stresses reach critical values. Mathematical continuum descriptions of elastic-plastic defor-

mations in crystals have been developed in [18] in the small-strain format and in [51, 29, 26] in the context of finite strains. The geometric basis of what is often called the macroscopic continuum slip theory of finite crystal elastoplasticity is a multiplicative decomposition of the local deformation gradient into a plastic part solely due to multislip on given crystallographic planes and a remaining part which describes elastic distortions and rigid rotations of the lattice. The constitutive equations which govern the slip resistance and the slip evolution can be motivated by micromechanical investigations of defects in crystals. The formulation of micromechanically motivated hardening laws for multislip is a cornerstone of the continuum slip theory and still an area of intensive research. We refer e.g. to the reviews [22, 4, 10] and references cited therein. Algorithmic representations of finite crystal plasticity models suitable for the numerical simulation of initial boundary value problems in context with the application of finite element methods were developed in the last two decades, e.g. [48, 10, 2, 32] and [42], which discuss alternative numerical schemes for the updates of the stresses and the active set of slip systems in the multisurface frameworks of rate-dependent and rate-independent crystal plasticity. The existence of a variational structure for the finite-step-sized algorithmic setting of crystal plasticity also has the important consequence of symmetry of consistent tangent operators.

The first part summarizes a current status of computational crystal plasticity at finite strains and its application to the analysis of polycrystalline microstructures. It starts with a successive set up of a local constitutive model of crystal plasticity that focuses on homogeneous elastic-plastic macrodistortions. This includes a precise definition of independent and dependent internal variables and the definition of kinetic laws for the evolution of plastic slip and the slip resistance. The formulation provides a framework for a wide range of applications in viscoplasticity with alternative hardening laws. For this class of local constitutive models we develop a unified family of stress update algorithms which exploit a possible variational structure. We discuss two possible definitions of the Schmid stress within the algorithmic setting at finite time steps. The standard definition adopted from the continuous formulation yields an algorithmically induced non-symmetry of the consistent tangents, however at low computational cost where only the first derivative of the exponential map is needed. A second definition, consistent with the algorithmic sensitivity of the lattice free energy with respect to the plastic slip, gives a symmetric formulation but at a computationally higher price that also needs the evaluation of second derivatives of the exponential map. The minimization structure of the incremental constitutive response is highly advantageous with respect to the detection of the currently active slip systems, which then result from a constrained optimization problem. The algorithmic setting of crystal plasticity is summarized Table 1.

With the algorithmic constitutive setting at hand the initial boundary value problem of crystal plasticity is recast into a sequence of incremental problems. Of particular interest in crystal plasticity is the boundary

value problem of homogenization that allows to perform two-scale computational analyses of the microstructure evolution, e.g the texture development and the resulting anisotropy, in polycrystals. Here, the existence of quasi-hyperelastic incremental stress potentials for standard dissipative materials allows the extension of homogenization approaches of elasticity such as outlined in [54, 44, 43, 50] to the incremental setting of inelasticity. Following the works [34, 35] and [41], we outline an incremental variational formulation of the homogenization problem for standard dissipative crystals, where a quasi-hyperelastic macro-stress potential is obtained from a global minimization problem with respect to the fine-scale fluctuation field. It is shown that this global minimization problem determines the state of a microstructure for finite increments of time. The understanding of the consequence of incremental energy minimization especially with respect to the simulation of evolving microstructures is an area of intensive research. This covers the development of microstructures in initially homogeneous solids as well as the microstructure evolution in microheterogeneous solids. Examples for the latter topic are the texture evolution in polycrystals as considered below or the resulting anisotropic yield surface evolution as investigated in [38]. Furthermore the incremental variational formulation of the homogenization problem allows for the investigation of material and structural instabilities on the micro- and the macro-scale as well as their interaction for the first time also for standard dissipative microheterogeneous materials as discussed in [36].

A further important field of research in crystal plasticity treats nonlocal formulations that allow to model size and boundary layer effects accompanied by additional hardening. Experimentally *size effects* have been already observed in *nonhomogeneously deforming metals* for decades ([16, 49, 13]). Micromechanical basis of such observations is the storage of *geometrically necessary dislocations, GNDs*, in connection with increasing deformation inhomogeneities and the associated strain gradients. These result in an overall hardening response where diminishing size entails an elevated dislocation storage and thus a scale dependent material behavior. A general view of the theory of continuous distributions of defects was developed in the works [45, 24, 6] and [25]. This lead to the definition of the *dislocation density tensor* as an incompatibility measure. Following the works [5] and [37] we develop a new formulation which extends the setting of local crystal plasticity discussed in the preceding to account for possible incompatibilities, the associated GND storage and thus a scale dependent material response in metal crystals. Here we concentrate on a precise incompatibility analysis, an interpretation of the dislocation density tensor as a mapping, the development of a micromechanically motivated strain gradient crystal plasticity model and the formulation of reliable and efficient algorithmic treatments. Other formulations capable of describing size effects in metal crystals are due to [13, 1, 53, 15, 55] on a rather phenomenological basis or [11] and [12] with a closer relation to the underlying micromechanics or [27, 58] based directly on discrete dislocations.

The subsequent four sections successively treat the incremental variational formulation for standard dissipative materials, the continuous and algorithmic settings of local crystal plasticity, alternative boundary value problems of homogenization and finally the modeling of size effects in metal crystals through a gradient enhanced crystal plasticity formulation. The performance of these formulations is demonstrated for representative numerical examples at the end of each section.

2 Variational Problem for Dissipative Materials

2.1 Internal Variable Formulation for Standard Materials

Let $\mathbf{F} \in GL_+(3)$ with $\det[\mathbf{F}] > 0$ be the homogeneous deformation of a material at time $t \in \mathcal{R}_+$. For mechanical problems, the homogeneous stress response is physically constrained by the Clausius-Planck inequality

$$\mathcal{D} := \mathbf{P} : \dot{\mathbf{F}} - \dot{\psi} \geq 0 \,, \tag{1}$$

where $\mathbf{P}$ denotes the nominal stress tensor. ψ is an objective *energy storage function* that is assumed to depend on the deformation $\mathbf{F}$ and a generalized vector $\boldsymbol{\mathcal{I}} \in \mathcal{G}$ of internal variables. $\mathcal{G}$ might e.g. be constrained to the Lie group $SL(3)$. A standard argument gives through (1) the constitutive equation

$$\mathbf{P} = \partial_{\mathbf{F}} \psi(\mathbf{F}, \boldsymbol{\mathcal{I}}) \tag{2}$$

and the reduced dissipation inequality

$$\mathcal{D} = \boldsymbol{\mathcal{F}} \cdot \dot{\boldsymbol{\mathcal{I}}} \geq 0 \qquad \text{with} \qquad \boldsymbol{\mathcal{F}} := -\partial_{\boldsymbol{\mathcal{I}}} \psi(\mathbf{F}, \boldsymbol{\mathcal{I}}) \,, \tag{3}$$

where $\boldsymbol{\mathcal{F}} \in \mathcal{R}^n$ is a generalized vector of internal forces conjugate to the internal variables $\boldsymbol{\mathcal{I}}$. A broad spectrum of materials is covered by the so-called standard dissipative materials where the evolution $\dot{\boldsymbol{\mathcal{I}}}$ is governed by a scalar *dissipation function* ϕ, depending on the flux $\dot{\boldsymbol{\mathcal{I}}}$ of the internal variables and the internal variables $\boldsymbol{\mathcal{I}}$ themselves, see for example [60] and [17]. It governs the evolution of $\boldsymbol{\mathcal{I}}$ in time by the constitutive differential equation

$$0 \in \partial_{\boldsymbol{\mathcal{I}}} \psi(\mathbf{F}, \boldsymbol{\mathcal{I}}) + \partial_{\dot{\boldsymbol{\mathcal{I}}}} \phi(\dot{\boldsymbol{\mathcal{I}}}, \boldsymbol{\mathcal{I}}) \quad \text{with} \quad \boldsymbol{\mathcal{I}}(0) = \boldsymbol{\mathcal{I}}_0 \,, \tag{4}$$

known as Biot's equation of standard dissipative systems, referring to [7]. The constitutive equations (2) and (4) determine the stress response of a normal-dissipative material. The dissipation functions are positively homogeneous of degree one with respect to the flux $\dot{\boldsymbol{\mathcal{I}}}$, i.e. $\phi(\epsilon \dot{\boldsymbol{\mathcal{I}}}, \boldsymbol{\mathcal{I}}) = \epsilon \phi(\dot{\boldsymbol{\mathcal{I}}}, \boldsymbol{\mathcal{I}})$ for all $\epsilon \in \mathcal{R}_+$.

2.2 Incremental Variational Formulation of Inelastic Materials

We proceed with the construction of an integrated version of constitutive equations giving a consistent approximation of the continuous differential equation (4) in a finite increment $[t_n, t_{n+1}] \in \mathcal{R}_+$ of time. Conceptually following the works [46, 47], the key point is the definition of an *incremental stress potential function* W depending on the deformation $\mathbf{F}_{n+1} := \mathbf{F}(t_{n+1})$ that determines the stresses at t_{n+1} by the quasi-hyperelastic function evaluation

$$\mathbf{P}_{n+1} = \partial_{\mathbf{F}} W(\mathbf{F}_{n+1}) \,. \tag{5}$$

Clearly, this function must cover characteristics of the storage function ψ and the dissipation function ϕ. To this end, consider the *minimization problem*

$$W(\mathbf{F}_{n+1}) = \inf_{\boldsymbol{\mathcal{I}} \in \mathcal{G}} \int_{t_n}^{t_{n+1}} [\, \dot{\psi} + \phi \,] \, \mathrm{d}t \quad \text{with} \quad \boldsymbol{\mathcal{I}}(t_n) = \boldsymbol{\mathcal{I}}_n \tag{6}$$

for dissipative standard materials proposed in [34]. Starting with the given initial condition $\boldsymbol{\mathcal{I}}(t_n) = \boldsymbol{\mathcal{I}}_n$, the minimum problem defines an optimal path of the internal variables $\boldsymbol{\mathcal{I}}(t)$ for $t \in [t_n, t_{n+1}]$, including the right boundary value $\boldsymbol{\mathcal{I}}_{n+1} := \boldsymbol{\mathcal{I}}(t_{n+1})$. We refer to [30] for a discussion of extremum paths in linear and nonlinear plasticity. (5) and (6) provide an approximative variational counterpart of the continuous setting (2) and (4) of the constitutive equations in the current discrete time step $[t_n, t_{n+1}]$. In order to show the consistency, we at first recast (6) into the form

$$W(\mathbf{F}_{n+1}) = \inf_{\boldsymbol{\mathcal{I}} \in \mathcal{G}} \left\{ \, [\psi(\mathbf{F}, \boldsymbol{\mathcal{I}})]_{t_n}^{t_{n+1}} + \int_{t_n}^{t_{n+1}} \phi(\dot{\boldsymbol{\mathcal{I}}}, \boldsymbol{\mathcal{I}}) \, \mathrm{d}t \, \right\} . \tag{7}$$

The necessary condition for the minimum problem is that the variation with respect to the internal variables of the term in brackets vanishes. For smooth functions, integration by parts yields the expression

$$[(\, \partial_{\boldsymbol{\mathcal{I}}} \psi + \partial_{\dot{\boldsymbol{\mathcal{I}}}} \phi \,) \cdot \delta \boldsymbol{\mathcal{I}}]_{t_n}^{t_{n+1}} + \int_{t_n}^{t_{n+1}} [\, -\frac{\mathrm{d}}{\mathrm{d}t} (\partial_{\dot{\boldsymbol{\mathcal{I}}}} \phi) + \partial_{\boldsymbol{\mathcal{I}}} \phi \,] \cdot \delta \boldsymbol{\mathcal{I}} \, \mathrm{d}t = 0 \,. \tag{8}$$

Thus the variational problem (6) yields Biot's equation (4)

$$\partial_{\boldsymbol{\mathcal{I}}} \psi + \partial_{\dot{\boldsymbol{\mathcal{I}}}} \phi = 0 \quad \text{for} \quad t = t_{n+1} \tag{9}$$

at the discrete right boundary of the interval $[t_n, t_{n+1}]$. The minimizing path of the internal variables inside the interval is determined by the Euler equation

$$-\frac{\mathrm{d}}{\mathrm{d}t} (\partial_{\dot{\boldsymbol{\mathcal{I}}}} \phi) + \partial_{\boldsymbol{\mathcal{I}}} \phi = 0 \quad \text{for} \quad t \in [t_n, t_{n+1}] \,. \tag{10}$$

For the limit $t_{n+1} \to t_n$, the form of the minimization path becomes irrelevant, because the time increment degenerates to a discrete time t_{n+1}. Because (9)

still holds in this case, it is shown that the proposed variational formulation (6) represents a consistent point-wise approximation of Biot's normal-dissipative evolution equation (4). Furthermore, taking the derivative of the incremental potential function with respect to the deformation $\mathbf{F}_{n+1}$, we have

$$\partial_{\mathbf{F}} W(\mathbf{F}_{n+1}) = \partial_{\mathbf{F}} \psi(\mathbf{F}_{n+1}, \boldsymbol{\mathcal{I}}_{n+1}) \,, \tag{11}$$

where $\boldsymbol{\mathcal{I}}_{n+1}$ is considered to be given by the minimization problem (6). A comparison with (2) then shows the consistency of the potential equation (5) with the continuous setting.

3 Variational Formulation for Crystal Plasticity

3.1 Constitutive Model of Local Single Crystal Plasticity

Assuming a locally homogeneous elastic-plastic deformation of the crystal we consider a multiplicative decomposition

$$\mathbf{F} = \mathbf{F}^e \mathbf{F}^p \tag{12}$$

of the deformation *total gradient* $\mathbf{F}$ into an *elastic part* $\mathbf{F}^e$ and a *plastic part* $\mathbf{F}^p$. Following [26, 51] and [29], the evolution problem for $\mathbf{F}^p$ is specified by

$$\dot{\mathbf{F}}^p \mathbf{F}^{p-1} = \sum_{\alpha=1}^{m} \dot{\gamma}^\alpha \mathbf{m}^\alpha \quad \text{with} \quad \mathbf{m}^\alpha := \mathbf{S}^\alpha \otimes \mathbf{M}^\alpha \quad \text{and} \quad \mathbf{F}^p(0) = \boldsymbol{1} \tag{13}$$

in terms of the slip direction $\mathbf{S}^\alpha$, the slip plane normal $\mathbf{M}^\alpha$ and the *plastic slips* γ^α which describe the cumulative dislocation motion on crystallographic slip planes α and are considered as internal variables governed by a constitutive assumption for the evolution $\dot{\gamma}^\alpha$. From (13) we may conclude the *sensitivities*

$$\mathbf{F}^p_{,\alpha} = \mathbf{m}^\alpha \mathbf{F}^p \quad \text{and} \quad \mathbf{F}^p_{,\alpha\beta} = \mathbf{m}^\alpha \mathbf{m}^\beta \mathbf{F}^p \tag{14}$$

required in the following. Here $(\cdot)_{,\alpha}$ describes the sensitivity of the quantity $(\cdot)$ with respect to the slip γ^α. The free energy of the deformed crystal is

$$\psi = \hat{\psi}(\mathbf{F}\mathbf{F}^{p-1}, \gamma^1 \dots \gamma^m) \,. \tag{15}$$

A physically-based assumption is the decoupled representation of (15)

$$\psi = \hat{\psi}^e(\mathbf{F}\mathbf{F}^{p-1}) + \hat{\psi}^p(\gamma^1 ... \gamma^m) \,. \tag{16}$$

It constitutes the observation that the micro-defect-stress fields do not directly influence the macro-stress fields. Exploitation of the constitutive equation (2) gives the stress expression along with the reduced dissipation

$$\mathbf{P} = \partial_{\mathbf{F}} \hat{\psi}^e = \mathbf{P}^e \overset{2}{\circ} \mathbf{F}^{p-T} \quad \text{and} \quad \mathcal{D} = \sum_{\alpha=1}^{m} (\tau_\alpha - \tau_\alpha^{p,s}) \dot{\gamma}^\alpha \geq 0 \tag{17}$$

(Here, we use the notation $(\bullet) \overset{i}{\circ} \mathbf{A}$ for the composition of the i-th slot of $(\bullet)$ with $\mathbf{A}$). In the latter, we have introduced *per definition* the so-called *Schmid stress* τ_α and the contribution $\tau_\alpha^{p,s}$ to the *slip resistance* τ_α^p due to the storage of micro-stress fields as the sensitivities with respect to the plastic slip γ^α

$$\tau_\alpha := -\hat{\psi}^e_{,\alpha} \quad \text{with} \quad \tau_\alpha^{p,s} := \hat{\psi}^p_{,\alpha} \, . \tag{18}$$

The Schmid stress can be expressed in terms of the macro-stress $\mathbf{P}^e$ by exploiting the sensitivity (14) of the plastic deformation map

$$\tau_\alpha = -\mathbf{P}^e : \mathbf{F}^e_{,\alpha} \quad \text{with} \quad \mathbf{F}^e_{,\alpha} := \mathbf{F}\mathbf{F}^{p-1}_{,\alpha} = -\mathbf{F}^e \mathbf{m}^\alpha \, . \tag{19}$$

Note carefully that this definition of the Schmid stress follows from a thermodynamical argument as the driving force conjugate to the plastic slip γ^α. The evolution $\dot{\gamma}^\alpha$ is governed by the *slip resistance* $\hat{\tau}_\alpha^p$ which induces the definition of an *elastic domain* in the space of the thermodynamic driving forces

$$\mathcal{E} = \{\tau_\alpha | \tau_\alpha - \hat{\tau}_\alpha^p(\gamma^1, ..., \gamma^m) \leq 0 \quad \text{for } \alpha = 1 \dots m\} \tag{20}$$

in which the response of the crystal is assumed to be purely elastic. In a *rate-dependent* setting of *crystal viscoplasticity*, we define the kinetic slip evolution by the *loading-unloading conditions*

$$\dot{\gamma}^\alpha \geq 0 \quad \wedge \quad (\tau_\alpha - \hat{\tau}_\alpha^p - \hat{\tau}_\alpha^v) \leq 0 \quad \wedge \quad \dot{\gamma}^\alpha (\tau_\alpha - \hat{\tau}_\alpha^p - \hat{\tau}_\alpha^v) = 0 \tag{21}$$

in terms of the viscous overstress $\tau_\alpha^v = \tau_0^v (\dot{\gamma}^\alpha / \dot{\gamma}_0)^{1/\epsilon}$, where τ_0^v, $\dot{\gamma}_0$ and ϵ are positive material parameters that characterize the viscosity of the plastic flow. Equations (21) define an *active set of slip systems*

$$\mathcal{A} := \{\alpha | \dot{\gamma}^\alpha > 0\} \subset \{1 \dots m\} \tag{22}$$

with currently non-zero slip. An uncertain part of the constitutive model is the specification of the slip resistance function $\hat{\tau}_\alpha^p(\gamma^1, ..., \gamma^m)$ that governs the hardening of the crystal. Classical representations use the rate formulation

$$\dot{\tau}_\alpha^p = \sum_{\beta=1}^{m} h^{\alpha\beta} \dot{\gamma}^\beta \quad \text{with} \quad \tau_\alpha^p(0) = \tau_0 \, , \tag{23}$$

as suggested in [18]. Here τ_0 is the initial critical shear stress and $h^{\alpha\beta}$ a possibly state dependent hardening matrix which governs the interaction between different slip systems. The simplest assumptions are linear *isotropic hardening* $h^{\alpha\beta} = h$ proposed in [56], or a *purely self-hardening* mechanism $h^{\alpha\beta} = h\, \delta^{\alpha\beta}$ as considered in [23]. Latent hardening characteristics can be incorporated in

the simplest form as suggested by [21] with $h^{\alpha\beta} = h[\, q + (1-q)\, \delta^{\alpha\beta}\,]$ where $q > 1$ represents the typical off-diagonal dominant latent hardening effect. [48] extended this function by the nonlinear isotropic ansatz

$$h = h(\gamma) = h_s + (h_0 - h_s)\,\mathrm{sech}^2\left[\frac{h_0 - h_s}{\tau_s - \tau_0}\gamma\right] \quad \text{with} \quad \gamma = \sum_{\alpha=1}^{m} \gamma^\alpha \tag{24}$$

representing a saturation-type hardening that converges to the value h_s. More advanced constitutive hardening models distinguish between the three characteristic ranges of hardening, e.g. in the model of [4]. A classical micromechanically-based approach is given by the dependence

$$\tau_\alpha^p = c\mu b\,\sqrt{n^\alpha} \tag{25}$$

of the slip resistance on the density n^α of forest dislocations on the system α, see e.g. [22]. $c \approx 0.3$ is a numerical coefficient, μ the shear modulus, and b the Burgers vector length. The density n^α of forest dislocations can be related to the slip γ^α on the primary systems via

$$n^\alpha = \sum_{\beta}^{m} G^{\alpha\beta}\rho^\beta \quad \text{and} \quad \dot\rho^\alpha = \frac{\dot\gamma^\alpha}{\gamma^{\mathrm{sat}}}\,(\rho^{\mathrm{sat}} - \rho^\alpha) \quad \text{and} \quad \rho^\alpha(0) = \rho^{0,\alpha}\,, \tag{26}$$

where $\rho^{0,\alpha}$ and ρ^{sat} are initial and saturated densities and γ^{sat} a characteristic accumulated slip. The interaction matrix $G^{\alpha\beta}$ characterizes the interaction strength between different systems such as specified in [14]. The combination of (25) and (26) yields the hardening moduli

$$h^{\alpha\beta} = \frac{c\mu b}{2\sqrt{n^\alpha}}\, a^{\alpha\beta}\, \frac{1}{\gamma^{\mathrm{sat}}}\left(\rho^{\mathrm{sat}} - \rho^\beta\right)\,. \tag{27}$$

3.2 Stress Update Algorithms in Single Crystal Plasticity

Integration Algorithm for Internal Variables

The flow-rule (13) defines in a time step $t \in [t_n, t_{n+1}]$ the initial value problem

$$\dot{\mathbf{F}}^p\mathbf{F}^{p-1} = \sum_{\alpha=1}^{m} \dot\gamma^\alpha \mathbf{m}^\alpha \quad \text{with} \quad \mathbf{F}^p(t_n) = \mathbf{F}_n^p\,. \tag{28}$$

Following [59, 2] and [32], we consider a fully implicit integration scheme with exponential shift of the evolution equations (28)

$$\mathbf{F}^{p-1} = \mathbf{F}_n^{p-1}\mathbf{e} \quad \text{with} \quad \mathbf{e} = \exp[\mathbf{N}] \quad \text{and} \quad \mathbf{N} := -\sum_{\alpha=1}^{m}(\gamma^\alpha - \gamma_n^\alpha)\,\mathbf{m}^\alpha \tag{29}$$

which preserves the unimodular structure of Schmid-type isochoric flow. γ^α are the current accumulated slips on systems $\alpha = 1...m$, which are constrained

to lie in the plastic loading cones $\gamma^\alpha \in \mathcal{K}^\alpha := \{\ \gamma^\alpha \in \mathcal{R} | \gamma^\alpha \geq \gamma_n^\alpha\ \}$. In what follows, all variables without subscript are assumed to be evaluated at time t_{n+1}. Based on (29) we express the elastic deformation map

$$\mathbf{F}^e = \mathbf{F}^{e*}\mathbf{e} \quad \text{with} \quad \mathbf{F}^{e*} := \mathbf{F}\mathbf{F}_n^{p-1} \tag{30}$$

by the *trial elastic deformation* $\mathbf{F}^{e*}$ which is known in a deformation driven context and the *inverse incremental plastic deformation map* $\mathbf{e}$. In the following we also need the first $\mathbf{e}_{,\alpha}$ and second $\mathbf{e}_{,\alpha\beta}$ derivatives of the inverse incremental plastic map. The evaluation of the exponential map itself and of its derivatives can be based on a *Taylor series expansion* or on *spectral decomposition formulae*. In this respect we refer to the works [32] and [41]. A computationally very efficient approximative approach is proposed in [33] and [40]. With the derivatives of the exponential map at hand, we finally obtain the first and second derivatives of the elastic deformation map from $(30)_1$

$$\begin{aligned} &\mathbf{F}^e_{,\alpha} = -\mathbf{F}^e\,\mathbf{m}^\alpha \text{ and } \mathbf{F}^e_{,\alpha\hat{\beta}} = -\mathbf{F}^{e*}\mathbf{e}_{,\beta}\,\mathbf{m}^\alpha \\ \text{or} \quad &\mathbf{F}^e_{,\hat{\alpha}} = \mathbf{F}^{e*}\mathbf{e}_{,\alpha} \quad \text{and } \mathbf{F}^e_{,\hat{\alpha}\hat{\beta}} = \mathbf{F}^{e*}\mathbf{e}_{,\alpha\beta}\ . \end{aligned} \tag{31}$$

Here, the *overset symbol* $\hat{\alpha}$ on the index α distinguishes the above obtained *derivatives* $\mathbf{F}^e_{,\hat{\alpha}}$ and $\mathbf{F}^e_{,\hat{\alpha}\hat{\beta}}$ of the finite-step-sized algorithmic formulation from the *sensitivity* $\mathbf{F}^e_{,\alpha}$ of the continuous formulation defined in $(19)_2$.

Standard Update Algorithm for Frozen Active Set

Assume an active set $\mathcal{A}$ to be given. Then the current slips $\{\gamma^\alpha\}_{\alpha\in\mathcal{A}}$ on these systems are obtained from the pseudo-consistency conditions

$$r_\alpha := -\tau_\alpha + [\tau_\alpha^p + \tau_\alpha^v] = 0 \quad \text{for} \quad \alpha \in \mathcal{A}\ . \tag{32}$$

This condition balances the algorithmic expression for the Schmid-stress τ_α with the slip resistance τ_α^p due to hardening and the viscous overstress τ_α^v. The system (32) is solved by means of a local Newton iteration, i.e.

$$\gamma^\alpha \Leftarrow \gamma^\alpha - \sum_{\beta\in\mathcal{A}} H^{-1}_{\alpha\beta}\, r_\beta \quad \text{for} \quad \alpha \in \mathcal{A} \tag{33}$$

in terms of the algorithmic sensitivity

$$H_{\alpha\beta} := -\tau_{\alpha,\beta} + \left[\tau^p_{\alpha,\beta} + \tau^v_{\alpha,\beta}\right] \quad \text{for} \quad (\alpha,\beta) \in \mathcal{A}\ . \tag{34}$$

The Newton iteration is terminated if $\sqrt{\sum_{\alpha\in\mathcal{A}}(r^\alpha)^2} \leq \text{tol}$. The iteration matrix $H_{\alpha\beta}$ is generally unsymmetric. Clearly in terms of efficiency of the implicit algorithm, *symmetry* of $H^{\alpha\beta}$ would be highly desirable. This is achieved for the variational update algorithm discussed next.

Variational Update Algorithm for Frozen Active Set

The specification of the incremental variational formulation discussed in Sect. 2.2 for the present case of single crystal plasticity reads

$$W(\mathbf{F}) = \inf_{\gamma^\alpha \in \mathcal{K}^\alpha} \hat{W}(\mathbf{F}, \gamma^{1\dots m}) \; . \tag{35}$$

Again, the plastic slips are constrained to lie in the plastic loading cones $\mathcal{K}^\alpha := \{\gamma^\alpha \in \mathcal{R} | \gamma^\alpha \geq \gamma_n^\alpha\}$. Thereby, the local incremental variational problem specifies the plastic slips by minimization of the incremental potential function

$$\begin{aligned} \hat{W}(\mathbf{F}, \gamma^{1\dots m}) = \hat{\psi}^e(\mathbf{F}\cdot\mathbf{F}^{p-1}(\gamma^{1\dots m})) + \hat{\psi}^p(\gamma^{1\dots m}) - \psi_n^e - \psi_n^p \\ + \Delta t \hat{\phi}^r(\gamma^{1\dots m}) + \Delta t \hat{\phi}^v(\gamma^{1\dots m}) \end{aligned} \tag{36}$$

based on a fully implicit algorithmic approximation of the integral of the dissipation function. According to (32) and (34), the residual expression and the tangent matrix for the local Newton iteration are now defined by

$$r_\alpha := \hat{W}_{,\alpha} \quad \text{and} \quad H_{\alpha\beta} := \hat{W}_{,\alpha\beta} \; . \tag{37}$$

The necessary Karush-Kuhn-Tucker optimization conditions of the constrained minimization problem (35) read

$$r_\alpha \geq 0 \quad , \quad \gamma^\alpha \geq \gamma_n^\alpha \quad , \quad r_\alpha(\gamma^\alpha - \gamma_n^\alpha) = 0 \; . \tag{38}$$

If some violations of (38) occur, the active set $\mathcal{A}$ will be updated, see Sect. 3.2, and a new iterate is computed.

Algorithmic Expressions for Standard and Variational Formulation

The respective contributions to the residual expressions (32) or $(37)_1$ and the iteration matrices (34) or $(37)_2$ are discussed next.

The *Schmid stresses* τ^α of the standard update are given by (18) of the *continuous* setting. In turn, the algorithmic Schmid stresses of the variational formulation are derived within the *incremental* setting, hence

$$-\tau^\alpha = \mathbf{P}^e : \mathbf{F}^e_{,\odot} \tag{39}$$

in terms of the stresses $\mathbf{P}^e := \partial_{\mathbf{F}^e} \hat{\psi}^e$. $\odot$ distinguishes between the sensitivities of the continuous and finite-step-sized settings. For the *standard update* we set $\odot \equiv \alpha$ by inserting the continuous sensitivity $\mathbf{F}^e_{,\alpha}$. In the *variational update* we set $\odot \equiv \hat{\alpha}$ by inserting the algorithmic sensitivity $\mathbf{F}^e_{,\hat{\alpha}}$. With the definition of the *nominal elastic moduli* $\boldsymbol{\mathcal{A}}^e := \partial^2_{\mathbf{F}^e\mathbf{F}^e} \hat{\psi}^e$, the sensitivities of (39) read

$$-\tau_{\alpha,\beta} = \mathbf{F}^e_{,\odot} : \boldsymbol{\mathcal{A}}^e : \mathbf{F}^e_{,\hat{\beta}} + \mathbf{P}^e : \mathbf{F}^e_{,\odot\hat{\beta}} \; . \tag{40}$$

Obviously the algorithmic sensitivity $\tau_{\alpha,\beta}$ of the standard formulation is *non-symmetric* as a consequence of mixed continuous/algorithmic sensitivities while the one of the variational formulation is *symmetric*.

The *slip resistance* is governed by the *potential functions* $\hat{\psi}^p$ and $\hat{\phi}^r$ representative of storage and dissipative mechanisms related to the plastic slip. For most micromechanically based hardening laws such potential functions can not be constructed. In many cases however, e.g. hardening laws with a rate form (23) and symmetric hardening matrix $h^{\alpha\beta}$, incremental quadratic potentials can be specified as follows (with $\dot{\gamma}^\alpha = (\gamma^\alpha - \gamma_n^\alpha)/\Delta t = \Delta\gamma^\alpha/\Delta t$)

$$\hat{\psi}^p(\gamma^{1\dots m}) = \sum_{\alpha\in\mathcal{A}}(\tau_\alpha^{p,n} - \tau_\alpha^{p\,0})\Delta\gamma^\alpha + \tfrac{1}{2}\sum_{\alpha\in\mathcal{A}}\sum_{\beta\in\mathcal{A}}\Delta\gamma^\alpha\, h_n^{\alpha\beta}\,\Delta\gamma^\beta$$
$$\text{and}\quad \hat{\phi}^r(\gamma^{1\dots m}) = \sum_{\alpha\in\mathcal{A}}\tau_\alpha^{p\,0}\dot{\gamma}^\alpha \tag{41}$$

which bases on a standard assumption for the split into storage and dissipative mechanisms. Here $\tau_\alpha^{p,n}$ and $h_n^{\alpha\beta}$ are the slip resistance and hardening modulus evaluated at t_n and $\tau_\alpha^{p\,0}$ is the slip resistance evaluated at $t = t_0$. Thus (41) provides in some cases a potential structure for the slip resistance and its sensitivity needed in (32) and (34) or (37) through

$$\tau_\alpha^p = \hat{\psi}^p_{,\alpha} + \Delta t\hat{\phi}^r_{,\alpha} = \tau_\alpha^{p,n} + \sum_{\beta\in\mathcal{A}} h_n^{\alpha\beta}(\gamma^\beta - \gamma_n^\beta) \quad\text{and}\quad \tau^p_{\alpha,\beta} = h_n^{\alpha\beta}\,. \tag{42}$$

A potential function $\hat{\phi}^v$ for the contribution of the viscous effects to the dissipation potential can, e.g. for a Norton-Bayley-type overstress, be constructed in the following manner

$$\hat{\phi}^v(\gamma^{1\dots m}) = \tau^{v\,0}\,\dot{\gamma}_0\,\frac{\epsilon}{1+\epsilon}\sum_{\alpha\in\mathcal{A}}\left[\frac{\gamma^\alpha - \gamma_n^\alpha}{\dot{\gamma}_0\Delta t}\right]^{(1+\epsilon)/\epsilon}, \tag{43}$$

which provides a potential structure for the viscous overstress and their sensitivity needed in (32) and (34) or (37), i.e.

$$\tau_\alpha^v = \Delta t\hat{\phi}^v_{,\alpha} \quad\text{and}\quad \tau^v_{\alpha,\beta} = \Delta t\hat{\phi}^v_{,\alpha\beta}\,. \tag{44}$$

Algorithmic Update of the Active Set of Slip Systems

The solution of (38) may be obtained by a projected Newton algorithm in context with an active set updating strategy suggested in [47] that exploits the minimizing property (35). The results are then also used within the standard update. We introduce a working set $\mathcal{W}$ of active constraints for which equality constraints hold and a complementary set of active slip systems $\mathcal{A}$

$$\Delta\gamma^\alpha = 0 \quad\text{for}\quad \alpha\in\mathcal{W} \quad\text{and}\quad \mathcal{A} := \mathcal{S}\setminus\mathcal{W}\,, \tag{45}$$

where $\mathcal{S} := \{1, 2, ..., m\}$ is the set of possible slip systems. We start with $\mathcal{A} = \mathcal{A}_n$. Then for a given active set $\mathcal{A}$ an improved solution of incremental

slip is obtained by the *Newton update* (33). If some violations of the constraint $(38)_2$ occur in the sense of $\Delta\gamma^\alpha < 0$ for $\alpha \in \mathcal{A}$, we identify the most violated system (index α^-) and compute an associated scaling parameter θ

$$\alpha^- := \arg\left[\min_{\alpha\in\mathcal{A}}\left(1 - \frac{\Delta\gamma^\alpha}{\Delta\Delta\gamma^\alpha}\right)\right] \quad \text{and} \quad \theta := \min_{\alpha\in\mathcal{A}}\left(1 - \frac{\Delta\gamma^\alpha}{\Delta\Delta\gamma^\alpha}\right) \tag{46}$$

and perform a *scaling update* of the plastic slip parameters by

$$\Delta\gamma^\alpha \Leftarrow \Delta\gamma^\alpha - (1-\theta)\,\Delta\Delta\gamma^\alpha \quad \text{for} \quad \alpha \in \mathcal{A}\,, \tag{47}$$

with $\Delta\Delta\gamma^\alpha = -\sum_{\beta\in\mathcal{A}} H^{-1}_{\alpha\beta}\, r_\beta$ according to (33). As a consequence of this scaling update, all incremental parameters are admissible in the sense of $(38)_2$, where $\gamma^{\alpha^-} = 0$ for the originally most violated system α^-. The question whether the system α^- needs to be removed from the active set $\mathcal{A}$ or not is decided based on the sensitivity ΔW of the incremental potential W with respect to the incremental slip in the solution point (47). For

$$\Delta W := \sum_{\alpha\in\mathcal{A}} r_\alpha \Delta\Delta\gamma^\alpha < 0 \tag{48}$$

the Newton iteration is on a minimization path, but in the non-admissible range of negative incremental slip. Then we remove the system α^-, i.e.

$$\mathcal{A} \Leftarrow \{\, \mathcal{A}/(\alpha^- \in \mathcal{A}) \,\} \tag{49}$$

and proceed with a new Newton update (33). If the Newton iteration with the possible removals of slip systems from the active set has converged we have obtained a local minimum in the admissible space of incremental plastic parameters $\gamma^\alpha \in \mathcal{A}$. Then we check the condition $(38)_1$ by monitoring the Lagrange parameters of the constrained systems. For some violations in the sense $r_\alpha < 0$ with $\alpha \in \mathcal{S}/\mathcal{A}$, we identify the most violated system(s)

$$\alpha^+ := \arg[\min_{\alpha\in\mathcal{S}/\mathcal{A}}(r_\alpha)]\,, \tag{50}$$

add them to the set of active slip systems via

$$\mathcal{A} \Leftarrow \{\, \mathcal{A} \cup (\alpha^+ \in \mathcal{S}/\mathcal{A}) \,\} \tag{51}$$

and proceed with the local Newton iteration. Otherwise the minimizing point is found and the local iteration terminated. The complete algorithm is summarized for both the standard and the variational update in Table 1. For the derivation of the consistent elastic-plastic tangent moduli we refer to [39].

3.3 Numerical Example: Simple Shear of a Single Crystal

As a first benchmark of single crystal plasticity we investigate the response of two copper fcc single crystals of different orientation in a homogeneous

simple shear test. The goal is to compare the results obtained through the standard and the variational update summarized in Table 1. We focus on isotropic elastic behavior (bulk modulus $\kappa = 128000\,\mathrm{MPa}$, shear modulus $\mu = 45000\,\mathrm{MPa}$) and ideal elastoplasticity with constant flow stresses $\tau_\alpha^p = 1\,\mathrm{MPa}$ on all twelve slip systems. The shearing, defined by a deformation gradient of the form $\mathbf{F} = \Gamma\,\mathbf{e}_2 \otimes \mathbf{e}_3$, is increased in increments $\Delta\Gamma = 0.1$ and $\Delta\Gamma = 0.01$, respectively. Euler angle sets of $(0°/0°/0°)$ and $(0°/45°/0°)$ characterize the initial crystal orientations. During the deformation of the specimen, a rotation of the crystal lattices is observed, which is represented in Fig. 1 by stereographic $\{111\}$-pole figures in the $\mathbf{e}_2$-$\mathbf{e}_3$-plane up to a maximum shearing of $\Gamma = 2$. While the orientation of the initially unrotated $(0°/0°/0°)$-crystal constantly changes with increasing deformation, the unit cell of the $(0°/45°/0°)$-specimen tends towards a fixed final position. Both the classical as well as the variational formulation lead to the same result, a dependence on the step size could not be observed in these pole figures. Figure 2 reports on the development of the Kirchhoff stress components τ_{12} and τ_{11} and the number of currently active slip systems n versus the amount of shear. The diagrams

Table 1. Stress update algorithm of multislip crystal plasticity

1. Given database $\{\mathbf{F}, \mathbf{F}_n^{p-1}, \mathcal{A}_n\}$. Initialize $\gamma^\alpha = 0 \;\; \forall\, \alpha \in \{1, ..., m\}$, $\mathcal{A} = \mathcal{A}_n$.
2. Update $\mathbf{F}^{p-1}$ and $\mathbf{F}^e$. Determine sensitivities $\{\, \mathbf{F}^e_{,\alpha}\,;\, \mathbf{F}^e_{,\alpha\hat\beta}\,\}$ or $\{\, \mathbf{F}^e_{,\hat\alpha}\,;\, \mathbf{F}^e_{,\hat\alpha\hat\beta}\,\}$.
3. Evaluate the derivatives of the lattice storage function

$$\mathbf{P}^e := \partial_{\mathbf{F}^e}\psi^e(\mathbf{F}^e) \quad \text{and} \quad \boldsymbol{\mathcal{A}}^e := \partial^2_{\mathbf{F}^e\mathbf{F}^e}\psi^e(\mathbf{F}^e).$$

4. For active slip systems $\alpha, \beta \in \mathcal{A}$ residuals and iteration matrix

$$r_\alpha := \mathbf{P}^e : \mathbf{F}^e_{,\odot} + \tau^p_{\alpha n} + \textstyle\sum_{\beta\in\mathcal{A}} h_{\alpha\beta n}\,(\gamma^\beta - \gamma^\beta_n) + \tau^v_0[\frac{\gamma^\alpha - \gamma^\alpha_n}{\dot\gamma_0 \Delta t}]^{1/\epsilon}$$
$$H_{\alpha\beta} := \mathbf{F}^e_{,\odot} : \boldsymbol{\mathcal{A}}^e : \mathbf{F}^e_{,\hat\beta} + \mathbf{P}^e : \mathbf{F}^e_{,\odot\hat\beta} + h^p_{\alpha\beta n} + \tfrac{\tau^v_0}{\epsilon\dot\gamma_0\Delta t}[\tfrac{\gamma^\alpha - \gamma^\alpha_n}{\dot\gamma_0\Delta t}]^{(1-\epsilon)/\epsilon}\delta_{\alpha\beta}$$

   with $\odot \equiv \alpha$ for *standard update* and $\odot \equiv \hat\alpha$ for *variational update*.
5. Convergence check: If $([\sum_{\alpha\in\mathcal{A}} r^2_\alpha]^{1/2} \le tol)$ go to 8.
6. Perform *Newton update* $\gamma^\alpha \Leftarrow \gamma^\alpha + \Delta\gamma^\alpha$ with $\Delta\gamma^\alpha := -\sum_{\beta\in\mathcal{A}}[\,H_{\alpha\beta}\,]^{-1}[\,r_\beta\,]$.
7. For violations $\gamma^\alpha < 0$ with $\alpha \in \mathcal{A}$: Identify $\alpha^- := \arg[\theta]$ with $\theta := \min_{\alpha\in\mathcal{A}}(1 - \gamma^\alpha/\Delta\gamma^\alpha)$, perform *scaling update* $\gamma^\alpha \Leftarrow \gamma^\alpha - (1-\theta)\Delta\gamma^\alpha$. For $\Delta W := \sum_{\alpha\in\mathcal{A}} r_\alpha \Delta\gamma^\alpha < 0$, perform *dropping update* $\mathcal{A} \Leftarrow \{\mathcal{A}/\alpha^-\}$. Go to 2.
8. For violations $r_\alpha < 0$ with $\alpha \in \mathcal{S}/\mathcal{A}$: Identify $\alpha^+ := \arg[\min_{\alpha\in\mathcal{S}/\mathcal{A}}(r_\alpha)]$, perform *system adding update* $\mathcal{A} \Leftarrow \{\mathcal{A} \cup \alpha^+\}$. Go to 2.
9. Compute the algorithmic elastic-plastic tangent moduli

$$\boldsymbol{\mathcal{A}}^{ep} := \boldsymbol{\mathcal{A}}^e - \textstyle\sum_{\alpha\in\mathcal{A}}\sum_{\beta\in\mathcal{A}}[\,H_{\alpha\beta}\,]^{-1}[\boldsymbol{\mathcal{A}}^e : \mathbf{F}^e_{,\odot})] \otimes [\boldsymbol{\mathcal{A}}^e : \mathbf{F}^e_{,\hat\beta})].$$

   Determine nominal stresses and tangent moduli

$$\mathbf{P} = \{\mathbf{P}^e\} \overset{2}{\circ} \mathbf{F}^{p-T} \quad \text{and} \quad \boldsymbol{\mathcal{A}} = \{\, \boldsymbol{\mathcal{A}}^{ep}\,\} \overset{2}{\circ} \mathbf{F}^{p-T} \overset{4}{\circ} \mathbf{F}^{p-T}.$$

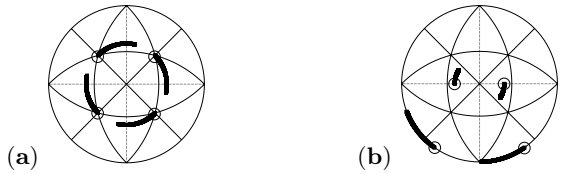


Fig. 1. Simple shear test. {111}-pole figures demonstrate the orientation evolution of fcc unit cell for initial (marked by circles) Euler angles (**a**) $(0°/0°/0°)$, (**b**) $(0°/45°/0°)$

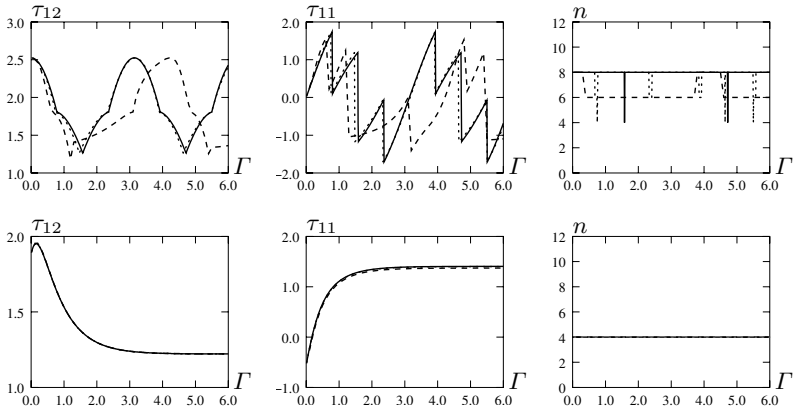


Fig. 2. Simple shear test. Kirchhoff stresses τ_{12} and τ_{11} and number n of active slip systems. Variational (- - - $\Rightarrow \Delta\Gamma = 0.1$, $\cdots \Rightarrow \Delta\Gamma = 0.01$) versus standard formulation (— $\Rightarrow \Delta\Gamma = 0.1$) for orientations $0°/0°/0°$ (top) and $0°/45°/0°$ (bottom)

at the top correspond to the $(0°/0°/0°)$-crystal, those at the bottom to the $(0°/45°/0°)$-orientation. Solid lines represent the standard update formulation with a load step $\Delta\Gamma = 0.1$, dashed curves reflect results of the variational update formulation with an identical step size, while the dotted lines indicate variational updates with a reduced load step size of $\Delta\Gamma = 0.01$. For the $(0°/0°/0°)$-crystal the standard return formulation yields reasonable results, which are symmetrical with respect to the amount of loading and, in a wide range, independent of the load incrementation. Therefore, only the curve for $\Delta\Gamma = 0.1$ is given as a reference solution. For this special crystal orientation the calculations based on the variational update show a distinct dependence on the load step. It has to be reduced to $\Delta\Gamma = 0.01$ in order to achieve similar results. In contrast to this first simulation, a crystal in $(0°/45°/0°)$-orientation gives smooth stress curves and the number of active slip systems remains constant throughout the deformation. As a result of this rather smooth material response, both algorithms and load step sizes give identical results.

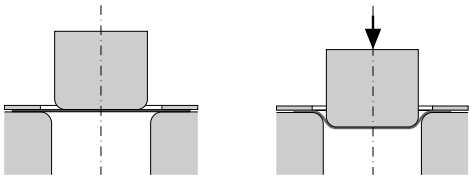


Fig. 3. Experimental setup of deep drawing

3.4 Numerical Example: Earing in Deep Drawing

We simulate a typical effect in deep drawing of fcc single crystal sheets. When drawing cups out of such sheets, the upper edges of the cups do not show a constant height, as one would expect for an isotropic material. Instead a periodic sequence of troughs and ears can be observed which is caused by the plastic anisotropy of the blank. The configuration in Fig. 3 is axisymmetric with respect to the punch direction. The circular sheet is fixed to the die by a ring. When pushing down the punch, the specimen is deformed until it finally drops out of the machine. In the calculation we concentrate on two types of single crystalline copper sheet metals, where number and form of the developing ears differ depending on the crystal orientation. The first one has a [111]-orientation, the latter one a [100]-orientation, meaning that those crystalline directions are perpendicular to the sheet surface. The orientation of the $\{111\}$-slip planes and their corresponding $\langle 110\rangle$-slip directions within the specimen are visualized in Fig. 4. Due to the symmetry of material behavior and loading, the deformation gives cups with six and four ears, respectively. The experimental results in Fig. 4 were published in [57], that analyzed earing in deep drawing of fcc aluminum sheets. The hardening law applied and the employed material parameters are reported upon in [39]. Exploiting the symmetry conditions of the problem, sectors of 60° and 45° are discretized, see Fig. 4. The results of the simulations are presented in Figs. 5 and 6. The plot sequence in Fig. 5 shows for the [111] aligned specimen the development of six ears. Figure 6 shows for an initial [100]-orientation the development of only four ears. Both simulations correspond well with experimental observations.

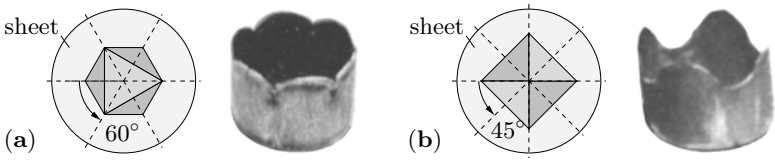


Fig. 4. Deep drawing of single crystal cups. Orientation of $\{111\}$-planes and experimental results. Symmetry of (**a**) 60° in [111]-orientation, (**b**) 45° in [100]-orientation

Notice, that in the latter case all ears have exactly the same shape, while for the [111]-sheet there are three equal pairs of ears.

4 Variational Problem of Homogenization

4.1 Fluctuation Fields on Heterogeneous Microstructures

Let $\boldsymbol{\varphi} : \mathcal{B} \times \mathcal{R}_+ \mathcal{R}^3$ denote the deformation map of a heterogeneous microstructure $\mathcal{B} \subset \mathcal{R}^3$ at $\mathbf{X} \in \mathcal{B}$ and time $t \in \mathcal{R}_+$. The deformation is assumed to be driven by a prescribed macro-deformation $\bar{\mathbf{F}}(t)$

$$\boldsymbol{\varphi}(\mathbf{X}, t) = \bar{\mathbf{F}}(t)\mathbf{X} + \mathbf{w}(\mathbf{X}, t) \tag{52}$$

consisting of a linear part $\bar{\mathbf{F}}(t)\mathbf{X}$ and a superimposed fine-scale fluctuation field $\mathbf{w} : \mathcal{B} \times \mathcal{R}_+ \mathcal{R}^3$. Following [19], the macro-deformation $\bar{\mathbf{F}}(t)$ is assumed to be governed by surface data of the microscopic deformation field $\bar{\mathbf{F}}(t) = \frac{1}{|\mathcal{B}|} \int_{\partial\mathcal{B}} \boldsymbol{\varphi}(\mathbf{X}, t) \otimes \mathbf{n}(\mathbf{X}) \, \mathrm{d}A$. Then, as a consequence of the decomposition (52), the fluctuation field $\mathbf{w}$ has to satisfy the constraint $\frac{1}{|\mathcal{B}|} \int_{\partial\mathcal{B}} \mathbf{w}(\mathbf{X}, t) \otimes \mathbf{n}(\mathbf{X}) \, \mathrm{d}A$. This constraint is satisfied for the three classes of constraints

$$\left.\begin{aligned} \bar{C}_D(\mathbf{w}; \mathbf{t}) &:= -\frac{1}{|\mathcal{B}|} \int_{\partial\mathcal{B}} \mathbf{t} \cdot \mathbf{w} \, \mathrm{d}A &&= 0 \\ \bar{C}_P(\mathbf{w}; \mathbf{t}^+) &:= -\frac{1}{|\mathcal{B}|} \int_{\partial\mathcal{B}^+} \mathbf{t}^+ \cdot (\mathbf{w}^+ - \mathbf{w}^-) \, \mathrm{d}A &&= 0 \\ \bar{C}_S(\mathbf{w}; \bar{\mathbf{P}}) &:= -\frac{1}{|\mathcal{B}|} \int_{\partial\mathcal{B}} (\bar{\mathbf{P}}\mathbf{n}) \cdot \mathbf{w} \, \mathrm{d}A &&= 0 \end{aligned}\right\} \tag{53}$$

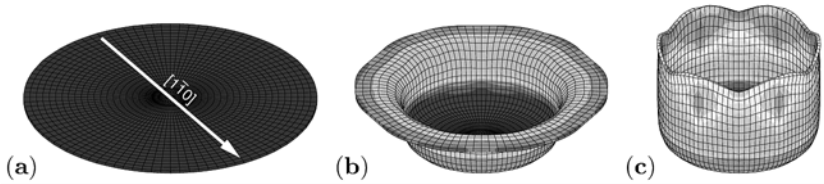


Fig. 5. Deep drawing of a single crystal sheet with [111]-orientation. (**a**) Initial geometry and crystalline orientation. (**b**)–(**c**) Accumulated plastic slip ($\gamma_{\max} = 1.99$)

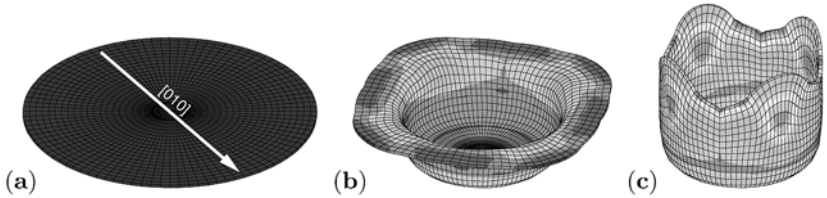


Fig. 6. Deep drawing of a single crystal sheet with [100]-orientation. (**a**) Initial geometry and crystalline orientation. (**b**)–(**c**) Accumulated plastic slip ($\gamma_{\max} = 2.22$)

associated with the surface $\partial\mathcal{B}$ of the micro-structure. The first constraint $\bar{C}_D$ in (53) demands *homogeneous deformation* with zero fluctuations $\mathbf{w} = \boldsymbol{0}$ on the boundary $\partial\mathcal{B}$ of the micro-structure, where the traction field $\mathbf{t}(\mathbf{X}, t)$ is considered to be a Lagrangian multiplier. The second constraint $\bar{C}_P$ states a *non-trivial periodicity* $\mathbf{w}^+ = \mathbf{w}^-$ of the superimposed fluctuation field on $\partial\mathcal{B}$. Here, the boundary $\partial\mathcal{B}$ is understood to be decomposed into two parts $\partial\mathcal{B} = \partial\mathcal{B}^- \cup \partial\mathcal{B}^+$ with outward normals $\mathbf{n}^+ = -\mathbf{n}^-$ at two associated points $\mathbf{X}^- \in \partial\mathcal{B}^-$ and $\mathbf{X}^+ \in \partial\mathcal{B}^+$. $\mathbf{t}^+(\mathbf{X}, t)$ is a Lagrangian multiplier field that characterizes the antiperiodic tractions on $\partial\mathcal{B}^+$. The third constraint $\bar{C}_S$ in (53) is associated with *homogeneous stresses* $\mathbf{t} = \bar{\mathbf{P}}\mathbf{n}$ on the boundary $\partial\mathcal{B}$, where the Lagrangian multiplier $\bar{\mathbf{P}}(t)$ is the macro-stress dual to $\bar{\mathbf{F}}(t)$. The discrete formulation of the boundary constraints (53) for finite element discretizations of microstructures is visualized in Fig. 7, see [35] for further details.

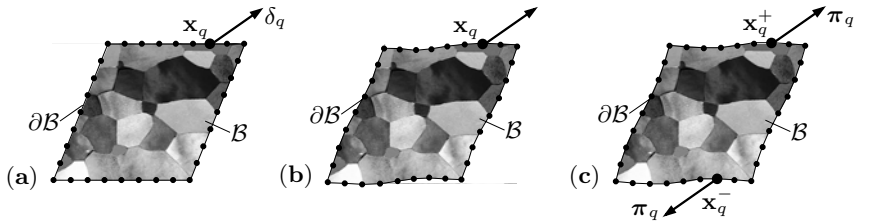


Fig. 7. Boundary constraints of discretized micro-structures with Lagrangian multipliers $\boldsymbol{\delta}_q$, $\boldsymbol{\sigma}$ and $\boldsymbol{\pi}$. (**a**) *Prescribed deformations* $C_D = -\frac{1}{|\mathcal{B}|}\sum_{q=1}^{M} \boldsymbol{\delta}_q \cdot [\mathbf{x}_q - \bar{\mathbf{F}}\mathbf{X}_q]$ with boundary forces $\boldsymbol{\delta}_{q=1\ldots M}$. (**b**) *Uniform tractions* $C_S = -\frac{1}{|\mathcal{B}|}\sum_{q=1}^{M} (\boldsymbol{\sigma}\mathbf{A}_q)\cdot\mathbf{x}_q + \boldsymbol{\sigma} : \bar{\mathbf{F}}$ with average stress $\boldsymbol{\sigma}$. (**c**) *Periodic fluctuations* $C_P = -\frac{1}{|\mathcal{B}|}\sum_{q=1}^{P} \boldsymbol{\pi}_q \cdot [(\mathbf{x}_q^+ - \mathbf{x}_q^-) - \bar{\mathbf{F}}(\mathbf{X}_q^+ - \mathbf{X}_q^-)]$ with forces $\boldsymbol{\pi}_q$ at $q = 1 \ldots P$ corresponding node pairs on boundary

4.2 Incremental Variational Formulation of Homogenization

Based on the constitutive minimization problem for the incremental response of standard dissipative materials defined in (6) we obtain the current stresses by the quasi-hyperelastic function evaluation (5). Hence, the micro-stress $\mathbf{P}_{n+1}$ and micro-moduli $\boldsymbol{\mathcal{C}}_{n+1}$ at a local point $\mathbf{X} \in \mathcal{B}$ of the heterogeneous micro-structure at time t_{n+1} are defined by

$$\mathbf{P}_{n+1} = \partial_{\mathbf{F}} W(\mathbf{F}_{n+1}; \mathbf{X}) \quad \text{and} \quad \boldsymbol{\mathcal{C}}_{n+1} = \partial^2_{\mathbf{FF}} W(\mathbf{F}_{n+1}; \mathbf{X}) \,. \tag{54}$$

The analogy to finite elasticity induces the following incremental variational formulation of the homogenization problem for normal-dissipative inelastic solids. As the key homogenization condition, we consider the *principle of minimum average incremental energy* proposed in [34]

$$\bar{W}(\bar{\mathbf{F}}_{n+1}) = \inf_{\mathbf{w}_{n+1}} \frac{1}{|\mathcal{V}|} \int_{\mathcal{B}} W(\bar{\mathbf{F}}_{n+1} + \nabla \mathbf{w}_{n+1}; \mathbf{X}) \, dV \tag{55}$$

for heterogeneous micro-structures with constituents of standard dissipative materials. The minimization problem is understood to be subject to the three alternative constraints defined in (53). The minimization problem determines the fluctuation field $\mathbf{w}_{n+1}$ on the micro-structure at discrete time t_{n+1}. The macro-stress potential $\bar{W}$ is assumed to define the macro-stresses $\bar{\mathbf{P}}_{n+1}$ and the macro-moduli $\bar{\boldsymbol{\mathcal{C}}}_{n+1}$ by the quasi-hyperelastic function evaluations

$$\bar{\mathbf{P}}_{n+1} = \partial_{\bar{\mathbf{F}}} \bar{W}(\bar{\mathbf{F}}_{n+1}) \quad \text{and} \quad \bar{\boldsymbol{\mathcal{C}}}_{n+1} = \partial^2_{\bar{\mathbf{F}}\bar{\mathbf{F}}} \bar{W}(\bar{\mathbf{F}}_{n+1}) \tag{56}$$

in complete analogy to (54). Thus the minimization problem (55) provides a shift of associated variables from the micro-scale to the macro-scale, often denoted as micro-to-macro transition. It extends the so-called *average variational principle* of nonlinear elasticity, outlined by [43] and [50], to the incremental formulation of finite inelasticity. The attractive feature of the proposed formulation (55) is that the structure of the average variational principle is preserved, i.e. the energy storage function ψ of finite elasticity is replaced by the incremental stress potential W of finite inelasticity. A finite element discretization of the minimization problem (55) of homogenization is outlined in Table 2. For further details we refer to the recent papers [34, 35] and [41].

Table 2. Discretization of minimization problem of homogenization

1. Given Database $\{\bar{\mathbf{F}}, \mathbf{d}_n\}$. Initialize nodal displacements $\mathbf{d}_{n+1} = \mathbf{d}_n$.
2. Determine gradient of fluctuation field $\mathbf{w}^h_{n+1} \in \mathcal{W}$ by the ansatz

$$\nabla \mathbf{w}^h_{n+1}(\mathbf{X}) = \mathbf{B}_e(\mathbf{X}) \mathbf{d}^e_{n+1} \quad \text{in} \quad \mathcal{B}^e \subset \mathcal{B}.$$

3. Evaluate minimization function

$$\bar{W}^h(\bar{\mathbf{F}}_{n+1}, \mathbf{d}_{n+1}) = \frac{1}{|\mathcal{V}|} \mathsf{A}_{e=1}^{n_e} \int_{\mathcal{B}^e} W(\bar{\mathbf{F}}_{n+1} + \mathbf{B}_e \mathbf{d}^e_{n+1}; \mathbf{X}) \mathrm{d}V$$

   and its derivatives $\bar{W}^h_{,\bar{\mathbf{F}}}$, $\bar{W}^h_{,\mathbf{d}}$, $\bar{W}^h_{,\bar{\mathbf{F}}\bar{\mathbf{F}}}$, $\bar{W}^h_{,\mathbf{dd}}$ and $\bar{W}^h_{,\mathbf{d}\bar{\mathbf{F}}}$.
4. Convergence check: If $([\sum_{\alpha \in \mathcal{A}} \bar{W}^{h\,2}_{,\mathbf{d}}]^{1/2} \leq tol)$ go to 6.
5. Perform Newton update of nodal displacements

$$\mathbf{d}_{n+1} \Leftarrow \mathbf{d}_{n+1} + \Delta \mathbf{d}_{n+1} \quad \text{with} \quad \Delta \mathbf{d}_{n+1} := -[\bar{W}^h_{,\mathbf{dd}}]^{-1} [\bar{W}^h_{,\mathbf{d}}].$$

6. Set nominal macro–stresses and macro–moduli

$$\bar{\mathbf{P}}_{n+1} = \bar{W}^h_{,\bar{\mathbf{F}}} \quad \text{and} \quad \bar{\boldsymbol{\mathcal{C}}}_{n+1} = \bar{W}^h_{,\bar{\mathbf{F}}\bar{\mathbf{F}}} - [\bar{W}^h_{,\bar{\mathbf{F}}\mathbf{d}}][\bar{W}^h_{,\mathbf{dd}}]^{-1} [\bar{W}^h_{,\mathbf{d}\bar{\mathbf{F}}}].$$

4.3 Macroscopic Incremental Variational Formulation

The existence of the homogenized stress potential (55) allows the introduction of an incremental minimization formulation of the macroscopic boundary-

value problem for standard dissipative solids. Consider a functional of the macro-deformation $\bar{\boldsymbol{\varphi}}_{n+1}$ at the right boundary of the increment $[t_n, t_{n+1}]$

$$\bar{I}(\bar{\boldsymbol{\varphi}}_{n+1}) = \int_{\bar{\mathcal{B}}} \bar{W}(\bar{\mathbf{F}}_{n+1})\, \mathrm{d}\bar{V} - [\, \bar{\Pi}_{\mathrm{ext}}(\bar{\boldsymbol{\varphi}}_{n+1}) - \bar{\Pi}_{\mathrm{ext}}(\bar{\boldsymbol{\varphi}}_n)\,] \tag{57}$$

with the global load potential function $\bar{\Pi}_{\mathrm{ext}}(\bar{\boldsymbol{\varphi}}) = \int_{\bar{\mathcal{B}}} \bar{\boldsymbol{\varphi}} \cdot \bar{\boldsymbol{\gamma}}\, \mathrm{d}\bar{V} + \int_{\partial\bar{\mathcal{B}}_t} \bar{\boldsymbol{\varphi}} \cdot \bar{\mathbf{t}}\, \mathrm{d}\bar{A}$ of dead body forces $\bar{\boldsymbol{\gamma}}(\bar{\mathbf{X}}, t)$ in $\bar{\mathcal{B}}$ and surface tractions $\bar{\mathbf{t}}(\bar{\mathbf{X}}, t)$ on $\partial\bar{\mathcal{B}}_{\bar{t}}$. The current macroscopic deformation map can then be determined by a *principle of minimum incremental energy for standard dissipative solids*

$$\bar{I}(\bar{\boldsymbol{\varphi}}^*_{n+1}) = \inf_{\bar{\boldsymbol{\varphi}}_{n+1} \in \bar{\mathcal{W}}} \bar{I}(\bar{\boldsymbol{\varphi}}_{n+1}) \tag{58}$$

subject to $\bar{\boldsymbol{\varphi}}_{n+1} \in \bar{\mathcal{W}} := \{\bar{\boldsymbol{\varphi}} \in \bar{\mathcal{W}}^{1,p}(\bar{\mathcal{B}}) | \bar{\boldsymbol{\varphi}} = \hat{\bar{\boldsymbol{\varphi}}}(\bar{\mathbf{X}})$ on $\partial\bar{\mathcal{B}}_{\bar{\varphi}}\}$ associated with prescribed deformations $\hat{\bar{\boldsymbol{\varphi}}}$ at $\bar{\mathbf{X}} \in \partial\bar{\mathcal{B}}_{\bar{\varphi}}$ on the boundary. The discretization of the minimization principle (58) can be performed in a straightforward manner by a displacement-type finite element method.

4.4 Numerical Example: Micro-structure Evolution in Polycrystals

We demonstrate the performance of the homogenization technique summarized in Table 2 by means of a texture analysis for a polycrystalline aggregate, as conceptually visualized in Fig. 8. The micro-structure considered is a copper polycrystal discretized by $7^3 = 343$ mixed finite elements. Individual crystal orientations, determined by a random distribution of quaternions as suggested in [42], are assigned to each integration point in order to simulate a polycrystalline aggregate of 2744 fcc single crystal grains. For the material parameters chosen, we refer to [42] and [41]. We consider a plane strain compression test of the polycrystalline micro-structure where the

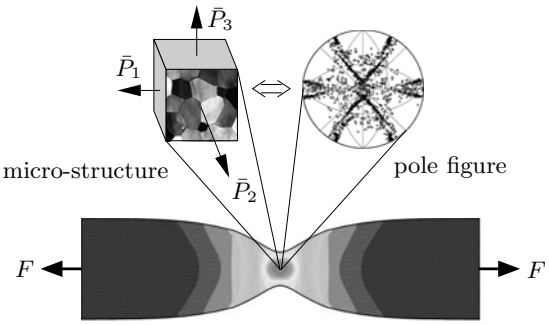


Fig. 8. Necking of a bar with an attached polycrystalline micro-structure. The texture plot visualizes the developed deformation micro-structure

macroscopic deformation gradient is prescribed by $\bar{\mathbf{F}} = [\bar{F}_{11}; \bar{F}_{22}; \bar{F}_{33}; ...]^T = [1; \exp(vt); \exp(-vt); 0; 0; 0; 0; 0; 0]^T$ with $v = 1.0s^{-1}$. During the simulation the time is linearly increased in an interval $t \in [0, 1.54\text{s}]$ in equal time steps $\Delta t = 2.5{\cdot}10^{-4}$s. The final state at time $t = 1.54$s corresponds to a compression of the micro-structure of 21% of its initial height. The plot sequences in Fig. 9 describe the deformed geometry for the three different constraints. The uniform traction condition represents the softest response of the polycrystalline aggregate, resulting in severely distorted finite element meshes in Fig. 9 (a). The texture development in the specimen is reported upon for the three types of micro-to-macro transitions in Fig. 10. We plotted the $\langle 111 \rangle$-pole figures with an equal-area projection. The results of the simulations are in a good qualitative agreement with the textures observed in experiments performed in [8]. The linear displacement condition in Fig. 10 (b) gives the sharpest texture and thus represents an upper bound of the pole figures. With increasing relaxation of the deformation constraint on the boundary the textures become more smeared. A lower bound is provided by the uniform stress condition in Fig. 10 (a). Note that during the final stage of deformation the texture becomes increasingly diffuse due to the heavily distorted finite element mesh depicted in Fig. 9 (a). The periodic solution in Fig. 10 (c) is very close to the experimental results in Fig. 10 (d).

5 Incompatibility Based Strain Gradient Plasticity

5.1 Incompatibility Induced Dislocation Storage

Within the local plasticity formulations discussed so far, a length scale is not included in the constitutive laws and hence size effects can not be predicted by such theories. However, several experiments, e.g. micro-bending, -torsion or -indentation tests, show that the real material behavior exhibits under certain circumstances a significant size dependence. These experiments generally result in plastically inhomogeneous deformations and consequently plastic strain

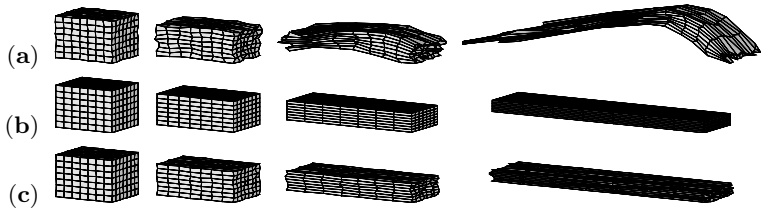


Fig. 9. Plane strain compression of a polycrystalline aggregate for different micro-to-macro transitions. Deformed micro-structures at 18%, 41%, 63% and 79% compression. (**a**) Uniform tractions, (**b**) linear displacements, (**c**) periodic displacements and antiperiodic tractions. Micro-structure stiffness (**c**) is bounded by (**a**) and (**b**)

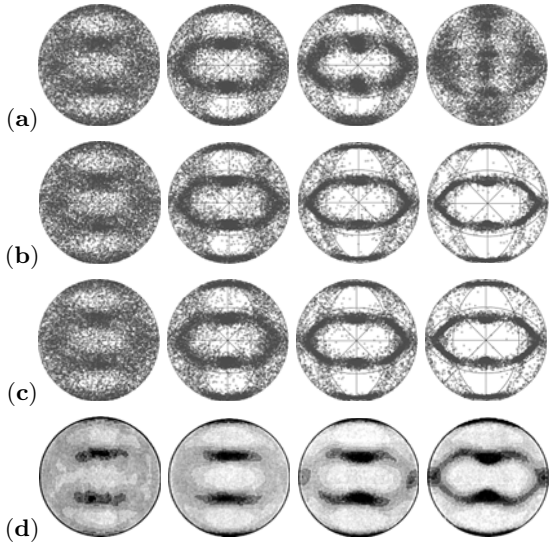


Fig. 10. Plane strain compression test. ⟨111⟩-pole figures in 12-plane for different micro-to-macro transitions corresponding to deformation states of figure 9. (**a**) Uniform tractions, (**b**) linear displacements, (**c**) periodic displacements and antiperiodic tractions, (**d**) experimental results of [8]. Pole figure (**c**) is bounded by (**a**) and (**b**)

gradients which, for compatibility reasons, require the storage of dislocations. Since these dislocations are *necessary* in order to accommodate non-uniform strains, they are denoted as *Geometrically Necessary Dislocations* (GNDs) in contrast to the *Statistically Stored Dislocations* (SSDs) which accumulate during uniform straining of the crystal. The GND density plays a predominant role at sufficiently small length scales where it might largely exceed the SSD density. An empirical description of size effects in polycrystals is given through the extended *Hall-Petch relation* ([16, 49]) for the yield strength σ

$$\sigma(\epsilon) = \sigma_0(\epsilon) + k\,(\epsilon) d^{-n} \tag{59}$$

with the yield strength $\sigma_0(\epsilon)$ of a single crystal and the slope $k(\epsilon)$ at an applied strain level ϵ and the exponent n. If it comes to plastically inhomogeneous deformations a purely local description in terms of $\mathbf{F}^e$ and $\mathbf{F}^p$, see Sect. 3.1, is insufficient since the corresponding intermediate configuration $\bar{\mathcal{B}}$ would be incompatible while the total deformation state remained compatible.

5.2 A Direct Approach to the Compatibility Analysis

To start the incompatibility analysis consider the integrability conditions

$$F^a{}_{A,B} = F^a{}_{B,A} \quad \text{and} \quad F^{A\,-1}{}_{a,b} = F^{A\,-1}{}_{b,a} \tag{60}$$

for the deformation gradient $\mathbf{F}$ and its inverse $\mathbf{F}^{-1}$ in Cartesian components. These conditions are *necessary* and on *simply connected* domains also *sufficient* for the existence of a continuous single valued primitive. Thus conditions (60) are the requirements for *global compatibility* of the spatial and the material configuration, respectively. An equivalent representation of (60) is given by

$$F^a{}_{A,B}\epsilon^{ABC} = 0 \quad \text{and} \quad F^{A\,-1}{}_{a,b}\epsilon^{abc} = 0\,, \tag{61}$$

or in direct notation

$$\mathrm{Curl}^T\mathbf{F} = \boldsymbol{0} \quad \text{and} \quad \mathrm{curl}^T\mathbf{F}^{-1} = \boldsymbol{0}\,. \tag{62}$$

Integrating (62) over an area $\mathcal{A}_\mathcal{B}$ ($\mathcal{A}_\mathcal{S}$) surrounded by a closed curve $\mathcal{C}_\mathcal{B}$ ($\mathcal{C}_\mathcal{S}$) and equipped with a surface normal $\mathbf{N}$ ($\mathbf{n}$) and applying Stoke's theorem gives

$$\int_{\mathcal{A}_\mathcal{B}} \mathrm{Curl}^T\mathbf{F}\mathbf{N}\,\mathrm{d}A = \oint_{\mathcal{C}_\mathcal{S}} \mathrm{d}\mathbf{x} = \boldsymbol{0} \quad \text{and} \quad \int_{\mathcal{A}_\mathcal{S}} \mathrm{curl}^T\mathbf{F}^{-1}\mathbf{n}\,\mathrm{d}a = \oint_{\mathcal{C}_\mathcal{B}} \mathrm{d}\mathbf{X} = \boldsymbol{0}\,. \tag{63}$$

If the Eulerian or Lagrangian configuration is incompatible the line integrals (63) are path dependent, i.e. reveal a closure failure, conditions (60) are violated, and $(62)_1$ and $(62)_2$ represent the associated incompatibility measures.

5.3 Dislocation Density Tensors

Interpreting $\mathbf{F}^p$ and $\mathbf{F}^{e\,-1}$ as *Pfaffian anholonomic transformations* $\mathrm{d}\bar{\mathbf{X}} = \mathbf{F}^p\mathrm{d}\mathbf{X}$ and $\mathrm{d}\bar{\mathbf{X}} = \mathbf{F}^{e\,-1}\mathrm{d}\mathbf{x}$, the *intermediate* closure failure is identified with the macroscopic *real* Burgers vector $\boldsymbol{\beta}^{\bar{\mathcal{B}}}$ representative of stored dislocations

$$\boldsymbol{\beta}^{\bar{\mathcal{B}}} := \oint_{\mathcal{C}_\mathcal{B}} \mathbf{F}^p\mathrm{d}\mathbf{X} = \int_{\mathcal{A}_\mathcal{B}} \mathrm{Curl}^T\mathbf{F}^p\mathbf{N}\,\mathrm{d}A = \oint_{\mathcal{C}_\mathcal{S}} \mathbf{F}^{e\,-1}\mathrm{d}\mathbf{x} = \int_{\mathcal{A}_\mathcal{S}} \mathrm{curl}^T\mathbf{F}^{e\,-1}\mathbf{n}\,\mathrm{d}a\,. \tag{64}$$

The corresponding incompatibility measures are denoted as *dislocation density tensors* measuring the density of excess dislocations of one sign. The fundamental plastic $\mathbf{A}^p_P$ and elastic $\mathbf{A}^e_{si}$ dislocation density tensors are defined by

$$\mathbf{A}^p_P := \mathrm{Curl}\mathbf{F}^p \quad \text{and} \quad \mathbf{A}^e_{si} := \mathrm{curl}\mathbf{F}^{e\,-1}\,. \tag{65}$$

$\mathbf{A}^p_P$ determines in formal analogy to the Cauchy theorem the *real* local Burgers vector $\bar{\mathbf{B}}_\mathcal{B}$ in the intermediate configuration

$$\mathbf{A}^{p\,T}_P : \{T^*_{\bar{\mathbf{X}}}\mathcal{B} \to T_{\bar{\mathbf{X}}}\bar{\mathcal{B}}, \quad \mathbf{N} \mapsto \bar{\mathbf{B}}_\mathcal{B} = \mathbf{A}^{p\,T}_P\mathbf{N}\}\,. \tag{66}$$

To point out this coincidence we equip the dislocation tensors with corresponding subscripts. Unlike $\mathbf{A}^p_P$, $\mathbf{A}^e_{si}$ does not correlate with a stress analogon and maps between the *s*-patial cotangent and the *i*-ntermediate tangent space

$$\mathbf{A}^{e\,T}_{si} : \{T^*_{\mathbf{x}}\mathcal{S} \to T_{\bar{\mathbf{X}}}\bar{\mathcal{B}}, \quad \mathbf{n} \mapsto \bar{\mathbf{B}}_\mathcal{S} = \mathbf{A}^{e\,T}_{si}\mathbf{n}\}\,. \tag{67}$$

Obviously $\mathbf{B}_{\bar{\mathcal{B}}}$ and $\bar{\mathbf{B}}_{\mathcal{S}}$ lie in the intermediate configuration but relate to a surface element $\mathrm{d}A$ or $\mathrm{d}a$ in the reference or the spatial configuration, respectively. Following [9], the *true* dislocation density tensor should measure the local Burgers vector in the intermediate configuration, per unit area in that configuration. Reformulation of (64) gives

$$\boldsymbol{\beta}^{\bar{\mathcal{B}}} := \int_{\mathcal{A}_{\bar{\mathcal{B}}}} \mathrm{Curl}^T \mathbf{F}^p \frac{1}{J^p} \mathbf{F}^{p\,T} \bar{\mathbf{N}}\, \mathrm{d}\bar{A} = \int_{\mathcal{A}_{\bar{\mathcal{B}}}} \mathrm{curl}^T \mathbf{F}^{e\,-1} J^e \mathbf{F}^{e-T} \bar{\mathbf{N}}\, \mathrm{d}\bar{A} \tag{68}$$

which implies the alternative representations purely in terms of $\mathbf{F}^p$ or $\mathbf{F}^{e\,-1}$

$$\mathbf{A} := \frac{1}{J^p} \mathbf{F}^p \mathrm{Curl} \mathbf{F}^p = \frac{1}{J^p} \mathbf{F}^p \mathbf{A}_P^p = J^e \mathbf{F}^{e\,-1} \mathrm{curl} \mathbf{F}^{e\,-1} = J^e \mathbf{F}^{e\,-1} \mathbf{A}_{si}^e \,. \tag{69}$$

Through (69) the relations $\mathbf{A}_P^p = J\mathbf{F}^{-1}\mathbf{A}_{si}^e$ and $\mathbf{A}_S^p := \mathbf{A}_P^p \mathbf{F}^{p-T}$ can be established where $\mathbf{A}_S^p$ represents the fully Lagrangian dislocation tensor. In analogy to the homogeneous equilibrium conditions, the continuity conditions demand that the divergences of the dislocation density tensors vanish

$$\mathrm{Div}[\mathbf{A}_P^{p\,T}] = J\,\mathrm{div}[\mathbf{A}_{si}^{e\,T}] = 0 \tag{70}$$

which states that a dislocation line can never end inside the crystal.

A more profound setting of the incompatibility analysis is provided through differential geometry and bases on the classical works [24, 6, 25] and more recently [53] and [1]. Thereby the continuum theory of dislocations is identified with a *Cartan geometry* where a non-vanishing *torsion tensor* measures in analogy to the dislocation density tensor the incompatibility and the associated GND storage and the *Riemann-Christoffel curvature tensor* has to vanish as long as only dislocations are considered as defects.

5.4 The Dislocation Density Tensor for Separate Single Slip

Evaluation of the true dislocation density tensor for the assumption of separate single slip on system α gives with $\mathbf{F}^{p\,\alpha} = \mathbf{1} + \gamma^\alpha \mathbf{S}^\alpha \otimes \mathbf{M}^\alpha$

$$\mathbf{A}^\alpha = (\nabla_{\mathbf{X}} \gamma^\alpha \mathbf{S}^\alpha)\, \mathbf{T}^\alpha \otimes \mathbf{S}^\alpha - (\nabla_{\mathbf{X}} \gamma^\alpha \mathbf{T}^\alpha)\, \mathbf{S}^\alpha \otimes \mathbf{S}^\alpha \,. \tag{71}$$

Thereby lattice geometric considerations show that a slip gradient $\nabla_{\mathbf{X}} \gamma^\alpha$ is directly related to the storage of a density of geometrically necessary edge $\rho^\alpha_{\mathrm{GND}\perp}$ and screw $\rho^\alpha_{\mathrm{GND}\odot}$ dislocations through

$$\rho^\alpha_{\mathrm{GND}\perp} = \frac{1}{b} \frac{\partial \gamma^\alpha}{\partial X_1} = \frac{1}{b} \nabla_{\mathbf{X}} \gamma^\alpha \mathbf{S}^\alpha \quad \text{and} \quad \rho^\alpha_{\mathrm{GND}\odot} = \frac{1}{b} \frac{\partial \gamma^\alpha}{\partial X_3} = \frac{1}{b} \nabla_{\mathbf{X}} \gamma^\alpha \mathbf{T}^\alpha \,. \tag{72}$$

Obviously a slip gradient perpendicular to the slip plane does not result in a GND storage. Insertion of (72) into (71) gives the following relation between the incompatibility measure $\mathbf{A}^\alpha$ and the GND density for a single slip system

$$\mathbf{A}^{\alpha} = b\,\rho^{\alpha}_{\mathrm{GND}\perp}\,\mathbf{T}^{\alpha}\otimes\mathbf{S}^{\alpha} - b\,\rho^{\alpha}_{\mathrm{GND}\odot}\,\mathbf{S}^{\alpha}\otimes\mathbf{S}^{\alpha} \tag{73}$$

which is the transpose of Nye's original representation or the negative of Kröner's dislocation density tensor. For multiple slip the relation (73) is generally not unique. To overcome possible redundancies, [3] proposed two alternative minimization techniques.

5.5 Dislocation Density Based Gradient Crystal Plasticity

To account for possible incompatibilities we extend the local crystal plasticity model outlined in Sect. 3.1 through inclusion of an isotropic hardening contribution due to the GND storage. Hence we extend (25) and (26) to

$$\tau^{p}_{\alpha} = c\mu b\sqrt{\sum\nolimits_{\beta} G^{\alpha\beta}\,(\rho^{\beta}_{\mathrm{SSD}} + \rho^{\beta}_{\mathrm{GND}})}\,. \tag{74}$$

Through (74) both, the SSDs and the GNDs equally contribute to the forest dislocation density. A possible kinematic hardening contribution such as considered in [53, 55] or [12] is not considered at this point but can be motivated through residual stresses resulting from the GND storage. In the present model we specify one GND density $\rho^{\alpha}_{\mathrm{GND}}$ for each slip system through the norm of the true dislocation tensor, i.e.

$$\rho^{\alpha}_{\mathrm{GND}} := \frac{1}{b}\,|\mathbf{A}^{\alpha}| = \sqrt{(\rho^{\alpha}_{\mathrm{GND}\perp})^2 + (\rho^{\alpha}_{\mathrm{GND}\odot})^2} \tag{75}$$

which is the SRSS-mean of edge and screw dislocations. Here the fact that screw dislocations may cross slip is neglected. This might be circumvented by an extended crystallographic dislocation basis (see [27]).

5.6 Numerical Implementation of Gradient Crystal Plasticity

Due to the incorporation of the slip gradients into the model the formulation becomes (*weakly*) nonlocal. For the implementation within a finite element context we developed two alternative algorithmic formulations denoted as *global* and *extended local formulation* in the following. Similar approaches are also discussed in [28] and [31]. Within the *monolithic* fully implicit treatment of the global formulation we consider both the deformation field $\boldsymbol{\varphi}$ and the plastic slips γ^{α} as *primary variables*. These fields are determined on a *global level* throughout the solution domain by solving simultaneously the *equilibrium subproblem* and the *loading unloading conditions* (21). The complementary restrictions $(21)_3$ are incorporated through a decomposition of the solution domain $\mathcal{B}$ into inactive elastic $\mathcal{B}^{\alpha}_{e}$ and active elastoplastic subdomains $\mathcal{B}^{\alpha}_{p}$ for each slip system α. This requires a non-trivial active set search at the node point level in line with the local format of Sect. 3.2. Therein it is also checked

for admissible plastic slips in the sense of $(21)_2$. For frozen active set this gives after linearization and discretization the system of fully coupled equations

$$\begin{aligned} \textstyle\sum_J [\mathbf{K}_{DD}]^{IJ} \Delta\mathbf{D}^J + \sum_J \sum_{\beta\in\mathcal{A}^J} [\mathbf{K}^{\beta}_{D\Gamma}]^{IJ} \Delta\Gamma^{\beta\, J} &= -[\mathbf{F}^{\rm int}_D - \mathbf{F}^{\rm ext}_D]^I \\ \textstyle\sum_J [\mathbf{K}^{\alpha}_{\Gamma D}]^{IJ} \Delta\mathbf{D}^J + \sum_J \sum_{\beta\in\mathcal{A}^J} [K^{\alpha\beta}_{\Gamma\Gamma}]^{IJ} \Delta\Gamma^{\beta\, J} &= -[F^{\alpha}_{\Gamma}]^I \quad \forall\, \alpha \in \mathcal{A}^I \text{at node } I \end{aligned} \tag{76}$$

in terms of the nodal displacement $\Delta\mathbf{D}$ and slip $\Delta\boldsymbol{\Gamma}$ increments and the tangent stiffnesses $\mathbf{K_{DD}}, \mathbf{K_{D\Gamma}}, \mathbf{K_{\Gamma D}}, \mathbf{K_{\Gamma\Gamma}}$ and right-hand sides $\mathbf{F}_D$ and $\mathbf{F_\Gamma}$. After each iteration step the active set is checked at each node. This results in continuously changing dimensions of the system (76) thus requires non-standard FE-structures and is especially for multislip rather costly. Alternatively we developed a much more efficient extended local formulation where the plastic slip distribution throughout the solution domain is determined on the basis of a *patch recovery* or a *L_2-projection* in the sense of an *operator split*. Thus in this approach the GND density is determined in each integration point on basis of the projected slips of the last time step

$$\rho^{\alpha}_{\rm GND} := \hat{\rho}^{\alpha}_{\rm GND}(\nabla_{\mathbf{X}} \boldsymbol{\Gamma}^{\alpha}_n). \tag{77}$$

5.7 Numerical Example: Simple Shear of a Crystalline Strip

As a first example we investigate the benchmark of constraint simple shear of an infinitely wide single crystalline copper strip under plane strain conditions for planar double slip, see Fig. 11 a). For the material parameters, the discretization with 1×100 quadrilaterals and further details we refer to [5] and [37]. Due to the constrained slip pronounced boundary layers with reduced shear deformation develop. The SSD density retains its initial value at the horizontal boundaries, where the plastic slip is constrained and accumulates towards the middle of the strip, see Fig. 12 a). In contrast the GND density accumulates in the range of large slip gradients, i.e. within the boundary layers, see Fig. 12 b). For sufficiently small strip heights the maximum

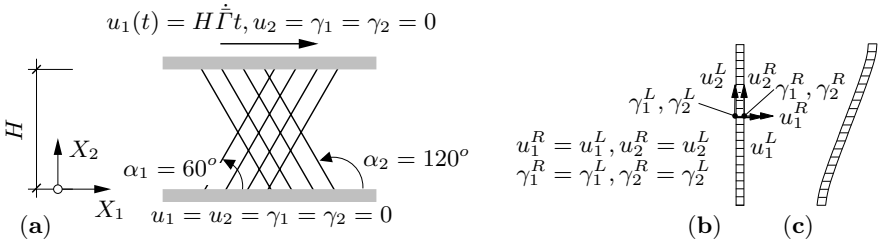


Fig. 11. Simple shear of an infinitely wide strip. (**a**) Geometry, boundary conditions and slip system orientations, (**b**) exemplary discretization and periodic boundary conditions, (**c**) deformed configuration for $H = 20\,\mu$m at $\bar{\Gamma} = 3\%$ (scaled by 10)

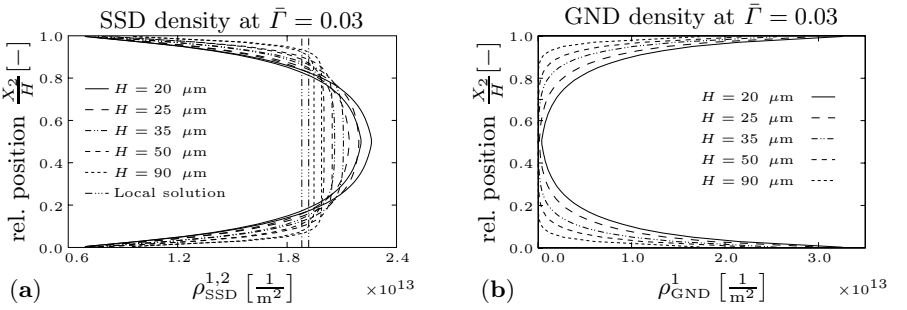


Fig. 12. Distribution of the SSD and GND density along the dimensionless strip height. (**a**) ρ_{SSD} profile (thick lines indicate the density ρ^1_{SSD} on system 1 and thin lines the density ρ^2_{SSD} on system 2, respectively) and (**b**) ρ_{GND} profile

GND density exceeds the SSD density and thus induces additional hardening. The resulting size dependence of the overall stress response $\bar{\tau}_{12}$ is visualized in Fig. 13 a). With the onset of plastic deformation a significant size effect is observable where the smaller strip heights exhibit a stiffer response. To quantify the size effect consider the *extended Hall-Petch relation* (59) i.e. $\bar{\tau}_{12} = \bar{\tau}_{12\,0}(\bar{\Gamma}) + k(\bar{\Gamma})H^{-n}$. A fit to the numerically obtained values is presented in Fig. 13 a) for applied shear levels of $\bar{\Gamma} = 0.01$ and $\bar{\Gamma} = 0.03$ and gives a *Hall-Petch exponent* of $n = 1.0$ along with a *Hall-Petch slope* which increases from $k(\bar{\Gamma} = 0.01) = 80$ to $k(\bar{\Gamma} = 0.03) = 175$. Qualitatively similar, slightly nonlinear, Hall-Petch relations have also been observed in [52] and [12].

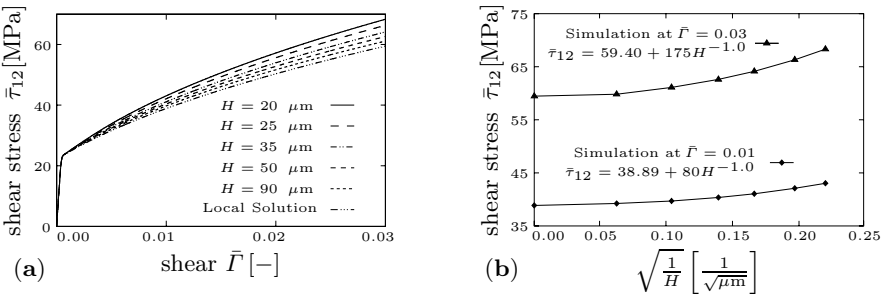


Fig. 13. Size effect. (**a**) Shear stress $\bar{\tau}_{12}$ vs. applied shear $\bar{\Gamma}$ for various heights H. (**b**) Flow ($\hat{=}$ shear) stress vs. Hall-Petch ratio $\sqrt{1/H}$ in current simulation (dots) and extended Hall-Petch fit (solid lines) at $\bar{\Gamma} = 0.01$ and $\bar{\Gamma} = 0.03$

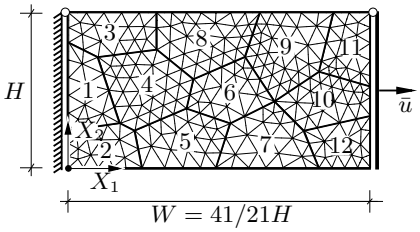


Fig. 14. Tension test of an aluminum polycrystal with 12 arbitrary oriented grains, boundary conditions and discretization with 472 quadratic triangles

Numerical Example: Tension Test for an FCC Polycrystal

As a second example we analyze an aluminum polycrystal subjected to uniaxial plane strain tension, see Fig. 14. The 12 single crystalline are characterized by arbitrary initial orientations and otherwise identical material parameters representative of aluminum, see [5] and [37]. The specimen is deformed up to a total deformation of 8%. Figure 15 a) shows the effective (i.e. over all 12 slip systems averaged) SSD density contours for $H = 84\,\mu$m. The SSDs accumulate inside the single crystalline grains while the GNDs accumulate along the grain boundaries, see Fig. 15 b). Again the GND density significantly increases with decreasing average grain size while the SSD density is almost insensitive to the overall specimen dimensions. In order to analyze the model capacity to predict the experimental measurements of [20, p. 238] we compute an average grain size $\bar{d}$ for the present setup through

$$A_{\text{grain}} = \frac{W \times H}{12} = \frac{41}{252} H^2 \quad \Rightarrow \quad \bar{d} = \sqrt{\frac{41}{252}} H \approx 0.403 H \,. \tag{78}$$

Then we fit the model parameters for an average grain size of $\bar{d} = 500\,\mu$m to the experimental data, see Fig. 16 a). With these parameters we simulate three specimens of smaller dimensions which result in the size effect shown in Fig. 16 a). Obviously the experimental observations can be recovered quite well, especially for $\bar{d} = 500\,\mu$m and $\bar{d} = 34\,\mu$m. The extended Hall-Petch

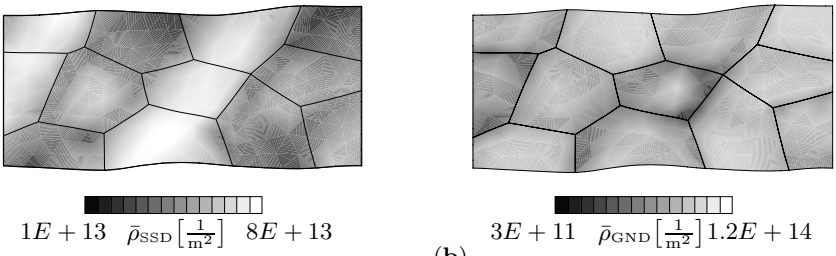


Fig. 15. Effective dislocation densities for $\bar{\epsilon} = 8\%$, $H = 84\,\mu$m. (**a**) $\bar{\rho}_{\text{SSD}}$, (**b**) $\bar{\rho}_{\text{GND}}$

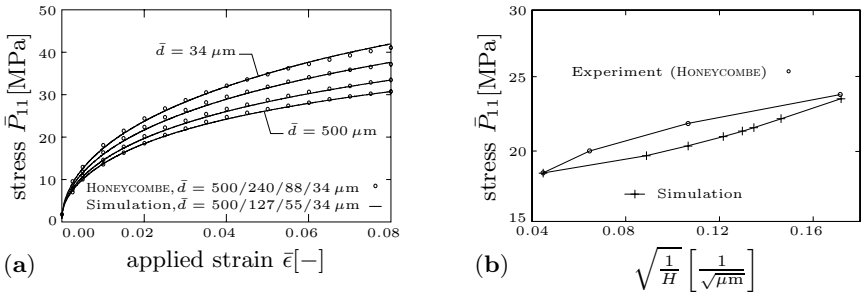


Fig. 16. Size effect, comparison between the simulation results and experimental measurements of Honeycombe [20]. (**a**) Overall response in terms of normal stress $\bar{P}_{11}$ versus applied strain $\bar{\epsilon}$. (**b**) Extended Hall-Petch relations at $\bar{\epsilon} = 2\%$

relations plotted at an applied strain level of $\bar{\epsilon} = 2\%$ in Fig. 16 b) show some differences for the intermediate grain sizes where the experiment reveals a Hall-Petch exponent of $n < 0.5$ while the present approach predicts an exponent of $n > 0.5$ close to $n = 1$. Thereby it is noteworthy that within the numerical simulations for the smaller specimen dimensions *no* further material parameters than the ones specified already for $\bar{d} = 500\,\mu$m had to be adopted.

6 Conclusions

We outlined recent results for the modeling of anisotropies in single crystals and polycrystals. The key contributions are (i) the development of an incremental variational formulation for standard dissipative materials, (ii) its application to the theoretical and computational treatment of crystal plasticity, (iii) the exploitation with respect to the homogenization of polycrystalline aggregates, e.g. for the prediction of texture evolution in polycrystals, (iv) the prediction of size effects through a physically motivated gradient enhanced crystal plasticity model. Focus within the recent research was a comparison of variational-type update algorithms in crystal plasticity with standard update schemes and the development of the new nonlocal formulation of crystal plasticity. The framework developed allows for the analysis of a broad spectrum of engineering applications such as the texture evolution in polycrystals, earing in deep drawing or size effects in polycrystals.

References

1. A. Acharya and J. L. Bassani. Lattice incompatibility and a gradient theory of crystal plasticity. *Journal of the Mechanics and Physics of Solids*, 48:1565–1595, 2000.
2. L. Anand and M. Kothari. A computational procedure for rate-independent crystal plasticity. *Journal of the Mechanics and Physics of Solids*, 44:525–558, 1996.
3. A. Arsenlis and D. M. Parks. Crystallographic aspects of geometrically-necessary and statistically-stored dislocation density. *Acta Materialia*, 47:1597–1611, 1999.
4. J. L. Bassani. Plastic flow of crystals. *Advances in Applied Mechanics*, 30:191–258, 1993.
5. M. Becker. *Incompatibility and Instability Based Size Effects in Crystals and Composites at Finite Elastoplastic Strains*. PhD thesis, Institut für Mechanik (Bauwesen), Report No. I-18, Universität Stuttgart, 2006.
6. B. A. Bilby, R. Bullough, and E. Smith. Continuous distributions of dislocations: a new application of the methods of non-riemannian geometry. *Proceedings of the Royal Society London A*, 231:263–273, 1955.
7. M. A. Biot. *Mechanics of Incremental Deformations*. John Wiley & Sons, Inc., New York, 1965.
8. C. A. Bronkhorst, S. R. Kalidindi, and L. Anand. Polycrystalline plasticity and the evolution of crystallographic texture in f.c.c. metals. *Philosophical Transactions of the Royal Society London*, 341:443–477, 1992.
9. P. Cermelli and M. E. Gurtin. On the characterization of geometrically necessary dislocations in finite plasticity. *Journal of the Mechanics and Physics of Solids*, 49:1539–1568, 2001.
10. A. M. Cuitiño and M. Ortiz. Computational modeling of single crystals. *Modelling and Simulation in Materials Science and Engineering*, 1:225–263, 1992.
11. H. Dai and D. M. Parks. Geometrically-necessary dislocation density and scale-dependent crystal plasticity. *Plasticity '97*, pages 17–18, 1997.
12. L. P. Evers, W. A. M. Brekelmans, and M. G. D. Geers. Non-local crystal plasticity model with intrinsic ssd and gnd effects. *Journal of the Mechanics and Physics of Solids*, 52:2379–2401, 2004.
13. N. A. Fleck, G. M. Müller, M. F. Ashby, and J. Hutchinson. Strain gradient plasticity: theory and experiment. *Acta Materialia*, 42:475–487, 1994.
14. P. Franciosi and A. Zaoui. Multislip in f.c.c. crystals: a theoretical approach compared with experimental data. *Acta Metallurgica*, 30:1627–1637, 1982.
15. M. E. Gurtin. A gradient theory of single-crystal viscoplasticity that accounts for geometrically necessary dislocations. *Journal of the Mechanics and Physics of Solids*, 50:5–32, 2002.
16. E. O. Hall. The deformation and aging of mild steel. part iii: discussion and results. *Proceedings of the Physical Society of London*, 64:747–753, 1951.
17. B. Halphen and Q. S. Nguyen. Sur les matéraux standards généralisés. *Journal de Mécanique*, 40:39–63, 1975.
18. R. Hill. Generalized constitutive relations for incremental deformation of metal crystals by multislip. *Journal of the Mechanics and Physics of Solids*, 14:95–102, 1966.

19. R. Hill. On constitutive macro-variables for heterogeneous solids at finite strain. *Proceedings of the the Royal Society of London, Series A*, 326:131–147, 1972.
20. R. W. K. Honeycombe. *The plastic deformation of metals.* Edward Arnold, London, 2nd edition, 1984.
21. J. W. Hutchinson. Elastic-plastic behaviour of polycrystalline metals and composites. *Proceedings of the the Royal Society of London, Series A*, 319:247–272, 1970.
22. U. F. Kocks. A statistical theory of flow stress and work-hardening. *The Philosophical Magazine*, 13:541–566, 1966.
23. W. T. Koiter. *General theorems of elasto-plastic solids,* Progress in Solid Mechanics. I. N. Sneddon; R. Hill (editors), North Holland Publishing Company, 1960.
24. K. Kondo. On the geometrical and physical foundations of the theory of yielding. *Proceedings Japan National Congress of Applied Mechanics*, 2:41–47, 1952.
25. E. Kröner. Allgemeine kontinuumstheorie der versetzungen und eigenspannungen. *Archive for Rational Mechanics and Analysis*, 4:273–334, 1960.
26. E. Kröner and C. Teodosiu. Lattice defect approach to plasticity and viscoplasticity. In A. Sawzuk, editor, *Problems in Plasticity.* Nordhoff International Publishing, 1972.
27. L. P. Kubin, G. Canova, M. Condat, B. Devincre, V. Pontikis, and Y. Brechet. Dislocation microstructures and plastic flow: a 3d simulation. *Solid State Phenomena*, 23/24:455–472, 1992.
28. T. Liebe and P. Steinmann. Theory and numerics of a thermomechanically consistent framework for geometrically linear gradient plasticity. *International Journal for Numerical Methods in Engineering*, 51:1437–1467, 2001.
29. J. Mandel. Plasticité clasique et viscoplasticité. In *CISM Courses and Lectures, No. 97.* Springer, 1972.
30. J. B. Martin. *Plasticity. Fundamentals and General Results.* MIT press, Cambridge, Massachusetts, 1975.
31. A. Menzel, R. Denzer, and P. Steinmann. On the comparison of two approaches to compute material forces for inelastic materials. application to single-slip crystal-plasticity. *Computer Methods in Applied Mechanics and Engineering*, 193:5411–5428, 2004.
32. C. Miehe. Exponential map algorithm for stress updates in anisotropic multiplicative elastoplasticity for single crystals. *International Journal for Numerical Methods in Engineering*, 39:3367–3390, 1996.
33. C. Miehe. Multisurface thermoplasticity for single crystals at large strains in terms of eulerian vector updates. *International Journal of Solids and Structures*, 33:3103–3130, 1996.
34. C. Miehe. Strain-driven homogenization of inelastic microstructures and composites based on an incremental variational formulation. *International Journal for Numerical Methods in Engineering*, 55:1285–1322, 2002.
35. C. Miehe. Computational micro-to-macro transitions for discretized microstructures of heterogeneous materials at finite strains based on the minimization of averaged incremental energy. *Computer Methods in Applied Mechanics and Engineering*, 192:559–591, 2003.
36. C. Miehe and M. Becker. Homogenization analysis in finite plasticity. Material and structural instabilities on the micro- and macro–scales and their interaction. *submitted to International Journal of Solids and Structures*, 2006.

37. C. Miehe and M. Becker. Incompatibility based strain gradient crystal plasticity. *submitted to Computer Methods in Applied Mechanics and Engineering*, 2006.
38. C. Miehe, M. Becker, and E. Gürses. Numerical computation of anisotropically evolving yield surfaces for polycrystals. *submitted to Acta Mechanica*, 2006.
39. C. Miehe and J. Schotte. Anisotropic finite elastoplastic analysis of shells: Simulation of earing in deep-drawing of single- and polycrystalline sheets by taylor-type micro-to-macro transitions. *Computer Methods in Applied Mechanics and Engineering*, 193:25–57, 2004.
40. C. Miehe and J. Schotte. Crystal plasticity and evolution of polycrystalline microstructure. In E. Stein, R. de Borst, and J. R. Hughes, editors, *Encyclopedia of Computational Mechanics*, chapter 8, pages 267–289. John Wiley & Sons, 2004.
41. C. Miehe, J. Schotte, and M. Lambrecht. Homogenization of inelastic solid materials at finite strains based on incremental minimization principles. Application to the texture analysis of polycrystals. *Journal of the Mechanics and Physics of Solids*, 50:2123–2167, 2002.
42. C. Miehe, J. Schröder, and J. Schotte. Computational homogenization analysis in finite plasticity. simulation of texture development in polycrystalline materials. *Computer Methods in Applied Mechanics and Engineering*, 171(3–4):387–418, 1999.
43. S. Müller. Homogenization of nonconvex integral functionals and cellular elastic materials. *Archive of Rational Mechanics and Analysis*, 99:189–212, 1987.
44. S. Nemat-Nasser and M. Hori. *Micromechanics: overall properties of heterogeneous materials*, volume 36 of *North-Holland series in applied mathematics and mechanics*. Elsevier Science Publisher B.V., 2. edition, 1999.
45. J. F. Nye. Some geometrical relations in dislocated crystals. *Acta Metallurgica*, 1:153–162, 1953.
46. M. Ortiz and E. A. Repetto. Nonconvex energy minimization and dislocation structures in ductile single crystals. *Journal of the Mechanics and Physics of Solids*, 47:397–462, 1999.
47. M. Ortiz and L. Stainier. The variational formulation of viscoplastic constitutive updates. *Computer Methods in Applied Mechanics and Engineering*, 171:419–444, 1999.
48. D. Peirce, R. Asaro, and A. Needleman. An analysis of nonuniform and localized deformation in ductile single crystals. *Acta Metallurgica*, 30:1087–1119, 1982.
49. N. J. Petch. The cleavage strength of polycrystals. *Journal of the Iron and Steel Institute*, 174:25–28, 1953.
50. P. Ponte Castañeda and P. Suquet. Nonlinear composites. *Advances in Applied Mechanics*, 34:171–303, 1998.
51. J. Rice. Inelastic constitutive relations for solids: an internal-variable theory and its application to metal plasticity. *Journal of the Mechanics and Physics of Solids*, 19:433–455, 1971.
52. J. Y. Shu, N. A. Fleck, E. Van der Giessen, and A. Needleman. Boundary layers in constrained plastic flow: comparison of nonlocal and discrete dislocation plasticity. *Journal of the Mechanics and Physics of Solids*, 49:1361–1395, 2001.
53. P. Steinmann. Views on multiplicative elastoplasticity and the continuum theory of dislocations. *International Journal of Engineering Science*, 34:1717–1735, 1996.

54. P. Suquet. Elements of homogenization for inelastic solid mechanics. In E. Sanchez-Palenzia and A. Zaoui, editors, *Lecture Notes in Physics: Homogenization Techniques for Composite Materials*, pages 193–278. Springer–Verlag, vol. 272 edition, 1987.
55. B. Svendsen. Continuum thermodynamic models for crystal plasticity including the effects of geometrically-necessary dislocations. *Journal of the Mechanics and Physics of Solids*, 50:1297–1329, 2002.
56. G. I. Taylor. Plastic strain in metals. *Journal of the Institute of Metals*, 62:307–324, 1938.
57. G. E. G. Tucker. Texture and earing in deep drawing of aluminum. *Acta Metallurgica*, 9:275–286, 1961.
58. E. Van der Giessen and A. Needleman. Discrete dislocation plasticity: a simple planar model. *Modelling and Simulation in Materials Science and Engineering*, 3:689–735, 1995.
59. G. Weber and L. Anand. Finite deformation constitutive equations and a time integration procedure for isotropic, hyperelastic-viscoplastic solids. *Computer Methods in Applied Mechanics and Engineering*, 79:173–202, 1990.
60. H. Ziegler. Some extremum principles in irreversible thermodynamics with application to continuum mechanics. In I. N. Sneddon and R. Hill, editors, *Progress in Solid Mechanics, Vol. IV*. North–Holland Publishing Company, Amsterdam, 1963.

Fluid-Structure Interaction of Incompressible Flows and Thin-Walled Structures*

Christiane Förster[1], Steffen Genkinger[1], Malte Neumann[1], Wolfgang A. Wall[2], and Ekkehard Ramm[1]

[1] Institute of Structural Mechanics, University of Stuttgart, Pfaffenwaldring 7, 70569 Stuttgart, Germany
eramm@statik.uni-stuttgart.de
neumann@statik.uni-stuttgart.de
genkinger@statik.uni-stuttgart.de
foerster@statik.uni-stuttgart.de

[2] Chair of Computational Mechanics, Technical University of Munich, Boltzmannstrasse 15, 85747 Garching, Germany
wall@lnm.mw.tum.de

Summary. A numerical approach to simulate the coupled problem of slender structures and incompressible Newtonian flows is described. The formulation is based on a partitioned scheme while finite elements are employed on the single fields. The structural field is discretized in space by finite elements based on a seven-parameter shell formulation capable of dealing with the complex dynamics of very slender structures. An efficient solver for the structural system of equations exploiting an algebraic multigrid approach has been developed. A stabilized fluid element following an ALE formulation of the incompressible Navier-Stokes equations is used to model the flow on a time dependent domain. The geometric conservation law is built in into the formulation to obtain a stable scheme which is second order accurate in time. The accuracy and reliability of the approach with respect to critical parameters such as very small time steps, distorted meshes or steep gradients in particular within the flow field has been analyzed. The approach is able to also deal with free fluid surfaces.

The inherent instability of weakly coupled approaches is analyzed which reveals the devastating influence of the so-called 'artificial added mass' effect. The convergence of the strongly coupled approach depends on an appropriately chosen relaxation parameter. Ways to automatically adjust the proper amount of relaxation are given. Special emphasis is also put on the efficient implementation particularly of the fluid problem. Vectorization is employed to significantly speed up the time spent for calculating element matrices on vector machines.

Keywords: Fluid-structure interaction, finite elements, stabilization methods, geometric conservation law, vectorization

* Research Project B4 "Fluid-Structure Interaction of Shells"

1 Introduction

The project is concerned with the modeling and discretization of interaction phenomena of incompressible flows and thin structures. It aims at a broad spectrum of applications from engineering problems such as liquid filled containers exposed to earthquake acceleration to biomedical questions like physiological blood flow. Within previous periods of the project the general methodology and a computational environment to treat coupled multi field problems had been developed. With regard to fluid-structure interaction partitioned strategies for surface coupled problems have been addressed. A three field fluid-structure interaction solver has been developed which can be decomposed into a computational fluid dynamics (CFD) module, a nonlinear structural mechanics part (CSD) and a mesh dynamics solver (CMD). The CFD approach to model incompressible viscous flows is based on a stabilized finite element method formulated in an Arbitrary Lagrangian-Eulerian (ALE) approach which accounts for the time dependent fluid domain. The geometrically nonlinear structural dynamics is based on a three-dimensional shell formulation, mixed hybrid finite elements and direct time integration with controllable numerical damping. Quasi-elastic pseudo-structural approaches are used to model the mesh motion.

At the beginning of the final period of the project further improvements had been developed among which there are two- and three-level approaches to solve the Navier-Stokes equations and free surface formulations which had been implemented for two-dimensional fluid and fluid-structural problems. Further robust, strongly coupled partitioned solution approaches basing on an iterative Dirichlet-Neumann substructuring approach equipped with relaxation methods which accelerate or even enable convergence had been developed. Additional effort had been put into the development of iterative and parallel solution approaches for thin shells.

2 Governing Equations and Basic Assumptions

The simulation of a wide variety of complex fluid-structure interaction problems requires generally applicable models of the different physical fields along with stable, efficient and accurate discretization and solution approaches.

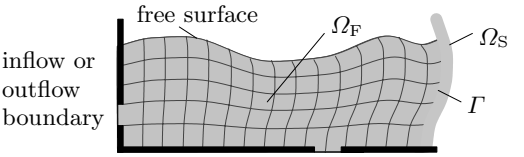


Fig. 1. Sketch of general fluid-structure interaction problem

2.1 Structural Domain

Within the considered fluid-structure interaction problems large structural deformations have to be taken into account. Thus geometrical nonlinearities are incorporated. However the strains of slender structures can be assumed to be small for many applications and the consideration of linear material responses suffices at many instants while nonlinear material behavior can be incorporated as well. The structural displacements $\mathbf{d}$ are governed by the geometrically nonlinear elastodynamic equations

$$\rho^S \frac{\mathrm{D}^2\mathbf{d}}{\mathrm{D}t^2} = \boldsymbol{\nabla} \cdot (\mathbf{F} \cdot \mathbf{S}) + \rho^S \mathbf{f}^S \qquad \text{in } \Omega_\mathrm{S} \times (0,T) \,, \tag{1}$$

where ρ^S and $\mathbf{f}^S$ represent the structural density and specific body force, respectively. The differential D denotes the material time derivative. The second Piola-Kirchhoff stress tensor $\mathbf{S}$ is related to the Green-Lagrangian strains via

$$\mathbf{S} = \mathbf{C} : \mathbf{E} \quad \text{with} \quad \mathbf{E} = \frac{1}{2}\left(\mathbf{F}^T \cdot \mathbf{F} - \mathbf{I}\right) \,,$$

where $\mathbf{C}$ denotes the material tensor and represents the deformation gradient. The problem (1) is subject to appropriate initial and boundary conditions.

2.2 Fluid Domain

A fluid is considered Newtonian governed by the incompressible Navier-Stokes equations. The flow equations determining the unknown velocity $\mathbf{u}$ and the kinematic pressure p read

$$\left.\frac{\partial \mathbf{u}}{\partial t}\right|_{\mathbf{x}} + \boldsymbol{\nabla}\left(\mathbf{u} \otimes \mathbf{u}\right) - 2\nu \boldsymbol{\nabla} \cdot \boldsymbol{\varepsilon}(\mathbf{u}) + \boldsymbol{\nabla} p = \mathbf{f}^F \qquad \text{in } \Omega_\mathrm{F} \times (0,T) \,, \tag{2}$$

$$\boldsymbol{\nabla} \cdot \mathbf{u} = 0 \qquad \text{in } \Omega_\mathrm{F} \times (0,T) \,. \tag{3}$$

The parameter $\nu = \mu/\rho^F$ denotes the kinematic viscosity where μ represents the viscosity and ρ^F the density of the fluid. The vector field $\mathbf{f}^F$ denotes the specific body force on the fluid. The kinematic pressure is denoted by p while $\bar{p} = p\rho^F$ is the physical pressure value.

In order to formulate the balance of momentum in a deforming ALE frame of reference the coordinate system $\boldsymbol{\chi}$ is introduced which follows the motion of the respective boundaries while deforming arbitrary in between. The geometrical location of a mesh point is obtained from the unique mapping $\mathbf{x} = \varphi(\boldsymbol{\chi}, t)$. Employing the reference system $\boldsymbol{\chi}$ and Reynolds transport theorem equation (2) can be reformulated on moving grids as

$$\left.\frac{\partial (\mathbf{u} J_t)}{\partial t}\right|_{\boldsymbol{\chi}} + \left\{\boldsymbol{\nabla} \cdot \left(\mathbf{u} \otimes (\mathbf{u} - \mathbf{u}^G)\right) - 2\nu \boldsymbol{\nabla} \cdot \boldsymbol{\varepsilon}(\mathbf{u}) + \boldsymbol{\nabla} p\right\} J_t = \mathbf{b} J_t \,, \tag{4}$$

where $J_t = \det(\partial \mathbf{x}/\partial \boldsymbol{\chi})$ denotes the time dependent Jacobian of the mapping and $\mathbf{u}^G = \partial \mathbf{x}/\partial t|_{\boldsymbol{\chi}}$ represents the velocity of the reference system, i.e. in the discretized case the grid velocity.

It is very common to use the finite volume method for the discretization of flow problems. In this case equation (4) has to be used as point of departure for a discretization in space which results in a difficulty with the temporal discretization of the mass term. This difficulty can be seen in the weighted residual formulation where the mass term yields a time derivative of an integral over a temporally changing domain. The time discretization of this term is rather cumbersome. Thus extra effort has to be made in order to satisfy the geometric conservation law resulting in the need for temporal averaging of either geometries or fluxes within a time step when a finite volume discretization is applied [9, 10]. Depending on the temporal discretization scheme different versions of the geometric conservation law, i.e. different discrete geometric conservation laws (DGCL), have to be used [22] if (4) is discretized directly.

In view of a discretization by means of finite elements these difficulties can be avoided as shown in [13]. To this end the geometric conservation law

$$\frac{\partial J_t}{\partial t} = J_t \boldsymbol{\nabla} \cdot \mathbf{u}^G \tag{5}$$

is introduced into (4) which links the temporal change of the domain to the domain velocity. Equation (5) can be incorporated into (4) yielding a local form of the balance of momentum which can straightforwardly be discretized in time and space

$$\left.\frac{\partial \mathbf{u}}{\partial t}\right|_{\boldsymbol{\chi}} + \left(\mathbf{u} - \mathbf{u}^G\right) \cdot \boldsymbol{\nabla} \mathbf{u} - \frac{1}{\rho^F} \boldsymbol{\nabla} \cdot \boldsymbol{\sigma} = \mathbf{f}^F \quad \text{in } \Omega_{\mathrm{F}} \times (0, T) \,, \tag{6}$$

where the stress tensor of a Newtonian fluid is given by

$$\boldsymbol{\sigma} = -\bar{p}\mathbf{I} + 2\mu\boldsymbol{\varepsilon}(\mathbf{u}) \quad \text{with} \quad \boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}\left(\boldsymbol{\nabla}\mathbf{u} + \boldsymbol{\nabla}\mathbf{u}^T\right)$$

denoting the strain rate tensor. The local ALE form (6) avoids the difficulties inherent in (4) and allows to preserve the stability as well as the order of accuracy which a discretization scheme has on fixed grids. As the geometric conservation law (5) has been incorporated prior to discretization the need for different DGCLs is removed [13].

The partial differential equation (6) is subject to the initial and boundary conditions

$$\begin{aligned} &\mathbf{u} = \mathbf{u}_0 \qquad \text{in } \Omega_{\mathrm{F}} \qquad \text{at } t = 0 \\ &\mathbf{u} = \hat{\mathbf{u}} \qquad \text{on } \Gamma_{\mathrm{D}} \,, \qquad \boldsymbol{\sigma} \cdot \mathbf{n} = \hat{\mathbf{h}}^F \quad \text{on } \Gamma_{\mathrm{N}} \,. \end{aligned} \tag{7}$$

3 Modeling and Discretization

The coupled problem is approximated by finite elements where elements specificly designed to meet the respective requirements are used on the single fields.

3.1 Nonlinear Shell Dynamics

Finite Element Formulation

Thin-walled structures are of particular interest in the context of fluid-structure interaction. Thus shell elements are employed, where the formulation includes geometrical nonlinearities and allows for the use of fully three-dimensional material laws without modification [4]. This seven-parameter shell formulation is depicted in Fig. 2. In addition to the classical Reissner-Mindlin kinematics (yielding five degrees of freedom per node) it contains the thickness stretch leading to an additional degree of freedom and a seventh strain degree of freedom which allows for a linearly varying strain E_{33} in shell thickness direction.

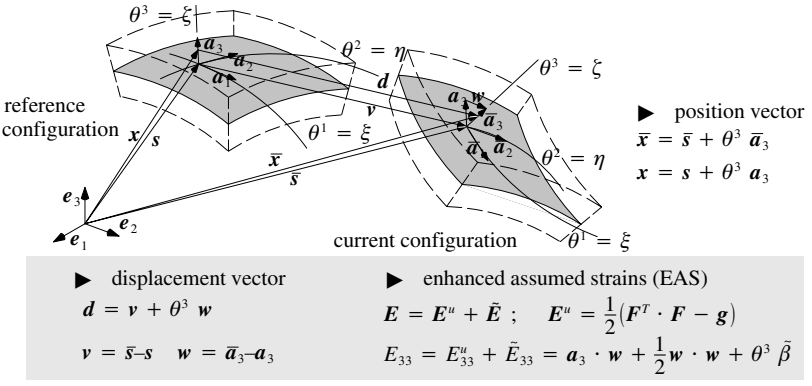


Fig. 2. Seven parameter shell formulation

This strain degree of freedom is locally introduced by means of the enhanced assumed strain method (EAS) [31] and condensed out on element base prior to assembly. This formulation is asymptotically correct and the lowest possible approach which allows a complete three-dimensional description. Locking inherent in low order elements is largely removed by means of enhanced assumed strains (EAS) and the Assumed Natural Strains method (ANS) [2, 4].

This spatial discretization yields the semi-discrete structural equation

$$\mathbf{M}^S\ddot{\mathbf{d}} + \mathbf{D}^S\dot{\mathbf{d}} + \mathbf{N}^S(\mathbf{d}) = \mathbf{f}^S \ , \tag{8}$$

where $\mathbf{M}^S$ and $\mathbf{D}^S$ represent the structural mass and damping matrix, respectively, while $\mathbf{N}^S$ and $\mathbf{f}^S$ denote the internal and external nodal force vectors. The nodal displacements, velocities and accelerations are given by $\mathbf{d}$, $\dot{\mathbf{d}}$ and $\ddot{\mathbf{d}}$, respectively. The temporal discretization of (8) is performed by means of the *generalized-α* time discretization [7]. Equations (8) are consistently linearized and solved iteratively by a Newton-Raphson procedure.

Solver for Thin Shell Structures

Implicit temporal discretization schemes require the solution of the effective structural system of equations

$$\mathbf{K}_{\text{eff}}\Delta\mathbf{d} = \mathbf{f}^{\alpha}_{\text{eff}} \tag{9}$$

several times within a Newton-Raphson iteration. Generalized-α time integration yields an effective coefficient matrix of

$$\mathbf{K}_{\text{eff}} = \left(\frac{1-\alpha_m}{\beta\Delta t^2}\mathbf{M}^S + \frac{(1-\alpha_f)\gamma}{\beta\Delta t}\mathbf{D}^S + (1-\alpha_f)\mathbf{K}_T(\mathbf{d}^n) \right) \ ,$$

where the parameters α_m and α_f allow to control the numerical dissipation of the scheme and $\mathbf{K}_T(\mathbf{d}^n)$ denotes the linearization of the internal force vector. Preconditioned iterative solvers which allow for parallelization and offer reasonable memory requirements and potentially small solution times are used to solve (9). However the condition number

$$c_K(\mathbf{K}_{\text{eff}}) = \frac{\lambda_{\max}}{\lambda_{\min}}(\mathbf{K}_{\text{eff}})$$

significantly influences the efficiency and stability of iterative solvers. It depends upon the slenderness of the structure as well as on the particular choice of the degrees of freedom. Compared to the classical shell formulation with rotational degrees of freedom neglecting the thickness change (five-parameter formulation, 5-PM) the seven-parameter formulation uses displacement differences. Thus it exhibits particularly bad conditioned systems. The diagram in Fig. 3 showing the spectrum of the eigenvalues of a thin plate highlights this effect. The seven-parameter formulation (7-PM) yields a number of very high eigenvalues and thus a particularly unfavorable condition number.

The problem is removed by means of a scaled director conditioning (SDC). To this end, a scaling factor $\mathcal{C}$ for the director difference displacement $\mathbf{w}$ as depicted in Fig.4 is introduced algorithmically increasing the shell thickness temporarily while the effect is condensed out afterwards, e.g. within the thickness integration of the material law [16, 38]. The underlying physics remain unchanged while the disadvantage of the difference displacement formulation is entirely removed as depicted in Fig. 3 c.

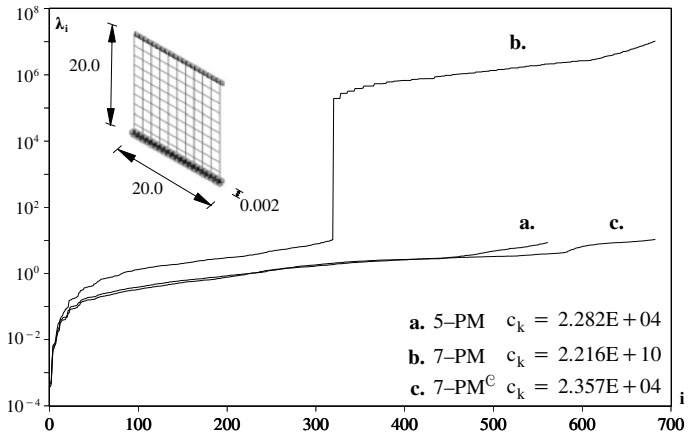


Fig. 3. Eigenvalue spectrum with scaled director conditioning

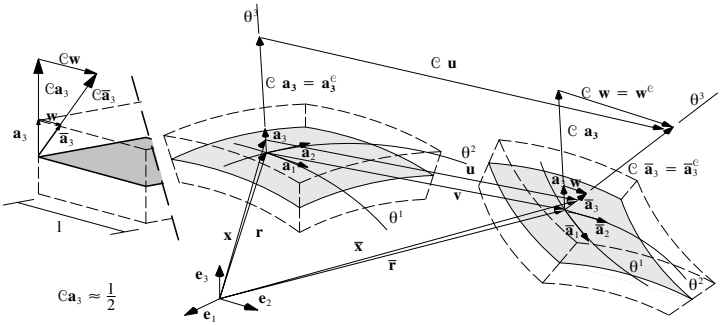


Fig. 4. Scaled director conditioning

The scaled director approach is implemented directly into the shell formulation and although similar it is not a preconditioner in the classical sense. The additional numerical effort is negligible while the approach can easily be combined with classical preconditioning. As the SDC approach almost exclusively effects the higher modes of the eigenvalue spectrum it can conveniently be combined with multi-level preconditioners which aim at smoothing low frequency modes [16].

3.2 Stabilized Fluid Formulation

Temporal Discretization

The fluid equations (6) and (3) are discretized in time and subsequently in space. One-step-θ time integration and second order backward differentiation (BDF2) are used for temporal discretization to obtain the spatial differential equation

$$\mathbf{u}^{n+1} + \delta \left[\left(\mathbf{u}^{n+1} - \mathbf{u}^{G,n+1} \right) \cdot \boldsymbol{\nabla} \mathbf{u}^{n+1} - 2\nu \boldsymbol{\nabla} \cdot \boldsymbol{\varepsilon}(\mathbf{u}^{n+1}) + \boldsymbol{\nabla} p^{n+1} \right] = \mathbf{r}^{n+1} \,, \tag{10}$$

where the superscript $n+1$ denotes the time level and δ represents a scalar coefficient depending upon the discretization scheme, i.e. $\delta_\theta = \theta \, \Delta t$ or $\delta_{BDF2} = 2\Delta t/3$ for one-step-θ or BDF2, respectively. The right-hand side vector $\mathbf{r}^{n+1}$ denotes the data at the new time level

$$\begin{aligned} \mathbf{r}_\theta^{n+1} &= \delta_\theta \mathbf{f}^{F,n+1} + (1-\theta)\Delta t \dot{\mathbf{u}}^n + \mathbf{u}^n \,, \\ \mathbf{r}_{BDF2}^{n+1} &= \delta_{BDF2} \mathbf{f}^{F,n+1} + \frac{4}{3}\mathbf{u}^n - \frac{1}{3}\mathbf{u}^{n-1} \,. \end{aligned}$$

Spatial Discretization

The semi discrete equation (10) is now discretized in space by finite elements.

To define the Galerkin weak form, we select the finite element spaces $\mathbf{V}_{0,t}^h \subset \mathbf{H}_0^1(\Omega_\mathrm{F})$ and $\mathbf{V}_t^h \subset \mathbf{H}^1(\Omega_\mathrm{F})$, where $\mathbf{V}_t^h$ is a space of functions satisfying the time dependent Dirichlet boundary conditions while all functions in $\mathbf{V}_{0,t}^h$ are zero on Γ_D. If not prescribed by Neumann boundary conditions the pressure is taken from the space $P_t^h \subset L_0^2(\Omega_\mathrm{F})$ of square integrable functions with prescribed mean to account for the free additive constant of the pressure variable. Explicitly the discrete velocity and pressure fields are given by

$$\mathbf{u} = \sum_j \varphi_j \, \mathbf{e}_i u_i^j \quad \text{and} \quad p = \sum_j \varphi_j \, p^j \,, \tag{11}$$

where φ_j is the basis function of the discrete space associated with the node j while u_i^j and p_j denote the velocity component in i-direction and the pressure at node j, respectively.

The discrete variational statement is as follows: Find $\mathbf{u} \in \mathbf{V}_{n+1}^h$, $p \in P_{n+1}^h$ such that

$$\begin{aligned} B_{gal}\left(\{\mathbf{u},p\},\{\mathbf{v},q\}\right) = \left(\mathbf{r}^{n+1},\mathbf{v}\right) + \left(\mathbf{h}^{n+1},\mathbf{v}\right)_{\Gamma_\mathrm{N},n+1} \\ \forall \, (\mathbf{v},q) \in (\mathbf{V}_{0,n+1}^h, P_{n+1}^h) \,, \end{aligned} \tag{12}$$

where the discrete operator $B_{gal}\left(\{\mathbf{u},p\},\{\mathbf{v},q\}\right)$ is given by

$$\begin{aligned} B_{gal}\left(\{\mathbf{u},p\},\{\mathbf{v},q\}\right) = (\mathbf{u},\mathbf{v}) + \left(\delta \, \left(\mathbf{u} - \mathbf{u}^{G,n+1}\right) \cdot \boldsymbol{\nabla}\mathbf{u}, \mathbf{v}\right) \\ + \left(\delta \, 2\nu \boldsymbol{\varepsilon}(\mathbf{u}), \boldsymbol{\varepsilon}(\mathbf{v})\right) - \left(\delta \, p, \boldsymbol{\nabla} \cdot \mathbf{v}\right) - \left(\delta \, \beta \, q, \boldsymbol{\nabla} \cdot \mathbf{u}\right) \,. \end{aligned}$$

Here $(\cdot,\cdot)$ denotes the L^2 inner product on the actual mesh configuration Ω_{n+1} if not indicated otherwise and $\mathbf{h}^{n+1}$ represents Neumann boundary forces at the time instant $t = t^{n+1}$. The parameter $\beta \in \{-1, 1\}$ carries the sign of the pressure test function q.

Stabilization

Residual based stabilization is used to circumvent the restrictions set by the classical inf-sup condition as well as numerical oscillations induced by a dominating convective term. The stabilized form is given by: Find $\mathbf{u} \in \mathbf{V}^h_{n+1}$, $p \in P^h_{n+1}$ such that

$$\begin{aligned} B_{\text{gal}}\left(\{\mathbf{u},p\},\{\mathbf{v},q\}\right) - \sum_k \tau_{Mk}\left(\mathbf{R}_M(\mathbf{u},p), \mathbf{L}^{\text{stab}}_M(\mathbf{u},\{\mathbf{v},q\})\right)_k \\ + \sum_k \tau_{Ck}(R_C(\mathbf{u}), L_C(\mathbf{v}))_k \qquad (13) \\ = \left(\mathbf{r}^{n+1},\mathbf{v}\right) + \left(\mathbf{h}^{n+1}_F,\mathbf{v}\right)_{\Gamma_{\text{N}},n+1} \quad \forall\,(\mathbf{v},q) \in (\mathbf{V}^h_{0,n+1}, P^h_{n+1})\,, \end{aligned}$$

where k counts all elements of the triangulation. $\mathbf{R}_M$ and R_C denote the residuals of the semi-discrete momentum and mass balance equation, respectively.

The general form of the stabilization operator $\mathbf{L}^{\text{stab}}_M(\mathbf{u},\{\mathbf{v},q\})$ is given by

$$\mathbf{L}^{\text{stab}}_M(\mathbf{u},\{\mathbf{v},q\}) = \eta\,\mathbf{v} + \delta\left[-\left(\mathbf{u}-\mathbf{u}^G\right)\cdot\boldsymbol{\nabla}\mathbf{v} - \alpha\,2\nu\boldsymbol{\nabla}\cdot\boldsymbol{\varepsilon}(\mathbf{v}) + \beta\boldsymbol{\nabla}q\right]\,,$$

where the additional parameters $\eta \in \{0,1\}$ and $\alpha \in \{-1,0,1\}$ allow to identify different stabilization methods. Setting $\alpha = -1$, $\eta = 0$ and $\beta = -1$ recovers the Galerkin Least Squares (GLS) stabilization of the stationary Navier-Stokes equations which is also used in the instationary case [37]. Neglecting the viscous term within the stabilization operator, i.e. $\alpha = 0$ (and $\eta = 0$) yields the Streamline-Upwind Petrov Galerkin/Pressure Stabilized Petrov Galerkin (SUPG/PSPG) approach which in the context of the Navier-Stokes equations goes back to Tezduyar et al. [34]. Using the parameters $\eta = \alpha = 1$ and $\beta = -1$ yields the so-called Unusual Stabilized Finite Element Method (USFEM) [15] which employs the adjoint of the original differential operator for stabilization purpose. Unusual stabilization terms can be derived by enriching the spaces $\mathbf{V}^h_t$ and $\mathbf{V}^h_{0,t}$ with bubble type functions and static condensation of the additional degrees of freedom. This virtual bubbles approach [1] can thus be regarded a first step towards a multiscale method. A three-level method for Large Eddy Simulation (LES) of turbulent flows has been developed in [18, 19, 20]. A zeroth order term within the stabilization operator (i.e. $\eta = 1$) ensures the advection stabilization properties of the method in the context of very small time steps but must not be used in conjunction with linear elements [12].

Irrespective of the particular configuration of the stabilization operator, residual based stabilization is consistent in the sense that a sufficiently smooth analytical solution ($\mathbf{u}$, p) satisfies the stabilized form (13) exactly.

A robust definition of the stabilization parameters is employed with

$$\tau_{Mk} = \frac{h_k^2}{h_k^2\,\xi(Pe_k) + \frac{4\,\nu\delta}{m_k}\xi(Re_k)} \quad \text{and} \quad \tau_{Ck} = \frac{\delta h_k|\mathbf{u}|}{2}\xi(Re_k)$$

with $\xi(x) = \max(1, x)$, where h_k denotes a characteristic element diameter and the dominating flow effect is estimated by $Pe_k = 4\,\delta\,\nu/\left(m_k\,h_k^2\right)$ and $Re_k = m_k|\mathbf{u}|h_k/\left(2\,\nu\right)$. The parameter is given by $m_k = \min(1/3, C_k)$, where C_k is the largest constant satisfying the inverse estimate

$$C_k h_k^2 ||\Delta \mathbf{v}||_k^2 \leq ||\nabla \mathbf{v}||_k^2 \quad \forall \mathbf{v} \in \mathbf{V}_t^h \ ,$$

which in the ALE case is time dependent. Stabilized formulations of the unusual type ($\alpha = 1$) require the correct m_k to yield a stable method. The stabilized form (13) allows the use of equal order interpolations for velocity and pressure and applies to higher order elements.

The discrete equations (13) are consistently linearized and solved iteratively by means of a Newton-Raphson procedure.

3.3 Computational Mesh Dynamics

It is due to the mesh motion to provide elements with positive Jacobians even at moderate deformations of the fluid domain and to keep the mesh quality as high as possible. To meet the demands in terms of generality of the application, pseudo structural mesh motion schemes are applied where the mesh domain Ω_{G} is treated as a continuum. This yields the linear system of equations

$$\mathbf{K}^G \mathbf{r} = \mathbf{F} \ ,$$

where the 'mesh stiffness' matrix $\mathbf{K}^G$ is assembled from elemental entities and $\mathbf{r}$ denotes the mesh node displacements which also determine the mesh velocity $\mathbf{u}^G$. The right-hand side forces $\mathbf{F}$ stem from Dirichlet boundary conditions on the mesh positions only. The stiffness is designed to maximize the robustness of the mesh motion.

4 Coupling Approaches

The algorithmic implementation of the coupled three-field problem of fluid-structure and mesh interaction is based on a non-overlapping partitioned solution approach. The wet structural surface is the natural interface or coupling surface Γ. Kinematic and dynamic continuity at Γ ensure the conservation of mass, impulse and energy at the interface.

4.1 Weakly Coupled Partitioned Approaches and Artificial Added Mass Effect

A particularly appealing way to solve the coupled problem is a sequentially staggered formulation which demands only one solution per field and time

step. While the fluid and structural domain are solved implicitly the coupling here introduces an explicit feature.

The time step from time level n to $n+1$ proceeds as follows where here the mesh movement is omitted for brevity and a variety of schemes differs by the specific accuracy of the particular steps [14].

i. Calculate an explicit predictor of the structural interface displacement at the new time level $\mathbf{d}_{\Gamma,P}^{n+1}$.
ii. Get fluid velocity at Γ to serve as Dirichlet boundary condition $\mathbf{u}_{\Gamma,P}^{n+1}\left(\mathbf{d}_{\Gamma,P}^{n+1}\right)$.
iii. Solve fluid equations by Newton iterations to obtain $\mathbf{u}^{n+1}$ and $\mathbf{p}^{n+1}$.
iv. Obtain fluid boundary traction along Γ $\mathbf{f}_{\Gamma}^{F,n+1}$.
v. Solve the structural field by Newton iterations for the new displacements $\mathbf{d}^{n+1}$ under consideration of the fluid load $\mathbf{f}_{\Gamma}^{F,n+1}$.
vi. Proceed to next time step.

While being very promising in the sense of efficiency sequentially staggered algorithms exhibit an inherent instability which increases with decreasing time step [26]. As the fluid forces have the effect of an extra mass on the structural interface degrees of freedom it has been termed *artificial added mass effect* [37]. An analysis shows that the instability is caused by too large eigenvalues of the amplification operator of the explicit step [6, 14].

Summarizing the above steps ii. to iv. allows to identify the dimensionless added mass operator $\boldsymbol{\mathcal{M}}_{\mathcal{A}}$. This operator directly transfers the predicted nodal accelerations $\dot{\mathbf{u}}_{\Gamma}$ at the interface Γ into the fluid forces $\mathbf{f}_{\Gamma}$ exerted on the structure by

$$\mathbf{f}_{\Gamma} = m^{F}\boldsymbol{\mathcal{M}}_{\mathcal{A}}\dot{\mathbf{u}}_{\Gamma} \ ,$$

where m^F denotes a characteristic fluid mass. Introducing this into the discrete linear and undamped structural equations and neglecting structural forces yields

$$\begin{bmatrix} \mathbf{M}_{II}^{S} & \mathbf{M}_{I\Gamma}^{S} \\ \mathbf{M}_{\Gamma I}^{S} & \mathbf{M}_{\Gamma\Gamma}^{S} \end{bmatrix} \begin{bmatrix} \ddot{\mathbf{d}}_{I} \\ \ddot{\mathbf{d}}_{\Gamma} \end{bmatrix} + \begin{bmatrix} \mathbf{K}_{II}^{S} & \mathbf{K}_{I\Gamma}^{S} \\ \mathbf{K}_{\Gamma I}^{S} & \mathbf{K}_{\Gamma\Gamma}^{S} \end{bmatrix} \begin{bmatrix} \mathbf{d}_{I} \\ \mathbf{d}_{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ -m^{F}\boldsymbol{\mathcal{M}}_{\mathcal{A}}\dot{\mathbf{u}}_{\Gamma} \end{bmatrix}, \qquad (14)$$

where the structural system of equations has been split into internal (subscript I) and interface (subscript Γ) degrees of freedom. Inserting the particular representation of the predictor and Dirichlet boundary condition allows to analyze the eigenvalue of the operator that transfers the interface displacements from time level n to $n+1$. Here two different cases can be distinguished where $\mathbf{d}_{\Gamma}^{n+1}$ is either a function of a limited number of old interface positions (for example $\mathbf{d}_{\Gamma}^{n+1} = f(\mathbf{d}_{\Gamma}^{n}, \mathbf{d}_{\Gamma}^{n-1}, \mathbf{d}_{\Gamma}^{n-2})$) or it depends upon all previously calculated interface positions $\mathbf{d}_{\Gamma}^{n+1} = f(\mathbf{d}_{\Gamma}^{n}, \mathbf{d}_{\Gamma}^{n-1}, \ldots, \mathbf{d}_{\Gamma}^{0})$. As shown in [14] this depends on the time discretization of the two fields as well as on the specific predictor and the way to obtain the Dirichlet boundary condition on the fluid field at Γ. In the first case an 'instability condition' of the form

$$\frac{m^F}{m^S} \max \mu_i > C_1 \tag{15}$$

can be obtained, where m^S represents a characteristic structural mass and μ_i denotes the ith eigenvalue of the added mass operator $\boldsymbol{\mathcal{M}}_{\boldsymbol{\mathcal{A}}}$. If (15) is satisfied the system is unstable. The limit C_1 depends upon the particular details of the temporal discretization and decreases with increasing accuracy [14]. The instability limits obtained for different structural predictors and backward Euler (BE) or second order backward differentiation (BDF2) time discretization of the fluid equations are given in Table 1.

Table 1. Instability limit C_1 for sequentially staggered fluid-structure interaction schemes depending on the structural predictor and fluid time discretization scheme

predictor	BE	BDF2
0th order	3	$\frac{3}{2}$
1st order	$\frac{3}{5}$	$\frac{3}{10}$
2nd order	$\frac{1}{3}$	$\frac{1}{6}$

The problem gets even worse in the second case where the interface displacements are a function of all previous positions. This situation yields an instability condition of the form

$$\frac{m^F}{m^S} \max \mu_i > \frac{C_2}{n}$$

where n is the number of the current time step. Thus the instability limit C_2/n decreases during the simulation and regardless of the density ratio a step will be reached at which the problem becomes unstable.

When stabilized fluid elements are considered the analysis gets more complicated and the simple instability limits given in Table 1 are not directly applicable any more. It can however be proven that for every sequentially staggered scheme a density ratio ρ^F/ρ^S exists at which the scheme becomes unstable [14]. Numerical investigations show that the instability limits are very restrictive when incompressible fluids are considered and stabilized finite elements are employed effectively preventing stable computations by means of sequential staggered algorithms.

4.2 Strongly Coupled Partitioned Approach

To avoid the instabilities inherent in sequential staggered schemes an iterative staggered Dirichlet-Neumann substructuring scheme is formulated which is based on the approach given in [25]. Every time step from time level n to

$n+1$ requires an iteration over the fields until convergence is obtained. The algorithm can be summarized as follows.

i. Compute an explicit predictor of the structural interface displacement at the new time level $\mathbf{d}_{\Gamma,0}^{n+1}$.
ii. *Fluid mesh:*
Solve fluid mesh domain for new nodal positions.
iii. *Fluid:*
Compute fluid velocity at Γ from mesh position to serve as Dirichlet boundary condition $\mathbf{u}_{\Gamma,i+1}^{n+1}\left(\mathbf{d}_{\Gamma,i+1}^{n+1}\right)$.
iv. Solve fluid equations on the new mesh to obtain $\mathbf{u}_{i+1}^{n+1}$ and $\mathbf{p}_{i+1}^{n+1}$.
v. Obtain fluid boundary traction along Γ $\mathbf{f}_{\Gamma,i+1}^{F,n+1}$.
vi. *Structure:*
Solve the structural field for the new displacements $\mathbf{d}_{i+1}^{n+1}$ including the interface displacements $\tilde{\mathbf{d}}_{\Gamma,i+1}^{n+1}$ under consideration of the fluid load $\mathbf{f}_{\Gamma,i+1}^{F,n+1}$.
vii. *Relaxation of the interface displacements* $\tilde{\mathbf{d}}_{\Gamma,i+1}^{n+1}$
Perform relaxation by $\mathbf{d}_{\Gamma,i+1}^{n+1} = \omega_i \tilde{\mathbf{d}}_{\Gamma,i+1}^{n+1} + (1-\omega_i)\mathbf{d}_{\Gamma,i}^{n+1}$.
viii. *Check convergence*
Restart iteration with new mesh calculation if not yet converged.

The approach can be formulated with overall second order accuracy in time by using second order accurate schemes to advance the structural and the fluid field in time and interpolating the mesh velocity highly enough [13], e.g.

$$\mathbf{u}_{i+1}^{G,n+1} = \frac{3\,\mathbf{r}_{i+1}^{n+1} - 4\,\mathbf{r}^{n} + \mathbf{r}^{n-1}}{2\,\Delta t} \,.$$

Conservation at the Interface

Geometric conservation demands that the Dirichlet boundary condition for the fluid domain at Γ is obtained consistently with the time discretization scheme employed on the fluid domain. Along with one-step-θ or BDF2 the interface velocity is obtained from

$$\mathbf{u}_{\Gamma,i+1}^{n+1} = 2\,\frac{\mathbf{r}_{\Gamma,i+1}^{n+1} - \mathbf{r}_{\Gamma}^{n}}{\Delta t} - \mathbf{u}_{\Gamma}^{n} \tag{16}$$

ensuring a correct mass balance within the incompressible fluid flow.

Particular care has also to be taken determining the fluid forces applied to the structure. The fluid force along the entire coupling interface Γ_{D} can be computed by

$$\mathbf{f}_{\sigma} = \int \left(2\,\mu\boldsymbol{\varepsilon}(\mathbf{u}) - \overline{p}\mathbf{I}\right)\cdot\boldsymbol{n}\,\mathrm{d}\Gamma_{\mathrm{D}} \,. \tag{17}$$

A discrete version of (17) is often used to obtain lift and drag or FSI coupling forces (e.g. [11]).

A more accurate and also more efficient way to obtain the coupling forces in a discretized setting is to employ consistent nodal forces [21]. Sharing the order of accuracy of the primary variables and naturally containing the viscous forces, consistent nodal fluid forces are a key ingredient for accurate coupling on moderately fine meshes. Consistent with the discretization used here, the nodal force component in direction i at the end of the time step $n+1$ on node j is given by

$$\begin{aligned} f_{\text{cons},i}^{n+1,j} = & \left(-\frac{\rho}{\delta}\mathbf{u}^{n+1} - \rho\left(\mathbf{u}^{n+1} - \mathbf{u}^{G,n+1}\right)\cdot\boldsymbol{\nabla}\mathbf{u}^{n+1} + \frac{\rho}{\delta}\mathbf{r}^{n+1}, \varphi_j\mathbf{e}_i \right) \\ & - \left(2\mu\,\boldsymbol{\varepsilon}\left(\mathbf{u}^{n+1}\right), \boldsymbol{\varepsilon}\left(\varphi_j\mathbf{e}_i\right)\right) + \left(\overline{p}^{n+1}, \boldsymbol{\nabla}\cdot\varphi_j\mathbf{e}_i\right) \\ & + \sum_k \tau_{Mk}\,\rho\left(\boldsymbol{\mathcal{R}}_M\left(\mathbf{u}^{n+1}, p\right), \left(\mathbf{u}^{n+1} - \mathbf{u}^{G,n+1}\right)\cdot\boldsymbol{\nabla}\varphi_j\mathbf{e}_i\right. \\ & + 2\,\mu\boldsymbol{\nabla}\cdot\left(\boldsymbol{\varepsilon}\left(\varphi_j\mathbf{e}_i\right)\right)_k \\ & - \sum_k \tau_{Ck}\,\rho\left(\mathcal{R}_C(\mathbf{u}^{n+1}), \boldsymbol{\nabla}\cdot\varphi_j\mathbf{e}_i\right)_k \ . \end{aligned} \tag{18}$$

Consistent nodal forces fit into a nodal based data structure and are thus easy and efficient to implement. They further allow to incorporate the effect of stabilization terms preserving consistency in the stabilized case.

The instability observed at the sequentially staggered scheme transfers to the iteratively staggered algorithm demanding an upper bound for the relaxation parameter ω_i to ensure convergence [6]. Two stable and robust ways of obtaining the relaxation parameter have been developed.

Iterative Substructuring Scheme Accelerated via Gradient Method

The first technique is an acceleration to the iterative substructuring scheme analog to the gradient method. This method also guarantees convergence. In every iteration a relaxation parameter ω_i is computed by

$$\omega_i = \frac{\mathbf{g}_i^T\mathbf{g}_i}{\mathbf{g}_i^T\mathbf{S}^{S^{-1}}\left(\mathbf{S}^F + \mathbf{S}^S\right)\mathbf{g}_i} = \frac{\mathbf{g}_i^T\mathbf{g}_i}{\mathbf{g}_i^T\left(\mathbf{S}^{S^{-1}}\mathbf{S}^F\mathbf{g}_i + \mathbf{g}_i\right)} \tag{19}$$

which is locally optimal with respect to the actual search direction, i.e. the residual

$$\mathbf{g}_i = \mathbf{S}^{S^{-1}}\left(\mathbf{f}_{\Gamma\,\text{ext}}^{\text{mod},n+1} - \left(\mathbf{S}^F + \mathbf{S}^S\right)\mathbf{d}_{\Gamma,i}^{n+1}\right) = \tilde{\mathbf{d}}_{\Gamma,i+1}^{n+1} - \mathbf{d}_{\Gamma,i}^{n+1} \ .$$

A procedure for evaluating equation (19) without explicitly computing the Schur complements $\mathbf{S}^F$ and $\mathbf{S}^S$ has been proposed in Wall et al. [41].

Iterative Substructuring Scheme Accelerated via the Aitken Method

A second technique for explicitly calculating a suitable relaxation parameter is the application of Aitken's acceleration scheme for vector sequences according

to Irons et al. [24]. To obtain ω_i the interfacial displacement difference is computed

$$\Delta \mathbf{d}_{\Gamma,i+1}^{n+1} := \mathbf{d}_{\Gamma,i}^{n+1} - \tilde{\mathbf{d}}_{\Gamma,i+1}^{n+1} \, .$$

The Aitken factor is obtained from

$$\mu_i^{n+1} = \mu_{i-1}^{n+1} + \left(\mu_{i-1}^{n+1} - 1\right) \frac{\left(\Delta \mathbf{d}_{\Gamma,i}^{n+1} - \Delta \mathbf{d}_{\Gamma,i+1}^{n+1}\right)^T \Delta \mathbf{d}_{\Gamma,i+1}^{n+1}}{\left(\Delta \mathbf{d}_{\Gamma,i}^{n+1} - \Delta \mathbf{d}_{\Gamma,i+1}^{n+1}\right)^2}$$

and yields the relaxation parameter

$$\omega_i = 1 - \mu_i^{n+1} \, .$$

Even though a rigorous analysis of its convergence properties does not exist, numerical studies have shown that the Aitken acceleration for vector sequences applied to the fluid-structure interaction problems shows a performance which is sometimes even better than acceleration via the gradient method. Furthermore the Aitken method yields an extremely cheap approach in terms of both CPU and memory and is simple to implement.

5 Free Surface

5.1 Free Surface – Dynamic Boundary Condition

The continuity of forces on the free surface Γ_{FS} implies the standard dynamic boundary condition (7). In general, the interfacial behavior is very complex and depends on the physiochemical properties of the fluids and the molecular structure of the interface. In a common simplification, the prescribed surface stress is assumed to consist of two parts, namely a normal stress component originating from the atmospheric pressure and the surface tension on a curved surface and the tangential stress component in the case of a surface tension gradient occurring along the surface:

$$\boldsymbol{\sigma} \cdot \mathbf{n} = -\left(p_a + \gamma \, \nabla_s \cdot \mathbf{n}\right) \mathbf{n} - \nabla_s \gamma \, . \tag{20}$$

where $\nabla_s \cdot \mathbf{n}$ denotes the curvature κ of the surface and ∇_s the surface gradient operator (Slikkerveer et al. [32]).

In this work it is assumed that the tangential stress component and the atmospheric pressure vanishes, so it is possible to simplify the continuity of forces on the free surface to one term depending on the local curvature of the free surface and a single material parameter γ, the isotropic and isothermal surface tension coefficient:

$$\boldsymbol{\sigma} \cdot \mathbf{n} = \mathbf{h} = \gamma \kappa \mathbf{n} \quad \text{on} \quad \Gamma_{FS} \tag{21}$$

5.2 Free Surface – Kinematic Boundary Condition

In the particular case of free surfaces and moving no-flux interfaces, the only requirement for the velocity of the domain boundary, i.e. also the mesh boundary, is

$$\mathbf{u} \cdot \mathbf{n} = \mathbf{u}^G \cdot \mathbf{n} \quad \text{on} \quad \Gamma_{FS} \tag{22}$$

where $\mathbf{u}^G = \partial \mathbf{r} / \partial t$ denotes the mesh velocity. As (22) is a scalar condition further assumptions have to be made to close the problem.

Local Lagrange Approach

In a local Lagrangian framework the mesh is assumed to be attached to the fluid particles on the free surface and equation (22) reduces to

$$\mathbf{u} - \mathbf{u}^G = \mathbf{0} \, . \tag{23}$$

Grid velocity and position are related in the same way as on the FSI coupling interface, i.e. according to (16).

The approach is simple and generally applicable. However it suffers robustness as great mesh distortions corresponding to the tangential surface velocities quickly occur demanding for undesired re-meshing.

Height Function Approach

To overcome the disadvantages of the local Lagrangian approach the description of the free surface via a height function is a common alternative. In this case it is assumed that the grid at the free surface moves along a prescribed direction only, i.e. $\mathbf{r} = \phi \mathbf{e}$. In the case of vertical nodal displacement ($r_1 = r_2 = 0$) the base vector $\mathbf{e}$ is chosen as the unit vector in vertical direction, and then $\phi(x_1, x_2, t)$ is a regular scalar valued function representing the position of the free surface. For this specific case the normal vector in (22) can be expressed by

$$\mathbf{n} = \frac{1}{||\tilde{\mathbf{n}}||} \tilde{\mathbf{n}} \quad \text{with} \quad \tilde{\mathbf{n}} = \begin{bmatrix} \dfrac{\partial \phi}{\partial x_1} & \dfrac{\partial \phi}{\partial x_2} & 1 \end{bmatrix}^T \tag{24}$$

and finally the free surface position can be determined by solving

$$\frac{\partial \phi}{\partial t} = -u_1 \frac{\partial \phi}{\partial x_1} - u_2 \frac{\partial \phi}{\partial x_2} + u_3 \quad \text{on} \quad S \tag{25}$$

where the domain S is the lower-dimensional region over which ϕ is determined, typically obtained from Ω_{F} by projection.

5.3 Partitioned Implicit Free Surface Approach

In [17] and [39] we discussed the disadvantages of a pure explicit or implicit free surface approach in combination with partitioned strongly coupled FSI solution algorithms. Thus a *partitioned* implicit algorithm has been developed for free surface flows. For this purpose the entire integration domain Ω_{F} is divided (or *partitioned*) into the internal domain Ω_{int} and a boundary domain close to the free surface Ω_{FS}, as displayed in figure 5:

$$\Omega_{\mathrm{F}} = \Omega_{\mathrm{int}} \cup \Omega_{\mathrm{FS}} \; . \tag{26}$$

The basic idea then is to include only the free surface position $\mathbf{r}$ or the mesh velocities $\mathbf{u}^G$ on the boundary as additional unknowns in the fluid solver and solve them in a monolithic way. By this we inherit all the benefits from a fully implicit approach but only add very little costs to the overall solution procedures. In simple cases the domain close to the free surface Ω_{FS} is chosen to one layer of elements adjacent to Γ_{FS}. When large deformations have to be considered the size of Ω_{FS} can be determined by an estimation based on the element size and the expected maximum grid velocity at the boundary Γ_{FS}, and the time increment Δt. Then a structured boundary mesh is used with simple constraints reaching from Γ_{FS} to the interface between Ω_{FS} and Ω_{int}. Thus also in this case the number of unknowns and costs remain small as before.

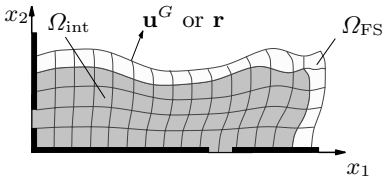


Fig. 5. Decomposition of fluid domain

This approach is independent of the free surface formulation. It can be combined with a local Lagrangian approach, with a height function approach or with the generalized elevation equation.

5.4 Generalized Free Surface Description

In many cases of interest, the height function equation (25) cannot be applied, since the prescribed direction of the mesh at the free surface is not vertical (see e.g. [3]). This is especially crucial when the walls are deformable themselves. More general ways to close the set of equations at the free surface have to be pursued since. The key point is that the free surface nodes should be able to distribute more or less freely at the free surface without following any simple

and a priori prescribed direction. One way to achieve this is the line-tracked interface update technique, which was introduced in [33]. It calculates the motion of each surface node along a locally-selected direction (typically in the surface-normal direction).

The key idea here is to use a dimensionally reduced pseudo structural approach. A descriptive interpretation of this would be to attach a shell (or rather a membrane) to the free surface in order to govern the tangential motion (i.e. two components in 3D) of the surface grid nodes.

An elegant and straightforward way to realize this would be to combine the general elevation equation (22) with a membrane equation on the free surface. Another rather simple approach is to pose the solution in a way that is directly based on the domain decomposition introduced for the partitioned implicit free surface approach. Only on the already defined domain Ω_{FS} mesh displacements $\mathbf{r}$ are introduced as additional unknowns. These mesh displacements are assumed to be governed by a quasi-static momentum equation $\nabla \cdot \boldsymbol{\sigma} = 0$ where $\boldsymbol{\sigma}$ is the Cauchy stress tensor. This idea has already been applied for determining the mesh displacements on Ω_G. However this newly introduced pseudo structure only exists within the boundary area at the free surface and its displacements are determined together with the fluid unknowns within the nonlinear iteration scheme. We then apply (22) as a boundary condition acting on the equations for the mesh displacements of Ω_{FS}.

The boundary value problem for the unknowns $\mathbf{r}^k$ at node k can be expressed by the weighted residual of the quasi-static momentum equation:

$$R_i^k = \int_{\Omega_{\mathrm{FS}}} \nabla \left(\varphi_k \mathbf{e}_i\right) : \boldsymbol{\sigma} \, \mathrm{d}\Omega_{\mathrm{FS}} = 0 \, , \tag{27}$$

where φ_k denotes the basis function associated with node k as defined in (11). At every node k (27) yields a scalar condition per spatial direction $\mathbf{e}_i$. Since one does not want to prescribe the distribution of the nodes at the free surface, but only the position of this free boundary, the kinematic boundary condition is introduced as a so-called distinguishing condition. The effectiveness of this procedure has been demonstrated in [30] for different two-dimensional applications. For each node k on the free surface boundary Γ_{FS} one of the three mesh equations defined by (27) is replaced by the weak form of the kinematic condition

$$R^k = \int_{\Gamma_{FS}} \varphi_k \left(\mathbf{u} - \mathbf{u}^G\right) \cdot \mathbf{n} \, \mathrm{d}\Gamma_{FS} = 0 \, . \tag{28}$$

After discretizing (28) in time, e.g. using a one-step-θ scheme

$$\mathbf{u}^{G,n+1} = \frac{\mathbf{r}^{n+1} - \mathbf{r}^n}{\theta \Delta t} - \frac{1-\theta}{\theta} \mathbf{u}^{G,n} \tag{29}$$

it is possible to solve for the unknown free surface displacements $\mathbf{r}$.

To replace the correct equation in a general case, the components of the governing equations are locally transfered into the directions normal and tangential to the free surface. Then the distinguishing condition is chosen to constrain the normal motion of the free surface. The remaining components of (27) allow a shear-free redistribution of the nodes in tangential direction. Cairncross et al. [5] successfully applied this method to three-dimensional stationary fluid problems.

Finally a short remark on the computation of the boundary normals is given. For a standard finite element implementation using 2D or 3D Lagrangian elements the normal at a boundary node is not defined uniquely. Possible choices are the uniquely defined normals on the element edge away from the nodes or an averaged normal from purely geometric considerations. However, numerical experiments have shown best results with respect to mass conservation when the mass-consistent normal described in [21] is used. In this case the unit normal at node k is derived from the discrete weak continuity equation for incompressible flow. After some algebra the following expression for the normal component in j-direction at node k is obtained

$$n_j^k = \frac{1}{||\tilde{\mathbf{n}}^k||}\tilde{n}_j^k \quad , \text{with} \quad \tilde{n}_j^k = \int\limits_{\Omega_\mathrm{F}} \frac{\partial \varphi_k}{\partial x_j}\, \mathrm{d}\Omega_\mathrm{F} \; . \tag{30}$$

5.5 Embedding the Partitioned Implicit Free Surface Approach into the Fluid-Structure Interaction Solver

Embedding a free surface description into the iterative Dirichlet-Neumann substructuring scheme for fluid-structure interaction, described in Sec. 4.2, represents an additional challenge. Inclusion of a simple explicit free surface approach is not advisable here, because its explicit fraction would cause a stability limit for the implicit overall approach. Increasing numerical costs due to a substantial increase of calls to CMD and the possible deterioration of convergence of the entire problem are the main drawbacks of monolithic and iterative staggered free surface approaches in conjunction with a strong coupling approach. Hence it is especially this class of problems where the partitioned implicit free surface approach shows its suitability and quality. A detailed overview of the algorithm can be found in [39].

6 Vectorization of Finite Element Integration

Besides the solution of the system of linear equations the element evaluation and assembly of stabilized fluid elements is often a main time consuming part of a finite element calculation. Thus, still a large amount of computing time can be saved by an expert implementation of the element routines.

6.1 Computational Efficiency

The partitioned fluid-structure interaction algorithm described has been implemented into the in-house research code `ccarat`. As the repeated solutions required to solve such problems are rather time consuming computational efficiency is of great interest. The NEC SX-6+ and SX-8 parallel vector computers are the main target architectures of this project. Vector optimization achieved by code modifications includes the reordering or fusion of loops to increase the vector length or the usage of temporary variables to break data dependencies in loops.

It is further shown that this kind of performance optimization also has a positive effect on the performance of the code on other architectures.

6.2 Sets of Elements

The main idea of this concept is to group computationally similar elements into sets and then perform all calculations necessary to set up the element matrices simultaneously for all elements in one set. *Computationally similar* in this context means that all elements in one set require exactly the same operations to integrate the element matrix, i.e. each set consists of elements with the same topology and the same number of nodes and integration points.

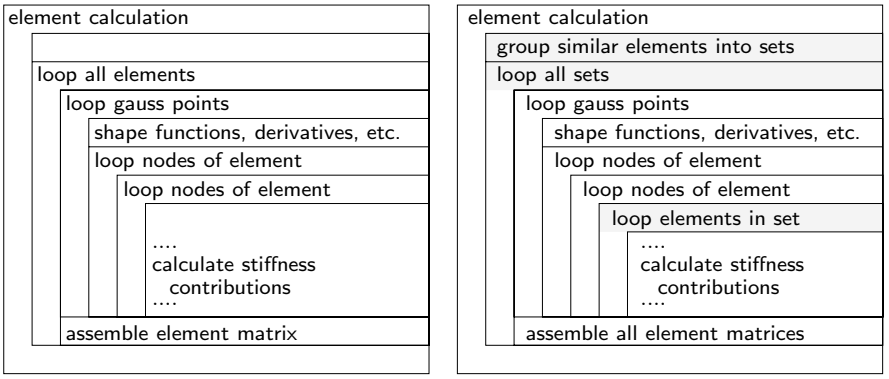


Fig. 6. Old (left) and new (right) structure of algorithm to evaluate element matrices

The changes necessary to implement this concept are visualized in the structure charts in Fig. 6. Instead of looping all elements and calculating the element matrix individually, now all sets of elements are processed. For every set the usual procedure to integrate the matrices is carried out, except on the lowest level, i.e. as the innermost loop, a new loop over all elements in the current set is introduced. This loop is especially suited for vector machines, as the calculations inside are quite simple and, most importantly, consecutive

steps do not depend on each other. In addition the length of this loop, i.e. the size of the element sets, can be chosen freely in order to fill the processor's vector pipes.

The only limitation for the size of the sets are additional memory requirements, as now intermediate results have to be stored for all elements in one set.

6.3 Programming Language and Array Management

It is well known that the programming language can have a large impact on the performance of a scientific code. Despite considerable effort on other languages [35, 36] Fortran is still considered the best choice for highly efficient codes [29] whereas some features of modern programming languages, like pointers in C or objects in C++, make vectorization more complicated or even impossible [28]. Especially the very general pointer concept in C makes it difficult for the compiler to identify data-parallel loops, as different pointers might alias each other. There are a few remedies for this problem like compiler flags or the *restrict* keyword. The latter is quite new in the C standard and it seems that it is not yet fully implemented in every compiler.

Table 2. Influences on performance, properties of five different variants and their relative time for calculation of stiffness contributions

	orig	var1	var2	var3	var4	var5
language	C	C	C	C	C	Fortran
array dimensions	multi	multi	multi	one	one	multi
restrict keyword			restrict		restrict	
SX-6+[1]	1.000	0.024	0.024	0.016	0.013	0.011
Itanium2[2]	1.000	1.495	1.236	0.742	0.207	0.105
Pentium4[3]	1.000	2.289	1.606	1.272	1.563	0.523

We have implemented the proposed concept for the calculation of the element matrices in five different variants. The first four of them are implemented in C, the last one in Fortran. Further differences are the array management and the use of the restrict keyword. For a detailed description of the variants see Table 2. Multi-dimensional arrays denote the use of 3- or 4-dimensional

[1] NEC SX-6+, 565 MHz; NEC C++/SX Compiler, Version 1.0 Rev. 063; NEC FORTRAN/SX Compiler, Version 2.0 Rev. 305.

[2] Hewlett Packard Itanium2, 1.3 GHz; HP aC++/ANSI C Compiler, Rev. C.05.50; HP F90 Compiler, v2.7.

[3] Intel Pentium4, 2.6 GHz; Intel C++ Compiler, Version 8.0; Intel Fortran Compiler, Version 8.0.

arrays to store intermediate results, whereas one-dimensional arrays imply a manual indexing.

The results in Table 2 give the CPU time spent for the calculation of some representative element matrix contributions standardized by the time used by the original code. The positive effect of the grouping of elements can be clearly seen for the vector processor. The calculation time is reduced to less than 3% for all variants. On the other two processors the grouping of elements does not result in a better performance for all cases. The Itanium architecture shows only an improved performance for one dimensional array management and the variant implemented in Fortran and the Pentium processor performs in general worse for the new structure of the code. Only for the last variant the calculation time is cut in half.

It can be clearly seen that the effect of the restrict keyword varies for the different compilers/processors and also for one-dimensional and multi-dimensional arrays. Using restrict on the SX-6+ results only in small improvements for one-dimensional arrays, on the Itanium architecture the speed-up for this array management is considerable though. In contrast to this on the Pentium architecture the restrict keyword has a positive effect on the performance of multi-dimensional arrays and a negative effect for one-dimensional ones.

The most important result of this analysis is the superior performance of Fortran. This is the reason why we favor Fortran for our performance critical scientific code and use the last variant for our further examples.

6.4 Results

The positive effect of the proposed concept is demonstrated for the calculation of element matrices on a full CFD simulation. The flow is the Beltrami-Flow (for details see [8]). The unit-cube was discretized by 32768 stabilized 8-noded hexahedral elements.

In Fig. 7 the total calculation time for 32 time steps of this example and the fractions for the element calculation and the solver on the SX-6+ are given for the original code and the full implementation of variant 5. The time spent for the element calculation, formerly the major part of the total time, could be reduced by a factor of 24.

This considerable improvement can also be seen in the sustained performance given in Table 3 as percentage of peak performance. The original code not written for any specific architecture has only a poor performance on the SX-6+ and a moderate one on the other platforms. For the complete element calculation the new code, designed for a vector processor, achieves an acceptable efficiency of around 30% and for several subroutines, like the calculation of some stiffness contributions, even a superior efficiency of above 70%. It has to be noted that these high performance values come along with a vector length of almost 256 and a vector operations ratio of above 99.5%.

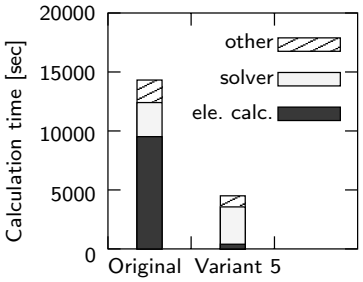


Fig. 7. Split-up of total calculation time for 32 time steps of the Beltrami Flow on the SX-6+

Table 3. Efficiency of original and new code in % of peak performance

	element calc.		stiffness contr.	
	original	var5	original	var5
SX-6+	0.95	29.55	0.83	71.07
Itanium2	8.68	35.01	6.59	59.71
Pentium4	12.52	20.16	10.31	23.98

But also for the Itanium2 and Pentium4 processors, which were not the main target architectures, the performance was improved significantly and for the Itanium2 the new code reaches around the same efficiency as on the vector architecture.

7 Numerical Simulations and Applications

Two- and three-dimensional simulations have been performed in order to highlight the capabilities of the described fluid-structure interaction formulation.

7.1 Two-Dimensional Bridge Cross Section

Wind belongs to the relevant load cases for an increasing number of slender wide spanned bridges. In order to show the effect of wind induced vibrations a H-shaped bridge cross section of about 12 m width has been subjected to horizontal flow. The cross section is similar to the one of the Tacoma Narrows bridge which collapsed due to wind induced vibrations in November 1940. The principal problem is sketched in Fig. 8 where u_∞ denotes the surrounding fluid velocity. The problem data has been taken from [23].

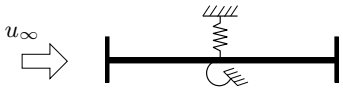


Fig. 8. Sketch of two-dimensional bridge cross section problem

The cross section supported by a vertical and torsional spring which represent the bending and torsional stiffness of the bridge building, respectively, starts to oscillate vertically as soon as the vortex shedding develops. As the cross section is shaped such that once a rotation occurs the overall wind

moment works in the same direction, the motion is dominated by torsional vibration soon afterwards which eventually yields failure. The displaced bridge profile at different time instants is depicted in Fig. 9.

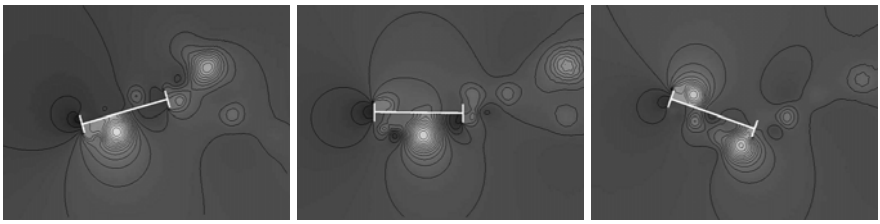


Fig. 9. Cross section at different time instants on the pressure field with isolines

7.2 Snap Through of a Gasket

The snap through of a gasket depicted in Fig. 10 is analyzed. The machine part consists of a rigid inner cylinder and external parts at which a thin flexible cap is attached at the bottom. The fluid entering the gasket at the top leaves at the circumferential outflow boundary.

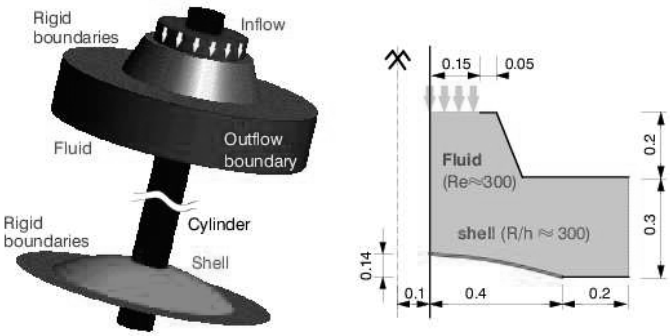


Fig. 10. Machine part with gasket: blown-up drawing and cross section

Three to five iterations of the iterative staggered Dirichlet-Neumann algorithm have been required to obtain a converged solution within a time step. Eventually the fluid pressure causes a snap through of the shell structure. While the full three-dimensional simulation allows to obtain an almost non-symmetric buckling mode an axisymmetric failure of the shell is obtained even if imperfections of the undeformed shell are considered. The snap-through

shown in Fig. 11 is highly influenced by the viscous damping and the inertia of the fluid preventing the non-symmetric failure one would expect otherwise.

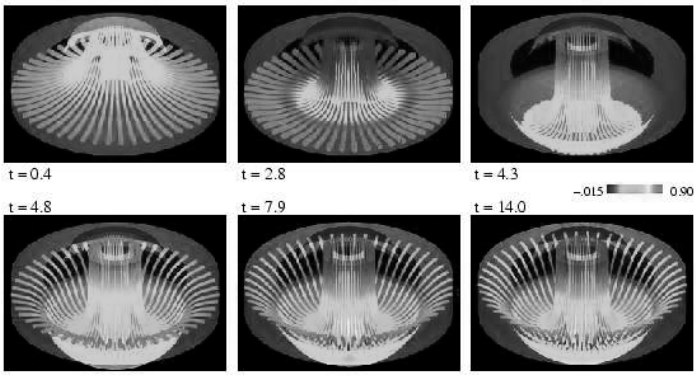


Fig. 11. Machine part with gasket: streamtubes and pressure

7.3 Vibrating U-Pipe

Coriolis flowmeters are an elegant way to measure the mass flow rate in a pipe. The measuring unit inside such a flowmeter is a flexible tube which classically is U-shaped, clamped at both ends and passed by the flow. The pipe is subject to a forced vibration at frequency f which induces opposite and time dependent Coriolis forces within the fluid in the inflow and outflow part of the tube. Thus the resulting vibration is not just the enforced bending but accompanied by an amount of torsion depending upon the frequency ratio f_t/f where f_t denotes the eigenfrequency of the torsional mode.

The problem description of a rubber pipe filled with water is given in Fig. 12 where gravity points in negative z-direction. The rubber is modeled by a compressible Neo-Hookean material. In Fig. 13 the displaced shell is depicted at different time instants where the view on the x-z-plane is shown. The overall motion consists of an oscillation in the first bending mode accompanied by the enforced oscillation at higher frequency which also induces torsional vibrations.

7.4 Flow in a Collapsible Tube

This example describes the problem of viscous flow in an elastic tube. Elastic tubes collapse, i.e. buckle non-axisymmetrically when the transmural pressure (internal minus external pressure) falls below a critical value. The tube's large

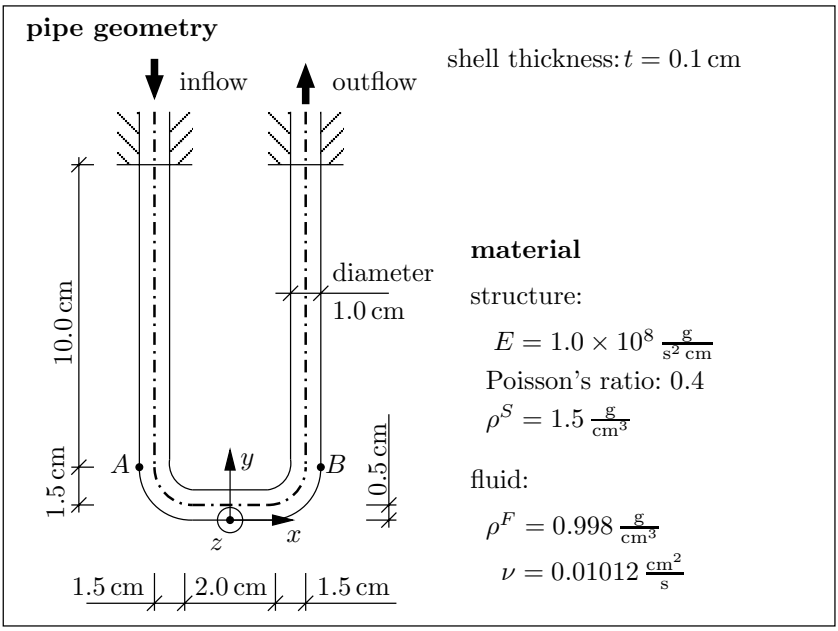


Fig. 12. U-pipe: Geometry and problem description

Fig. 13. U-pipe at different time instants exhibiting torsional oscillation

deformation during the buckling causes a strong interaction between fluid and solid. To illustrate the response of the tube we consider its deformation in a procedure in which the flow rate is prescribed by means of a volumetric pump attached to the upstream end of the tube, see Fig. 14.

The material of the tube has a Young's modulus of $E = 8.75 \times 10^8\,\text{N/cm}^2$, Poisson's ratio of $\nu^S = 0.3$ and a mass density of $\rho^S = 7.5\,\text{g/cm}^3$. The tube is passed by water with a density of $\rho^F = 1.0\,\text{g/cm}^3$, a kinematic viscosity of $\nu^F = 0.01\,\text{cm}^2/\text{s}$ and the prescribed inflow velocity is $\hat{u}_3 = 1.0\,\text{cm/s}$.

We keep the fluid pressure at the far downstream end of the tube constant, $p_{\text{down}} = 0.0$, and induce its collapse by increasing the chamber pressure. As the chamber pressure is increased, the transmural pressure decreases and first becomes negative (compressive) at the downstream end. When the compressive load exceeds a critical value, the axisymmetric deformation loses its stability and the tube buckles non-axisymmetrically. Figure 15 shows the wall deformation of the tube as the non-axisymmetric collapse progresses.

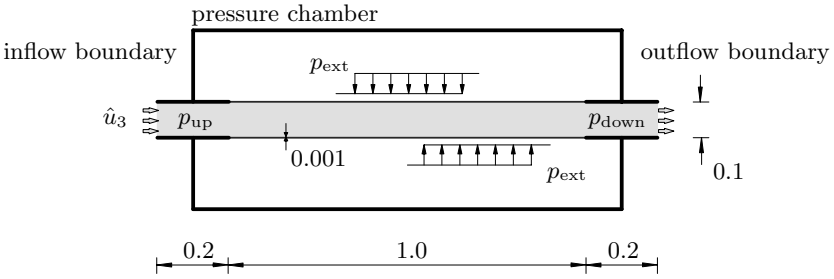


Fig. 14. Problem definition collapsing tube

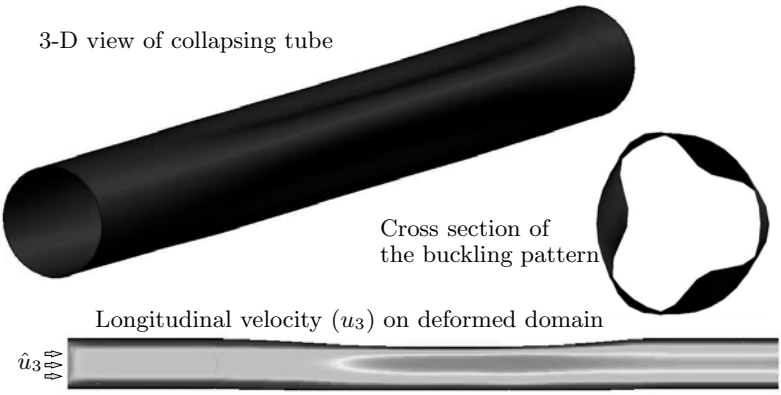


Fig. 15. Tube deformation

7.5 Basin with Collapsing Arch

The following example involves a two-dimensional basin as depicted in Fig. 16. It demonstrates the capability of the generalized free surface description in our FSI solver. The lateral walls and parts of the bottom are rigid. The center part of the bottom is closed by an elastic arch. The initial fluid depth is 20.0 cm. The basin is filled through two lateral channels with $\hat{u}_2 = 5.0\,\text{cm/s}$.

The elastic arch has a thickness of 0.1 cm. The structural material parameters were chosen to $\rho^S = 500\,\text{g/cm}^3$, $E = 9.0 \cdot 10^8\,\text{N/cm}^2$ and $\nu^S = 0.3$. The fluid density is $\rho^F = 1.0\,\text{g/cm}^3$ and the kinematic viscosity is $\nu^F = 9.0\,\text{cm}^2/\text{s}$. At both lateral inclined walls slip boundary conditions were applied. The fluid domain was discretized with 3200 Q1Q1 stabilized fluid elements. The time increment was chosen to $\Delta t = 0.025\,\text{s}$.

Due to the increasing fluid height the pressure on the elastic arch reaches a critical value and the structure collapses. As shown in Fig. 17, the present approach is able to reproduce this highly transient coupled buckling process exhibiting large structural and free surface deformations. Due to the damp-

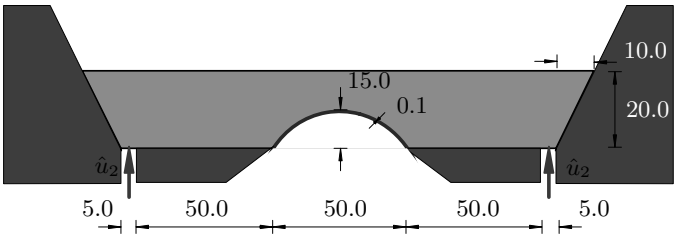


Fig. 16. Basin with elastic arch: initial configuration

ing of the fluid the system reaches a new steady equilibrium position after approximately $t = 72.0\,\mathrm{s}$.

8 Conclusions

The final period of the project focused on three different aspects. Firstly further insight into the theoretical basis for the fluid formulation and the interaction with thin-walled structures has been obtained. Clarifying the aspect of geometric conservation in the context of finite element formulations for fluid flow based on an ALE formulation allowed to develop a stable and second order accurate in time numerical scheme on deforming ALE domains. The accuracy of a stabilized finite element formulation with respect to the mesh distortion inherent in ALE schemes has been investigated showing that reliable and accurate stabilized elements can be designed [12, 40]. Further basic work regards the analysis of the instability of sequential staggered coupling schemes in the context of slender structures and incompressible flow. Special attention has been put on the influence of the stabilization incorporated into the fluid elements. It could however been shown that irrespective of the stabilization or the details of the formulation there is a critical ratio between the mass densities of fluid and structure at which every sequential staggered scheme becomes unstable due to the artificial added mass effect [14].

As a second focus free surface approaches have been improved and extended to three dimensions [39]. The free surface formulation has been incorporated into the partitioned three field problem while the convergence properties are preserved. Particular emphasis has been put on accurately satisfying mass conservation in the context of free surface flows.

For complex three-dimensional fluid structure interaction problems the computing time is still a limiting issue. Thus as a third focus a vectorization concept has been developed to optimize the code using a large portion of the actual peak performance. Significant improvements of the computing

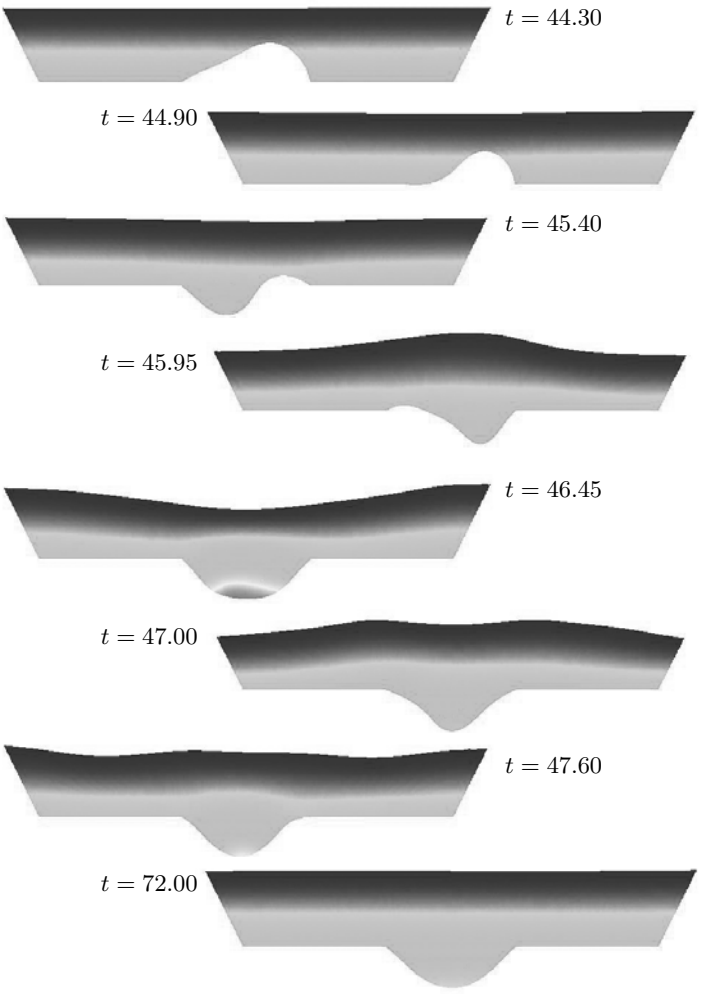


Fig. 17. Basin with elastic arch: Collapsing arch with computed pressure solution

times could be achieved allowing for moderately large long time FSI computations [27].

Acknowledgement

The authors thank Dr. M. Gee and Dr. D. P. Mok for their valuable contributions to the success of the project. The present study is supported by a grant of the foundation "Deutsche Forschungsgemeinschaft" (DFG) under

Project B4 of the Collaborative Research Center SFB 404 'Multifield Problems in Continuum Mechanics'. This support is gratefully acknowledged.

References

1. C. Baiocchi, F. Brezzi, and L. P. Franca. Virtual bubbles and Galerkin-least-squares type of methods (Ga.L.S.). *Comput. Methods Appl. Mech. Engrg.*, 105:125–141, 1993.
2. K.-J. Bathe and E. N. Dvorkin. A four-node plate bending element based on Mindlin/Reissner theory. *Int. J. Numer. Meth. Engrg.*, 21:367–383, 1985.
3. M. Behr and F. Abraham. Free-surface flow simulations in the presence of inclined walls. *Comput. Methods Appl. Mech. Engrg.*, 191:5467–5483, 2002.
4. M. Bischoff and E. Ramm. On the physical significance of higher order kinematic and static variables in a three-dimensional shell formulation. *Int. J. Solids and Structures*, 37:6933–6960, 2000.
5. R. A. Cairncross, P. R. Schunk, T. A. Baer, R. R. Rao, and P. A. Sackinger. A finite element method for free surface flows of incompressible fluids in three dimensions. Part I. Boundary fitted mesh motion. *Int. J. Numer. Meth. Fluids*, 125:375–403, 2000.
6. P. Causin, J.-F. Gerbeau, and F. Nobile. Added-mass effect in the design of partitioned algorithms for fluid-structure problems. *Comput. Methods Appl. Mech. Engrg.*, 194:4506–4527, 2005.
7. 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.
8. C. R. Ethier and D. A. Steinman. Exact Fully 3d Navier Stokes Solution for Benchmarking. *Int. J. Numer. Meth. Fluids*, 19:369–375, 1994.
9. C. Farhat and P. Geuzaine. Design and analysis of robust ALE time-integrators for the solution of unsteady flow problems on moving grids. *Comput. Methods Appl. Mech. Engrg.*, 193:4073–4095, 2004.
10. C. Farhat, P. Geuzaine, and C. Grandmont. The discrete geometric conservation law and the nonlinear stability of ALE schemes for the solution of flow problems on moving grids. *J. Comp. Physics*, 174:669–694, 2001.
11. C. Farhat, M. Lesoinne, and P. Le Tallec. Load and motion transfer algorithms for fluid/structure interaction problems with non-matching discrete interfaces: Momentum and energy conservation, optimal discretization and application to aeroelasticity. *Comput. Methods Appl. Mech. Engrg.*, 157:95–114, 1998.
12. C. Förster, W. A. Wall, and E. Ramm. On residual based stabilisation methods for transient problems at small time increments. *Preprint SFB 404*, 2006. Institute of Structural Mechanics, University of Stuttgart.
13. C. Förster, W. A. Wall, and E. Ramm. On the geometric conservation law in transient flow calculations on deforming domains. *Int. J. Numer. Meth. Fluids*, (50):1369–1379, 2006.
14. C. Förster, W. A. Wall, and E. Ramm. Artificial added mass instabilities in sequential staggered coupling of nonlinear structures and incompressible flow. *Comput. Methods Appl. Mech. Engrg.*, Submitted.
15. L. P. Franca and C. Farhat. Bubble functions prompt unusual stabilized finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 123:299–308, 1995.

16. M. Gee, W. A. Wall, and E. Ramm. Parallel multilevel solution of nonlinear shell structures. *Comput. Methods Appl. Mech. Engrg.*, 194:2513–2533, 2005.
17. S. Genkinger, W. A. Wall, and E. Ramm. A parallel computational approach for free surface flows with structural interactions. *Proc. Appl. Math. Mech.*, 3:276–277, 2003.
18. V. Gravemeier. *The variational multiscale method for laminar and turbulent incompressible flow.* PhD thesis, Institute of Structural Mechanics, University of Stuttgart, 2003.
19. V. Gravemeier, W. A. Wall, and E. Ramm. A three-level finite element method for the instationary incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 193:1323–1366, 2004.
20. V. Gravemeier, W. A. Wall, and E. Ramm. Large eddy simulation of turbulent incompressible flows by a three-level finite element method. *Int. J. Numer. Meth. Fluids*, 48:1067–1099, 2005.
21. P. M. Gresho and R. L. Sani. *Incompressible flow and the finite element method.* Wiley, New York, 1998.
22. H. Guillard and C. Farhat. On the significance of the geometric conservation law for flow computations on moving meshes. *Comput. Methods Appl. Mech. Engrg.*, 190:1467–1482, 2000.
23. B. Hübner. *Simultane Analyse von Bauwerk-Wind-Wechselwirkungen.* PhD thesis, Institut für Statik, Technische Universität Braunschweig, 2003.
24. B. Irons and R. C. Tuck. A version of the Aitken acceleratior for computer implementation. *Int. J. Numer. Meth. Engrg.*, 1:275–277, 1969.
25. P. Le Tallec and J. Mouro. Fluid structure interaction with large structural displacements. *Comput. Methods Appl. Mech. Engrg.*, 190:3039–3067, 2001.
26. D. P. Mok and W. A. Wall. Partitioned analysis schemes for the transient interaction of incompressible flows and nonlinear flexible structures. In W. A. Wall, K.-U. Bletzinger, and K. Schweizerhof, editors, *Trends in Computational Structural Mechanics*, pages 689–698. CIMNE, Barcelona, 2001.
27. M. Neumann, S. R. Tiyyagura, W. A. Wall, and E. Ramm. Robustness and efficiency aspects for computational fluid structure interaction. In E. Krause, Y. Shokin, M. Resch, and N. Shokina, editors, *Proc. of the Second Russian-German Advanced Research Workshop on Computational Science and High Performance Computing. IN: Notes on Numerical Fluid Mechanics and Multidisciplinary Design (NNFM), Vol.91*, Stuttgart, Germany, March 2005. Springer. in press.
28. L. Oliker, A. Canning, J. Carter, J. Shalf, D. Skinner, S. Ethier, R. Biswas, J. Djomehri, and R. van der Wijngaart. Evaluation of Cache-based Superscalar and Cacheless Vector Architectures for Scientific Computations. In *Proceedings of the ACM/IEEE Supercomputing Conference 2003, Phoenix, Arizona, USA*, 2003.
29. T. Pohl, F. Deserno, N. Thürey, U. Rüde, P. Lammers, G. Wellein, and T. Zeiser. Performance Evaluation of Parallel Large-Scale Lattice Boltzmann Applications on Three Supercomputing Architectures. In *Proceedings of the ACM/IEEE Supercomputing Conference 2004, Pittsburgh, USA*, 2004.
30. P. A. Sackinger, P. R. Schunk, and R. R. Rao. A Newton-Raphson pseudo-solid domain mapping technique for free and moving boundary problems: a finite element implementation. *J. Comp. Physics*, 125:83–103, 1996.

31. J. C. Simo and S. Rifai. A class of mixed assumed strain methods and the method of incompatible modes. *Int. J. Numer. Meth. Engrg.*, 29:1595–1638, 1990.
32. E. Slikkerveer, P. van Lohuizen, and S. B. G. O'Brien. An implicit surface tension algorithm for Picard solvers of surface-tension-dominated free and moving boundary problems. *Int. J. Numer. Meth. Fluids*, 22:851–865, 1996.
33. T. E. Tezduyar. Finite element methods for fluid dynamics with moving boundaries and interfaces. In E. Stein, R. de Borst, and T. Hughes, editors, *Encyclopedia of Computational Mechanics: Fluids (Ch. 17, Vol 3)*, pages 545–577. John Wiley & Sons, West Sussex, 2004.
34. T. E. Tezduyar, S. Mittal, S. E. Ray, and R. Shih. Incompressible flow computations with stabilized bilinear and linear equal-order-interpolation velocity-pressure elements. *Comput. Methods Appl. Mech. Engrg.*, 95:221–242, 1992.
35. T. L. Veldhuizen. Scientific Computing: C++ Versus Fortran: C++ has more than caught up. *Dr. Dobb's Journal of Software Tools*, 22(11):34, 36–38, 91, Nov. 1997.
36. T. L. Veldhuizen and M. E. Jernigan. Will C++ be Faster than Fortran? In *Proceedings of the 1st International Scientific Computing in Object-Oriented Parallel Environments (ISCOPE'97)*, Lecture Notes in Computer Science. Springer-Verlag, 1997.
37. W. A. Wall. *Fluid-Struktur-Interaktion mit stabilisierten Finiten Elementen.* PhD thesis, Institute of Structural Mechanics, University of Stuttgart, 1999.
38. W. A. Wall, M. Gee, and E. Ramm. The challenge of a three-dimensional shell formulation – the condition problem. In *Proceedings of 'IASS-IACM 2000', 4th Int. Coll. on Computation of Shell and Spatial Structures.*, Plenary paper, Chania, Crete, Greece, 2000.
39. W. A. Wall, S. Genkinger, and E. Ramm. A strong coupling partitioned approach for fluid-structure interaction with free surfaces. *Computers and Fluids*, in press, 2006.
40. W. A. Wall, A. Gerstenberger, P. Gamnitzer, C. Förster, and E. Ramm. Large deformation fluid structure interaction - advances in ale methods and new fixed grid approaches. In *Fluid-Structure-Interaction: Modelling - Simulation - Optimization. LN SCI Series*. Springer, 2006. in press.
41. W. A. Wall, D. P. Mok, and E. Ramm. Partitioned analysis approach of the transient coupled response of viscous fluids and flexible structures. In W. Wunderlich, editor, *Solids, Structures and Coupled Problems in Engineering, Proceedings of the European Conference on Computational Mechanics ECCM '99*, Munich, August/September 1999.

Large-Scale Simulations of Acoustic-Structure Interaction Using the Fast Multipole BEM*

Lothar Gaul and Matthias Fischer

Institute of Applied and Experimental Mechanics, University of Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany
gaul@iam.uni-stuttgart.de
fischer@iam.uni-stuttgart.de

Summary. For the simulation of acoustic-structure interaction problems, the coupled field equations must be solved. The structure is commonly discretized using finite elements, whereas for the acoustic field the boundary element method (BEM) is favorable. A mortar BEM-FEM coupling algorithm is developed that allows the combination of non-conforming meshes. The high flexibility for the choice of discretizations offers a high efficiency, since specialized shape functions and adaptive mesh refinement can be used in the subdomains. The mortar coupling algorithm yields a saddle point problem that is solved using a preconditioned inexact Uzawa algorithm. The iterative solver enables the use of the fast multipole BEM and thus coupled simulations on large boundary element models. Uncertain model parameters can be represented by fuzzy numbers and subsequently fuzzy arithmetic based on the transformation method can be used to evaluate BEM-FEM models in the presence of fuzzy-valued parameters.

Keywords: Acoustic-structure interaction, fast multipole BEM, BEM-FEM coupling, fuzzy BEM

1 Introduction

Very often one can perform structural-acoustic simulations neglecting the influence of the acoustic field on the vibration behavior of the structure. However, this is not acceptable for thin and flexible structures that are easily excited by the acoustic pressure. This situation is encountered e.g. for payloads in aerospace vehicles, partitioning walls, or piping systems. In such applications, the acoustic field must be fully coupled to the vibrating structure. For the simulation of structural vibrations, the FEM is the method of choice in engineering practice. For the acoustic field, the BEM offers the advantage that

* Research Project B5 "Acoustic and Hydro-Acoustic Radiation of Vibrating Structures"

only the boundary of the domain must be discretized and that the Sommerfeld radiation condition is inherently fulfilled for exterior problems. Thus, a BEM-FEM coupling algorithm is required for the simulation of the acoustic-structure interaction. The mortar method proposed by Bernardi et al. [1] can be adapted to BEM-FEM coupling to yield a high flexibility in the choice of discretizations [11]. The algorithm introduces a Lagrange multiplier on the coupling interface and thus allows the combination of non-conforming meshes what permits the use of specialized ansatz functions and adaptive mesh refinement in the subdomains.

The limiting factor of classical BEM-FEM coupling approaches are the fully populated system matrices of the BEM. Computing time and memory requirements increase quadratically with the number of boundary elements which restricts BEM simulations to rather small models. Using the fast multipole algorithm that was originally developed by Greengard and Rokhlin [14] for particle simulations in physics the numerical cost for the evaluation of the BEM matrix-vector product can be reduced to quasi-linear [9, 10]. The large savings of the fast multipole BEM are achieved by series expansion of the fundamental solution and a hierarchical multi-level scheme.

When using the multipole BEM for acoustic-structure interaction, the coupled system of equations must be solved iteratively. An inexact Uzawa type algorithm proves to be suitable for the saddle point problem arising from the mortar coupling. The nested iteration scheme consists of an outer GMRES iteration applied on the reduced equation for the Lagrange multiplier and preconditioned inner CG and GMRES iterations for the FEM and BEM subsystems, respectively. For the inner iterations a relaxation scheme can be applied to yield a further improvement of the efficiency.

As precise input data is often not available for all parameters for real-world applications, uncertainties should be included into the modeling procedure. A very practical approach to simulation and analysis of uncertain systems is the numerical implementation of the uncertain parameters as fuzzy numbers, and subsequently, the use of fuzzy arithmetic based on the transformation method, which has been proposed by Hanss in [16, 17].

The presented methodology provides a complete package for efficient simulation of structure-acoustic field interaction that can be applied to industrial-scale problems [8]. In Sect. 2 the mortar coupling algorithm is derived. For the stability of the formulation, special care is paid to the discretization of the Lagrange multiplier. The inexact Uzawa solver is presented in Sect. 3 including the discussion of a suitable relaxation scheme. In Sect. 4, the multipole evaluation of the BEM matrix-vector products is presented and Sect. 5 covers an approximate inverse approach for preconditioning of the BEM subsystem. The potential of the proposed methods is demonstrated in Sect. 6 with two numerical examples. The paper concludes with a short presentation of the transformation method as a tool for the evaluation of uncertain systems and its implementation into the BEM-FEM simulation environment in Sect. 7.

2 Mortar Coupling Algorithm

Before developing the mortar coupling scheme, the model problem for acoustic-structure interaction is shortly reviewed. A structure is assumed to be fully submerged in an acoustic fluid as displayed in Fig. 1. The structure is modeled as a thin Kirchhoff plate on the interaction boundary Γ^{int}. The out-of-plane displacement is denoted by w and the loading $f = f^{\mathrm{f}} + f^{\mathrm{e}}$ consists of surface forces due to the acoustic field f^{f} and externally applied forces f^{e}. For time-harmonic behavior $\exp(-\mathrm{i}\omega t)$, the pressure p in the acoustic field Ω^{f} is governed by the Helmholtz equation $\triangle p + \kappa^2 p = 0$ with the circular wavenumber $\kappa = \omega / c^{\mathrm{f}}$. The acoustic flux on the boundary is defined as $q = \partial p / \partial n^{\mathrm{f}}$. The boundary $\partial\Omega^{\mathrm{f}} = \Gamma = \Gamma^{\mathrm{int}} \cup \Gamma^{\mathrm{D}} \cup \Gamma^{\mathrm{N}}$ is composed of acoustic-structure interface, Dirichlet boundary, and Neumann boundary. The Sommerfeld radiation condition is inherently fulfilled by the BEM for computations on exterior domains. On the acoustic-structure interface Γ^{int}, the coupling conditions enforce equilibrium $p = f^{\mathrm{f}}$ and continuity $q = -\rho^{\mathrm{f}}\omega^2 w$ where the different orientation of the normal vectors in the structure and fluid subdomains is considered.

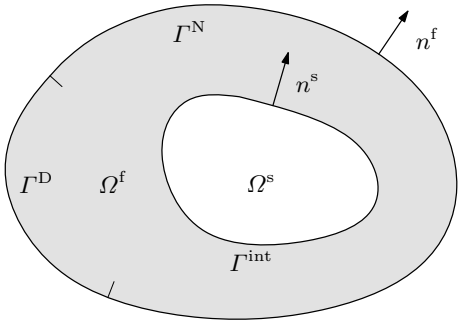


Fig. 1. Acoustic-structure interaction model problem

The choice of the Lagrange multiplier is the essential step for the formulation of the mortar coupling algorithm. In the presented approach, the pressure on the acoustic-structure interface is employed as Lagrange multiplier, i.e. $\lambda = p^{\mathrm{int}} = f^{\mathrm{f}}$.

For the simulation of the Kirchhoff plate, a finite element formulation is chosen. For time-harmonic vibration, the bilinear form for a plate of thickness t is given by

$$a(w,v) = \int_{\Gamma^{\mathrm{int}}} \left[D(1-\nu) \left(\frac{\partial^2 v}{\partial x_1^2} \frac{\partial^2 w}{\partial x_1^2} + 2 \frac{\partial^2 v}{\partial x_1 \partial x_2} \frac{\partial^2 w}{\partial x_1 \partial x_2} + \frac{\partial^2 v}{\partial x_2^2} \frac{\partial^2 w}{\partial x_2^2} \right) + D\nu \triangle v \triangle w - \rho^{\mathrm{s}} t \omega^2 w v \right] \mathrm{d}s_x \,. \quad (1)$$

with the bending stiffness $D = Et^3/\left[12(1-\nu^2)\right]$, Young's modulus E, Poisson's ratio ν and density ρ^{s}. The displacement field of the fluid-loaded plate is obtained by the solution of

$$a(w, v^w) - \int_{\Gamma^{\mathrm{int}}} v^w \lambda \, \mathrm{d}s_x = \int_{\Gamma^{\mathrm{int}}} v^w f^{\mathrm{e}} \, \mathrm{d}s_x \; . \tag{2}$$

For the acoustic domain the BEM is used. Starting point is the representation formula

$$p(x) = \int_{\Gamma} P^*(x,y)\, q(y) \, \mathrm{d}s_y - \int_{\Gamma} \frac{\partial P^*(x,y)}{\partial n_y} \, p(y) \, \mathrm{d}s_y \, , \quad x \in \Omega^{\mathrm{f}} \, , \tag{3}$$

yielding the acoustic pressure at an arbitrary field point x. For details on the derivation of boundary integral formulations, it is referred to [13]. Taking the limit of (3) on the smooth boundary $\Omega^{\mathrm{f}} \ni x \to \Gamma$, one obtains the boundary integral equation

$$p(x) = \frac{1}{2} p(x) + \underbrace{\int_{\Gamma} P^*(x,y)\, q(y) \, \mathrm{d}s_y}_{(Vq)(x)} - \underbrace{\int_{\Gamma} \frac{\partial P^*(x,y)}{\partial n_y} \, p(y) \, \mathrm{d}s_y}_{(Kp)(x)} \, , \quad x \in \Gamma \, , \tag{4}$$

where $(Vq)(x)$ and $(Kp)(x)$ are the single and double layer potential, respectively. The hyper-singular boundary integral equation is obtained by taking the normal derivative of (3)

$$q(x) = \frac{1}{2} q(x) + \underbrace{\int_{\Gamma} \frac{\partial P^*(x,y)}{\partial n_x} \, q(y) \, \mathrm{d}s_y}_{(K'q)(x)} - \underbrace{\int_{\Gamma} \frac{\partial^2 P^*(x,y)}{\partial n_x \partial n_y} \, p(y) \, \mathrm{d}s_y}_{-(Dp)(x)} \, , \quad x \in \Gamma \, , \tag{5}$$

where $(K'q)(x)$ and $(Dp)(x)$ are the adjoint double layer potential and the hyper-singular operator, respectively.

The pressure and flux fields on the boundary are decomposed to yield a symmetric formulation: $p = p^{\mathrm{int}} + \tilde{p} + \bar{p}$ and $q = q^{\mathrm{int}} + \tilde{q} + \bar{q}$, where $\bar{p}$ and $\bar{q}$ are the prescribed Dirichlet and Neumann boundary conditions, respectively. The extensions are $\tilde{p} = 0$ on $\Gamma^{\mathrm{int}} \cup \Gamma^{\mathrm{D}}$ and $\tilde{q} = 0$ on $\Gamma^{\mathrm{int}} \cup \Gamma^{\mathrm{N}}$.

On the structure-acoustic field interface Γ^{int}, the boundary integral equation (4) is weighted with test functions v^q and the term $p^{\mathrm{int}} - \lambda$ is added to enforce the equilibrium coupling condition

$$\begin{aligned} \int_{\Gamma^{\mathrm{int}}} v^q (Vq^{\mathrm{int}})(x) \, \mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} v^q (V\tilde{q})(x) \, \mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} v^q \left(-\frac{1}{2} I - K \right) p^{\mathrm{int}}(x) \, \mathrm{d}s_x \\ - \int_{\Gamma^{\mathrm{int}}} v^q (K\tilde{p})(x) \, \mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} v^q (p^{\mathrm{int}} - \lambda)(x) \, \mathrm{d}s_x \\ = - \int_{\Gamma^{\mathrm{int}}} v^q (V\bar{q})(x) \, \mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} v^q \left(\frac{1}{2} I + K \right) \bar{p}(x) \, \mathrm{d}s_x \; . \end{aligned} \tag{6}$$

Likewise, on the Dirichlet boundary

$$\int_{\Gamma^{\mathrm{D}}} v^q (V q^{\mathrm{int}})(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{D}}} v^q (V\tilde{q})(x)\,\mathrm{d}s_x - \int_{\Gamma^{\mathrm{D}}} v^q (K p^{\mathrm{int}})(x)\,\mathrm{d}s_x$$
$$- \int_{\Gamma^{\mathrm{D}}} v^q (K\tilde{p})(x)\,\mathrm{d}s_x = -\int_{\Gamma^{\mathrm{D}}} v^q (V\bar{q})(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{D}}} v^q \left(\frac{1}{2}I + K\right)\bar{p}(x)\,\mathrm{d}s_x\,. \quad (7)$$

The hyper-singular boundary integral equation (5) is tested on the interface Γ^{int}

$$\int_{\Gamma^{\mathrm{int}}} v^p (D p^{\mathrm{int}})(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} v^p (D\tilde{p})(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} v^p \left(-\frac{1}{2}I + K'\right) q^{\mathrm{int}}(x)\,\mathrm{d}s_x$$
$$+ \int_{\Gamma^{\mathrm{int}}} v^p (K'\tilde{q})(x)\,\mathrm{d}s_x = -\int_{\Gamma^{\mathrm{int}}} v^p \left(D\bar{p}\right)(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} v^p \left(\frac{1}{2}I - K'\right)\bar{q}(x)\,\mathrm{d}s_x\,, \quad (8)$$

as well as on the Neumann boundary Γ^{N}

$$\int_{\Gamma^{\mathrm{N}}} v^p (D p^{\mathrm{int}})(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{N}}} v^p (D\tilde{p})(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{N}}} v^p (K' q^{\mathrm{int}})(x)\,\mathrm{d}s_x$$
$$+ \int_{\Gamma^{\mathrm{N}}} v^p (K'\tilde{q})(x)\,\mathrm{d}s_x = \int_{\Gamma^{\mathrm{N}}} v^p \left(\frac{1}{2}I - K'\right)\bar{q}(x)\,\mathrm{d}s_x - \int_{\Gamma^{\mathrm{N}}} v^p \left(D\bar{p}\right)(x)\,\mathrm{d}s_x\,. \quad (9)$$

The continuity condition is enforced by

$$\int_{\Gamma^{\mathrm{int}}} v^\lambda \left(\rho^{\mathrm{f}}\omega^2 w + q^{\mathrm{int}}\right)\,\mathrm{d}s_x = 0\,. \quad (10)$$

For the discretization of the Kirchhoff plate, the ansatz functions must provide C^1 continuity. A four-node finite plate element proposed by Zienkiewicz [28] is chosen. On each node, it possesses three degrees of freedom: the out-of-plane displacement and two rotations. Incomplete bi-cubic ansatz functions $\boldsymbol{\varphi}^w$ are used for interpolation. They do not guarantee continuity of the slope normal to the element edges, however, convergence is well established for rectangular elements.

$$w_h(x) = \boldsymbol{\varphi}^w(x)^{\mathrm{T}}\mathbf{w}\,. \quad (11)$$

A consistent discretization of distributed pressure loads on the plate would induce nodal forces on the rotational degrees of freedom. For the load vector in the coupling algorithm, thus, simplified shape functions $\tilde{\boldsymbol{\varphi}}^w(x)$ are used where the contribution of the rotational degrees of freedom is set to zero.

The acoustic pressure and flux are interpolated on the boundary element triangulation using linear ansatz functions $\boldsymbol{\varphi}^p$ and constant ansatz functions $\boldsymbol{\varphi}^q$, respectively.

$$p_h(x) = \boldsymbol{\varphi}^p(x)^{\mathrm{T}}\mathbf{p}\,, \qquad q_h(x) = \boldsymbol{\varphi}^q(x)^{\mathrm{T}}\mathbf{q}\,. \tag{12}$$

For a unique solution of the saddle point problem consisting of (2) and (6)–(10), a stability requirement – the discrete Babuška-Brezzi or inf-sup condition – has to be established [3]. Thus, special attention must be paid to the discretization of the Lagrange multiplier. Choosing the boundary element mesh as the slave or mortar triangulation and using linear ansatz functions for the Lagrange multiplier, the ansatz space is rich enough. However, at the edges of the coupling interface, the linear ansatz space must be restricted, since the Lagrange multiplier, i.e. the fluid load on the plate, is not unique where structural Dirichlet boundary conditions are applied. A one-dimensional illustration of the modified ansatz functions $\boldsymbol{\varphi}^\lambda$ is shown in Fig. 2.

$$\lambda_h(x) = \boldsymbol{\varphi}^\lambda(x)^{\mathrm{T}}\boldsymbol{\lambda}\,. \tag{13}$$

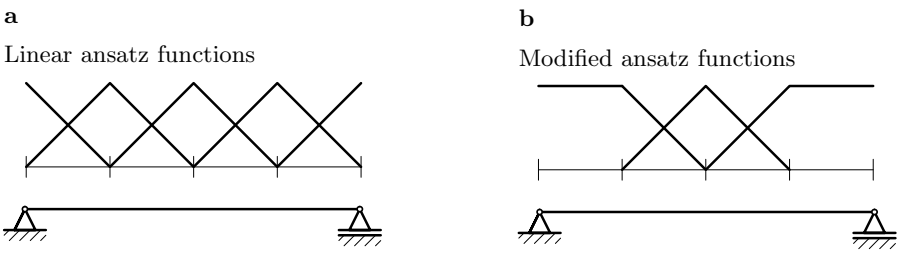


Fig. 2. Modification of the ansatz space for the Lagrange multiplier

Using the shape functions (11)–(13) and isoparametric test functions, one obtains the block skew-symmetric system of equations

$$\begin{pmatrix} \rho^{\mathrm{f}}\omega^2\mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{C}_{\mathrm{FEM}} \\ \mathbf{0} & \mathbf{V}_{\mathrm{i,i}} & \mathbf{V}_{\mathrm{i,D}} & \frac{1}{2}\mathbf{I} - \mathbf{K}_{\mathrm{i,i}} & -\mathbf{K}_{\mathrm{i,N}} & -\mathbf{C}_{\mathrm{BEM}} \\ \mathbf{0} & \mathbf{V}_{\mathrm{D,i}} & \mathbf{V}_{\mathrm{D,D}} & -\mathbf{K}_{\mathrm{D,i}} & -\mathbf{K}_{\mathrm{D,N}} & \mathbf{0} \\ \mathbf{0} & -\frac{1}{2}\mathbf{I}^{\mathrm{T}} + \mathbf{K}_{\mathrm{i,i}}^{\mathrm{T}} & \mathbf{K}_{\mathrm{i,D}}^{\mathrm{T}} & \mathbf{D}_{\mathrm{i,i}} & \mathbf{D}_{\mathrm{i,N}} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{\mathrm{i,N}}^{\mathrm{T}} & \mathbf{K}_{\mathrm{D,N}}^{\mathrm{T}} & \mathbf{D}_{\mathrm{N,i}} & \mathbf{D}_{\mathrm{N,N}} & \mathbf{0} \\ \mathbf{C}_{\mathrm{FEM}}^{\mathrm{T}} & \mathbf{C}_{\mathrm{BEM}}^{\mathrm{T}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \mathbf{q}^{\mathrm{int}} \\ \tilde{\mathbf{q}} \\ \mathbf{p}^{\mathrm{int}} \\ \tilde{\mathbf{p}} \\ \boldsymbol{\lambda} \end{pmatrix}$$

$$= \begin{pmatrix} \rho^{\mathrm{f}}\omega^2 \int_{\Gamma^{\mathrm{int}}} \tilde{\boldsymbol{\varphi}}^w f^{\mathrm{e}}\,\mathrm{d}s_x \\ -\int_{\Gamma^{\mathrm{int}}} \boldsymbol{\varphi}^q (V\bar{q})(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} \boldsymbol{\varphi}^q \left(\frac{1}{2}I + K\right)\bar{p}(x)\,\mathrm{d}s_x \\ -\int_{\Gamma^{\mathrm{D}}} \boldsymbol{\varphi}^q (V\bar{q})(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{D}}} \boldsymbol{\varphi}^q \left(\frac{1}{2}I + K\right)\bar{p}(x)\,\mathrm{d}s_x \\ -\int_{\Gamma^{\mathrm{int}}} \boldsymbol{\varphi}^p \left(D\bar{p}\right)(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{int}}} \boldsymbol{\varphi}^p \left(\frac{1}{2}I - K'\right)\bar{q}(x)\,\mathrm{d}s_x \\ -\int_{\Gamma^{\mathrm{N}}} \boldsymbol{\varphi}^p \left(D\bar{p}\right)(x)\,\mathrm{d}s_x + \int_{\Gamma^{\mathrm{N}}} \boldsymbol{\varphi}^p \left(\frac{1}{2}I - K'\right)\bar{q}(x)\,\mathrm{d}s_x \\ \mathbf{0} \end{pmatrix}, \tag{14}$$

where $\mathbf{A}$ is the dynamic FEM stiffness matrix and $\mathbf{V}$, $\mathbf{K}$, and $\mathbf{D}$ are the BEM system matrices evaluated on the boundary sections corresponding to their

subscripts. The coupling matrices $\mathbf{C}_{\mathrm{FEM}}$ and $\mathbf{C}_{\mathrm{BEM}}$ are defined by integration of the respective shape functions over the coupling interface

$$\mathbf{C}_{\mathrm{FEM}} = \rho^{\mathrm{f}}\omega^2 \int_{\Gamma^{\mathrm{int}}} \tilde{\boldsymbol{\varphi}}^{w}(x)\boldsymbol{\varphi}^{\lambda}(x)^{\mathrm{T}}\,\mathrm{d}s_x\,, \quad \mathbf{C}_{\mathrm{BEM}} = \int_{\Gamma^{\mathrm{int}}} \boldsymbol{\varphi}^{q}(x)\boldsymbol{\varphi}^{\lambda}(x)^{\mathrm{T}}\,\mathrm{d}s_x\,. \quad (15)$$

3 Inexact Uzawa Solver

Immediate iterations on the system (14) converge very poorly, thus, an Uzawa type algorithm is employed. In the Uzawa algorithm, the system (14) is formally solved for $\boldsymbol{\lambda}$ and iterations are applied on the reduced equation

$$\underbrace{(\mathbf{C}_{\mathrm{FEM}}^{\mathrm{T}}\ \mathbf{C}_{\mathrm{BEM}}^{\mathrm{T}}) \begin{pmatrix} (\rho^{\mathrm{f}}\omega^2\mathbf{A})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{\mathrm{FEM}} \\ \mathbf{C}_{\mathrm{BEM}} \end{pmatrix}}_{\mathbf{S}} \boldsymbol{\lambda} = \underbrace{\begin{pmatrix} \rho^{\mathrm{f}}\omega^2\mathbf{f}_{\mathrm{FEM}} \\ \mathbf{f}_{\mathrm{BEM}} \end{pmatrix}}_{\mathbf{f}}\,, \quad (16)$$

where the BEM matrix $\mathbf{B}$ is introduced to simplify the notation and represents the BEM system of (14). The matrix inverses in (16) are not evaluated explicitly, but conjugate gradient and GMRES iterations are applied on the FEM and BEM subsystems, respectively. Instead of evaluating the exact matrix-vector product $\mathbf{S}\boldsymbol{\lambda}_k$ at each outer iteration step k, an approximation with a relative precision η_k is computed as

$$\tilde{\mathbf{S}}_k\boldsymbol{\lambda}_k = \mathbf{S}\boldsymbol{\lambda}_k + \mathbf{g}_k\,, \quad (17)$$

with $||\mathbf{g}_k||_2 \leq \eta_k||\mathbf{S}\boldsymbol{\lambda}_k||_2$. For most applications of inexact Uzawa algorithms, i.e. nested inner-outer iterations, Newton type schemes are employed for the outer iterations. They offer the advantage that for the inner iterations a coarse approximation is sufficient in the beginning. Only with shrinking outer residual, the approximation of the inner systems must be improved.

For the acoustic-structure interaction system (16), standard Newton methods show a very poor convergence behavior and GMRES is thus employed for the outer iterations. GMRES offers quick convergence, however, the intuitive scheme – improving the precision of the inner approximation when the outer residual becomes small – breaks down. On the contrary, the precision of the inner approximation must be high in the beginning and can be relaxed subsequently.

An empirical relaxation strategy for a wide range of application is developed by Bouras and Frayssé [2]. Their procedure is supported by the analyses of Simoncini and Szyld [25] and van den Eshof and Sleijpen [26]. For the inexact evaluation of the matrix-vector product (17), the Arnoldi algorithm of the GMRES on (16) creates an orthogonal basis that does not correspond to the Krylov space $\mathcal{K}_k(\mathbf{S}, \mathbf{r}_0)$, and the value $|\gamma_{k+1}|$ which is used to check the convergence of GMRES according to [24] does not correspond to the true residual $\mathbf{r}_k = \mathbf{f} - \mathbf{S}\boldsymbol{\lambda}_k$. This residual gap must be controlled by the employed relaxation scheme.

For the solution of (16) with a target residual $||\mathbf{f} - \mathbf{S}\boldsymbol{\lambda}_k||_2 \leq \epsilon^{\text{outer}}$, a relative precision of

$$\eta_k = \frac{\epsilon^{\text{outer}}}{|\gamma_k|} \tag{18}$$

is required for the evaluation of the approximate matrix-vector product $\tilde{\mathbf{S}}_k\boldsymbol{\lambda}_k$ according to the analysis of van den Eshof and Sleijpen [26]. Thus, the inner systems $(\rho^{\text{f}}\omega^2\mathbf{A})^{-1}\mathbf{C}_{\text{FEM}}\boldsymbol{\lambda_k}$ and $\mathbf{B}^{-1}\mathbf{C}_{\text{BEM}}\boldsymbol{\lambda_k}$ must be solved with a relative residual of

$$\epsilon_k^{\text{inner}} = \underbrace{\frac{\left|\left|\tilde{\mathbf{S}}_k\boldsymbol{\lambda_k}\right|\right|_2}{\left|\left|\begin{pmatrix}\mathbf{C}_{\text{FEM}}\\ \mathbf{C}_{\text{BEM}}\end{pmatrix}\boldsymbol{\lambda_k}\right|\right|_2}}_{c_k} \frac{\epsilon^{\text{outer}}}{|\gamma_k|} \,. \tag{19}$$

The value of the first ratio c_k in (19) is not known prior to actually computing the approximations of the inner systems. Numerical experiments show that the value of the ratio usually does not change significantly from one outer iteration to the next. Thus, the ratio of the previous step can be used as a first guess to determine the required precision of the current approximation. Using the computed result, the true ratio is calculated and compared to the previous one. Only if there is a significant deviation, i.e. if the true ratio is significantly smaller than the guess, the approximation must be computed again up to the correct precision.

The nested iterative solution scheme requires a frequent evaluation of the matrix-vector products of the BEM and FEM subsystems. Compared to standard BEM the computing cost for large boundary element models can be largely reduced using the fast multipole BEM. However each evaluation of the matrix-vector product is still rather costly. Thus, preconditioning of the subsystems is essential for the efficiency of the solver. As an advantage of the presented approach, standard preconditioners for the FEM and BEM parts can be used. In the numerical examples, focus is on the boundary element part. The approximate inverse approach presented in Sect. 5 proves to work well for the coupled simulations.

4 Multipole Evaluation of BEM Matrix-Vector Products

As mentioned before, the matrix-vector products of large BEM models can be evaluated efficiently using the fast multipole BEM. The numerical cost is reduced significantly compared to the standard BEM due to the series expansion of the fundamental solution and a multilevel scheme. The appropriate theory of the multipole expansion for acoustics and its diagonal translation operators was developed by Rokhlin. Following [23], the multipole expansion of the fundamental solution around the center z as depicted in Fig. 3 reads

$$\frac{\mathrm{e}^{\mathrm{i}\kappa|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|} = \frac{\mathrm{e}^{\mathrm{i}\kappa|\mathbf{D}+\mathbf{d}|}}{|\mathbf{D}+\mathbf{d}|}$$
$$= \mathrm{i}\kappa \sum_{l=0}^{\infty} (2l+1)(-1)^l j_l(\kappa|\mathbf{d}|)\, h_l^{(1)}(\kappa|\mathbf{D}|)\, P_l(\hat{\mathbf{D}}\cdot\hat{\mathbf{D}}), \quad |\mathbf{D}| > |\mathbf{d}| \,. \tag{20}$$

$j_l(\cdot)$ denotes spherical Bessel functions, $h_l(\cdot)$ Hankel functions and $P_l(\cdot)$ the Legendre polynomials. Normalized vectors are indicated by $\hat{(\cdot)} = (\cdot)/|\cdot|$. Using the orthonormality of the Legendre polynomials on the unit sphere $\mathbb{S}^2$ and the expansion of spherical waves

$$4\pi \mathrm{i}^l j_l(\kappa|\mathbf{d}|) P_l(\hat{\mathbf{D}}\cdot\hat{\mathbf{D}}) = \int_{\mathbb{S}^2} \mathrm{e}^{\mathrm{i}\kappa\mathbf{d}\cdot\hat{\mathbf{s}}} P_l(\hat{\mathbf{s}}\cdot\hat{\mathbf{D}})\,\mathrm{d}s \,, \tag{21}$$

one obtains a diagonal form of the multipole expansion

$$\frac{\mathrm{e}^{\mathrm{i}\kappa|\mathbf{D}+\mathbf{d}|}}{|\mathbf{D}+\mathbf{d}|} = \frac{\mathrm{i}\kappa}{4\pi} \sum_{l=0}^{\infty} (2l+1)\mathrm{i}^l h_l^{(1)}(\kappa|\mathbf{D}|) \int_{\mathbb{S}^2} \mathrm{e}^{\mathrm{i}\kappa\mathbf{d}\cdot\hat{\mathbf{s}}} P_l(\hat{\mathbf{s}}\cdot\hat{\mathbf{D}})\,\mathrm{d}s \,. \tag{22}$$

The summation over l in (22) must be truncated, since the Hankel function diverges for large l. For the truncated series, summation and integration can be interchanged and one defines the translation operators

$$M_L(s, D) = \sum_{l=0}^{L} (2l+1)\mathrm{i}^l h_l^{(1)}(\kappa|\mathbf{D}|) P_l(\hat{\mathbf{s}}\cdot\hat{\mathbf{D}}) \,, \tag{23}$$

which solely depend on the distance vector $\mathbf{D}$ and allow the implementation of an efficient multilevel scheme.

The boundary elements of the triangulation are assigned to a cluster tree with levels $\ell = 0 \,..\, \ell_{\max}$. The cluster on the highest level $\ell = 0$ represents a parallelepiped containing all boundary elements. Subsequent child clusters are constructed by bisection of the parent cluster. The clusters are denoted by C_ℓ^γ where $\gamma = 1 \,..\, 2^\ell$. Two clusters on the same level are in each others nearfield when the distance D between their centers z_ℓ^γ fulfills the condition

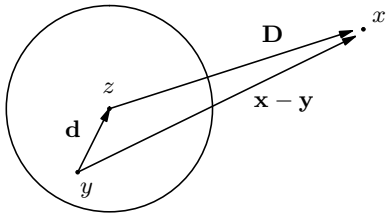


Fig. 3. Definition of distance vectors for translation operators

$$D < c\frac{d_\ell}{2} \,, \tag{24}$$

where d_ℓ is the diameter of the clusters at level ℓ and c is a suitable constant. Clusters, whose father clusters fulfill the nearfield condition, but themselves are not in each others nearfield, form the interaction list.

Using these definitions, the multipole algorithm is formulated for the ν-th component of the matrix-vector product $v_\nu = (\mathbf{Vu})_\nu$ of the single layer potential evaluated for constant shape functions. The matrix-vector products of the other BEM matrices in (14) can be treated in a similar way. For convenience, μ and ν are used as indices for the element vectors as well as to denote the elements μ and ν, respectively. The nearfield part of the boundary integral operator is evaluated directly—i.e. the matrix entries are computed and stored when field and load point are close to each other—whereas for the farfield part, the multipole method is applied. The double surface integral is substituted by Gauss quadrature on the elements

$$\begin{aligned} v_\nu &= (\mathbf{V}_{\text{nearfield}}\mathbf{u})_\nu + \int_\nu \sum_{\mu \in \text{farfield}} u_\mu \int_\mu P^*(x,y)\ \mathrm{d}s_y\ \mathrm{d}s_x \\ &\approx (\mathbf{V}_{\text{nearfield}}\mathbf{u})_\nu + \sum_{j=1}^{G_\nu} \omega_{\nu,j} \Delta_\nu \sum_{\mu \in \text{farfield}} u_\mu \sum_{i=1}^{G_\mu} \omega_{\mu,i} \Delta_\mu P^*(x_{\nu,j}, y_{\mu,i}) \,, \end{aligned} \tag{25}$$

where Δ_μ, $\omega_{\mu,i}$, and $y_{\mu,i}$ are the Jacobi determinant, Gauss weight, and integration point for element μ, respectively. The terms are defined analogous for the element ν. The expression

$$\Phi(x_{\nu,j}) = \sum_{\mu \in \text{farfield}} u_\mu \sum_{i=1}^{G_\mu} \omega_{\mu,i} \Delta_\mu P^*(x_{\nu,j}, y_{\mu,i}) \tag{26}$$

can be evaluated efficiently using the multipole method, whereas the remaining operations in (25) are local at element ν. The multipole algorithm consists of the following steps:

(i) Compute the farfield signature $F^\gamma_{\ell_{\max}}(s)$ for all clusters on the lowest level $\ell = \ell_{\max}$

$$F^\gamma_{\ell_{\max}}(s) = \sum_{\mu \in C^\gamma_{\ell_{\max}}} u_\mu \Delta_\mu \sum_{i=1}^{G_\mu} \omega_{\mu,i} \mathrm{e}^{\mathrm{i}\kappa(\mathbf{y}_{\mu,i} - \mathbf{z}^\gamma_{\ell_{\max}})\cdot\hat{\mathbf{s}}} \,. \tag{27}$$

(ii) Translate $F^\gamma_\ell(s)$ to the interaction list using the translation operators M_L

$$N^\gamma_\ell(s) = \sum_{\text{interaction list}} M_L(s,D)\ F^\gamma_\ell(s) \,, \tag{28}$$

where D is the distance between z^γ_ℓ and the center of the respective interaction list cluster.

(iii) Shift $F_\ell^\gamma(s)$ to the center of the father cluster.
(iv) Repeat last two steps upwards until the interaction list is empty.
(v) In the downward pass, shift the nearfield signatures $N_\ell^\gamma(s)$ in interaction lists to the child clusters.
(vi) On the lowest level, recover solution in integration points $x_{\nu,j}$

$$\Phi(x_{\nu,j}) = \frac{\mathrm{i}\kappa}{4\pi} \int_{\mathbb{S}^2} \mathrm{e}^{\mathrm{i}\kappa(\mathbf{z}^\gamma_{\ell_{\max}} - \mathbf{x}_{\nu,j})\cdot\hat{\mathbf{s}}} N^\gamma_{\ell_{\max}}(s)\,\mathrm{d}s \ . \tag{29}$$

Finally, the matrix-vector product (25) is evaluated by summation of the element contributions $\Phi(x_{\nu,j})$ and addition of the nearfield contribution which is calculated directly using the standard BEM.

The expansion length L plays a crucial role for the accuracy and the efficiency of the multipole method. Coifman et al. [5] give the semi-empirical fit

$$L_\ell(\kappa d_\ell) = \kappa d_\ell + p \ln(\kappa d_\ell + \pi) \ , \tag{30}$$

where the expansion length depends on the wave number κ and the cluster diameter d_ℓ. The parameter p does not translate directly to the number of accurate digits, since the error in the truncated multipole expansion does not only depend on the cluster size but also on the distance between two clusters where the translation operator is applied. A detailed discussion of the truncation error is given in the papers by Koc et al. [21] and Darve [6].

A consequence of the choice (30) is that the expansion length changes from level to level and, thus, efficient algorithms for interpolation and filtering are needed when shifting the expansion to the father and child cluster, respectively. A suitable scheme based on fast Fourier transforms is proposed by Gyure and Stalzer [15].

For the complete evaluation of a BEM matrix-vector product the computing cost can be estimated as $\mathcal{O}(N \log^2 N)$ for computations at $\kappa h = \text{const}$ what corresponds to the engineering practice of a constant number of elements per wavelength [9].

5 Approximate Inverse Preconditioning

For preconditioning of BEM systems in acoustics, approximate inverse approaches yield favorable results [12]. The conditioning of the BEM system

$$\mathbf{B}\mathbf{x} = \mathbf{f} \tag{31}$$

is improved by multiplying the system with a preconditioning matrix $\mathbf{M}^{-1}$ that is an approximation of the inverse of the system matrix $\mathbf{B}^{-1}$.

The basic idea of the approximate inverse approach is operator splitting. The boundary integral operator $B = B_0 + \tilde{B}$ is divided into a bounded contribution B_0 and the remaining $\tilde{B}$. It can be shown that the eigenvalues of

$B_0^{-1}B = I + B_0^{-1}\tilde{B}$ cluster around one which yields an improved convergence of the iterative solver [4].

The application of the operator splitting on the boundary integral operators is straight forward. For the single layer potential, the boundary integral is partitioned

$$(Vq)(x) = \int_{\Gamma_0} P^*(x,y)\; q(y)\,\mathrm{d}s_y + \int_{\tilde{\Gamma}} P^*(x,y)\; q(y)\,\mathrm{d}s_y\;, \tag{32}$$

where $\Gamma = \Gamma_0 \cup \tilde{\Gamma}$, defining V_0 and $\tilde{V}$, respectively. For the double layer potential and the hyper-singular operator, the procedure is equivalent and the operator for preconditioning the coupled system in Sect. 2 is set up accordingly. The choice of the size of Γ_0 is a trade-off between effect and numerical cost of the preconditioner. A two-element layer around the element containing the load point as depicted by the hatched elements in Fig. 4 turns out to yield good results at a negligible computing cost and is employed for the numerical examples in Sect. 6.

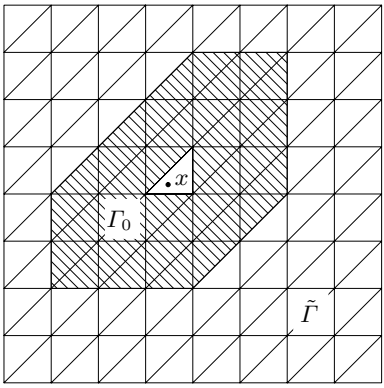


Fig. 4. Two-element layer around load point for operator splitting

The operator B_0 is sparse, however, the exact application of B_0^{-1} would be too expensive. Thus, an approximate inverse of B_0 is computed as preconditioner. The preconditioning matrix $\mathbf{M}^{-1}$ is computed by a column oriented algorithm that minimizes $||\mathbf{B}_0(\mathbf{M}^{-1})_j - \mathbf{e}_j||$, conserving the sparse pattern of $\mathbf{B}_0$ for $\mathbf{M}^{-1}$. For approximate inverse preconditioning of GMRES, the system (31) is simply replaced by

$$\mathbf{B}\mathbf{M}^{-1}\tilde{\mathbf{x}} = \mathbf{f}\,, \qquad \mathbf{x} = \mathbf{M}^{-1}\tilde{\mathbf{x}}\,. \tag{33}$$

6 Numerical Examples

The proposed coupling algorithm is demonstrated on two examples. First, the response of an elastic plate backed by a closed acoustic cavity is examined. The second example is the simulation of the sound field emitted from a submerged plate in a water basin. The numerical results are compared to experiments conducted in the hydro-acoustic lab at the Institute of Applied and Experimental Mechanics.

6.1 Plate Backed by a Closed Acoustic Cavity

The elastic plate considered in the first example has the dimensions $1\,\mathrm{m} \times 1\,\mathrm{m}$ and a thickness of $t = 0.01\,\mathrm{m}$. It is made from steel ($E = 2.1 \times 10^{11}\,\mathrm{N/m^2}$, $\nu = 0.3$, $\rho^{\mathrm{s}} = 7900\,\mathrm{kg/m^3}$) and is simply supported on all edges. The panel is coupled to a closed acoustic cavity with dimensions $1\,\mathrm{m} \times 1\,\mathrm{m} \times 1\,\mathrm{m}$. The remaining surfaces of the cavity are reverberant walls, i.e. homogeneous Neumann boundary conditions ($\bar{q} = 0$) are applied. The acoustic fluid is water ($c^{\mathrm{f}} = 1481\,\mathrm{m/s}$, $\rho^{\mathrm{f}} = 1000\,\mathrm{kg/m^3}$). For this simple model problem, an analytic series solution is developed by Pretlove [22] that is used as a reference solution.

For the simulation of the frequency response of the cavity backed panel, a boundary element mesh with 316 elements on the interface and a 20×20 finite plate element mesh is used. The computed frequency response function at the point $(0.2,\, 0.3)\,\mathrm{m}$ on the plate due to a force of $F = 1\,\mathrm{N}$ at the same position is plotted in Fig. 5. For comparison, the analytical series solution is plotted as dashed line. The BEM-FEM results agree completely with the analytic series solution in the lower frequency regime and one notices increasing deviation for higher frequencies.

The vibration modes of the plate close to the resonance frequencies are shown in Fig. 6. Comparing the vibration modes to the eigenmodes of an uncoupled plate, the effect of the acoustic cavity is particularly noticeable for mode shapes that have a non-zero average flux over the interface. Then, the stiffness effect of the cavity plays a dominant role. For example, the first uncoupled eigenfrequency of the plate is at 49 Hz whereas its mode shape can be identified at a frequency of 132 Hz for the coupled system. Mode shapes with zero average flux are found at frequencies slightly below their uncoupled counterparts due to the added mass effect of the acoustic fluid.

The convergence of the mortar coupling algorithm is studied at a frequency of 180 Hz. The acoustic pressure on the surface of the cavity and the plate displacement obtained for a simulation using the boundary element mesh with 316 elements on the interface and 20×20 finite plate elements are displayed in Fig. 7. By inspection, one finds the discretization to resolve the fields sufficiently and a typical engineering error of a few per cent is expected.

The flexibility of the mortar coupling algorithm allows an independent refinement of the FEM and BEM meshes. The convergence behavior of the

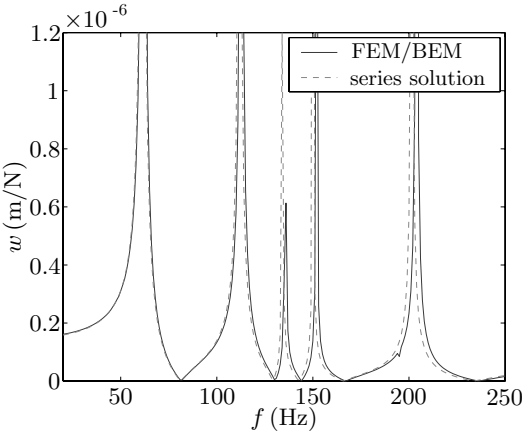


Fig. 5. Frequency response of cavity backed panel. Plate displacement w at position (0.2, 0.3) m

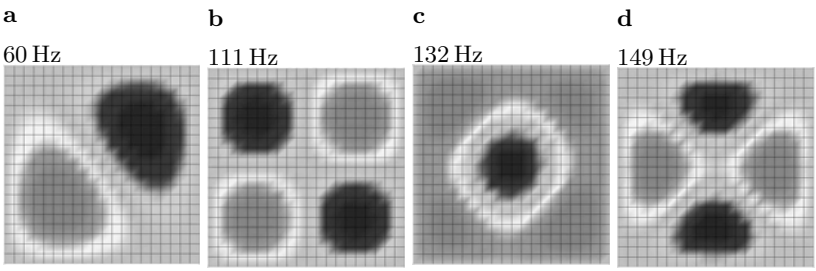


Fig. 6. Vibration modes of cavity backed plate

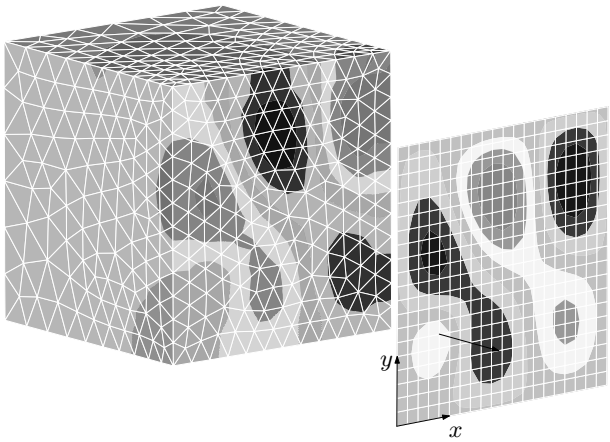


Fig. 7. Surface pressure distribution and plate displacement at 180 Hz

structure error e_2^{s} and the fluid error e_2^{f} is documented in Table 1 and Fig. 8. The errors are defined as

$$e_2^{\mathrm{s}} = \frac{||\mathbf{w}_{\mathrm{FEM}} - \mathbf{w}_{\mathrm{series}}||_2}{||\mathbf{w}_{\mathrm{series}}||_2} \qquad \text{and} \qquad e_2^{\mathrm{f}} = \frac{||\mathbf{p}_{\mathrm{BEM}}^{\mathrm{int}} - \mathbf{p}_{\mathrm{series}}^{\mathrm{int}}||_2}{||\mathbf{p}_{\mathrm{series}}^{\mathrm{int}}||_2} \,. \tag{34}$$

The simulation on the finest BEM grid, i.e. with 2936 boundary elements on the interface and a total of 11,680 boundary elements, is computed using the fast multipole BEM to take advantage of the largely reduced memory requirements. A coupled simulation of this large model using the standard BEM is not possible on the available hardware. For the multipole algorithm, a nearfield parameter of $c = 4$ and an expansion parameter of $p = 5$ are employed according to (24) and (30), respectively. This parameter combination is chosen to yield a very accurate multipole evaluation of the inner BEM matrix-vector products to satisfy the requirements of the inexact Uzawa algorithm for a true residual of the Lagrange multiplier of 10^{-4}.

The fluid error, that corresponds to the error of the Lagrange multiplier which is discretized on the BEM mesh, decreases with refinement of the FEM mesh until it approaches a value that corresponds to the discretization error

Table 1. Error of displacement and pressure field on the interface at 180 Hz

		outer iterations	BEM matrix-vector products	e_2^{s}	e_2^{f}
box 12	plate 20×20	28	1678	$6.28 \cdot 10^{-2}$	$8.19 \cdot 10^{-2}$
(316 interface	plate 30×30	28	1689	$5.54 \cdot 10^{-2}$	$7.48 \cdot 10^{-2}$
elements)	plate 40×40	28	1629	$4.05 \cdot 10^{-2}$	$6.71 \cdot 10^{-2}$
	plate 60×60	28	1628	$3.54 \cdot 10^{-2}$	$6.54 \cdot 10^{-2}$
	plate 80×80	28	1638	$3.40 \cdot 10^{-2}$	$6.50 \cdot 10^{-2}$
box 16	plate 20×20	36	2440	$4.79 \cdot 10^{-2}$	$5.79 \cdot 10^{-2}$
(572 interface	plate 30×30	36	2448	$2.72 \cdot 10^{-2}$	$4.17 \cdot 10^{-2}$
elements)	plate 40×40	36	2367	$2.85 \cdot 10^{-3}$	$4.16 \cdot 10^{-2}$
	plate 60×60	36	2371	$1.82 \cdot 10^{-2}$	$3.76 \cdot 10^{-2}$
	plate 80×80	36	2370	$1.66 \cdot 10^{-2}$	$3.72 \cdot 10^{-2}$
box 24	plate 20×20	42	3732	$3.93 \cdot 10^{-2}$	$4.54 \cdot 10^{-2}$
(1274 interface	plate 30×30	43	3683	$1.87 \cdot 10^{-2}$	$2.35 \cdot 10^{-2}$
elements)	plate 40×40	43	3687	$1.24 \cdot 10^{-2}$	$1.83 \cdot 10^{-2}$
	plate 60×60	43	3686	$8.57 \cdot 10^{-3}$	$1.62 \cdot 10^{-2}$
	plate 80×80	43	3688	$6.54 \cdot 10^{-3}$	$1.56 \cdot 10^{-2}$
box 36	plate 20×20	49	5860	$3.65 \cdot 10^{-2}$	$4.22 \cdot 10^{-2}$
(2936 interface	plate 30×30	50	5919	$1.71 \cdot 10^{-2}$	$1.94 \cdot 10^{-2}$
elements)	plate 40×40	50	5922	$9.74 \cdot 10^{-3}$	$1.12 \cdot 10^{-2}$
	plate 60×60	50	5921	$5.14 \cdot 10^{-3}$	$7.00 \cdot 10^{-3}$
	plate 80×80	50	5921	$1.95 \cdot 10^{-3}$	$6.25 \cdot 10^{-3}$

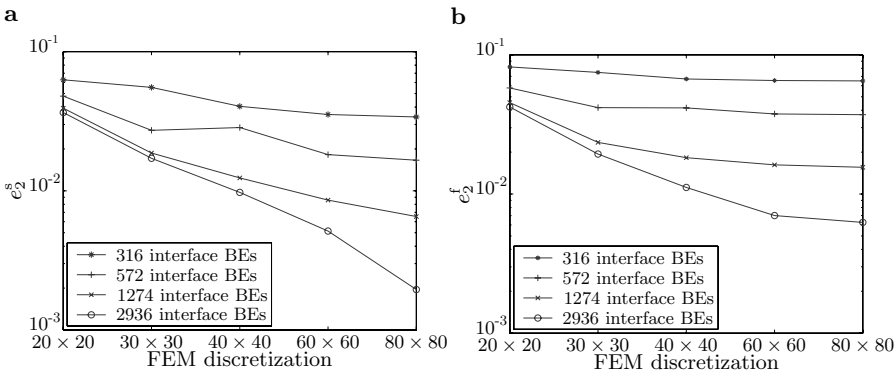


Fig. 8. (**a**) Convergence of structure error e_2^{s} ; (**b**) Convergence of fluid error e_2^{f}

of the chosen BEM mesh. The behavior of the structure error also depends on the accuracy of the acoustic model. For the finest BEM mesh, one observes the quadratic convergence rate that is expected for the FEM plate formulation. The acoustic field is approximated well enough for plate discretizations up to 80×80 elements. For the coarser BEM meshes, the convergence of the structure error is affected by the fluid errors.

In the simulations presented in Table 1, a threshold residual of 10^{-4} is chosen for the outer GMRES iterations. For the approximation of the inner FEM system, the residual of the conjugate gradient solver is set to 10^{-7}. At this precision, the effect of inexact evaluation of the FEM part can be neglected. For the inner BEM system, GMRES is applied with a relaxation strategy according to (19) and diagonal preconditioning. The required number of outer iterations does not depend on the discretization of the finite element model, but only on the discretization of the boundary element model, i.e. the discretization of the Lagrange multiplier. With mesh refinement, a moderate increase of outer iterations can be observed. The total number of required BEM matrix-vector products increases significantly, since the conditioning of the inner systems deteriorates with boundary element mesh refinement. The influence of the relaxation strategy and preconditioning of the inner system on the efficiency of the solver is discussed in the following.

In Table 2 the numerical cost and the true residual are documented for the solution of the system arising from the simulation at 180 Hz using the boundary element mesh with 316 elements on the interface and 20×20 finite plate elements. The target residual of the outer GMRES is set to 10^{-4}. Using the relaxation scheme, the numerical cost is reduced from 2018 BEM matrix-vector products to 1678 BEM matrix-vector products, i.e. the computing time is reduced by more than 15%.

The relation (19) is found to be a sharp limit for restricting the residual gap of GMRES. As shown in Table 2, reducing the required precision for the inner BEM systems by a factor of two and four, respectively, the true

residual increases to $1.29 \cdot 10^{-4}$ and $3.22 \cdot 10^{-4}$. The number of required BEM matrix-vector products is reduced accordingly, when allowing a residual gap. However, this scheme is not recommended, since it is usually more efficient to set a higher outer residual and to restrict the residual gap.

The convergence of the outer GMRES iterations for the fixed and the relaxation strategy are compared in Fig. 9. The achieved residuals are identical for the two strategies approving the relaxation according to (19). The reduction of required matrix-vector products for the relaxation strategy takes place as the outer residual decreases. The overall saving is moderate, since GMRES shows the typical superlinear convergence behavior, i.e. the majority of outer GMRES iterations is required while the residual is not yet reduced significantly.

Table 2. Choice of residual $\epsilon_k^{\text{inner}}$ for solution of BEM system. Target residual for inexact Uzawa algorithm $\epsilon^{\text{outer}} = 10^{-4}$

$\epsilon_k^{\text{inner}}$	outer iterations	BEM matrix-vector products	GMRES residual	true residual
fixed strategy				
$c_k \epsilon^{\text{outer}}$	28	2018	$9.06 \cdot 10^{-5}$	$1.03 \cdot 10^{-4}$
relaxation strategy				
$c_k \epsilon^{\text{outer}} / \lvert\gamma_k\rvert$	28	1678	$9.06 \cdot 10^{-5}$	$1.01 \cdot 10^{-4}$
$2c_k \epsilon^{\text{outer}} / \lvert\gamma_k\rvert$	28	1604	$9.09 \cdot 10^{-5}$	$1.29 \cdot 10^{-4}$
$4c_k \epsilon^{\text{outer}} / \lvert\gamma_k\rvert$	28	1520	$9.09 \cdot 10^{-5}$	$3.22 \cdot 10^{-4}$
$\epsilon^{\text{outer}} = 3 \cdot 10^{-4}$				
$c_k \epsilon^{\text{outer}} / \lvert\gamma_k\rvert$	26	1487	$2.09 \cdot 10^{-4}$	$2.92 \cdot 10^{-4}$

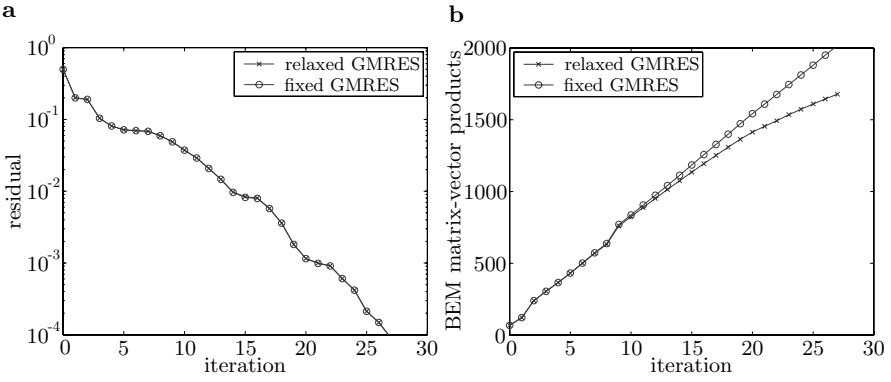


Fig. 9. Convergence (**a**) and BEM matrix-vector products (**b**) of inexact Uzawa solver

The acceleration of the solution of the inner BEM systems by preconditioning is essential for the efficiency of the proposed solver. The required number of matrix-vector products for approximating the inner BEM system with GMRES without, with diagonal, and with approximate inverse preconditioning is plotted in Fig. 10. Without preconditioning, a very high number of matrix-vector products is observed, which is prohibitive for a practical applications. Diagonal preconditioning and approximate inverse preconditioning, both yield a significant improvement. The approximate inverse approach performs superior and is recommended for the solution of coupled BEM systems.

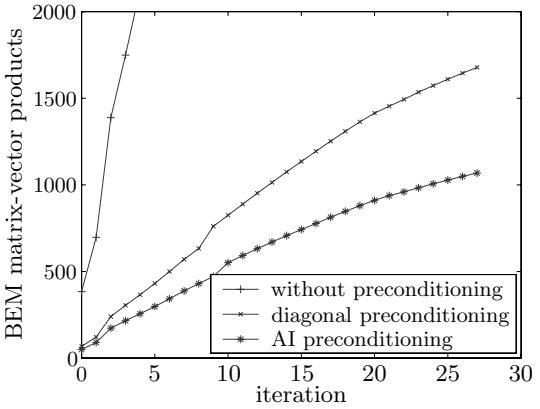


Fig. 10. Preconditioning of inner BEM system for inexact Uzawa algorithm with relaxed GMRES iterations

The importance of the modification of the Lagrange multiplier ansatz space as illustrated in Fig. 2 can be demonstrated very clearly for simulations on a regular grid. In Fig. 11, the pressure on the interface at 180 Hz is plotted for computations using a linear ansatz space for the Lagrange multiplier on the left and using the modified ansatz space on the right. The original Lagrange multipliers yield a rotated checkerboard pattern that is typical for a violated inf-sup condition.

6.2 Sound Radiation by a Submerged Plate

A practical application of the mortar FEM-BEM coupling algorithm is the simulation of experiments on active structural-acoustic control in the hydroacoustic lab at the Institute of Applied and Experimental Mechanics [7]. For the chosen example, a rectangular plate ($0.475\,\mathrm{m} \times 0.48\,\mathrm{m}$, $t = 1.5\,\mathrm{mm}$) is mounted on the box depicted in Fig. 12 and submerged in the water pool.

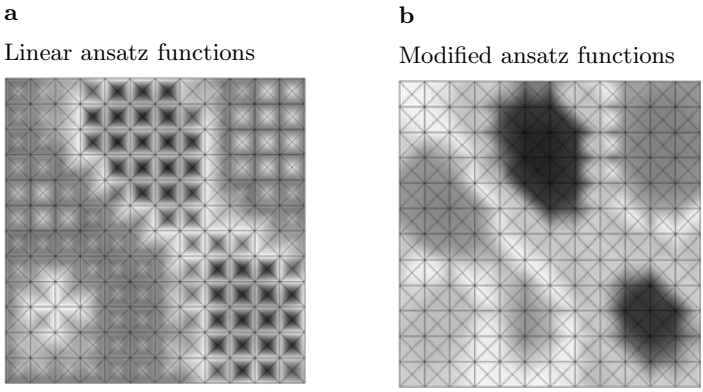


Fig. 11. Lagrange multiplier on coupling interface

Plate vibrations are excited by a shaker inside the box, and the acceleration on the plate as well as the acoustic pressure in the pool are recorded.

For the simulations, a BEM mesh of the pool and the box is set up that consists of 4726 triangular elements and that is shown without the top-surface elements in Fig. 13. The lower left hand corner of the plate is chosen as origin for the displayed coordinate system. Homogeneous Dirichlet boundary conditions are applied on the pressure-free surface of the pool, whereas the pool walls are modeled as rigid, i.e. homogeneous Neumann boundary conditions are applied. The plate is discretized using 16×16 finite elements and is assumed to be clamped onto the rigid box.

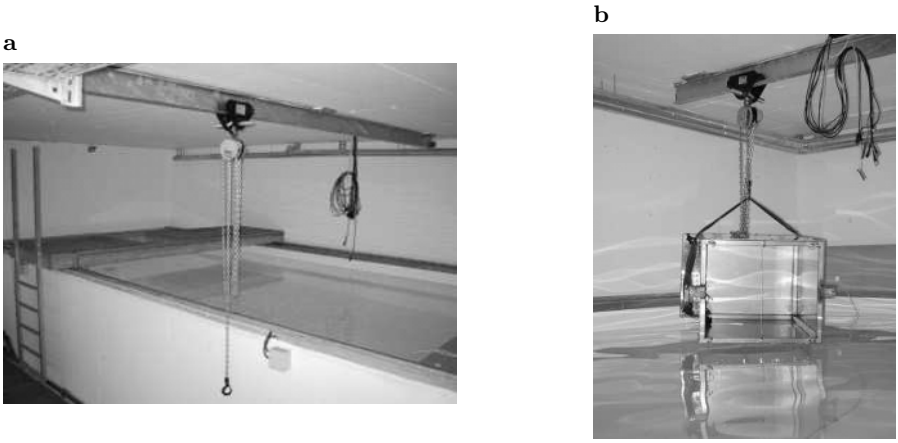


Fig. 12. (**a**) Hydro-acoustic lab; (**b**) Submergeable box

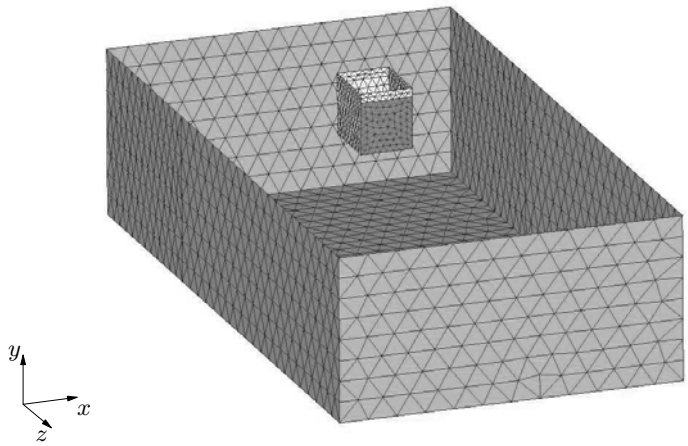


Fig. 13. BEM model of pool and hydro-acoustic box

In Fig. 14 the simulated and experimental frequency response functions for the plate displacement at position (0.42, 0.32) m and the acoustic pressure in the pool at position (0.15, −0.50, 0.51) m are plotted. The shaker is mounted at the position (0.20, 0.11) m on the plate and the forcing amplitude is used to normalize the plotted values. One notices that the principal behavior of simulation and experiment correlates. The differences can be traced back to various sources. The peaks in the experimental frequency response functions at 55 Hz, 71 Hz, and 113 Hz are due to resonance frequencies of the pool's front wall that is modeled as rigid in the FEM-BEM simulation. The influence of mounting the plate onto the box plays a dominant role as well: sealing material and fixing screws introduce a vast amount of model uncertainties. Finally, the acoustic boundary conditions of the pool walls must be determined in more detail. In this context, the simulation results are satisfying, however, more effort should be placed on the modeling of the system pool-box-plate.

7 Fuzzy BEM

The methodology presented in the previous sections enables the efficient simulation of structure-acoustic field interaction for industrial-scale applications. However, to achieve precise results for the simulation of real-world systems, exact values for all input parameters should be available. In practice, these exact values can often not be provided, and the model parameters, such as material properties and boundary conditions, commonly exhibit a rather high degree of uncertainty. Furthermore, uncertainties are introduced through unmodeled effects and model simplification like in the example of the submerged

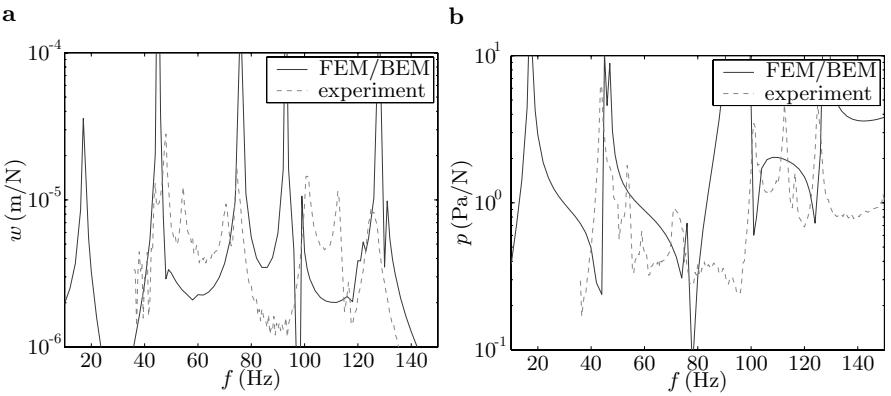


Fig. 14. Experimental and computed frequency response of submerged plate

plate in Sect. 6.2. Therefore, a methodology is presented to include uncertainties into the modeling procedure.

A very practical approach to modeling, simulation and analysis of uncertain systems is the numerical implementation of the uncertain model parameters as fuzzy numbers, and subsequently, the use of fuzzy arithmetic based on the transformation method, which has been proposed by Hanss in [16, 17]. By this technique, the complete information about the uncertainties is included in the model, and one can demonstrate how the uncertainties are propagated through the calculation procedure. The transformation method avoids the possibly serious drawbacks of conventional fuzzy arithmetic [16, 18] and allows the determination of degrees of influence of each fuzzy parameter, quantifying the proportion to which the uncertainty of each model parameter contributes to the overall uncertainty of the model output.

The transformation method has already been applied to various engineering applications in different disciplines [18]. In particular, it has been used to evaluate complex FE-models [19, 20], but to the knowledge of the authors it has not yet been applied on BE-models.

7.1 Fuzzy Sets and Fuzzy Numbers

The theory of fuzzy sets was introduced by Zadeh [27] as a generalization of classical set theory. While elements x of a universal set U do either entirely belong or not belong to a specific crisp set A, they are allowed to be contained to a certain degree in a specific fuzzy set $\tilde{A}$. For the description of fuzzy sets, the characteristic function μ_A of a crisp set A, where $\mu_A(x) \in \{0, 1\}$ with 1 indicating membership and 0 non-membership, can be generalized to a membership function $\mu_{\tilde{A}}$ for a fuzzy set $\tilde{A}$, where $\mu_{\tilde{A}}(x) \in [0, 1]$.

Fuzzy numbers are a special class of fuzzy sets defined on the universal set $\mathbb{R}$ showing some specific properties. Fuzzy numbers are convex fuzzy sets, for whom $\mu_{\tilde{A}}(x) = 1$ is true only for one single value $x = \bar{x} \in \mathbb{R}$, the so-called

center value or nominal value $\bar{x}$. Frequently used fuzzy numbers (Fig. 15) are fuzzy numbers of triangular (linear) shape $\tilde{p}_1$, which can be defined by the membership function

$$\mu_{\tilde{p}_1}(x) = \begin{cases} 0 & \text{for} \quad x \le \bar{x}_1 - \alpha_\mathrm{l} \\ 1 + (x - \bar{x}_1)/\alpha_\mathrm{l} & \text{for} \quad \bar{x}_1 - \alpha_\mathrm{l} < x < \bar{x}_1 \\ 1 - (x - \bar{x}_1)/\alpha_\mathrm{r} & \text{for} \quad \bar{x}_1 \le x < \bar{x}_1 + \alpha_\mathrm{r} \\ 0 & \text{for} \quad x \ge \bar{x}_1 + \alpha_\mathrm{r} \end{cases} \quad \text{or} \tag{35}$$

$$\mu_{\tilde{p}_1}(x) = \min\left\{ \max\left[0, 1 - (\bar{x}_1 - x)/\alpha_\mathrm{l}\right], \max\left[0, 1 - (x - \bar{x}_1)/\alpha_\mathrm{r}\right] \right\} \quad \forall\, x \in \mathbb{R} \tag{36}$$

and fuzzy numbers of Gaussian shape $\tilde{p}_2$, which can be defined by the membership function

$$\mu_{\tilde{p}_2}(x) = \begin{cases} \exp\left[-(x - \bar{x}_2)^2/(2\,\sigma_\mathrm{l}^2)\right] & \text{for} \quad x < \bar{x}_2 \\ \exp\left[-(x - \bar{x}_2)^2/(2\,\sigma_\mathrm{r}^2)\right] & \text{for} \quad x \ge \bar{x}_2 \end{cases} \quad \forall\, x \in \mathbb{R}\,. \tag{37}$$

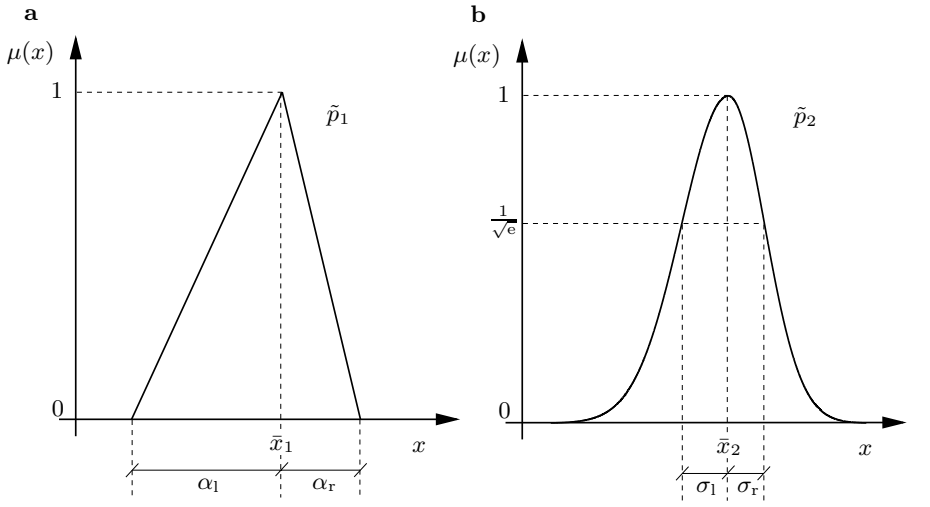


Fig. 15. (**a**) Triangular fuzzy number $\tilde{p}_1$; (**b**) Gaussian fuzzy number $\tilde{p}_2$

7.2 The Transformation Method

For the evaluation of systems with uncertain, fuzzy-valued model parameters, the transformation method can be used. The method is available in a general, a reduced, and an extended form [16, 17]. Assuming the uncertain system to

be characterized by n fuzzy-valued model parameters $\tilde{p}_i$, $i = 1, 2, \ldots, n$, the major steps of the method can briefly be described as follows [18]:

In the first step, each fuzzy number $\tilde{p}_i$ is discretized into a number of intervals $X_i^{(j)} = [a_i^{(j)}, b_i^{(j)}]$, corresponding to the levels μ_j, $j = 0, 1, \ldots, m$, of membership that result from subdividing the range of membership equally spaced by $\Delta\mu = 1/m$ (Fig. 16). In a second step, the input intervals $X_i^{(j)}$, $i = 1, 2, \ldots, n$, $j = 0, 1, \ldots, m$, are transformed to arrays $\hat{X}_i^{(j)}$ that are obtained from the upper and lower interval bounds after the application of a well-defined combinatorial scheme. Each of the columns of these arrays represents a specific sample of possible parameter combinations and serves as a crisp input parameter set to the problem to be evaluated. As a result of the evaluation of the model for the input arrays $\hat{X}_i^{(j)}$, output arrays $\hat{Z}^{(j)}$ are obtained which are then retransformed to the output intervals $Z^{(j)} = [a^{(j)}, b^{(j)}]$ for each membership level μ_j and finally recomposed to the fuzzy-valued output $\tilde{q}$ of the system. This fuzzy-valued output quantifies the overall uncertainty of

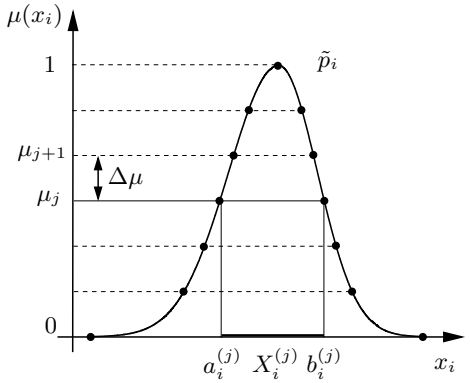


Fig. 16. Decomposition of a fuzzy number $\tilde{p}_i$

the simulation result generated by the interaction of the different uncertain input parameters. In addition, the transformation method can be used to analyse the influence of each fuzzy-valued input parameter $\tilde{p}_i$ on the overall fuzziness of the model output. For this purpose, so-called gain factors $\eta_i^{(j)}$ have been introduced that express the contribution of the uncertainty of the ith model parameter $\tilde{p}_i$ to the overall uncertainty of the model output $\tilde{q}$ at the membership level μ_j.

7.3 Implementation and Outlook

One advantage of the transformation method is its characteristic property of reducing fuzzy arithmetic to multiple crisp number evaluations of the simu-

lation model. This entails that the method can be implemented without major problems into the existing BEM-FEM simulation environment. The steps of decomposition and transformation of the fuzzy input as well as retransformation and recomposition of the fuzzy output can be coupled to existing software by a separate pre- and postprocessing tool. For this purpose, the program FAMOUS (**F**uzzy **A**rithmetical **M**odeling **O**f **U**ncertain **S**ystems) has been developed, which carries out the transformation method and provides an interface to the BEM-FEM simulation software.

As a first step, the fuzzy simulation tool will be used for the simulation and analysis of different real-world acoustic-structure interaction problems with uncertain input parameters. Later on, the tool can be extended to cover also the inverse problem which consists in calculating the exciting structural vibration and structural properties from the acoustic response. For this generally ill-posed problem, uncertainties have to be handled with special care. An approach to inverse fuzzy arithmetic, which consists basically of an appropriate combination of the forward simulation and the analysis part of the transformation method, has been presented in [18]. Advancing and generalizing this approach might render possible to identify structural uncertainties starting from uncertain data for the acoustic field variables.

8 Conclusions

For the simulation of acoustic-structure interaction, a mortar BEM-FEM coupling algorithm is derived. The mortar scheme is based on Lagrange multipliers and allows the coupling of non-conforming discretizations. This can improve the efficiency of the coupled simulation since specialized ansatz functions and adapted element sizes can be used in the subdomains. For the solution of the mortar saddle point problem, a modified Uzawa algorithm is implemented. A GMRES method is employed for the outer iterations on the reduced system for the Lagrange multiplier whereas the inner FEM and BEM systems are approximated by conjugate gradient and relaxed GMRES iterations, respectively. The BEM system is efficiently preconditioned using an approximate inverse approach. The solution method allows the use of the multipole BEM for coupled problems which reduces the memory requirements significantly. However, due to the nested iteration scheme, the iteration time can be rather long. Further work should thus be devoted to the development of suitable preconditioners and solvers for the coupled system that reduce the number of required matrix-vector products.

The transformation method is presented as a tool for the simulation and analysis of acoustic-structure interaction models with uncertain, fuzzy-valued parameters. The method should thoroughly be tested on real-world applications of this field and further effort should be put into the development of an inverse algorithm.

The presented work is seen as a contribution to promote fast BEM for industrial applications. It is shown that the use of advanced numerical algorithms can significantly improve the efficiency of structural-acoustic simulations. However, besides the crucial point of computing cost, reliability and simplicity of use must also be kept in mind for a successful application. It is reminded that even the standard BEM is not a black-box simulation tool which yields results on a simple click, but a fair amount of insight into the method is required to apply it to engineering problems. So the additional complexity when employing the fast multipole BEM is assessed small compared to the savings offered by the method and its practical application can be encouraged.

9 Acknowledgement

Section 7 has been coauthored by Astrid Pieringer.

References

1. C. Bernardi, Y. Maday, and A. T. Patera. A new nonconforming approach to domain decomposition: the mortar element method. In H. B. et al., editor, *Nonlinear partial differential equations and their applications*, pages 13–51. Pitman, 1994.
2. A. Bouras and V. Frayssé. A relaxation strategy for inexact matrix-vector products for Krylov methods. CERFACS Technical Report TR/PA/00/15, 2000.
3. F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, New York, 1991.
4. K. Chen and P. Harris. Efficient preconditioners for iterative solution of the boundary element equations for the three-dimensional Helmholtz equation. *Applied Numerical Mathematics*, 36:475–489, 2001.
5. R. Coifman, V. Rokhlin, and S. Wandzura. The fast multipole method for the wave equation: A pedestrian prescription. *IEEE Antennas and Propagation Magazine*, 35:7–12, 1993.
6. E. Darve. The fast multipole method I: error analysis and asymptotic complexity. *SIAM Journal of Numerical Analysis*, 38(1):98–128, 2000.
7. O. Fein and L. Gaul. Reduction of sound emission from submerged structures using piezoelectric sensors and actuators. In *Proceedings of the 9th International Congress on Sound and Vibration*, pages 316–323, Orlando, Fl, 2002.
8. M. Fischer. The Fast Multipole Boundary Element Method and its Application to Structure-Acoustic Field Interaction. PhD thesis, Institut A für Mechanik, Universität Stuttgart, 2004.
9. M. Fischer, U. Gauger, and L. Gaul. A multipole Galerkin boundary element method for acoustics. *Engineering Analysis with Boundary Elements*, 28(2):155–162, 2004.
10. M. Fischer and L. Gaul. Application of the fast multipole BEM for structural-acoustics simulations. *Journal of Computational Acoustics*, 13(1):87–98, 2005.

11. M. Fischer and L. Gaul. Fast BEM-FEM mortar coupling for acoustic-structure interaction. *International Journal for Numerical Methods in Engineering*, 62(12):1677–1690, 2005.
12. M. Fischer, H. Perfahl, and L. Gaul. Approximate inverse preconditioning for the fast multipole BEM in acoustics. *Computing and Visualization in Science*, 8:169–177, 2005.
13. L. Gaul, M. Kögl, and M. Wagner. *Boundary Element Methods for Engineers and Scientists*. Springer-Verlag, Berlin, 2003.
14. L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73:325–348, 1987.
15. M. F. Gyure and M. A. Stalzer. A prescription for the multilevel Helmholtz FMM. *IEEE Computational Science & Engineering*, 5(3):39–47, 1998.
16. M. Hanss. The transformation method for the simulation and analysis of systems with uncertain parameters. *Fuzzy Sets and Systems*, 130(3):277–289, 2002.
17. M. Hanss. The extended transformation method for the simulation and analysis of fuzzy-parameterized models. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(6):711–727, 2003.
18. M. Hanss. *Applied Fuzzy Arithmetic – An Introduction with Engineering Applications*. Springer, Berlin, 2005.
19. M. Hanss, U. Gauger, and L. Gaul. Modeling and simulation of vibrating automotive components with uncertain parameters using fuzzy arithmetic. In *Proceedings of the 23rd International Modal Analysis Conference (IMAC XXIII)*, Orlando, Fl, 2005.
20. M. Hanss, A. Pieringer, J. Becker, M. Maess, and L. Gaul. Fuzzy analysis of actively damped piezoelectric structures with uncertainties. In *Proceedings of the 24rd International Modal Analysis Conference (IMAC XXIV)*, St.Louis, Mo, 2006.
21. S. Koc, J. M. Song, and W. C. Chew. Error analysis for the numerical evaluation of the diagonal forms of the scalar spherical additon theorem. *SIAM Journal of Numerical Analysis*, 36(3):906–921, 1999.
22. A. Pretlove. Forced vibrations of a rectangular panel backed by a closed rectangular cavity. *Journal of Sound and Vibration*, 3(3):252–261, 1966.
23. V. Rokhlin. Diagonal forms of translation operators for the Helmholtz equation in three dimensions. *Applied and Computational Harmonic Analysis*, 1:82–93, 1993.
24. Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, second edition, 2003.
25. V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM Journal of Scientific Computing*, 25(2):454–477, 2003.
26. J. van den Eshof and G. L. G. Sleijpen. Inexact Krylov subspace methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 26(1):125–153, 2004.
27. L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.
28. O. C. Zienkiewicz and R. L. Taylor. *The Finite Element Method*. Butterworth-Heinemann, Oxford, UK, fifth edition, 2000.

Dynamics of Poured Polyhedra of Different Shape*

Peter Eberhard and Beate Muth

Institute of Engineering and Computational Mechanics, University of Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany
eberhard@itm.uni-stuttgart.de
muth@itm.uni-stuttgart.de

Summary. In this work different models for the simulation of granular material are investigated. Due to the fact that a detailed simulation considering complex shapes of the bodies is very time consuming, methods from molecular dynamics (MD) are used. By means of granular matter simulations large systems like bulk solids, silicic or sand e.g. in silos, or other systems like carriage systems can be investigated. The MD method is combined here with ideas from multibody systems, so that it is possible to determine the collisions and the motion of differently shaped bodies. Here, particles are modeled as rigid bodies, where the interaction is described by means of elastic contact forces. The particles considered are convex or non-convex three-dimensional polyhedra.

After neighboring body pairs are found, the very time consuming collision detection for polygonal bodies can be accomplished. In order to check whether there is a collision between two neighboring body pairs, it has to be checked, whether there is a vertex of one body inside the other body. It will be focused also on the calculation of the contact forces in normal and tangential direction.

Keywords: Molecular dynamics, contact force, neighborhood search, collision detection

1 Introduction

In many industrial processes particles are stored, transported and processed, where a great number of particles can be involved. For many such processes, as e.g. filling mechanisms or conveyers, the procedure can be optimized by means of a skillful particle flow and an ideal configuration of the facilities. For such an optimization an accurate simulation of the process is required. In order to determine the dynamical behavior of systems consisting of many objects or particles, several fully developed approaches exist. The main differences are

* Research Project B7 "Hybrid Multibody Simulations of Poured Particles"

the assumptions about the particle shapes and their behavior on collisions. In a simulation bodies can be treated as (i) perfectly rigid objects, as (ii) non-deformable objects with (small) overlaps at the contacts, or as (iii) deformable bodies with a peculiar contact dynamics.

Systems consisting of bodies with negligible deformations, case (i), can be described by means of the so-called multibody system method (MBS) [19, 26], and mass point systems may be regarded as a special case of the MBS. Using a so-called molecular dynamics (MD) approach [1, 4, 24], case (ii), fake-deformable bodies can be represented by a collection of non-deformable particles connected by springs [21]. For more advanced studies of flexible, deformable bodies, (iii), usually the Finite Element Method (FEM), see [5, 20] or the boundary element method are used. Each of these methods has its own advantages and disadvantages. While the MBS is in general characterized by relatively short computation times due to a small number of degrees of freedom, deformations usually are not handled. On the other hand, systems investigated using FEM have a large number of degrees of freedom that yield a rather extensive number of equations of motion, but deformations are taken into account. The MD method can be seen as a compromise in so far that the number of degrees of freedom is kept small by assuming overlap dependent contact force laws that rely on certain assumptions and do not take the eigen-modes of a particle into account.

2 Discrete Element Method

The discrete element method (DEM) was originally brought up in [4], which used the method for granular media consisting of spherical bodies. The modeled systems consist of many distinct particles which displace independently from one another. He called the model distinct element method , but today it is often referred to as discrete element method. Its basic ideas are the same as for the molecular dynamics (MD) method, which is usually used for point masses.

The method bases on the assumption that deformations of individual particles are small in comparison to the deformation of the total granular assembly. Hence, the exact deformation of the particles is not important in order to obtain a good approximation of the mechanical behavior. In this method, the particles are allowed to overlap at contact points, [4]. Then, the size of the overlap is related to the contact force but is small compared to the particle size. For these small overlaps, which are sometimes interpreted to model the elasticity in the area of contact and, therefore, of the contacting materials, constitutive equations may be used, [33].

Very efficient methods were developed for molecular dynamics simulations, and are, e.g., applied to the dynamic and static behavior of granular matter [10, 22, 30]. The MD method is often used for different kinds of particle systems consisting of solids, molecules, charge carriers and also for gas or

fluid, [1, 13, 22]. Here it can be dealt with motions and contacts of many thousands of particles. Besides body forces like the gravitational forces, one typically has contact forces resulting from the boundaries of the system and from other particles within the system. The formulation of the contact forces between the different bodies is based on simple models in order to keep the calculation times within a feasible range.

Apart from the above mentioned repulsive, and dissipative contact forces also attractive forces may occur due to cohesion, gravity, or electro-static interactions. Long-range forces are far from the scope of this study, where we focus on short range, contact interactions only. Here, the MD method is applied to polygonal bodies in 2D and polyhedral bodies in 3D systems.

3 Neighborhood Search

In the following, three methods are explained which allow for determination of neighboring body pairs. It is the goal of these methods to reduce the high effort needed for collision detection, which is of $O(n^2)$ for n bodies, if all possible body pairs of a system have to be checked. Using a neighborhood search method, detected neighboring particle pairs are stored in a neighbor data structure (NDS). After this procedure of pre-sorting neighboring body pairs, the very time consuming collision detection only has to be done for stored particle pairs.

3.1 Verlet List

The fundamental idea of neighborhood search comes from Verlet, [29], who described the Verlet method (VL) in its basic form. Main idea of this method is to provide the information of neighboring bodies for each particle, [1, 29]. By means of this method particles are stored as neighboring pairs if the gap between them does not exceed a certain distance.

Thereby, the costs to investigate neighboring particles is still of order $O(n^2)$. However, these created neighborhood lists only have to be updated at certain time intervals, since special neighborhood zones, a kind of buffer areas of sufficient size, are used while creating these lists. Often circles or balls are used around the respective particles, see Fig. 1. The smaller the radius r_l of these areas, the smaller the NDS, which is provided for each particle. However, the greater these zones, the less often these lists have to be updated. To achieve a high efficiency, a compromise has to be found between the time distances, when the lists have to be updated and the size of the neighborhood zones. The size of r_l has to be chosen depending on the size and velocity of the bodies and the density of the system. Values for the update time step and the optimal size of r_l are often chosen intuitively. One possibility to choose these values is explained in [6]. There, the total motion of each body is summed

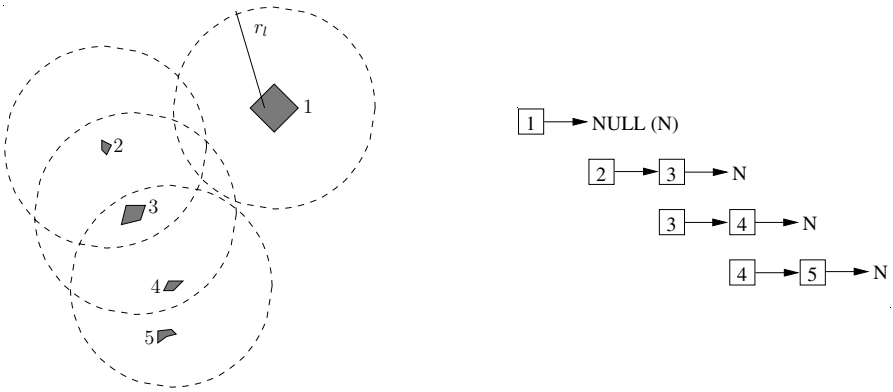


Fig. 1. Verlet circles and storage of bodies in lists, see [1, 15]

up vectorially. Then, as soon as the total sum of the motion of two bodies is greater than r_l, the lists have to be updated.

At the right-hand side of Fig. 1 the lists are shown, which are obtained for each particle. Each particle pair has to be stored only once. To ensure that each body pair is stored only once, all neighboring bodies with a higher number than the body itself are stored in its list. Therefore, for the "last" body these lists will always be empty and can be omitted.

One possibility to store the neighboring body pairs is a linked matrix, [1, 15]. The basic idea is, to store the information of a body pair i, j at position i, j of the matrix. Since each body pair is stored only once this will lead to an upper triangular matrix. By means of linked lists in C it is possible to reduce the necessary memory space by ignoring non-colliding body pairs and the lower triangular matrix. This is denoted at the right-hand side of Fig. 1 and is clarified in Fig. 2.

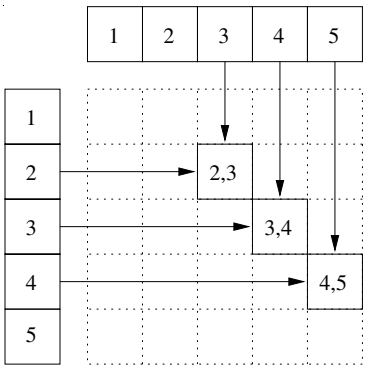


Fig. 2. Memory space allocation of the NDS for the information of neighboring body pairs at their respective location

3.2 Linked Cell Method

An alternative method is the linked cell method (LC), see [23]. There, the system is divided into a regular lattice of (e.g. for a cubic system) $m \times m \times m$ cells, see [1]. Non-cubic systems can be used, too, then the cell size has to be chosen such that the entire domain is covered with cells. Similar to the optimal neighborhood size r_l of Sect. 3.1, the optimal cell size depends on the different system parameters, e.g. the velocity of the particles, the density of the system etc. However, the chosen cell size must be at least greater than the largest particle of the system. Often the cell size is chosen in such a way that approximately three particles fit into one cell.

In contrast to the neighborhood zones of the VL, the cells here are fixed in space. Thus, the number of cells and the number of particles are not the same, but bodies are contained in cells temporarily. A two dimensional example is shown in Fig. 3, consisting of 70 particles. However, these cells, too, meet the demand of a kind of large buffer area around a special body.

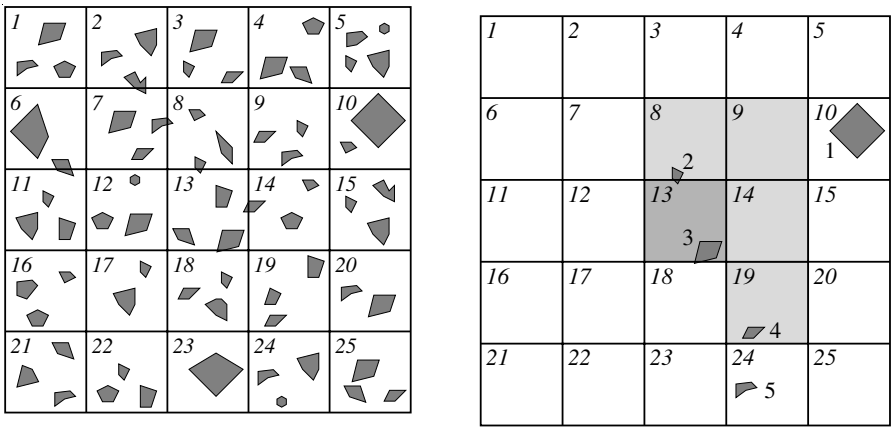


Fig. 3. System consisting of 70 particles, which is divided into cells (left-hand side) and five of these particles, where those cells are marked that are important for the neighbor list of bodies of cell 13 (right-hand side)

Here, bodies are treated as neighboring bodies, if they are located in the same cell or in the directly neighboring cells. For a particle of cell 13 that means that cells 7-9,12-14 and 17-19 contain neighboring particles. Therefore, for a planar system, there are 9 cells, for a three dimensional system 27 cells that contain neighbors. However, since each particle pair is ought to be stored only once, only half of the neighboring cells have to be searched for neighboring bodies, see right-hand side of Fig. 3. Then, for a d-dimensional system $(3^d + 1)/2$ cells have to be analyzed. To show this more clearly, at the right-hand side of Fig. 3 five bodies are shown exemplary. Therefore, this is now

the system of Fig. 1. Here, neighbors of body 3 are then bodies 2 and 4, which are inside the shaded area. To ensure that neighboring body pairs are stored only once only half of the neighboring cells are searched, as already mentioned before. Now we concentrate on what is going on within the single cell, here cell 13. For bodies within the same cell only bodies of a higher body-number are stored in the neighborhood list. This is shown in Fig. 3, where for each particle neighbors are searched inside the same cell and inside the cells to the right and top. Then, for this example bodies 1 and 2 will have no neighboring bodies, for particle three the pairs 2,3 and 3,4 are found, for particle 4 there will be no further partner and for particle 5 it is pair 4,5. So in this case there will be found the same neighboring body pairs as for the search in Sect. 3.1.

3.3 Linked Linear List

The third method described in the following is called the linked linear list (LLL), [2, 27]. The main idea here is different from the previously described methods. In a first step, bounding boxes are applied to each body, where the edges of the boxes are parallel to the coordinate axes. The test, whether two boxes overlap can be done with less effort than the collision detection for complex bodies. Bodies, whose bounding boxes overlap are stored as neighboring bodies. At the left-hand side of Fig. 4, the system of Fig. 1 is shown again. Here it can be seen directly that for this method none of the body pairs are regarded as neighbors. At the right-hand side of this Figure, the two particles of cell 10 of Fig. 3 are depicted. The bounding boxes of these two particles do overlap, so this body pair is stored in the NDS. It is clear, that the NDS for this method contains fewer body pairs, since no buffer zones exist between particles. Therefore, the NDS has to be updated in each time step. This update, however, can be done by means of special lists, so that nearly a linear behavior of the effort over the number of particles can be obtained.

In order to do the updating efficiently, in a next step the created bounding boxes are projected onto the coordinate axes, see the right-hand side of Fig. 4. After the projection only the order of the beginnings "b" and endings "e" of these projections is of interest. For all three axes these sequences are then stored in separate lists. For a three dimensional system, three such lists are created, the length of such a list is two times the number of particles of the system. The basic idea is, that as soon as there is the beginning or the ending of a body inbetween the beginning and the ending of another body, the bounding boxes of these two bodies are overlapping and the pair is stored in the NDS.

For the first time step of the simulation these lists have to be created. This is a very time consuming procedure and despite using efficient sorting methods this is of order $O(n\log n)$, see [2]. However, further on in the simulation, the already existing lists only have to be updated, which can be done very efficiently since from one time step to the next there are usually only

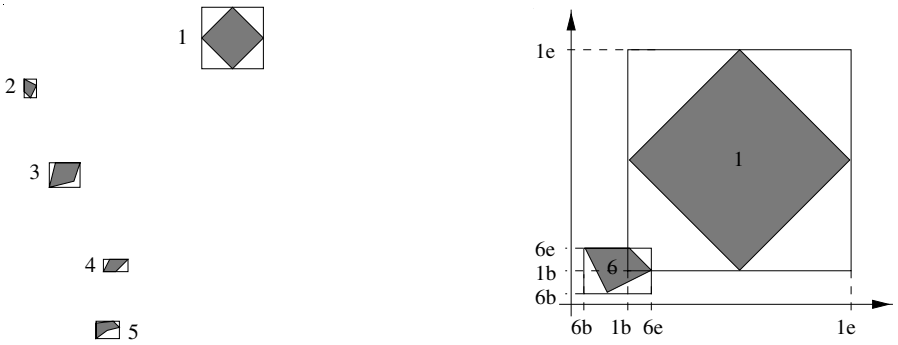


Fig. 4. Bounding boxes of bodies 1 and 6 provide no neighbors for the example of Fig. 1, however, neighbors are the two bodies of cell 10 from the left-hand side of Fig. 3. See also [2, 27]

few permutations in the sequences of the lists, see [2, 27]. That means, it is a matter of sorting a nearly sorted list. While sifting through the lists linearly the actual positions of the bodies are compared to the storage position in the list and eventually exchanged. By means of this procedure it is possible to sort the lists with an effort proportional to $O(n + s)$, see [2], where s is the number of required permutations. If the sequence has to be changed for a certain particle, four cases have to be distinguished, [27].

1. Two beginnings have to be exchanged, i.e. the bounding boxes of these two bodies have already been overlapping and continue to overlap, so there will be no change in the NDS. In the example of Fig. 4 this would happen, if e.g. body 6 is moving to the right.
2. Two endings have to be exchanged, which means again that the bounding boxes of these two bodies have already been overlapping and continue to overlap, so there will be no change in the NDS. In the example of Fig. 4 this would happen, if body 6 has been moving to the right a little further.
3. A beginning and a subsequent ending of two different bodies have to be exchanged, i.e. the bounding boxes have been overlapping but are now not overlapping any more. For this case the body pair has to be removed from the NDS. Here, body 6 in the example of Fig. 4 has moved a little to the left.
4. An ending and a subsequent beginning of another body have to be exchanged, i.e. the bounding boxes of two bodies are now overlapping and an entry of this body pair has to be made to the NDS.

3.4 Comparison of the Methods

The three techniques introduced shall in the following be compared with respect to the simulation times, see [15]. In order to keep the influence of the

different computers, of the different programming languages and compilers, as well as the influence resulting from the different programming styles as small as possible, different test series are used for the comparisons. The goal of these series is to keep some system properties as constant as possible and to change only some well defined influencing factors.

The program for the LC was programmed in C++ while the VL and LLL were programmed in C. For this reason and since the program runs were performed on different Linux PC's the results are normalized in such a way that for the smallest investigated number of particles for each system (where the influence of the used method is nearly negligible) scaling factors have been computed. Therefore, all curves in the following figures have the same starting point, see [15].

In a first example, a spatial monodisperse system is investigated. A series of 3D systems with an increasing number of particles of equal size, $R = 5.0 \times 10^{-4}\,\mathrm{m}$, and a proportionally growing space around the particles is analyzed. For the stiffness of the interactions it was chosen $k = 10^5\,\mathrm{N/m}$, the damping coefficient was set to zero, the density of the particles was $\rho = 10^{10}\,\mathrm{kg/m^3}$, and the time step for the integration was chosen as $\Delta t = 10^{-5}\,\mathrm{s}$. Here, systems of 512, 2197, 5832, 12167, 21952, 35937, and 59320 particles are investigated, where the linked cell size was kept equal $l_c = 2.0 \times 10^{-3}\,\mathrm{m}$ for each system. The behavior of the computation time with respect to the number of particles within the particular system is shown in Fig. 5 as log-log plot.

The more the number of particles rises (and with it the number of contacts during the simulation), the more the performance for VL worsens in compar-

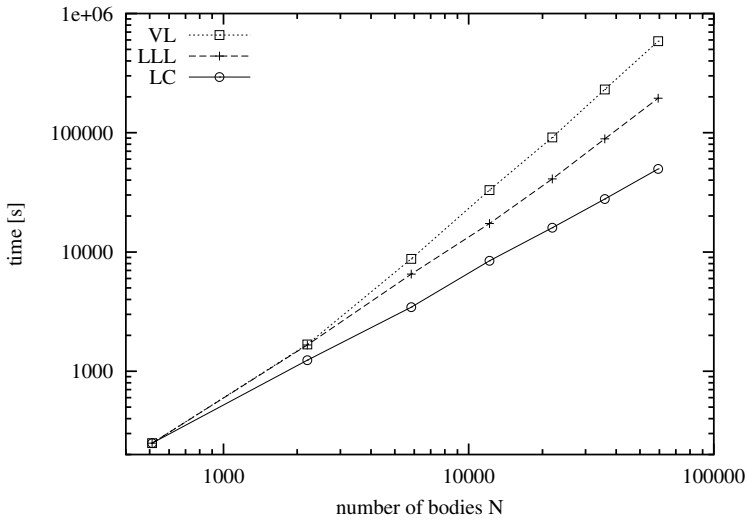


Fig. 5. Comparison results for an increasing 3D example

ison to LLL and LC. Here LC has the best performance. Clearly it can be seen, that the gradient is lower than for LLL. Due to clearness of the plot the approximating curves, $y \approx (0.027\ x)^{1.8}$ for VL, $y \approx (0.12\ x)^{1.36}$ for LLL and the polynomial for LC, $y \approx (0.2\ x)^{1.15}$, are not plotted in Fig. 5. Therefore, it can be said, that the behavior of LLL and LC both remain close to linear in contrast to the behavior for VL, which behaves almost quadratic.

In a next comparison, a spatial polydisperse system with a fracturing system size is investigated. In the previous comparison the system size is growing. In contrast to that, here the number of particles is increased, but no change of either the system or cell-size is undertaken, $l_c = 0.033\,\mathrm{m}$. Chosen parameters for this system are the stiffness $k = 4.0 \times 10^6\,\mathrm{N/m}$, damping coefficient $d = 0\,\mathrm{Ns/m}$, density of the bodies $\rho = 7000\,\mathrm{kg/m^3}$, and the time step for the integration $\Delta t = 4 \times 10^{-7}\,\mathrm{s}$. The systems can be seen as a series of fracture of some of the particles, where neither the volume enclosed in the system boundaries nor the volumetric content of the system is changed. This means that the volume fraction and the density of the system are unchanged, while the number of particles within the system is increased. In the first system 1000 particles are situated, with equal radii $R = 0.01\,\mathrm{m}$. Approximately half of these particles are now successively fractured in the next systems: There are about 500 particles of radius R, but approximately eight times 500 particles of radius $R/2$ (and thus an eighth of the original particle volume). The systems therefore contain

1000 particles (see Fig. 6 on the left),
4451 particles, about 4000 smaller bodies of $r = R/2$,
14676 particles, about 14000 particles of $r = R/3$,
32374 particles, about 31000 particles of $r = R/4$, and
65480 particles, about 65000 particles of $r = R/5$ (see Fig. 6 on the right).

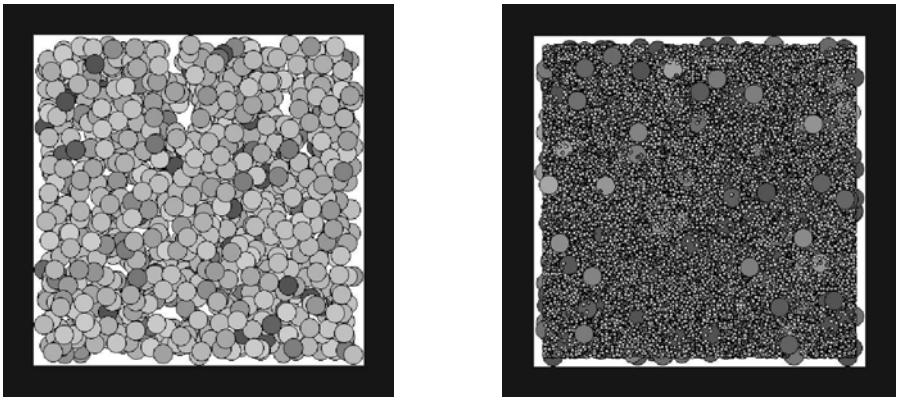


Fig. 6. A monodisperse system and a polydisperse system with $r = R/5$ of the same volume and density (volume fraction $\nu = 0.12$)

Here, R is the original radius of all particles of the first system and r the smaller one. For this system the computation time needed per particle is presented in Fig. 7 over the ratio of the radii that is a measure of the polydispersity of the system.

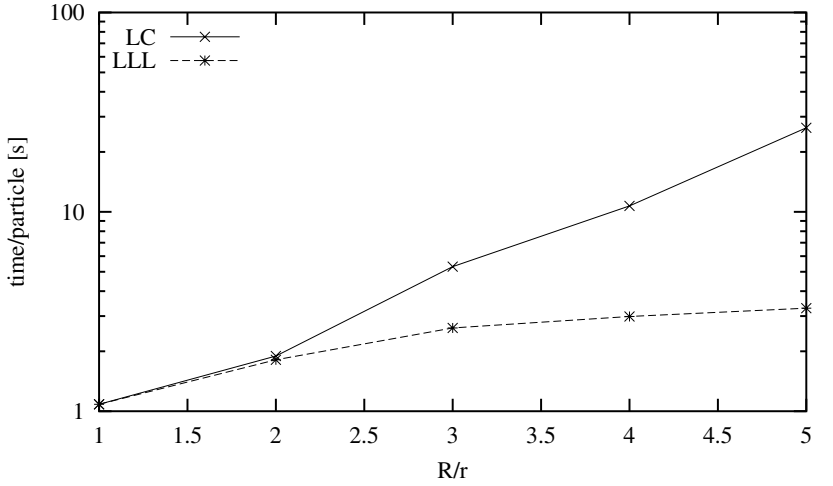


Fig. 7. Computation time for a fracturing system of constant density

It can be seen that the gradient of the curve over the polydispersity of the LC calculation is a lot higher than for LLL. Before scaling, so that both curves have the same starting point in the picture, the starting point of LC (for the monodisperse system) was underneath the starting point for LLL, and the curves were intersecting at about $R/r = 2.5$. This means that the LC method has advantages for quite monodisperse systems, while the LLL shows its advantages for polydisperse systems.

4 Collision Detection

If a neighboring body pair is found within the system, the respective bodies have to be checked for collision. Spherical bodies do not have to be discretized. For these bodies, the collision detection can be done very easily by simply comparing the distance of the centers with the sum of the radii. Also the contact geometry does not have to be determined, since the point for contact force application is the point which is on the connecting line of the two centers of masses and which is in the middle of the overlapping zone. The normal contact force is aligned parallel to that line, whereas the tangential contact force is lying on the plane which is rectangular to that line.

However, for polygonal bodies the collision detection is more complex. There exist several possibilities for an efficient collision detection. Two methods will be explained in the following [8, 16, 18], the ray crossing method and the fast multipole boundary element method. Both are methods where a point in polyhedron detection is done. Such methods search for discrete points within bodies, i.e. if such a method is used, there have to be enough discretization points at the surface of the bodies to ensure a reliable collision detection, but easily complex bodies can be investigated.

4.1 Ray Crossing Method

In this section, the ray crossing method is explained, which is a method in order to detect whether points are inside polyhedra, see [18]. The basic idea of this method is to generate an arbitrary ray from the respective point to infinity. Then, by means of the number of intersections of this ray with the surface of the body it can be detected whether the point is in or outside the body. A point P is located within the polyhedron, when the number of intersections is odd, see Fig. 8. In the same way it can be said that a point Q is outside a body, when the number of intersections of this point with the surface is even.

For this investigation the ray has to be tested against each part of the surface of the body, i.e. for a triangular surface discretization against each

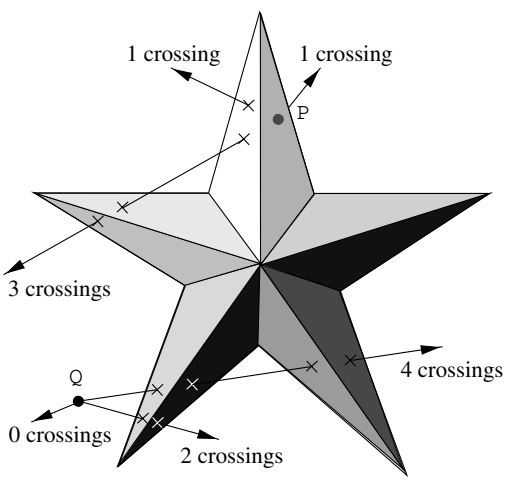


Fig. 8. Collision detection by means of checking a ray for each observed point of another body. Here, three possible rays are shown for a point P that is inside the body (1 or 3 crossings), and for a point Q that is located outside the body (0, 2, and 4 crossings)

surface triangle f. A problem of counting these ray crossings is the wide variety of possible degeneracies that could occur, see [18]. These degeneracies could be, e.g., a ray laying in a face, hitting a vertex or an edge, lying collinearly to an edge, etc. However, the treatment of such degeneracies must not be taken into account for the implementation. Since the direction of the ray is completely arbitrary and the creation of another ray is not time consuming, it is simply checked for each ray, whether such degeneracies exist, and if so, a new ray is created with random direction. Such a ray to infinity can be substituted by a line segment PR that is long enough, if its endpoint R is definitely outside the body.

The collision test then is accomplished in several steps. In a first step, all points of the tested body 1 that are not within the bounding box of the other polyhedron 2 are discarded. For the remaining points, the random rays have to be generated and tested against all faces of body 2. So all faces f of body 2 are tested whether the created ray crosses them. Therefore, it has to be checked whether the ray crosses the plane in which that face is located, and, whether it crosses within the boundary of the face. This is done in a complex procedure of several steps, see [16, 18, 32]. In the end, the total number of intersections of the ray with these faces f has to be added up, in order to detect whether the point is inside or outside the polyhedron.

4.2 Double Layer Potential and Fast Multipole Boundary Element Method

Another possible approach that can be used for collision detection is based on a special property of the double layer potential of the boundary element method (BEM) and the fast multipole BEM (FMM) providing a fast calculation of boundary integral operators, [16]. The main idea is to compute a potential of a certain surface or boundary in evaluation points, which are here the surface nodes of other bodies [8]. Therefore, an evaluation point P, being e.g. the surface point of a body X, is investigated, and is tested for collision with a body Y, whose surface is discretized by a boundary mesh containing triangles.

The double layer potential is the boundary integral of the normal derivate of the Laplacian fundamental solution. Usually, it is applied to a density function u, which is chosen as $u = 1$ here. The main idea of this method is, that the double layer potential $J(P)$ then only takes the values

$$J(P) = \begin{cases} 1 & \text{for } P \in Y \ , \\ 0 & \text{for } P \notin Y \ , \\ \frac{1}{2} & \text{for } P \in \Gamma \ . \end{cases} \tag{1}$$

Therefore, this functional $J(P)$ can be used to decide how a point P is located with respect to the domain Y. The potential $J(P)$ can either be computed exactly in the case that the boundary of Y is described by plane

triangles. Then, this computation of the functional $J(P)$ has to be done for each of these N elements in order to know whether a point P is inside or outside the domain. If body X contains M points, the functional $J(P)$ is evaluated in M evaluation points P. Then, the corresponding effort will be of order $\mathcal{O}(NM)$ as the functional J has to be evaluated for all combinations. However, the FMM can be used in order to reduce this effort, see e.g. [25]. The combination of both methods is applied to collision detection methods in [16]. There, also comparisons are shown between this method, using the fast multipole approach, and the ray crossing method. These comparisons show basically, that for very complex bodies, the FMM is very efficient. However, for simpler bodies, the behavior of both methods concerning simulation time methods is very similar, see [16].

5 Contact Forces

For the calculation of the contact forces different approaches and interpretations exist. They basically consist of two ideas. Principle of the first idea is, e.g. for normal contact, the non-penetration condition between contacting bodies. For this case the dynamics of the system is independent of the modeling of the contact forces. For the contact force calculation, often methods are used where this non-penetration condition is fulfilled exactly. Therefore, in case of this first idea it is the duty of the contact treatment to minimize or even eliminate the unphysical penetration, see e.g. [5].

The totally different second idea is based on the concept that bodies are never totally rigid but at least their surface asperities always somewhat flexible. This flexibility is dependent on the material properties. In this approach the bodies are modeled as rigid bodies, where the penetration between the bodies is interpreted as elastic deformation of the surface asperities of the two contacting bodies, its depth being restricted by the material behavior of both bodies. This is done, e.g., in molecular dynamics (MD), [1, 13, 15, 24] but can also be used in constitutive equations for contact interfaces, see e.g. [9, 33]. Mathematically this corresponds to a penalty of homogenization approach.

5.1 Normal Contact Forces

In this first part, the forces in normal direction are considered. There are repulsive contact forces and attractive contact forces. The here described forces, however, occur only in the case of collision, i.e. long-range forces are not considered. For non-adhesive materials only repulsive contact forces have to be applied to the bodies. For adhesive materials, there has to be applied a possibly sticking contact force, [12]. In the following, two models for repulsive contact forces with dissipation are described and one method for the application of adhesive contact forces is mentioned.

Model for Dissipation by Means of the Coefficient of Restitution

The contact forces are calculated by means of the Kelvin-Voigt model , [9, 11]. There, the contact force consists of an elastic and a viscous part. The elastic part shows a direct relationship between the elastic contact force and the overlap of the two bodies. The absolute value of the viscous part may not be larger than the elastic part, since the total value of the contact force could then become negative for negative relative velocity. Therefore, in this model the viscous part of the contact force is dependent on the overlap as well. The contact force in normal direction can then be written as

$$\mathbf{f}_N = \underbrace{K\delta^n\,\mathbf{n}}_{\text{elastic force}} + \underbrace{K\delta^n\,\frac{3\,(1-e^2)}{4}\frac{\dot\delta}{\dot\delta_{\text{beg}}}\,\mathbf{n}}_{\text{viscous force}}. \tag{2}$$

Here, δ and $\dot\delta$ are the overlap between the two bodies and their relative velocity at the contact point, respectively. The exponent n can e.g. be 1, which leads to a linear, or 3/2, which leads to a contact law based on Hertz contact . The relative velocity at the beginning of the contact is denoted with $\dot\delta_{\text{beg}}$ and $e \leq 1$ is a constant that corresponds to the coefficient of restitution, but this law is only applicable if its value is close to 1. That means, the damping coefficient d is here depending on the stiffness parameter K, and on the overlap between the bodies

$$d = \frac{3\,K\,(1-e^2)}{4\,\dot\delta_{\text{beg}}}\,. \tag{3}$$

The contact force distribution then arises from this as depicted in Fig. 9. As mentioned, this model is only valid for e close to 1.

For more energy dissipation, a smaller coefficient of restitution is essential. An extension of the above model has been done by [7]. There, the relation of (3) between the damping factor d and the effective coefficient of restitution is re-evaluated. In [14] it is shown that the relation between e and $\dot\delta_{\text{beg}}$, $e = 1 - \alpha\dot\delta_{\text{beg}}$, used in [9, 11], can be achieved as a special case from the equations of motion and the contact force from the above model as

$$\mathbf{f}_N = K\delta^n\,\mathbf{n} + K\delta^n\,d\,\dot\delta\,\mathbf{n}\,. \tag{4}$$

However, in [7] it is shown how this can be solved more generally. There, the obtained relation between d and e states as

$$\frac{1+d/e}{1-d} = \exp\left[d(1+1/e)\right]\,. \tag{5}$$

Inserting different values for e between 0 and 1, this can be solved for $d(e)$ and a table can be made in a pre-processing state. Then, during calculation, the adequate value for d can be obtained from the table. By means of this method, it is, therefore, possible to apply all values for e between 0 and 1, see Fig. 10.

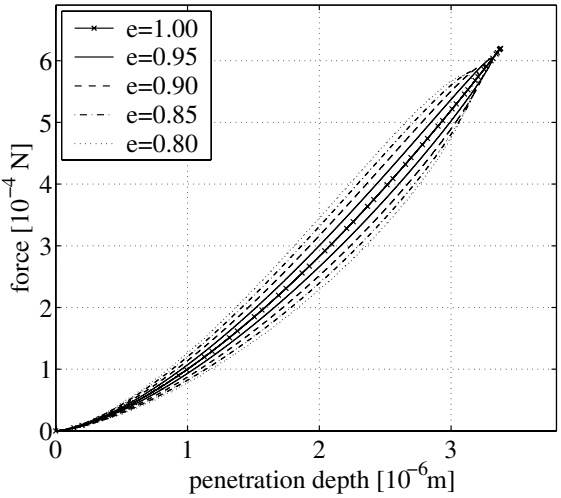


Fig. 9. Contact force over penetration depth, for different but large values of the coefficient of restitution, see [11]

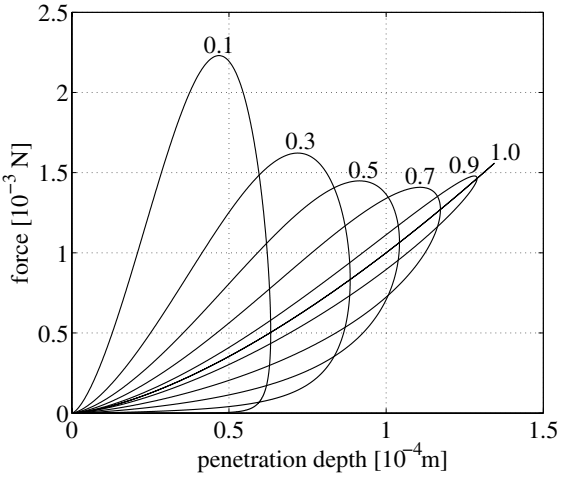


Fig. 10. Contact force over penetration depth for different values of the coefficient of restitution, see also [7]

Energy Dissipation by Means of Different Stiffness Parameters

Another possibility for the modeling of stronger dissipative forces is described, e.g., in [12, 31]. Here, the force between the bodies is simplified by a piece-wise relationship between the overlap of the bodies and the normal contact force. The piece-wise description can, e.g., consist of linear contact forces or can be applied to the Hertz contact law. Then, also a hysteretic contact force

is achieved, e.g. for the piece-wise linear relationship, as

$$\mathbf{f}_N^{\text{hys}} = \begin{cases} k_1\delta & \text{during loading ,} \\ k_2(\delta - \delta_0) & \text{during un-/reloading ,} \\ 0 & \text{during unloading .} \end{cases} \tag{6}$$

Therefore, k_1 has to be smaller than k_2, see Fig. 11. During the initial loading process the force increases with respect to the overlap δ, Fig. 11. The proportionality factor for loading is k_1 until the maximum overlap $\delta_{\max}$ is reached. This actually highest reached overlap value has to be stored. During unloading (when the two contacting bodies move apart from each other) the force decreases following a second line with the constant factor k_2, until it gets zero. Then there is no contact force applied to the bodies anymore. If however a reloading occurs, then there will be an inner loop, with constant k_2, see in Fig. 11 the dash-dotted loops. It is important to notice that a reloading can only happen due to contacts with other bodies. That is the reason for the gain of energy at the left-hand side of Fig. 11.

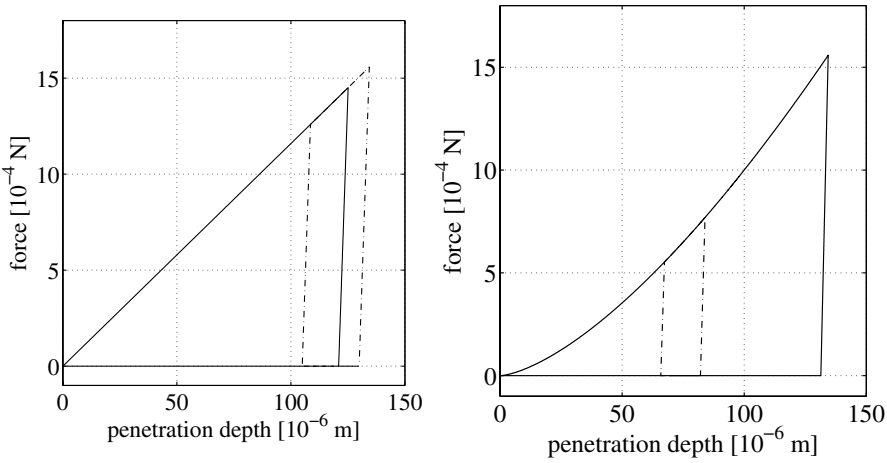


Fig. 11. Hysteretic repulsive contact force, with the linear model (left) and the Hertz contact law (right), see [12]

Attractive Contact Forces

In this section, attractive contact forces are described that are applied to bodies during contact, i.e. when the surfaces of both bodies are in touch, [12]. Such contact forces are often called adhesive or cohesive contact forces, where

adhesion stands for contacts between different materials and cohesion for contacting bodies of the same material. The model used for these forces is similar to the model of Sect. 5.1. Again, three cases are distinguished, where here three stiffness parameters have to be chosen. The force can then be written as

$$f_{ij}^{\text{hys}} = \begin{cases} k_1\,\delta & \text{for loading ,} \\ k_2\,(\delta - \delta_0) & \text{for un-/reloading ,} \\ k_c\,\delta & \text{for unloading ,} \end{cases} \tag{7}$$

where the subscripts i and j indicate two different bodies. During the initial loading process the force increases with respect to the overlap δ, Fig. 12. During unloading, the x-axis may be crossed at δ_0. Then, the algebraic signs of the forces between both bodies change. The force between both bodies then is becoming an attractive force. The qualities k_1, k_2 and k_c are the material parameters which have to be chosen, δ_0 and $\delta_{\max}$ follow from the computation.

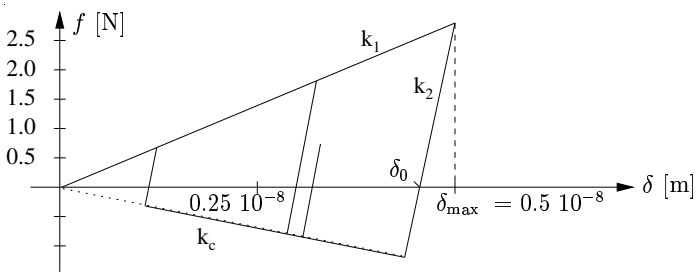


Fig. 12. Force law for two contacting bodies with overlap δ and three stiffnesses; two of them for initial loading and subsequent un-/reloading and the third stiffness for the cohesion strength k_c, see [12, 28]

If a strong adhesion is applied to two contacting bodies, the negative contact force can lead to a reloading effect. Then the reloading will take place along the k_2-line. For a reloading effect due to a strong adhesive behavior this will lead to an oscillatory behavior of un-/reloading along the k_2-line. However, the reloading can also be caused by another body contacting with one of the two bodies. Then the contact force for the actual surface point in contact is not calculated along the line with gradient k_1 again before the overlap has increased to the last stored value of $\delta_{\max}$. However, if the achieved attractive force is not strong enough to change the direction of motion of the bodies, the attractive force is increasing. The maximum value of the attractive force is reached, when the un-/reloading line is crossing the third, cohesive line of gradient k_c. Further unloading forces are calculated along this third line, that limits the maximum possible attractive forces between the particles.

Still this may lead to a reloading process, if the attractive force is now strong enough to change the direction of motion of the bodies. In that case the force has to be determined along another line with gradient k_2, see Fig. 12.

Otherwise, in the case that no reloading will happen, the contact force calculation is continued along line k_c until the overlap δ is zero, which means that the particles do not overlap anymore. For a new contact phase between the same two bodies, the whole calculation is starting again with a contact force proportional to the overlap along the initial line with gradient k_1.

An example of an adhesive system can be seen at the left-hand side of Fig. 13. The dark color shows bodies that are fixed in space, whereas the other bodies are free and subject to gravitational forces, but stick to the dark obstacles as well as together due to adhesion.

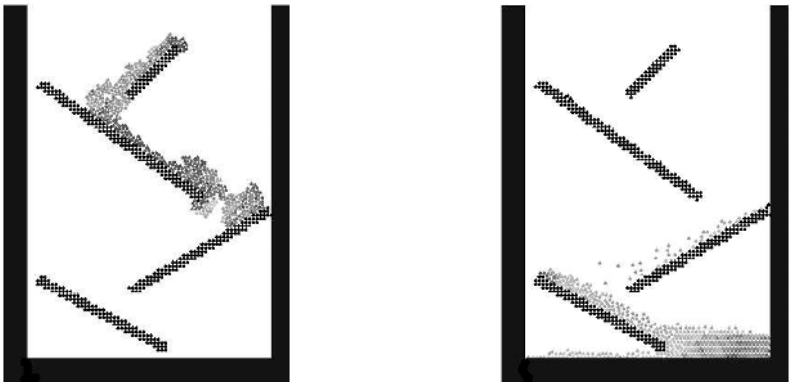


Fig. 13. Two different situations of a system one with adhesion (on the left-hand side) and one without adhesion (on the right)

Here, the bodies have been falling from above, but instead of slipping through the spaces between the obstacles (as they do on the right-hand side of Fig. 13), the bodies stick to them. The used parameters of the materials are listed in Table 1. The bodies are quite gluey and hardly move after a while, see left-hand side of Fig. 13. The different brightnesses of the bodies show the little kinetic energy that is left in the bodies. The darker they are, the less kinetic energy is left, therefore, the obstacles are completely dark.

5.2 Tangential Contact Forces

For the contact force in tangential direction again a simple model such as the Cundall-Strack model [3, 4] has to be used, where a small displacement between the bodies is also accepted during sticking. For the tangential contact force, generally two phases can be distinguished. The friction process is then divided into a static sticking phase and a dynamic phase that can be called

Table 1. Parameters of adhesion used for the two examples in Fig. 13, with the total simulation time t_{sim} and the time step size Δt

t_{sim}	Δt	density ρ	k_1	k_2	k_c	damping d
25 s	10^{-5} s	10000 kg/m^3	5e8 N/m	5e13 N/m	-5e8 N/m	2000 Ns/m
25 s	10^{-5} s	10000kg/m^3	5e8 N/m	5e8 N/m	0 N/m	2000 Ns/m

sliding or slipping, see [4, 33], respectively. The well-known Coulomb friction model distinguishes between sticking and sliding in such a way, that during sticking, no tangential motion is possible. The contact force during sticking is not given explicitly, but it has to fulfill the condition that no relative motion between the bodies can occur, which leads to an undifferentiability. In order to avoid this, some authors introduced associated models, see e.g. [17, 33], where the frictional contact force is based on a functional form which leads to a smooth transition from sticking to sliding. For instance, an arctan-function is used sometimes. For these associated models, the friction force is dependent on the relative velocity between the contacting bodies

$$\mathbf{f_t} = -\frac{\mathbf{v_t}}{\|\mathbf{v_t}\|} \min(k\|\mathbf{v_t}\|, \mu\mathbf{f_n}) \; . \tag{8}$$

The problem with such a force law is, however, that it leads to a contact force that is zero for no relative tangential velocity, and hence, e.g. a block on an inclined plane can never come to rest.

In the following, we want to use the Cundall-Strack model [4]. That model also distinguishes between a sticking and a sliding phase, see Fig. 14 but the contact force is not dependent on the relative velocity in the tangential plane but to its corresponding relative displacement

$$\mathbf{f_t} = -\frac{\boldsymbol{\xi_t}}{\|\boldsymbol{\xi_t}\|} \min(k\|\boldsymbol{\xi_t}\|, \mu\mathbf{f_n}) \; . \tag{9}$$

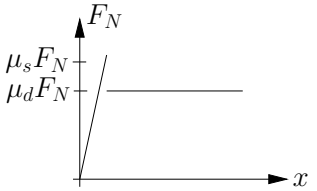


Fig. 14. Friction force, Cundall-Strack model

It is the approach used in this model, that the friction force acts in opposite direction to a small tangential motion that is allowed here also in the sticking phase, and has to be chosen such that it can characterize surface roughness. The sticking force is then influenced by that tangential displacement in its direction and also in its value. In Fig. 15 the tangential plane is shown. The figure shows a cut through the friction cone, i.e. a cut perpendicular to the normal direction at the level of the actual normal contact force $\mathbf{f}_N$ of Fig. 14. Therefore, the applied force is shown in this figure in the tangential plane. The circle shows the border between the sticking and the sliding zone, where sticking occurs inside and sliding outside.

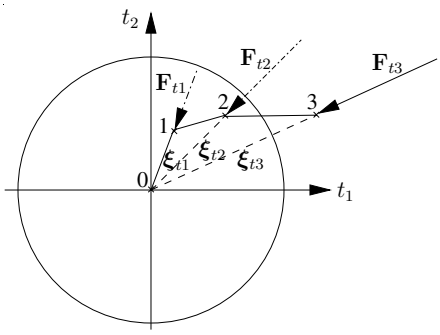


Fig. 15. Cut of friction cone at actual normal force, parallel to the tangential plane

The beginning of the contact is at the origin of the coordinate system. In this example the line from 0 to 1 shows the tangential displacement of two bodies, $\boldsymbol{\xi}_{t1}$, due to some acting forces. Point 1 is clearly within the sticking zone. Therefore, the contact force that has to be applied, $\mathbf{F}_{t1}$, is a sticking-force, its size being proportional to the distance between 0 and 1, see Fig. 15. In a next time step, the tangential displacement may change, e.g., due to the impulse of another body in the system. Then the point that is in contact has moved and its projection onto the tangential plane is here, e.g., point 2. If this point is still within the circle, then still sticking occurs. The sticking force in this case is proportional to the tangential displacement $\boldsymbol{\xi}_{t2}$ in its total direction and value, see $\mathbf{F}_{t2}$ of Fig. 15. That means, in each time step, the displacement $\boldsymbol{\xi}_t$ in tangential direction has to be re-calculated and added up vectorially. The force is then applied in opposite direction of that total displacement. As soon as sliding occurs, e.g. in the time step from point 2 to 3, the value of the friction force will be restricted to the sliding force. That means, its magnitude is restricted to $\mu_d \mathbf{F}_N$ which corresponds to the radius of the circle. The direction of this force still changes with each time step and will always be directed to the origin of the tangential system.

If, in a later time step when the contact still exists, the relative motion stops and changes its direction, a sticking force has to be applied to the bodies. By means of the conventional model, this is not possible, see [3]. If the elongation of $\boldsymbol{\xi_t}$ has been updated in each time step, after a long sliding phase it will take a long backward motion in opposite direction to release this "spring". Therefore, [3] propose not to update the elongation of $\boldsymbol{\xi_t}$ during sliding, but to keep it constant, i.e. at the value that corresponds to the particular radius of the circle, until sticking occurs again. The solution for the contact force in a planar example is given in [3], and is extended here to a spatial statement,

$$\begin{aligned} \mathbf{f_t} &= -k_t \boldsymbol{\xi_t} \,, \\ \boldsymbol{\xi_t} &= \int_{t_0}^{t} \mathbf{v_t} \Theta \left(\tfrac{\mu \|\mathbf{f_n}\|}{k_t} - \|\boldsymbol{\xi_t}\| \right) dt' \,, \end{aligned} \tag{10}$$

where $\Theta(\bullet)$ is the Heaviside function. Therefore, an update of $\boldsymbol{\xi_t}$ will only be carried out for $\mu \|\mathbf{f_n}\| > k_t \|\boldsymbol{\xi_t}\|$.

Results Obtained by Means the of Cundall-Strack Model

In this example the advantage of the above friction model is shown. Contrary to models, where the friction force is detected by means of the relative velocity between two bodies, here e.g. a block on an inclined plane can come to rest. Such an example is shown here, with a non-convex block of volume $0.001\,\mathrm{m}^3$. It has the density of $\rho = 2700\,\mathrm{kg/m}^3$, therefore its mass is $m = 2.7\,\mathrm{kg}$. The angle of the inclined plane on which the block is located is $\alpha = 30^\circ$ and the critical value for the friction coefficient is $\mu = \tan(\alpha) \approx 0.57735$. Here, the block is located in its balanced state in normal direction. Tangentially to the plane, an initial velocity $v_0 = 5\,\mathrm{m/s}$ is given to the block. So, the block is sliding upward first, until the relative velocity is zero. This time can be determined as $t = v_0/(g\ \sin(\alpha) + \mu g\ \cos(\alpha))$. In the following, a simulation with a time step size of $\Delta t = 5\mathrm{e}-6\,\mathrm{s}$ is carried out for three different values of μ around the critical value. For this example, time t_1 at which the velocity of the block gets zero for $\mu_1 = 0.5$ is $t_1 \approx 0.546\,\mathrm{s}$. For a friction coefficient of $\mu_2 = 0.58$, $t_2 \approx 0.508\,\mathrm{s}$, and finally for $\mu_3 = 0.6$ we get $t_3 \approx 0.499\,\mathrm{s}$. At the left-hand side of Fig. 16 the velocity of the block is shown over the time. As an inset, the assembly of the example, i.e. the non-convex block on the plane is shown. The block is sliding upwards, while its velocity drops until it gets zero at a certain point. At this point, the block comes to rest.

For $\mu_1 < \tan(\alpha)$, the velocity drops further, it changes its sign and gets negative, i.e., the block glides down the plane. A more detailed image of the scene is given in the right-hand side of Fig. 16. Here, points t_1, t_2, and t_3, are specified, at which the velocity is zero. It can be seen clearly, that for all cases, the velocity decreases further for a short time, since the block moves downward during sticking. Then, after sticking is overridden, the velocity for

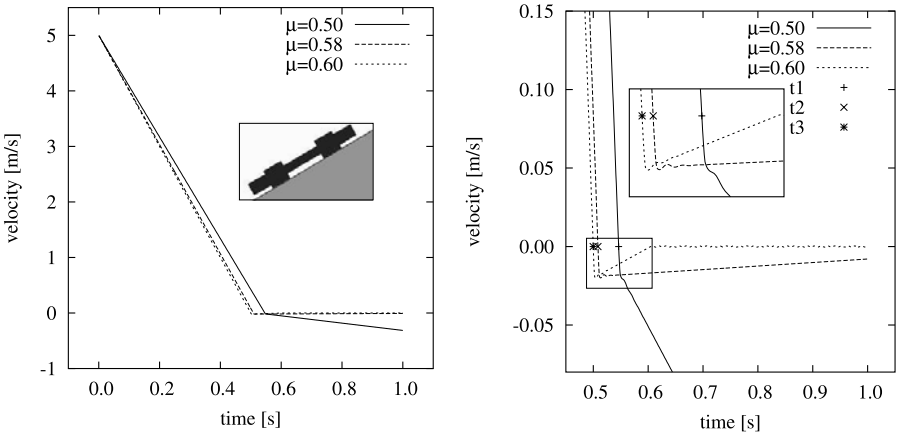


Fig. 16. Velocity of the non-convex block on an inclined plane for three different friction coefficients (left) and zoomed depiction of the velocity of the block

the small friction coefficient decreases further. For μ_2 and μ_3, the velocity rises until it gets zero and remains there.

6 Example

The explained methods are implemented in a simulation program in order to simulate particles of different shape. This program is a tool for the simulation of particle flow e.g. in production processes. Our basic interest is the simulation of the dynamical behavior of differently shaped polyhedral particles that appear e.g. in production processes in mechanical engineering. In the following, results from the motion of bulk solids transported into a silo are shown.

In this simulation polyhedral bodies consisting of 18 vertices and 32 edges are used. These goods are moving on a conveyor belt with a constant velocity of 10 m/s. At the end of the belt, the media is falling into a silo available e.g. for the purpose of storage. In this simulation, the bodies are subject to normal contact forces with strong dissipation, acting between the bodies and walls of the silo and between the bodies among each other. Frictional contact forces are applied to the bodies, as well as adhesional influences. In Fig. 17 and 18 this procedure is shown for two different points of time.

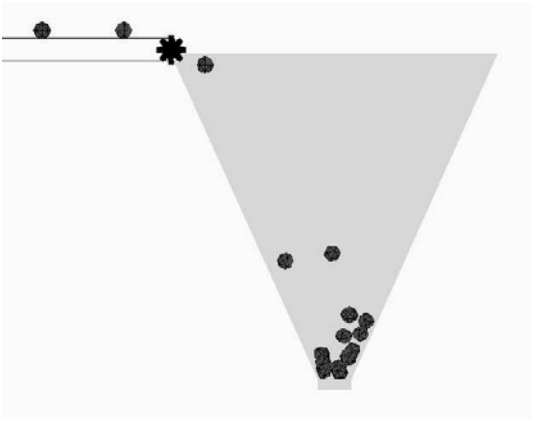


Fig. 17. Polyhedral bodies moving on a conveyor belt and gathering in a silo

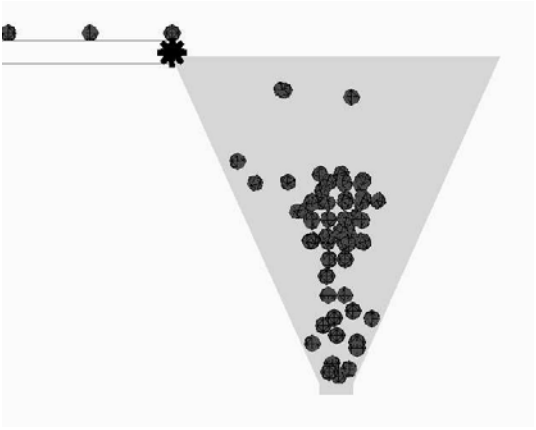


Fig. 18. Polyhedral bodies in the silo at a later point of time

7 Conclusion

In this work different tasks were described that are used to obtain a program for the simulation of granular media. The combination of ideas from MD and MBS makes it possible to simulate the dynamical behavior of systems of differently shaped bodies. In MD the material behavior is represented by the motion and forces acting on the bodies. In order to calculate contact forces effectively, overlaps between the particles are accepted and the value of the contact force depends on the distance between the colliding bodies. Another concept, which was used, is the neighborhood search. By means of that method it is possible to search for close body pairs in advance and to evaluate the necessary contact forces only for the colliding body pairs. By means of such

neighborhood search methods it is possible to reduce the necessary simulation time of order $O(n^2)$ in such a way that a linear behavior $O(n)$ can be obtained.

The equations of motion could be obtained easily, since simulations of granular material or bulk solids are based on free bodies in space. The very time consuming collision detection only had to be carried out for neighboring body pairs. For colliding bodies, the contact force calculation was conducted. For the calculation of the contact forces in normal direction different models were described. One possibility is the Kelvin-Voigt model. Another method was described, where different stiffness parameters are used. A similar method was used in order to describe adhesive contact forces. For the modeling of the tangential contact force, the Cundall-Strack model was used. There, the friction force is modeled dependently on the relative displacement between two contacting bodies. Finally, an example of bodies falling into a silo was presented.

References

1. M. P. Allen and D. J. Tildesley. *Computer Simulations of Liquids.* Clarendon Press, Oxford, 1989.
2. D. Baraff. *Dynamic Simulation of Non-Penetrating Rigid Bodies.* Cornell University Ithaka, Ph.D. Thesis 92-1275, 1992.
3. L. Brendel and S. Dippel. Lasting contacts in molecular dynamics simulations. In *Physics of Dry Granular Materials - NATO ASI Series E 350*, pages 313–318. Kluwer Academic Publishers, Dordrecht, 1998.
4. P. A. Cundall and O. D. L. Strack. A discrete numerical model for granular assemblies. *Geotechnique*, 29(1):47–65, 1979.
5. P. Eberhard. *Kontaktuntersuchungen durch hybride Mehrkörpersystem / Finite Elemente Simulationen.* Shaker, Aachen, 2000.
6. D. Fincham and B. J. Ralston. Molecular dynamics simulation using CRAY-1 vector processing computer. *Computer Physics Communications*, 23:127–134, 1981.
7. Y. Gonthier, J. McPhee, C. Lange, and J.-C. Piefboeuf. A regularized contact model with asymmetric damping and dwell-time dependent friction. *Multibody System Dynamics*, 11(3):209–233, 2004.
8. L. Greengard. *The Rapid Evaluation of Potential Fields in Particle Simulation.* MIT Press, Cambridge, 1987.
9. K. H. Hunt and F. R. E. Grossley. Coefficient of restitution interpreted as damping in vibroimpact. *ASME Journal of Applied Mechanics*, 7:440–445, 1975.
10. Y. Kishino, editor. *Powders & Grains*, Balkema, Rotterdam, 2001.
11. H. M. Lankarani and P. E. Nikravesh. Continuous contact force models for impact analysis in multibody systems. *Nonlinear Dynamics*, 5:193–207, 1994.
12. S. Luding. Collisions and contacts between two particles. In *Physics of Dry Granular Media - NATO ASI Series E 350*, pages 285–304. Kluwer Academic Publishers, Dordrecht, 1998.
13. S. Luding. *Die Physik trockener granularer Medien.* Logos, Berlin, 1998.

14. D. W. Marhefka and D. E. Orin. A compliant contact model with nonlinear damping for simulation of robotic systems. *IEEE Transaction on Systems, Man and Cybernetics – Part A: Systems and Humans*, 29(6):566–572, 1999.
15. B. Muth, M.-K. Müller, P. Eberhard, and S. Luding. Contacts between many bodies. *Machine Dynamics Problems*, 28(1):101–114, 2004.
16. B. Muth, G. Of, P. Eberhard, and O. Steinbach. Collision detection for complicated polyhedra using the multipole method or ray crossing (submitted 2006).
17. J. T. Oden and E. B. Pires. Nonlocal and nonlinear friction laws and variational principles for contact problems in elasticity. *Journal of Applied Mechanics*, 50(1):67–76, 1983.
18. J. O'Rourke. *Computational Geometry in C, 2. Edition.* Cambridge University Press, Cambridge, 1998.
19. F. Pfeiffer and C. Glocker. *Multibody Dynamics with Unilateral Contacts.* Wiley, New York, 1996.
20. J. Pfister and P. Eberhard. Frictional contact of flexible and rigid bodies. *Granular Matter*, 4(1):25–36, 2002.
21. T. Pöschel and V. Buchholtz. Static friction phenomena in granular materials: Coulomb law vs. particle geometry. *Physical Review Letters*, 71(24):3963–3974, 1993.
22. T. Pöschel and S. Luding, editors. *Granular Gases*, Springer, Berlin, 2001.
23. B. Quentrec and C. Brot. New method for searching for neighboring molecular dynamics. *Journal of Computational Physics*, 13:430–432, 1973.
24. D. C. Rapaport. *The Art of Molecular Dynamics Simulation.* Cambridge University Press, Cambridge, 1995.
25. V. Rokhlin. Rapid solution of integral equations of classical potential theory. *Journal of Computational Physics*, 60:187–207, 1985.
26. W. Schiehlen and P. Eberhard. *Technische Dynamik.* B.G. Teubner, Stuttgart, 2004.
27. A. Schinner. Fast algorithms for the simulations of polygonal particles. *Granular Matter*, 2(1):35–43, 1999.
28. R. Tykhoniuk, J. Tomas, S. Luding, M. Kappl, L. Heim, and H.-J. Butt. Adhesion, inelastic contact behavior and simulation of shear dynamics of ultrafine cohesive powder. In *Powders and Grains 2005*, pages 499–503. Balkema, Leiden, 2005.
29. L. Verlet. Computer experiments on classical fluids. I. Thermodynamical properties of Lennart-Jones molecules. *Physical Review*, 159(1):98–103, 1967.
30. P. A. Vermeer, S. Diebels, W. Ehlers, H. J. Herrmann, S. Luding, and E. Ramm, editors. *Continuous and Discontinuous Modelling of Cohesive Frictional Materials*, Springer, Berlin, 2001.
31. O. R. Walton and R. L. Braun. Viscosity, granular-temperature, and stress calculations for shearing assemblies of inelastic, frictional disks. *Journal of Rheology*, 30(5):949–980, 1986.
32. W. Wessel. *Kontakterkennung räumlicher polyhedraler Körper mit Hilfe von Methoden der Molekulardynamik.* University of Stuttgart, Student-Thesis STUD-212, 2004.
33. P. Wriggers. *Computational Contact Mechanics.* John Wiley & Sons, Chichester, 2002.

Multilevel Numerical Algorithms and Experiments for Contact Dynamics*

Peter Eberhard[1], Stefan Hüeber[2], Yu Jiang[1], and Barbara I. Wohlmuth[2]

[1] Institute of Engineering and Computational Mechanics (ITM), University of Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany
eberhard@itm.uni-stuttgart.de
jiang@itm.uni-stuttgart.de

[2] Institute of Applied Analysis and Numerical Simulation (IANS), University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
hueeber@ians.uni-stuttgart.de
wohlmuth@ians.uni-stuttgart.de

Summary. Nonlinear frictional contact problems are still a challenging task both from the mathematical and engineering point of view. These problems are of crucial importance in various applications. In this report we study dynamical contact problems on mathematical and experimental aspects. In the mathematical part we present a variationally consistent formulation based on mortar techniques with dual Lagrange multipliers for this type of problems. Furthermore, new optimal a priori and a posteriori error estimates were achieved, and numerical results for nearly incompressible materials are given. To solve the resulting nonlinear algebraic problem, we use a primal-dual active set strategy which can also be interpreted as semismooth Newton method. In combination with optimal multigrid methods, the inexact version of this approach can be regarded as a nonlinear multigrid method, and we end up with an efficient iterative solver. In the engineering part experiments to study the properties of the impact between a rotating disc and an elastic strip are presented. Experimental setup and methods are designed to release the disc with prescribed translational and rotational velocities. The impact event is captured by a high-speed digital camera system. Based on image processing, impact quantities, that is, coefficients of normal and tangential restitution, impulse ratio, rotational velocity change, incidence and rebound angles, are measured. A numerical model for interpreting the experimental data is developed, which can also give some insight into effects of strip flexibility. Results are also compared with those from finite element calculations.

Keywords: Contact experiments, frictional contact, mortar discretization, oblique impact, optimal a priori error estimates, primal-dual active set strategies

* Research Project B8 "Contact Dynamics with Multibody Systems and Multigrid Methods"

1 Introduction

The investigation of nonlinear frictional contact dynamics plays an important role for a wide range of engineering applications. On one hand, this project is mainly concerned with the mathematical analysis and efficient numerical algorithms, and, on the other hand, the design, realization and analysis of experiments for frictional contact problems.

For such problems, nonconforming domain decomposition techniques provide a powerful tool. In this project we work with variationally consistent formulations for this type of problems based on mortar techniques with dual Lagrange multipliers. Dual Lagrange multiplier spaces were introduced in [72], and they are based on a biorthogonal basis resulting in a diagonal mass matrix. Thus we obtain for each node on the slave side a local non-penetration condition and a local friction condition. Therefore, primal-dual can be applied to solve the discrete nonlinear problem.

New optimal a priori error estimates for the discretization error in the H^1-norm for the displacements and in the $H^{-1/2}$-norm for the Lagrange multiplier were achieved. Abstract error estimates for variational inequalities can be found, e.g., in [19, 22] and a priori bounds for the discretization error of unilateral contact problems are given, e.g., in [27]. Recently a lot of work has been done to analyze mortar formulations based on standard Lagrange multipliers. A priori error estimates for the displacements in the H^1-norm and for the Lagrange multiplier in the $H^{-1/2}$-norm of order $h^{0.75}$ have been established, see, e.g., [8, 9, 15, 51], under H^2-regularity assumption. Using additional regularity assumptions on the Lagrange multiplier, order h has been shown, see, e.g., [15, 29]. Although the order h is optimal, the regularity assumptions are quite strong and restrictive. These first a priori results have been considerably improved during the last couple of years, see, e.g., [7, 9, 30, 64]. Most of the theoretical results are obtained for standard Lagrange multipliers. Here, we have considered the case of linear and quadratic mortar finite elements based on dual Lagrange multipliers. Moreover, we show order $h^{0.5+\nu}$ a priori estimates for the displacements and the Lagrange multiplier if the solution is $H^{3/2+\nu}$-regular, $0 < \nu \leq 0.5$. The techniques are based on introducing locally defined truncation operators measuring the nonconformity of the discretization and on Sobolev–Slobodeckij norms.

To solve the nonlinear problem we use a primal-dual active set strategy to find the actual contact zone and slippy/sticky nodes. We will see that this active set strategy can be interpreted as a semismooth Newton method yielding superlinear convergence. For Newton methods in general and their application to contact problems we refer to [3, 13, 31, 66]. For Coulomb friction we present new efficient algorithms for the two-dimensional and three-dimensional case. Here, the superlinear convergence of the active set strategy also applies. Combining the active set approach with an optimal multigrid method as solver for the resulting linear problems and with inexact techniques leads to an efficient iterative solver for the nonlinear problems. Nonlinear material laws and large

deformation problems can be treated within the same framework. Moreover, the method can be easily extended to nearly incompressible materials by using a modified Hu–Washizu formulation. Furthermore, the proposed method can be generalized to the dynamic case. Using the time integration schemes proposed by [12, 50, 63], we get an energy-conserving time integrator for frictional two-body contact problems.

For low-regularity problems it is known that adaptive refined meshes perform much better than uniform refined meshes. We present a new error estimator for the 2D case based on equilibrated fluxes and the Arnold–Winther element yielding an optimal upper estimate of the error. Numerical examples including time-dependent problems will show the flexibility and efficiency of our methods.

From an engineering point of view, there is a great deal of results available about rigid multibody impact dynamics [16, 23, 57]. The investigation of elastic multibody impact problems has thus drawn more attention to gaining further understanding of impact transients and their coupling with flexible-body responses [16, 18, 41, 58, 62].

A number of works has been carried out for studying impact transients of compact bodies, where local elastic deformation is considered. The pioneering work is attributed to Hertz's contact theory for frictionless problems. Mindlin and Deresiewicz [55] presented a formulation of the tangential force under some normal loading scenarios with friction. The tangential force was shown to be dependent on load history. Maw et al. [53] developed a solution for the oblique impact of an elastic sphere on a half-space of similar materials by means of the small-increment method. This model indicated the necessity of considering tangential elastic contact compliance. An extended model for dissimilar materials was provided by Johnson [42]. Experiments were also conducted to verify these models [21, 54]. Moreover, a number of numerical methods can be found in [38, 39, 76]. In contrast to previous elastic continuum approaches, simplified models were proposed by Walton [69], Stronge [68] and Pfeiffer and Glocker [57], and were confirmed by experiments [6, 21, 24, 44, 52].

Problems of elastic impact of a compact body striking a slender body generally possess the characteristic of low velocity and dry friction. Concerning axial and transverse impact of compact-slender bodies, a combined approach including Hertz's contact law and wave theory was evidenced to be a good approximation [2, 23, 32, 59]. In more recent work, Christoforou and Yigit [14] investigated the dynamic response of beams and plates subject to low velocity transverse impact and showed that for a light mass striking a heavy and stiff structure the response was dominated by local contact deformation, while for a heavy mass striking a light and compliant structure a quasi-static mass-spring model may be used. For the cases in-between a full simulation may however be needed due to complexity. Sondergaard et al. [65] experimentally studied the influences of the clamped boundary and initial velocities on the response of sphere-plate impact, and reported some results on oblique impact.

Nevertheless, it should be noted that most of the studies on the compact-slender-body impact focused on normal longitudinal and transverse impact, where friction is negligible. Furthermore, for the oblique collision between a rotating compact body and a slender body, seldom results have been reported. In this study an experimental investigation of the oblique impact of a rotating disc hitting a flexible strip is performed. We start with an analysis of impact models, then describe the experimental setup and finally discuss the results with an emphasis on the effects of the strip flexibility and the disc initial velocity. A model combining local frictional contact relation and global vibrations of the slender body is also developed. This model is shown to be able not only to capture the experimental data, but also to provide some insight into the flexibility dependence.

In Sect. 2 we discuss the mathematical part of the project concerning the discretization methods and the numerical algorithms. Section 3 presents the engineering part dealing with the experimental setup and results.

2 Discretization and Numerical Simulation

In this section we present the results of the mathematical part of our joint project. In Sect. 2.1 we give the detailed mathematical problem setting. Section 2.2 contains the discretization methods and the optimal a priori error estimates. Next we present in Sect. 2.3 the primal-dual active set algorithm. Section 2.4 shows the extension to the dynamical case. Finally, we give in Sect. 2.5 the a posteriori error estimator. A various set of numerical examples during this section illustrates the flexibility of the proposed formulations and the methods.

2.1 Problem Setting

In this section, we state the contact problem with Coulomb friction. For a general introduction to contact problems with and without friction we refer to [45, 49, 71, 76]. For simplicity of notation, we restrict ourselves to the case of two bodies being in contact. The two bodies are modeled by Ω^s and Ω^m, where the superscript m stands for the master side and the superscript s for the slave side, as it is standard in the mortar framework. We recall that a Lagrange multiplier is defined on the slave side and that the displacements on the slave side depend on the ones on the master side. Let $\Omega^l \subset \mathbb{R}^d$, $l \in \{m, s\}$, $d = 2$ or $d = 3$, denote the two bodies. We assume the boundaries $\partial\Omega^l$ to be divided into three disjoint measurable parts $\Gamma_D^l, \Gamma_N^l, \Gamma_C^l$ with $\text{meas}(\Gamma_D^l) \neq 0$. Furthermore, we assume that $\overline{\Gamma_C^s}$ is a compact subset of $\partial\Omega^s \backslash \overline{\Gamma_D^s}$. We impose Dirichlet conditions $\mathbf{u}_D^l$ on Γ_D^l and Neumann data, i.e., surface traction $\mathbf{p}^l \in [L^2(\Gamma_N^l)]^d$ on Γ_N^l. Moreover, we denote by $\mathbf{f}^l \in [L^2(\Omega^l)]^d$ volume forces acting on Ω^l. Let $\mathbf{P}^l$ indicate the first Piola–Kirchhoff stress tensor of the considered material, then our equilibrium condition can be written as

$$\begin{aligned} -\operatorname{div} \mathbf{P}^l &= \mathbf{f}^l && \text{in} \quad \Omega^l \,, \\ \mathbf{u}^l &= \mathbf{u}_D^l && \text{on} \quad \Gamma_D^l \,, \\ \mathbf{P}^l \mathbf{n}^l &= \mathbf{p}^l && \text{on} \quad \Gamma_N^l \,. \end{aligned} \tag{1}$$

Here, $\mathbf{n}^l$ denotes the outer unit normal vector on the boundary $\partial\Omega^l$. To state the contact and friction conditions, we introduce for each point of Γ_C^s the vectors $\boldsymbol{\tau}_k$, $1 \le k \le d-1$, that span the tangential hyperplane. We assume that $\{\mathbf{n}, \boldsymbol{\tau}_k\}$, $\mathbf{n} := \mathbf{n}^s$, is an orthonormal basis in $\mathbb{R}^d$ for each point of Γ_C^s. In order to formulate the non-penetration condition of the two bodies, we use a predefined relation between the points of the possible contact zones Γ_C^l. This relation is realized by an injective and smooth mapping $R : \Gamma_C^s \to \Gamma_C^m$ satisfying $R(\Gamma_C^s) \subset \Gamma_C^m$. Here, the mapping R is assumed to map any $x \in \Gamma_C^s$ to the intersection of the normal on Γ_C^s at x with Γ_C^m. Then the contact conditions on Γ_C^s are given by

$$[\mathbf{u}]_n - g \le 0 \,, \quad \sigma_n \le 0 \,, \quad \sigma_n([\mathbf{u}]_n - g) = 0 \quad \text{on } \Gamma_C^s \,, \tag{2}$$

where $g \ge 0$ denotes the initial gap in normal direction between the two bodies, $\sigma_n := \mathbf{n}^\top \mathbf{P}^s \mathbf{n}$ the normal component of the boundary stress and $[\mathbf{u}]_n := (\mathbf{u}^s(x) - \mathbf{u}^m(R(x)))\mathbf{n}$. We remark that for large deformations we do not use this approach. Here we apply the final load in several small load steps, such that we obtain in each loadstep a contact problem with a small deformation. Finally, the Coulomb friction law states that

$$\begin{cases} \|\boldsymbol{\sigma}_\tau\| \le \mathcal{F}|\sigma_n| \,, \\ \|\boldsymbol{\sigma}_\tau\| < \mathcal{F}|\sigma_n| \Rightarrow [\mathbf{u}]_\tau = 0 \,, \\ \|\boldsymbol{\sigma}_\tau\| = \mathcal{F}|\sigma_n| \Rightarrow \boldsymbol{\sigma}_\tau = -\beta^2 [\mathbf{u}]_\tau \end{cases} \quad \text{on } \Gamma_C^s \,. \tag{3}$$

Here, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$, and $\boldsymbol{\sigma}_\tau = \mathbf{P}^s \boldsymbol{n} - \sigma_n \mathbf{n}$ the tangential component of the boundary stress, and $[\mathbf{u}]_\tau$ is the relative tangential displacement given by $[\mathbf{u}]_\tau := (\mathbf{u}^s(x) - \mathbf{u}^m(R(x))) - [\mathbf{u}]_n \mathbf{n}$. Furthermore, $\mathcal{F} : \Gamma_C^s \to \mathbb{R}$, $\mathcal{F} > 0$, is the friction coefficient.

2.2 Mortar Discretization and a priori Error Estimates

In this section, we are going to present the variational formulation of the contact problem (1)–(3). Furthermore we give a short overview on the mortar discretization technique, and we provide optimal a priori error estimates for the discretization error. We briefly discuss the main steps for obtaining the a priori error estimate for the frictionless two-body contact problem. For details we refer to [38] in case of linear finite elements and to [34] for quadratic finite elements. Optimal a priori error estimates in the case of a given friction bound are given in [35]. In this section, we consider a linearized Saint Venant–Kirchhoff material, where the first Piola–Kirchhoff stress tensor is given by

$$\boldsymbol{\sigma}^l := \lambda^l \operatorname{tr}(\boldsymbol{\varepsilon}^l)\mathrm{Id}_d + 2\mu^l \boldsymbol{\varepsilon}^l =: \mathcal{C}^l \boldsymbol{\varepsilon}^l \quad \text{in} \quad \Omega^l \,, \tag{4}$$

where $\boldsymbol{\varepsilon}^l := 1/2\big(\nabla \mathbf{u}^l + (\nabla \mathbf{u}^l)^\top\big)$ is the linearized strain tensor. The parameters λ^l and μ^l are the Lamé constants on Ω^l. Moreover, tr denotes the matrix trace operator and Id_d the identity matrix in $\mathbb{R}^d$.

2.2.1 Energy Minimization and Weak Formulation

To give the variational formulation of problem (1) with the contact conditions (2) for the frictionless case, where (3) is replaced by the condition $\boldsymbol{\sigma}_\tau = 0$, we introduce the product space $\boldsymbol{V} := [H^1_*(\Omega^m)]^d \times [H^1_*(\Omega^s)]^d$, equipped with the broken H^1-norm $\|\mathbf{v}\|^2_{1,\Omega} := \sum_{l=m,s} \|\mathbf{v}\|^2_{1,\Omega^l}$, where the spaces $[H^1_*(\Omega^l)]^d$ are defined by

$$[H^1_*(\Omega^l)]^d := \Big\{ \mathbf{v}^l \in \big[H^1\left(\Omega^l\right)\big]^d \,:\, \mathbf{v}^l|_{\Gamma^l_D} = \mathbf{0} \Big\} \,.$$

For $\mathbf{u} := (\mathbf{u}^m, \mathbf{u}^s) \in \mathbf{V}$ and $\mathbf{v} := (\mathbf{v}^m, \mathbf{v}^s) \in \mathbf{V}$, we define the bilinear form $a(\mathbf{u}, \mathbf{v})$ and the linear form $f(\mathbf{v})$ by

$$a(\mathbf{u}, \mathbf{v}) := \sum_{l=m,s} \int_{\Omega^l} \boldsymbol{\sigma}^l\left(\mathbf{u}^l\right) : \boldsymbol{\varepsilon}(\mathbf{u}^l)\,\mathrm{d}x \,, \quad f(\mathbf{v}) := \sum_{l=m,s} \Big(\int_{\Omega^l} \mathbf{f}^l \cdot \mathbf{v}^l \,\mathrm{d}x + \int_{\Gamma^l_N} \mathbf{p}^l \cdot \mathbf{v}^l \,\mathrm{d}s \Big) \,.$$

Then the weak solution of the contact problem can be obtained as solution of a minimization problem on the convex subset $\mathbf{K} := \{\mathbf{v} \in \mathbf{V} \,:\, [\mathbf{v}]_n \le g \ \text{ on } \ \Gamma^s_C\}$. Find $\mathbf{u} \in \mathbf{K}$ such that

$$J(\mathbf{u}) = \min_{\mathbf{v} \in \mathbf{K}} J(\mathbf{v}) \,, \qquad J(\mathbf{v}) := \frac{1}{2} a(\mathbf{v}, \mathbf{v}) - f(\mathbf{v}) \,, \tag{5}$$

see, e.g., [10, 26]. One can show that (5) is equivalent to a minimization problem on the unconstrained space $\mathbf{V}$. For any projection operator $P : \mathbf{V} \to \mathbf{K}$ satisfying $P\mathbf{v} = \mathbf{v}$ for $\mathbf{v} \in \mathbf{K}$, we define the energy functional $\tilde{J}(\cdot)$ by

$$\tilde{J}(\mathbf{v}) := \frac{1}{2} a(\mathbf{v} - P\mathbf{v},\, \mathbf{v} - P\mathbf{v}) + \frac{1}{2} a(P\mathbf{v},\, P\mathbf{v}) - f(P\mathbf{v}) \,.$$

Then the equivalence

$$\min_{\mathbf{v} \in \mathbf{V}} \tilde{J}(\mathbf{v}) = \min_{\mathbf{v} \in \mathbf{K}} J(\mathbf{v}) \tag{6}$$

holds. Using that P restricted to $\mathbf{K}$ is the identity, we find

$$\min_{\mathbf{v} \in \mathbf{V}} \tilde{J}(\mathbf{v}) \le \min_{\mathbf{v} \in \mathbf{K}} \tilde{J}(\mathbf{v}) = \min_{\mathbf{v} \in \mathbf{K}} J(\mathbf{v}) \,.$$

On the other hand, we can show that

$$\min_{\mathbf{v} \in \mathbf{V}} \tilde{J}(\mathbf{v}) \ge \min_{\mathbf{v} \in \mathbf{V}} \Big(J(P\mathbf{v}) + \frac{1}{2} a(\mathbf{v} - P\mathbf{v}, \mathbf{v} - P\mathbf{v}) \Big) \ge \min_{\mathbf{v} \in \mathbf{V}} J(P\mathbf{v}) = \min_{\mathbf{v} \in \mathbf{K}} J(\mathbf{v}) \,.$$

Furthermore, the minimization problem (5) is equivalent to a variational inequality on the convex subset $\mathbf{K} \subset \mathbf{V}$: find $\mathbf{u} \in \mathbf{K}$ such that

$$a(\mathbf{u}, \mathbf{v} - \mathbf{u}) \geq f(\mathbf{v} - \mathbf{u}) \,, \qquad \mathbf{v} \in \mathbf{K} \,. \tag{7}$$

We remark that in the frictional case, we get a quasivariational inequality. To give the saddle point formulation, we introduce a Lagrange multiplier space $\mathbf{M}$, being the dual space of the trace space $\mathbf{W}$ of $[H^1_*(\Omega_s)]^d$ restricted to Γ^s_C. By assumption $\overline{\Gamma^s_C}$ is a compact subset of $\partial\Omega^s \backslash \overline{\Gamma^s_D}$, and thus we have $\mathbf{W} = [H^{1/2}(\Gamma^s_C)]^d$. In the case $\overline{\Gamma^s_C} = \partial\Omega_s \backslash \Gamma^s_D$, we have to work with $[H^{1/2}_{00}(\Gamma^s_C)]^d$ instead of $[H^{1/2}(\Gamma^s_C)]^d$. We now define the following convex cone of Lagrange multipliers by

$$\mathbf{M}^+ := \left\{ \boldsymbol{\mu} \in \mathbf{M} \,:\, \langle \boldsymbol{\mu}, \mathbf{v} \rangle \geq 0, \;\; \mathbf{v} \in \mathbf{W}^+ \right\} \,,$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $\mathbf{M}$ and $\mathbf{W}$ on Γ^s_C, and $\mathbf{W}^+ := \{\mathbf{v} \in \mathbf{W} : \mathbf{v} \cdot \mathbf{n} \geq 0\}$. Defining the bilinear form $b(\cdot, \cdot)$ on the product space $\mathbf{V} \times \mathbf{M}$ by

$$b(\mathbf{v}, \boldsymbol{\mu}) := \langle \boldsymbol{\mu}, [\mathbf{v}] \rangle \,, \qquad \boldsymbol{\mu} \in \mathbf{M} \,, \;\; \mathbf{v} \in \mathbf{V} \,,$$

we get the saddle point formulation of the unilateral contact problem without friction, see, e.g., [27, Chap. 1.3]: find $\mathbf{u} \in \mathbf{V}$ and $\boldsymbol{\lambda} \in \mathbf{M}^+$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, \boldsymbol{\lambda}) &= f(\mathbf{v}) \,, && \mathbf{v} \in \mathbf{V} \,, \\ b(\mathbf{u}, \boldsymbol{\mu} - \boldsymbol{\lambda}) &\leq \langle g, (\boldsymbol{\mu} - \boldsymbol{\lambda}) \cdot \mathbf{n} \rangle \,, && \boldsymbol{\mu} \in \mathbf{M}^+ \,. \end{aligned} \tag{8}$$

The existence and uniqueness of $(\mathbf{u}, \boldsymbol{\lambda}) \in \mathbf{V} \times \mathbf{M}^+$ has been stated, e.g., in [27, Theorem 3.11 and Remark 3.10]. Moreover, $\mathbf{u}$ is also the unique solution of the minimization problem (5) and the variational inequality (7), and we find that $\boldsymbol{\lambda} = -\boldsymbol{\sigma}^s(\mathbf{u}^s)\mathbf{n}$.

2.2.2 Discretization

In this section, we consider a discrete formulation of the saddle point formulation (8). Mortar techniques with standard Lagrange multiplier spaces for contact problems have been considered and analyzed, e.g., in [8, 9, 15, 29, 30]. Here, we apply these techniques to dual Lagrange multiplier spaces, see, e.g., [72]. New optimal a priori error estimates for the discretization error in the H^1-norm and for the Lagrange multiplier in the $H^{-1/2}$-norm will be given. To approximate $\mathbf{V}$, we use standard conforming finite elements of lowest order on quasi-uniform simplicial, quadrilateral or hexahedral triangulations. The finite element space associated with the shape regular triangulation $\mathcal{T}_{h,\Omega^l}$ is denoted by $S_1(\Omega^l, \mathcal{T}_{h,\Omega^l})$. The meshsize h_l is defined by the maximal diameter of the elements in $\mathcal{T}_{h,\Omega^l}$ and the global meshsize h by $h := \max\{h_s, h_m\}$. We define the discrete product space $\mathbf{V}_h$ by

$$\mathbf{V}_h := \Big\{ \mathbf{v}_h \in \prod_{l \in \{s,m\}} \left[S_1\left(\Omega^l,\, \mathcal{T}_{h,\Omega^l}\right) \right]^d \,:\, \mathbf{v}_h|_{\Gamma_D^l} = \mathbf{0} \Big\} \subset \mathbf{V} \,.$$

As it is standard in the mortar context, the Lagrange multiplier space inherits its $(d-1)$-dimensional mesh from the d-dimensional triangulation on the slave side. We assume that Γ_C^s can be written as a union of edges in the two-dimensional case and faces in the three-dimensional case on the slave side. Here, we use discontinuous piecewise linear or bilinear nodal basis functions for the dual Lagrange multiplier. The discrete Lagrange multiplier space is denoted by $\mathbf{M}_h$ and can be spanned by $\{\psi_i \mathbf{e}_k,\ \ i = 1, \ldots, N_{\mathbf{M}_h},\ \ k = 1, \ldots, d\}$, where $\mathbf{e}_k$ denotes the k-th unit vector, ψ_i the i-th scalar dual basis function, and $N_{\mathbf{M}_h}$ is the number of nodes on the slave side of $\overline{\Gamma_C^s}$. In contrast to the general mortar setting, we do not have to remove the degrees of freedom of $\mathbf{M}_h$ on $\partial\Gamma_C^s$. We note that this has to be done if $\overline{\Gamma_D^s} \cap \overline{\Gamma_C^s} \neq \emptyset$ but not in our situation. Let $\mathbf{W}_h$ be the vector valued trace space of $[S_1\left(\Omega^s,\, \mathcal{T}_{h,\Omega^s}\right)]^d$ restricted to Γ_C^s. In the following, we define for each $\mathbf{v}_h = \sum_i \boldsymbol{\beta}_i \phi_i \in \mathbf{W}_h$ the discrete scalar product $\mathbf{v}_h \cdot \mathbf{n}_h$ by $\mathbf{v}_h \cdot \mathbf{n}_h := \sum_i (\boldsymbol{\beta}_i \cdot \mathbf{n}_i)\phi_i$, where $\mathbf{n}_i$ denotes the outer normal of the domain Ω^s at the node i. Similarly, we define the product $\boldsymbol{\mu}_h \cdot \mathbf{n}_h := \sum_i (\boldsymbol{\alpha}_i \cdot \mathbf{n}_i)\psi_i$ for each $\boldsymbol{\mu}_h = \sum_i \boldsymbol{\alpha}_i \psi_i \in \mathbf{M}_h$. We now define the discrete approximation in $\mathbf{M}^+$ by

$$\mathbf{M}_h^+ := \left\{ \boldsymbol{\mu}_h \in \mathbf{M}_h \,:\, \langle \boldsymbol{\mu}_h, \mathbf{v}_h \rangle \geq 0,\, \mathbf{v}_h \in \mathbf{W}_h^+ \right\} ,$$

where $\mathbf{W}_h^+ := \{\mathbf{v}_h \in \mathbf{W}_h \,:\, \mathbf{v}_h \cdot \mathbf{n}_h \geq 0\}$. In case of dual Lagrange multipliers, it is easy to verify that $\mathbf{M}_h^+$ is equal to

$$\left\{ \boldsymbol{\mu}_h = \sum_{i=1}^{N_{\mathbf{M}_h}} \boldsymbol{\alpha}_i \psi_i : \boldsymbol{\alpha}_i \in \mathbb{R}^d,\, \boldsymbol{\alpha}_i = \alpha_i^n \mathbf{n}_i,\, \alpha_i^n \in \mathbb{R},\, \alpha_i^n \geq 0,\, i = 1, \ldots, N_{\mathbf{M}_h} \right\} ,$$

see [38]. We note that this equivalence does not hold if standard Lagrange multiplier basis functions are used. We remark that $\mathbf{M}_h^+$ is not a conforming approach for the Lagrange multiplier space, i.e., $\mathbf{M}_h^+ \not\subset \mathbf{M}^+$. The discrete convex subset $\mathbf{K}_h \subset \mathbf{V}_h$ could be then defined by

$$\mathbf{K}_h := \left\{ \mathbf{v}_h \in \mathbf{V}_h \,:\, b(\boldsymbol{\mu}_h, \mathbf{v}_h) \leq \int_{\Gamma_C^s} g_h (\boldsymbol{\mu}_h \cdot \mathbf{n}_h)\, \mathrm{d}s,\ \ \boldsymbol{\mu}_h \in \mathbf{M}_h^+ \right\} ,$$

where g_h is a suitable approximation of g on $\mathbf{W}_h$. We remark that the non-penetration condition is fulfilled in a weak sense. Then the discrete analogon of the minimization problem (5) on the constrained space $\mathbf{K}_h$ is given by

$$J(\mathbf{u}_h) = \min_{\mathbf{v}_h \in \mathbf{K}_h} J(\mathbf{v}_h) \,. \tag{9}$$

Defining the discrete projection operator $P_h : \mathbf{V}_h \to \mathbf{K}_h$ by

$$\mathbf{v}_h = \sum_i \mathbf{v}_i \phi_i \mapsto P_h \mathbf{v}_h := \mathbf{v}_h - \frac{\max\{0,\, b(\mathbf{v}_h, \psi_i \mathbf{n}_i) - g_i\}}{b(\phi_i \mathbf{n}_i, \psi_i \mathbf{n}_i)} \mathbf{n}_i \phi_i \,,$$

where g_i denotes the weak distance given by $g_i := \int_{\Gamma_C^s} g_h \psi_i \, \mathrm{d}s$, it is easy to verify that $P_h \mathbf{v}_h = \mathbf{v}_h$ for $\mathbf{v}_h \in \mathbf{K}_h$ and in terms of (6), we get the equivalence of the discrete constrained minimization problem (9) and the discrete unconstrained minimization problem

$$\min_{\mathbf{v}_h \in \mathbf{V}_h} \tilde{J}_h(\mathbf{v}_h) \; , \tag{10}$$

where the discrete energy functional $\tilde{J}_h(\cdot)$ is defined by

$$\tilde{J}_h(\mathbf{v}_h) := \frac{1}{2} a(\mathbf{v}_h - P_h \mathbf{v}_h, \, \mathbf{v}_h - P_h \mathbf{v}_h) + \frac{1}{2} a(P_h \mathbf{v}_h, \, P_h \mathbf{v}_h) - f(P_h \mathbf{v}_h) \; .$$

We note that P_h is a projection because of the use of biorthogonal Lagrange multipliers. In the case of standard Lagrange multipliers, P_h is not a projection. Furthermore, the discrete mortar formulation of the saddle point problem (8) is now given by: find $\mathbf{u}_h \in \mathbf{V}_h$ and $\boldsymbol{\lambda}_h \in \mathbf{M}_h^+$ such that

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, \boldsymbol{\lambda}_h) &= f(\mathbf{v}_h) \; , && \mathbf{v}_h \in \mathbf{V}_h \; , \\ b(\mathbf{u}_h, \boldsymbol{\mu}_h - \boldsymbol{\lambda}_h) &\leq \langle g_h, (\boldsymbol{\mu}_h - \boldsymbol{\lambda}_h) \cdot \mathbf{n}_h \rangle \; , && \boldsymbol{\mu}_h \in \mathbf{M}_h^+ \; . \end{aligned} \tag{11}$$

Existence and uniqueness of a solution follow from a discrete inf-sup condition, see, e.g., [73, Chap. 1.2.3], for the spaces $\mathbf{M}_h$ and $\mathbf{V}_h$, see, e.g., [27, Chap. 2.4.2] or [30].

2.2.3 Optimal a priori Error Estimates

To achieve optimal a priori estimates for the formulation (11) which is equivalent to the minimization problems (9) and (10), we consider the simplified case that the two bodies in the initial configuration are in contact on their common boundary part, i.e., $\Gamma_C^m = \Gamma_C^s$ and $g = g_h = 0$, and that the normal vector $\mathbf{n} = \mathbf{n}^s = -\mathbf{n}^m$ is constant on Γ_C^s. To prove the new optimal a priori error estimate for the discretization errors in the primal and dual variables, the crucial point is the following abstract lemma, see, e.g., [30].

Lemma 1. *Let $(\mathbf{u}, \boldsymbol{\lambda}) \in \mathbf{V} \times \mathbf{M}^+$ be the solution of (8) and let $(\mathbf{u}_h, \boldsymbol{\lambda}_h) \in \mathbf{V}_h \times \mathbf{M}_h^+$ be the solution of the discrete formulation (11). Then there exists a constant $0 < C < \infty$ independent of the meshsize h, such that*

$$\begin{aligned} &\|\mathbf{u} - \mathbf{u}_h\|_{1,\Omega} + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C^s} \leq C \left\{ \inf_{\mathbf{v}_h \in \mathbf{V}_h} \|\mathbf{u} - \mathbf{v}_h\|_{1,\Omega} \right. \\ &\left. + \inf_{\boldsymbol{\mu}_h \in \mathbf{M}_h^n} \|\boldsymbol{\lambda} - \boldsymbol{\mu}_h\|_{-\frac{1}{2},\Gamma_C^s} + \max\Big(b(\mathbf{u}, \boldsymbol{\lambda}_h), \, 0\Big)^{\frac{1}{2}} + \max\Big(b(\mathbf{u}_h, \boldsymbol{\lambda}), \, 0\Big)^{\frac{1}{2}} \right\} \; . \end{aligned}$$

We remark that the term $\max\big(b(\mathbf{u}_h, \boldsymbol{\lambda}), \, 0\big)$ takes into account the discrete penetration of the two bodies on the actual contact set γ_a. The term

$\max\big(b(\mathbf{u},\boldsymbol{\lambda}_h),\,0\big)$ can be greater than zero if the discrete Lagrange multiplier $\boldsymbol{\lambda}_h \cdot \mathbf{n}$ is negative on a part of $\Gamma_C^s \backslash \gamma_a$. We recall that $\mathbf{M}_h^+$ is not a subspace of $\mathbf{M}^+$, and thus $\boldsymbol{\lambda}_h \cdot \mathbf{n}$, $\boldsymbol{\lambda}_h \in \mathbf{M}_h^+$, can be smaller than zero. The first two terms in the upper bound of Lemma 1 are the best approximation errors. They reflect the quality of the approximation of the spaces $\mathbf{V}_h$ and $\mathbf{M}_h$. The third and the fourth term are the consistency error terms and measure the nonconformity of the approach. In the following, we provide upper bounds for the consistency errors.

Lemma 2. *Let* $(\mathbf{u},\boldsymbol{\lambda}) \in \mathbf{V} \times \mathbf{M}^+$ *be the solution of (8) and let* $(\mathbf{u}_h,\boldsymbol{\lambda}_h) \in \mathbf{V}_h \times \mathbf{M}_h^+$ *be the solution of the discrete formulation (11). Under the regularity assumption* $\mathbf{u} \in \left[H^{\frac{3}{2}+\nu}(\Omega)\right]^d$, $0 < \nu \leq \frac{1}{2}$, *we then have the a priori error estimate*

$$b(\mathbf{u}_h,\boldsymbol{\lambda}) \leq C\left(h^{1+2\nu}|\mathbf{u}|^2_{\frac{3}{2}+\nu,\Omega} + h^{\frac{1}{2}+\nu}|\mathbf{u}|_{\frac{3}{2}+\nu,\Omega}\|\mathbf{u}-\mathbf{u}_h\|_{1,\Omega}\right)$$

for a positive constant $C < \infty$ *independent of* h. *Moreover, under a suitable regularity assumption on the actual contact zone, see [38, Assumption 3.1], the a priori estimate*

$$b(\mathbf{u},\boldsymbol{\lambda}_h) \leq Ch^{\frac{1}{2}+\nu}|\mathbf{u}|_{\frac{3}{2}+\nu,\Omega}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C^s}$$

holds for a positive constant $C < \infty$ *independent of* $h < h_0$.

Using the well-known approximation property for the spaces $\mathbf{V}_h$ and $\mathbf{M}_h$, Lemma 2 and Young's inequality, we find an optimal a priori estimate for multibody contact problems.

Theorem 1. *Under the assumptions of Lemma 2, we get the optimal a priori error estimate*

$$\|\mathbf{u}-\mathbf{u}_h\|_{1,\Omega} + \|\boldsymbol{\lambda}-\boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C^s} \leq Ch^{\frac{1}{2}+\nu}|\mathbf{u}|_{\frac{3}{2}+\nu,\Omega}$$

for a positive constant C *independent of* h.

Replacing $\mathbf{V}_h$ by quadratic finite elements, we obtain a higher order a priori estimate. Quadratic finite elements and linear Lagrange multipliers yield an order $h^{\frac{1}{2}+\nu}$, $0 < \nu < 1$, upper bound for the discretization error if the solution is $H^{\frac{3}{2}+\nu}$-regular. Replacing the linear Lagrange multiplier space by quadratic Lagrange multipliers does not give a higher order, see, e.g., [30]. For a proof and numerical results comparing quadratic finite elements with linear Lagrange multipliers and with quadratic Lagrange multipliers, we refer to [34]. For the case of curved boundaries, we refer to [20] carried out in the framework of the project C12. Furthermore, optimal a priori error estimates for a simplified frictional problem were achieved in the framework of this project in [35]. Combining Theorem 1 with the results in [35] we have an optimal a priori error estimate for a two-body contact problem with a given friction bound.

2.2.4 Numerical Results

In this section, we show numerical results for the discretization error of the displacement in the L^2- and the H^1-norm and of the Lagrange multiplier in a weighted L^2-norm on the contact boundary. Moreover, we present numerical results for nearly incompressible materials. Using standard Q1-elements shows volume locking. To overcome this effect, we propose an abstract framework based on generalized three-field formulations. Numerical results demonstrate the quality of the introduced discretizations. For the iterative algorithm to solve the nonlinear system, we refer to Sect. 2.3. The implementation of all numerical examples presented in Sects. 2.2–2.5 is based on the finite element toolbox UG, see [5]. We start with a coarse triangulation and use uniform refinement techniques.

Hertz Contact Problem

As an example, we consider the Hertz contact problem of a linear elastic circle with a plane. In general, a two-body contact problem does not admit an analytical solution. To evaluate the discretization errors, we compute a reference solution denoted by $\mathbf{u}_{\text{ref}}$ for the displacements and by $\boldsymbol{\lambda}_{\text{ref}}$ for the Lagrange multiplier corresponding to a finer mesh. We note that in our example, the meshsize h_{ref} for the reference solution satisfies $h_{\text{ref}} \leq 1/4h$.

An elastic disc is pressed by a single point load $\mathbf{f}^s = (0, -100)^\top$ on the top to a rigid plane. In our setting, we replace the rigid plane by a linear elastic rectangle. Young´s modulus of the rectangle is set to be larger than the one of the disc. We apply homogeneous Dirichlet boundary conditions at the bottom and at the two sides of the rectangle. On the disc with radius $r = 1$, we set the material parameters to be $E^s = 7000$ and $\nu^s = 0.3$. For the rectangle with height one, we use $E^m = 10^6$ and $\nu^m = 0.45$. The problem definition and the geometry are shown in Fig. 1(a). We replace the point load by a surface load to avoid a strong singularity on the upper part of the disc, which is assumed to be the slave side. The maximal normal contact stress of the Hertz contact problem is then given by $\boldsymbol{\lambda}_n^{\max} = 494.8$, see, e.g., [28, 45]. Figure 1(b) illustrates the numerical approximation of the contact stress. From

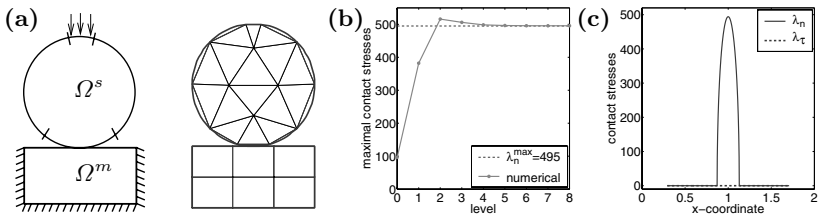


Fig. 1. **(a)** Problem definition and initial triangulation, **(b)** maximal contact stresses, **(c)** contact stresses

level 4 on, the discrete maximal normal stress is a very good approximation of the analytical one. In Fig. 1(c), the normal and tangential contribution of the Lagrange multiplier are given.

Table 1. Relative $L^2(\Omega)$-, relative $H^1(\Omega)$-error of $\mathbf{u}_h$ with respect to $\mathbf{u}_{\text{ref}}$, weighted $L^2(\Gamma_C^s)$-error of $\boldsymbol{\lambda_h}$ with respect to $\boldsymbol{\lambda}_{\text{ref}}$ and the numerical convergence orders

level	$\|\mathbf{u}_h - \mathbf{u}_{\text{ref}}\|_{0,\Omega}/\|\mathbf{u}_{\text{ref}}\|_{0,\Omega}$		$\|\mathbf{u}_h - \mathbf{u}_{\text{ref}}\|_{1,\Omega}/\|\mathbf{u}_{\text{ref}}\|_{1,\Omega}$		$\|\boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{\text{ref}}\|_{-\frac{1}{2},h,\Gamma_C^s}$	
1	$5.227370e-02$	–	$4.663629e-01$	–	$5.845408+01$	–
2	$1.381026e-02$	1.92	$3.214708e-01$	0.54	$4.998951+01$	0.23
3	$4.844067e-03$	1.51	$1.807113e-01$	0.83	$2.120954+01$	1.24
4	$1.350194e-03$	1.84	$9.735769e-02$	0.89	$8.377665+00$	1.34
5	$4.175901e-04$	1.69	$5.111927e-02$	0.93	$3.269378+00$	1.36
6	$1.102450e-04$	1.92	$2.584385e-02$	0.98	$1.168388+00$	1.48

To compute the discretization errors, we use the mortar finite element solution on level 8 as reference solution. The errors are given in Table 1. Asymptotically, we observe optimal convergence rates. In the L^2-norm, the convergence rate tends to two with increasing number of refinement steps, whereas the convergence rate in the H^1-norm tends to one. We observe asymptotically a convergence rate of 1.5 for the Lagrange multiplier in the weighted L^2-norm on the contact zone. This results from the observation that the error in the energy norm restricted to a strip of width h can be bounded by $C\sqrt{h}\|\mathbf{u}\|_{1,\Omega}$.

Nearly Incompressible Materials

It is well-known that low-order finite elements based on four-noded quadrilaterals or eight-noded hexahedral show volume locking effects in the nearly incompressible case; in other words, they do not converge uniformly with respect to the Lamé parameter λ. Three-field mixed formulations in which the unknown variables are displacement, stress and strain are a popular approach to overcoming this problem. The corresponding weak formulation is known as the Hu–Washizu formulation, see [33, 70], and incorporates weak statements of the equation of equilibrium, the strain-displacement equation and the elasticity equation. The method of enhanced assumed strains is a special case of the Hu–Washizu formulation which also serves as the generating formulation for approaches such as the method of mixed enhanced strains, see [43]. In this section, we present a more general form of the Hu–Washizu formulation, the so-called α-formulation and the corresponding optimal a priori error estimate.

To introduce the modified Hu–Washizu formulation, we consider a three-field formulation depending on a parameter $\alpha \in \mathbb{R}$ and the three fields $(\mathbf{u}, \mathbf{d}, \boldsymbol{\sigma}) \in [H^1(\Omega)]^d \times [L^2(\Omega)]^{d\times d} \times [L^2(\Omega)]^{d\times d}$, reflecting the displacement, strain and stress, respectively. Then we have to find $(\mathbf{u}, \mathbf{d}, \boldsymbol{\sigma})$ such that

$$
\begin{aligned}
&a_\alpha((\mathbf{u},\mathbf{d}),(\mathbf{v},\mathbf{e})) + b_\alpha((\mathbf{v},\mathbf{e}),\boldsymbol{\sigma}) = f(\mathbf{v}) \,, \qquad (\mathbf{v},\mathbf{e}) \in [H^1(\Omega)]^d \times [L^2(\Omega)]^{d\times d} \,, \\
&b_\alpha((\mathbf{u},\mathbf{d}),\boldsymbol{\tau}) - \tfrac{(\lambda-\alpha\mu)}{(2\mu+d\lambda)^2}\, c(\boldsymbol{\sigma},\boldsymbol{\tau}) = 0 \,, \qquad \boldsymbol{\tau} \in [L^2(\Omega)]^{d\times d} \,,
\end{aligned}
$$

where the bilinear forms are defined by

$$
\begin{aligned}
a_\alpha((\mathbf{u},\mathbf{d}),(\mathbf{v},\mathbf{e})) &:= 2\mu\big(\mathbf{d},\mathbf{e}\big) + \alpha\mu\big(\mathrm{tr}(\mathbf{d}),\mathrm{tr}(\mathbf{e})\big) \,, \\
b_\alpha((\mathbf{v},\mathbf{e}),\boldsymbol{\sigma}) &:= \big(\boldsymbol{\varepsilon}(\mathbf{v}) - 2\mu\,\mathcal{C}^{-1}\mathbf{e},\boldsymbol{\sigma}\big) - \tfrac{\alpha\mu}{2\mu+d\lambda}\big(\mathrm{tr}(\boldsymbol{\sigma}),\mathrm{tr}(\mathbf{e})\big) \,, \\
c(\boldsymbol{\sigma},\boldsymbol{\tau}) &:= \big(\mathrm{tr}(\boldsymbol{\sigma}),\mathrm{tr}(\boldsymbol{\tau})\big) \,,
\end{aligned}
$$

where we denote by $(\cdot,\cdot)$ the L^2-scalar product. The standard Hu–Washizu formulation is given by $\alpha = \lambda/\mu$, and thus the factor in front of the bilinear form $c(\cdot,\cdot)$ is equal to zero. The discrete formulation is obtained by using four-noded quadrilateral or eight-noded hexahedral elements for the displacement field, as well as element-wise defined function spaces $S_h \subset D_h \subset L^2(\Omega)$ for the stress and strain fields, respectively. We refer to [48] for details. The following result obtained by using an equivalent displacement formulation

$$
\int_\Omega \mathcal{C}_h \Pi_h \boldsymbol{\varepsilon}(\mathbf{u}_h) : \Pi_h \boldsymbol{\varepsilon}(\mathbf{v}_h)\,\mathrm{d}s = f(\mathbf{v}_h) \,, \qquad \mathbf{v}_h \in \mathbf{V}_h \tag{12}
$$

gives an optimal and robust estimate for the displacement field, for the proof we refer to [48]. Here, Π_h is an element-wise projection of the strain $\boldsymbol{\varepsilon}(\mathbf{u}_h)$ onto the space S_h and $\mathcal{C}_h$ is a modification of Hooke's tensor (4). Under some appropriate assumptions on the spaces S_h and D_h and under the regularity assumption $\|\mathbf{u}\|_{2,\Omega}+\lambda\|\,\mathrm{div}\,(\mathbf{u})\|_{1,\Omega} \le C\,\|f\|_{0,\Omega}$, we obtain for the case without contact an optimal a priori estimate for the discretization error

$$
\|\mathbf{u} - \mathbf{u}_h\|_{1,\Omega} \le Ch\|f\|_{0,\Omega} \,,
$$

where $C < \infty$ is independent of λ and h. It is easy to show that in the case $S_h \subset D_h$ and $\mathcal{C}D_h \subset D_h$ the numerical solution is independent of the parameter α, see [48]. For extensions of these three-field formulations to geometrically nonlinear and general hyperelastic material laws, we refer to [11].

To show the behavior of the proposed formulation in comparison with the standard Q1-formulation, we consider the example of Sect. 2.2.4. with different material parameters on Ω^s and Ω^m and a point load $\mathbf{f}^s = (0,\,-10)^\top$. Table 2 shows the computed and analytical values for the maximal contact stress $\boldsymbol{\lambda}_h^{\max}$ and the contact radius for both formulations compared to the analytical ones. We investigate the behavior of the formulation for different values for the parameter $\nu^m \to 0.5$. The standard formulation shows a locking effect, whereas the proposed formulation shows a good behavior for $\nu^m \to 0.5$.

In a second example, we press a nearly incompressible punch Ω^s against two elastic blocks. The two elastic blocks are discretized by standard P1-elements whereas the punch is discretized by the standard Q1-formulation or by a modified Q1-formulation of the type (12). Figure 2 shows the effective von Mises stress on the distorted bodies and the distorted mesh on the contact zone

Table 2. Comparison of Numerical with Analytical results for $\nu^m \to 0.5$

ν^m	contact stress $\boldsymbol{\lambda}_h^{\max}$			contact radius		
	Numerical		Analytical	Numerical		Analytical
	Q1	Hu–Washizu		Q1	Hu–Washizu	
0.1	16.783952	16.770766	17.1563	0.3437	0.3750	0.37092
0.45	19.097267	19.045985	18.9238	0.3125	0.3125	0.33628
0.49	19.925502	19.568954	19.3374	0.2812	0.3125	0.32908
0.499	21.568993	19.700721	19.4394	0.2500	0.3125	0.32735
0.4999	31.045828	19.714365	19.4497	0.1875	0.3125	0.32718
0.49999	52.259179	19.715747	19.4507	0.1250	0.3125	0.32716
0.499999	58.462228	19.715886	19.4508	0.0937	0.3125	0.32714

for the modified Q1-formulation (a) and the Q1-formulation (b). As expected, we observe a locking effect for the standard Q1-formulation, whereas our new formulation does not exhibit locking effects.

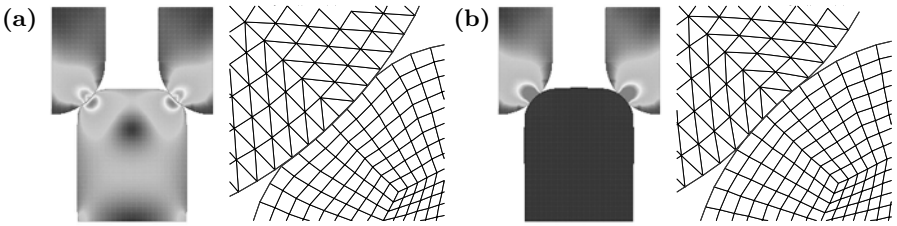


Fig. 2. (a) Modified Hu–Washizu formulation, **(b)** standard Q1-formulation for a nearly incompressible body pressed again two elastic obstacles

2.3 Inexact Primal-Dual Active Set Strategy

In this section, we present the inexact primal-dual algorithm for solving the nonlinear two-body contact problem with Coulomb friction. The essential point to apply the algorithm to the two-body case is a basis transformation of the finite element basis. This algorithm is based on a nonlinear complementarity function and can be interpreted as a semismooth Newton method. For primal-dual active set strategies we refer to [31]. The applications to frictionless contact problems are discussed in [38] and for the frictional case in [35, 37]. Newton methods for contact problems are also discussed in [3, 13].

2.3.1 Basis Transformation

Due to the use of dual Lagrange multipliers, we can apply locally a basis transformation of the nodal finite element basis. In the new constrained basis, a linear combination of the nodal basis functions on the slave side is glued to

the nodal basis functions of the master side. The construction is done such that the new basis functions on the master side are weakly continuous at the interface.

Thus, the basis functions on the slave side in the new basis describe the relative displacement between the contact interfaces, i.e., the jump in the displacement denoted by $[\mathbf{u}_h]$. In this new basis, the two-body contact problem has the same structure as a one-body problem. Moreover, all constraints at the contact zone are restricted to the degrees of freedom on the slave side. For details, we refer to [39, 74, 75]. We separate the nodes of the finite element mesh into three disjoint sets. The set $\mathcal{S}$ contains all nodes on Γ_C^s, the set $\mathcal{M}$ all nodes on Γ_C^m and the set $\mathcal{N}$ all remaining nodes. The finite element space $\mathbf{V}_h$ can be written in terms of the standard finite element basis $\boldsymbol{\phi}$ as $\mathbf{V}_h = \operatorname{span}\{\phi_p \mathbf{e}_k,\ p \in \mathcal{S} \cup \mathcal{M} \cup \mathcal{N},\ 1 \le k \le d\}$. Additionally to the basis $\boldsymbol{\phi}$, we denote the constrained finite element basis by $\hat{\boldsymbol{\phi}}$. To introduce these basis functions, we define the entries of the coupling matrices D and M between the finite element basis functions ϕ_p and the basis functions for the Lagrange multiplier space ψ_p by

$$\begin{aligned} D[p,q] &:= \langle \psi_p, \phi_q \rangle \, Id_d =: D_p \, Id_d \,, \quad && p, q \in \mathcal{S} \,,\ D_p \in \mathbb{R} \,, \\ M[p,q] &:= \langle \psi_p, \phi_q \rangle \, Id_d \,, && p \in \mathcal{S} \,,\ q \in \mathcal{M} \,. \end{aligned}$$

Due to the biorthogonality, the matrix D is diagonal, see, e.g., [72]. In terms of $\hat{M} := D^{-1}M$, we obtain the constrained basis $\hat{\boldsymbol{\phi}}$ of $\mathbf{V}_h$ from the nodal basis $\boldsymbol{\phi}$ of $\mathbf{V}_h$ by the transformation

$$\hat{\boldsymbol{\phi}} := \begin{pmatrix} \hat{\phi}_{\mathcal{N}} \\ \hat{\phi}_{\mathcal{M}} \\ \hat{\phi}_{\mathcal{S}} \end{pmatrix} := \begin{pmatrix} Id & 0 & 0 \\ 0 & Id & \hat{M}^\top \\ 0 & 0 & Id \end{pmatrix} \begin{pmatrix} \phi_{\mathcal{N}} \\ \phi_{\mathcal{M}} \\ \phi_{\mathcal{S}} \end{pmatrix} =: Q\boldsymbol{\phi} \,.$$

We note that only basis functions associated with a node $p \in \mathcal{M}$ are changed, and that by definition

$$b(\hat{\mathbf{u}}_q \hat{\phi}_q, \boldsymbol{\mu}_p \psi_p) = 0 \,, \quad \hat{\mathbf{u}}_q, \boldsymbol{\mu}_p \in \mathbb{R}^d, \quad p \in \mathcal{S} \,,\ q \in \mathcal{M} \,.$$

This equation states that the jump of the finite element basis functions on the master side vanishes in a weak sense. We note that we denote in the following the coefficient of $\mathbf{u}_h$ at the node p with respect to the new basis by $\hat{\mathbf{u}}_p$.

2.3.2 Nonlinear Complementarity Functions and the Algorithm

The second line of (11) results due to the use of dual Lagrange multipliers and the basis transformation in a point-wise condition

$$\hat{u}_{n,p,s} \le g_p \,, \quad \lambda_{n,p,sq} \ge 0 \,, \quad \lambda_{n,p,sq}(\hat{u}_{n,p,s} - g_p) = 0 \,, \qquad p \in \mathcal{S} \,, \tag{13}$$

where the scaled values at the node p are given by $\hat{u}_{n,p,s} := \mathbf{n}_p^\top D_p \hat{\mathbf{u}}_p$ and $\lambda_{n,p,sq} := \boldsymbol{\lambda}_p^\top D_p^2 \mathbf{n}_p$. We note that we use different scaling factors and that

the entries of D_p are proportional to the meshsize h. This is motivated by the fact that the $H^{-1/2}$-norm for the Lagrange multiplier and the $H^{1/2}$-norm for the displacements have the same error reduction but scale differently. The proposed scaling factors lead to better convergence rates in the inexact version of the algorithms. Defining at each node the nonlinear complementarity function

$$C_n(\hat{u}_{n,p,s}, \lambda_{n,p,sq}) := \lambda_{n,p,sq} - \max\{0,\, \lambda_{n,p,sq} + c_n(\hat{u}_{n,p,s} - g_p)\} \tag{14}$$

for any $c_n > 0$, then (13) can be equivalently written as $C_n(\hat{u}_{n,p,s}, \lambda_{n,p,sq}) = 0$. The discrete Coulomb friction conditions (3) are given by

$$\begin{cases} \|\boldsymbol{\lambda}_{\tau,p,s}\| \le \mathcal{F}|\lambda_{n,p,s}| \,, \\ \|\boldsymbol{\lambda}_{\tau,p,s}\| < \mathcal{F}|\lambda_{n,p,s}| \Rightarrow \hat{\mathbf{u}}_{\tau,p} = 0 \,, & \text{for all } p \in \mathcal{S} \,. \\ \|\boldsymbol{\lambda}_{\tau,p,s}\| = \mathcal{F}|\lambda_{n,p,s}| \Rightarrow \boldsymbol{\lambda}_{\tau,p,s} = \beta^2 \hat{\mathbf{u}}_{\tau,p} \,, \end{cases} \tag{15}$$

The scaled tangential vectors at the node p for $d = 3$ are given by

$$\boldsymbol{\lambda}_{\tau,p,s} := \begin{pmatrix} \boldsymbol{\lambda}_p^\top D_p \boldsymbol{\tau}_{1,p} \\ \boldsymbol{\lambda}_p^\top D_p \boldsymbol{\tau}_{2,p} \end{pmatrix} \in \mathbb{R}^2 \,, \qquad \hat{\mathbf{u}}_{\tau,p} := \begin{pmatrix} \hat{\mathbf{u}}_p^\top \boldsymbol{\tau}_{1,p} \\ \hat{\mathbf{u}}_p^\top \boldsymbol{\tau}_{2,p} \end{pmatrix} \in \mathbb{R}^2 \,.$$

Here, for simplicity, $\mathcal{F}$ is assumed to be constant; otherwise we have to replace the expression $\mathcal{F}|\lambda_{n,p,s}|$ in (15) by $|\lambda_{n,p}| \int_{\Gamma_C^s} \mathcal{F}\, \phi_p \, \mathrm{d}s$. Defining for any constant $c_\tau > 0$ at each node $p \in \mathcal{S}$ the nonlinear complementarity function

$$C_\tau(\hat{\mathbf{u}}_p, \boldsymbol{\lambda}_p) := \begin{cases} \boldsymbol{\lambda}_{\tau,p,s} & \text{if } \; b_{n,p} \le 0 \,, \\ \max\{b_{n,p}, \|\boldsymbol{\lambda}_{\tau,p,s} + c_\tau \hat{\mathbf{u}}_{\tau,p}\|\}\boldsymbol{\lambda}_{\tau,p,s} & \\ \quad + \max\{0, b_{n,p}\}(\boldsymbol{\lambda}_{\tau,p,s} + c_\tau \hat{\mathbf{u}}_{\tau,p}) & \text{if } \; b_{n,p} > 0 \end{cases} \tag{16}$$

with a suitable friction bound $b_{n,p}$, (15) can be equivalently written as $C_\tau(\hat{\mathbf{u}}_p, \boldsymbol{\lambda}_p) = 0$. Now, we consider two different solution strategies. In our first approach (fixpoint Newton), we combine a Newton approach for a given friction bound $b_{n,p}$ with a fixpoint iteration in the friction bound. Here we set $b_{n,p} := \lambda_{n,p,s}$ and handle it as a given constant. The second method (full Newton) is based on a full Newton scheme and $b_{n,p}$ is defined in terms of $\boldsymbol{\lambda}_p$ and $\hat{\boldsymbol{u}}_p$. For this approach $b_{n,p}(\hat{\mathbf{u}}_p, \boldsymbol{\lambda}_p) := \lambda_{n,p,s} + (c_n/D_p)(\hat{u}_{n,p,s} - g_p)$ is regarded as a function of $\hat{\mathbf{u}}_p$ and $\boldsymbol{\lambda}_p$.

Now we can formulate the inexact primal-dual active set algorithm based on a semismooth Newton method for (14) and (16). We note that both approaches can be implemented in the form of Algorithm 1.

Algorithm 1: Inexact primal-dual active set strategy for Coulomb friction problem in the three-dimensional case

(0) Set $l = 1$, $k = 1$ and choose $c_n > 0$, $c_\tau > 0$ and $m \in \mathbb{N}$.
Initialize $\mathbf{u}_h^{0,0}$ and $\boldsymbol{\lambda}_h^0$.

(1) Define the active and inactive sets by

$$\begin{aligned}
\mathcal{A}_n^k &:= \left\{p \in \mathcal{S} \,:\, \lambda_{n,p,sq}^{k-1} + c_n\big(\hat{u}_{n,p,s}^{k-1,0} - g_p\big) > 0\right\} \,, \\
\mathcal{I}_n^k &:= \left\{p \in \mathcal{S} \,:\, \lambda_{n,p,sq}^{k-1} + c_n\big(\hat{u}_{n,p,s}^{k-1,0} - g_p\big) \leq 0\right\} \,, \\
\mathcal{A}_\tau^k &:= \left\{p \in \mathcal{S} \,:\, \|\boldsymbol{\lambda}_{\tau,p,s}^{k-1} + c_\tau \hat{\mathbf{u}}_{\tau,p}^{k-1,0}\| - b_{n,p}^{k-1} \geq 0\right\} \,, \\
\mathcal{I}_\tau^k &:= \left\{p \in \mathcal{S} \,:\, \|\boldsymbol{\lambda}_{\tau,p,s}^{k-1} + c_\tau \hat{\mathbf{u}}_{\tau,p}^{k-1,0}\| - b_{n,p}^{k-1} < 0\right\} \,.
\end{aligned}$$

(2) For $i = 1, \ldots, m$, compute

$$\mathbf{u}_h^{k,i} = MG(\mathbf{u}_h^{k,i-1}, \mathcal{A}_n^k, \mathcal{I}_n^k, \mathcal{A}_\tau^k, \mathcal{I}_\tau^k, \mathbf{u}_h^{k-1,m}, \boldsymbol{\lambda}_h^{k-1}) \,.$$

(3) Compute the Lagrange multiplier as

$$\boldsymbol{\lambda}_h^k = D^{-1}\big(\mathbf{f}_{\mathcal{S}} - A_{\mathcal{S}\star}\mathbf{u}_h^{k,m}\big) \,.$$

(4) If $\|\mathbf{u}_h^{k,m} - \mathbf{u}_h^{k,0}\| \,/\, \|\mathbf{u}_h^{k,m}\| < \varepsilon_u$ stop.

(5) Set $\mathbf{u}_h^{k+1,0} = \mathbf{u}_h^{k,m}$ and $k = k + 1$ and go to step (1).

In Step (2), we denote by $MG(\cdots)$ one multigrid step for the linear problem obtained from the equilibrium condition resulting from the first equation in (11) and the additional boundary conditions on the active and inactive sets defined in Step (1) after eliminating the Lagrange multiplier $\boldsymbol{\lambda}_h^k$. We note that the two approaches only differ in Step (2). In case of the full Newton, we have $\mathcal{I}_n^k \subset \mathcal{A}_\tau^k$. Moreover, for a node $p \in \mathcal{I}_n^k \cap \mathcal{A}_\tau^k$, we find the Neumann boundary condition $\boldsymbol{\lambda}^k = 0$. For $p \in \mathcal{A}_n^k \cap \mathcal{I}_\tau^k$, i.e. a stick node, we get in normal direction Dirichlet conditions $\hat{u}_{n,p,s} = g_p$ and in tangential direction Robin boundary conditions $L_p^k \hat{\mathbf{u}}_{\tau,p}^k + R_p^k \lambda_{n,p}^k = r_p^k$ which tends in the limit case to the Dirichlet condition $\hat{u}_{\tau,p} = 0$. For $p \in \mathcal{A}_n^k \cap \mathcal{A}_\tau^k$, i.e. a slip node, we get again the Dirichlet condition $\hat{u}_{n,p,s} = g_p$ and in tangential direction the Robin boundary condition $G_p^k \hat{\mathbf{u}}_{\tau,p}^k + H_p^k \boldsymbol{\lambda}_p^k = h_p^k$. For the definitions of the matrices L_p^k, R_p^k, G_p^k, H_p^k and the vectors r_p^k, h_p^k, we refer to [37]. In contrast to the full Newton, the condition $\mathcal{I}_n^k \subset \mathcal{A}_\tau^k$ is not necessarily satisfied in the case of the fixpoint Newton, and for $p \in \mathcal{I}_\tau^k$ we obtain the boundary condition $\hat{\mathbf{u}}_\tau^k = 0$.

Algorithm 1 can be interpreted as a nonlinear multigrid method. For the choice $m = \infty$, we obtain an exact semismooth Newton method. In comparison with the inexact case, e.g., $m = 1$, this approach is rather costly.

2.3.3 Numerical Examples

First we consider the example presented in Fig. 3. The lower domain Ω^m, assumed to be the master side, models a spherical shell that is fixed at the outer boundary. Against this shell we press the body modeled by the domain

Ω^s which is assumed to be the slave side. At the top surface of Ω^s, we apply the surface traction $(0, 0, -150\exp(-100r^2))^\top$, where r denotes the distance to the midpoint of the top surface of Ω^s. The geometry is given by $r_i = 0.7$, $r_a = 1.0$, $r = 0.6$, $h = 0.5$ and $d = 0.3$. In Ω^s, we use the material law (4) with a Young modulus $E^s = 300$ and a Poisson ratio $\nu^s = 0.3$, while in Ω^m we have $E^m = 400$ and $\nu^m = 0\,3$.

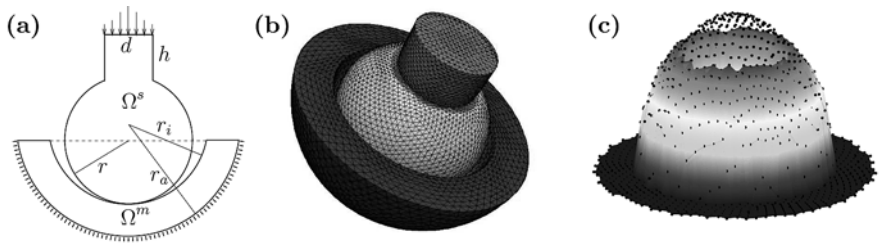


Fig. 3. (a) Problem definition, **(b)** deformed mesh with effective von Mises stress on level 3, **(c)** visualization of $\|\boldsymbol{\lambda}_\tau\|$ with the friction bound $\mathcal{F}\lambda_n$ (dots) for $\mathcal{F} = 0.7$

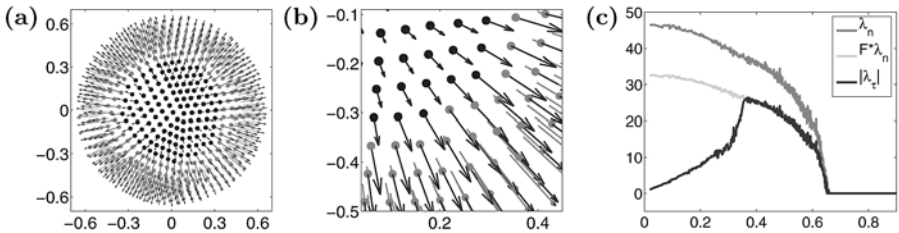


Fig. 4. (a) Visualization of the nodes being in contact together with the deformation and stress vectors, **(b)** a cutout, **(c)** a two-dimensional visualization of the Lagrange multipliers for $\mathcal{F} = 0.7$

The results for the friction coefficient $\mathcal{F} = 0.7$ are shown in Figs. 3 and 4. Figure 3(b) shows the deformed body with the effective von Mises stress and 3(c) the norm of the tangential part $\|\boldsymbol{\lambda}_\tau\|$. The dots represent the friction bound $0.7\lambda_n$. Figure 4 shows the nodes being in contact, their relative tangential slip $\hat{\mathbf{u}}_{\tau,p}$ (lines) and the tangential contact pressure $\|\boldsymbol{\lambda}_{\tau,p}\|$ (arrows). We remark that the nodes in the center with only an arrow are the sticky nodes, the others are slippy. In Fig. 3(c), we show the normal and tangential part of the Lagrange multiplier and the friction bound for all nodes $p \in \mathcal{S}$ over their distance to the midpoint of the contact zone on the surface of Ω^s. The small oscillations in the plot occur due to the fact that all nodes $p \in \mathcal{S}$ over the whole contact zone are presented and that we work with an unstructured mesh.

Table 3. Behavior of Algorithm 1: Necessary numbers $k_{\max}$ and K_l on each level l for $c_n = c_\tau = 1000$, $m = 1$, $\varepsilon_u = 10^{-9}$ and $\mathcal{F} = 0.7$

l	DOF	fixpoint approach				full Newton approach	
		K_l^1	$k_{\max}^1$	K_l^∞	$k_{\max}^\infty$	K_l^∞	$k_{\max}^\infty$
0	104	4	25	4	25	4	9
1	541	9	34	7	21	6	10
2	3384	8	26	6	19	7	11
3	23694	9	31	8	19	9	12

In Table 3, we investigate the behavior of Algorithm 1 for the first example. We show for the fixpoint Newton approach and the full Newton approach on the different levels l the numbers K_l^∞ and $k_{\max}^\infty$. Here K_l^m is the minimal iteration number in which the correct active sets are found and never changed in the following iterations. By $k_{\max}^m$ we denote the total number of iteration steps in Algorithm 1 to obtain the required accuracy. We recall that m stands for the number of multigrid iterations in Step (2) of our algorithm. While K_l^m is almost the same for all strategies, the necessary steps $k_{\max}^\infty$ for the fixpoint approach are approximately two times the steps of the full Newton approach. After finding the correct sets, the full Newton approach only needs about three to four steps to converge. In these steps, the algorithm adjusts the direction of the tangential traction to the tangential displacement. Furthermore, we compare for the fixpoint Newton approach the exact approach with the inexact one, i.e., $m = 1$. Comparing the numbers K_l^1 with K_l^∞, we observe that for the inexact approach we need one additional iteration step to obtain the correct active and inactive sets. Of course, the numbers $k_{\max}^1$ are larger compared with $k_{\max}^\infty$ but we point out that for the inexact case we have to do only one step of the multigrid iteration process in each step whereas in the exact case, we have to solve a full linear system in each step.

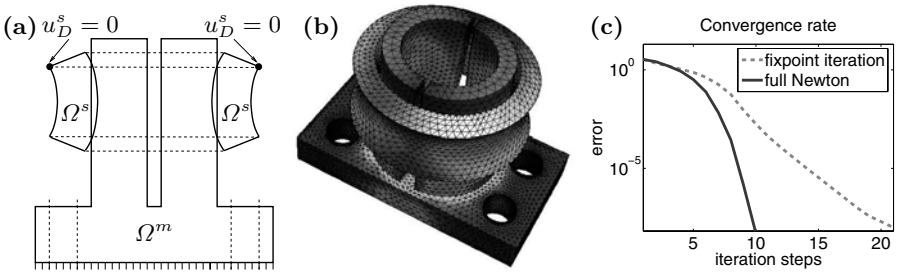


Fig. 5. **(a)** Problem definition, **(b)** distorted domains with the effective von Mises stress, **(c)** convergence rates for $c_n = c_\tau = 10^8$ and $\mathbf{u}_h^{0,0} = \boldsymbol{\lambda}_h^0 = 0$

In a second example, we compare the convergence rates of the fixpoint iteration with the ones of the full Newton approach. We consider the situation presented in Fig. 5. In Fig. 5(a) a two-dimensional cross section of the problem definition is shown. We use the linear material law (4) with a Young modulus $E^m = 8.13 \times 10^8$ and a Poisson ratio $\nu^m = 0.3$ in Ω^s and $E^m = 9 \times 10^7$, $\nu^s = 0.3$ in Ω^s. For the friction coefficient we use $\mathcal{F} = 0.7$. We remark that the bodies penetrate in their reference configuration. In Fig. 5(c) the relative error $\|\boldsymbol{\lambda}_h^k - \boldsymbol{\lambda}_h\|/\|\boldsymbol{\lambda}_h\|$ is shown for the case $m = \infty$. We remark that due to the superlinear convergence of the full Newton approach we need only half of the iteration steps compared with the fixpoint approach.

2.4 Energy Conserving Time Integrators

A central task to model the impact of a rotating disc onto an elastic ground is to use an energy conserving time integrator that does not increase or decrease the energy of the system in the case of zero volume forces, zero surface traction and $\Gamma_D^l = \emptyset$. To discretize the contact in time, we divide the time interval $[0, T]$ into N_T equidistant subintervals $[t_k, t_{k+1}]$, $k = 0, \ldots, N_T - 1$, with $t_k := k\,\Delta t$ and the time stepwidth $\Delta t := T/N_T$. Following [63], we use the midpoint rule based on the relation

$$\frac{\mathbf{u}_{k+1} - \mathbf{u}_k}{\Delta t} = \frac{\mathbf{v}_k + \mathbf{v}_{k+1}}{2} =: \mathbf{v}_{k+1/2} \ ,$$

where we denote by $\mathbf{v}_k$ the velocity at time t_k. Defining the kinetic energy $\mathcal{K}_k$ and the internal stress energy $\mathcal{E}_k^i$ at time t_k by

$$\mathcal{K}_k := \frac{1}{2} \int_\Omega \varrho \|\mathbf{v}_k\|^2 \, \mathrm{d}s \ , \qquad \mathcal{E}_k^i := \frac{1}{2} \int_\Omega \mathbf{P}(\mathbf{u}_k) : \nabla \mathbf{u}_k \, \mathrm{d}s \ ,$$

the energy conservation reads as

$$\mathcal{K}_k + \mathcal{E}_k^i + \mathcal{E}_k^c = \text{const} \ , \qquad \mathcal{E}_k^c := \sum_{j=0}^{k-1} \Delta t \, \mathcal{D}^c_{[t_j, t_{j+1}]} \ ,$$

where the power of the contact forces in the time interval $[t_j, t_{j+1}]$ is given by

$$\mathcal{D}^c_{[t_j, t_{j+1}]} := \int_{\Gamma_c^s} \boldsymbol{\lambda}_{j+1/2} \frac{\Delta[\mathbf{u}_j]}{\Delta t} \, \mathrm{d}s$$

with $\Delta[\mathbf{u}_j] := [\mathbf{u}_{j+1}] - [\mathbf{u}_j]$. In the dynamical case, we have to use in the contact conditions (2) and (3) the velocity instead of the displacement. Furthermore, we have to add a persistency condition $\sigma_n \frac{\mathrm{d}}{\mathrm{d}\,t}([\mathbf{u}]_n - g) = 0$. Due to the persistency condition, we find that

$$(\lambda_{n,p,sq})_{j+1/2} \, \Delta(\hat{u}_{n,p,s})_j = 0 \tag{17}$$

has to be satisfied by our algorithm. Thus the normal components of the displacement do not contribute to $\mathcal{D}^c_{[t_j,t_{j+1}]}$. Moreover, in the case that all nodes are sticky, then there is no energy dissipation due to the friction. Otherwise (3) guarantees that $\mathcal{D}^c_{[t_j,t_{j+1}]} \geq 0$. Following [12, 50], we have to modify our primal-dual active set strategy such that (17) and

$$\begin{aligned} (\boldsymbol{\lambda}_{\tau,p,s})_{j+1/2} &= 0 \quad \text{if} \quad p \in \mathcal{I}_\tau \ , \\ (\boldsymbol{\lambda}_{\tau,p,s})_{j+1/2}\, \Delta(\hat{\mathbf{u}}_{\tau,p,s})_j &\geq 0 \quad \text{if} \quad p \in \mathcal{A}_\tau \end{aligned}$$

is satisfied.

2.4.1 Numerical Example

We consider a rotating two-dimensional disc with a Young modulus $E^s = 2986 \times 10^6$ and a Poisson ratio $\nu^s = 0.38$ impacting a rectangle with a Young modulus $E^m = 210 \times 10^9$ and a Poisson ratio $\nu^m = 0.3$. The mass density on the disc is given by $\varrho^s = 1391$ and on the rectangle we set $\varrho^m = 7830$. The initial conditions for the disc are an initial angular velocity of $\omega = -150$ overlaying an initial translation velocity $\mathbf{v}^s = (50, -50)^\top$. The rectangle is subjected to homogeneous Dirichlet boundary conditions at its bottom. For the friction, we assume Coulomb's law with $\mathcal{F} = 0.2$. We compute the solution for the time interval [0, 0.004] with 200 time steps, i.e., $\Delta t = 2 \times 10^{-5}$. As material law, we assume for both bodies a nonlinear Saint Venant–Kirchhoff law, where we replace in (4) the linearized strain tensor $\boldsymbol{\varepsilon}^l$ by the nonlinear one given by $\mathbf{E}^l := 1/2\big(\nabla \mathbf{u}^l + (\nabla \mathbf{u}^l)^\top + (\nabla \mathbf{u}^l)^\top (\nabla \mathbf{u}^l)\big)$. Figure 6(a) shows the

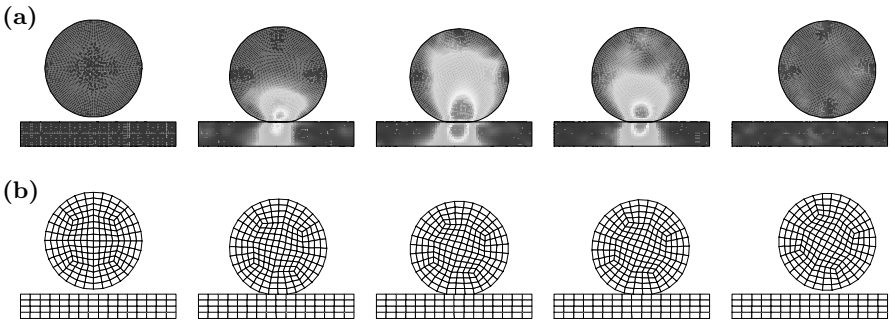


Fig. 6. (a) Distorted bodies with the effective von Mises stress **(b)** mesh on a coarse level for the time steps t_k with $k = 1$, $k = 70$, $k = 100$, $k = 130$ and $k = 200$

distorted bodies with the effective von Mises stress and Fig. 6(b) the coarse grid at the time steps t_k with $k = 1$, $k = 70$, $k = 100$, $k = 130$ and $k = 200$. Figure 7 shows the energy terms over the time interval [0, 0.004]. In Fig. 7(a), the total energy $\mathcal{K} + \mathcal{E}^i$ and its decomposition into the elastic part $\mathcal{E}^i$ and the

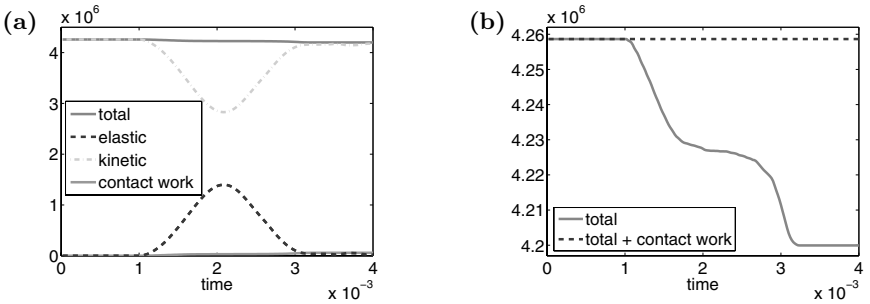


Fig. 7. (a) Total Energy $\mathcal{K} + \mathcal{E}^i$, elastic energy $\mathcal{E}^i$, kinetic energy $\mathcal{K}$ and contact work $\mathcal{E}^c$, **(b)** total energy $\mathcal{K} + \mathcal{E}^i$ with the contact work $\mathcal{E}^c$ for the example in Fig. 6

kinetic part $\mathcal{K}$ is shown. Since slip occurs, the total energy is not constant over the time. The dissipative part reflected in the contact work $\mathcal{E}^c$ is also shown. As one can see in Fig. 7(b), the total energy plus the contact work is constant, i.e., $\mathcal{K} + \mathcal{E}^i + \mathcal{E}^c = \text{const}$. In Fig. 8, we show the behavior of the disc

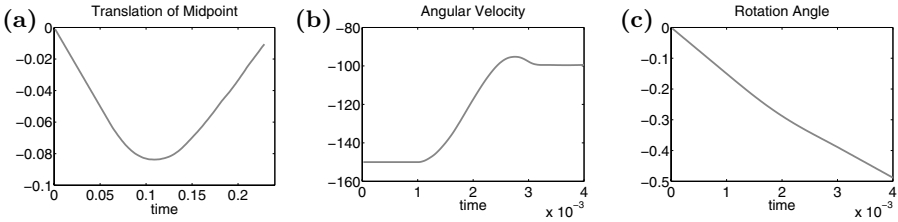


Fig. 8. (a) Translation of the midpoint of the disc, **(b)** angular velocity of the disc, **(c)** rotation angle of the disc

during the impact. In Fig. 8(a), the translation of the midpoint is shown, in Fig. 8(b) the angular velocity and in Fig. 8(c) the rotation angle with respect to the time.

As second example, we consider an inhomogeneous disc with a nearly incompressible layer, see Fig. 9, with the nonlinear Saint Venant–Kirchhoff law. For the inner part of the disc and the rectangle, we use the same material parameters as in the example given before. For the nearly incompressible layer, we use Young modulus $E^s = 2986 \times 10^6$, Poisson ratio $\nu^s = 0.499$ and mass density $\varrho^s = 1391$. For the discretization, we use a Q1P0-element, which is also covered by the general framework of the modified Hu–Washizu formulation given in Sect. 2.2.4. The results are presented in Figs. 9–10.

2.5 A posteriori Error Estimates

In this section, we present an a posteriori error estimator for contact problems in the two-dimensional case. We only give the basic ideas for the simplicial

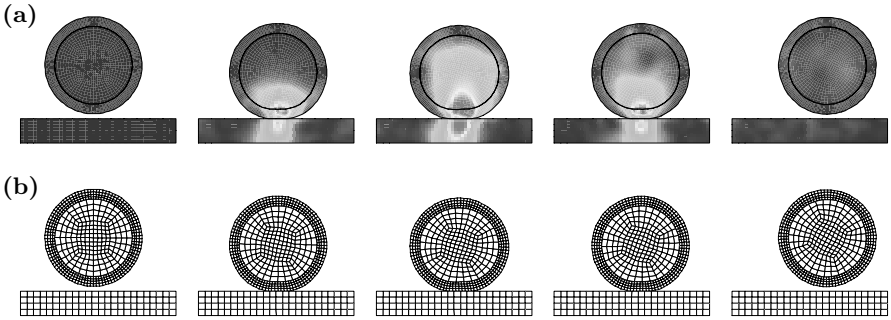


Fig. 9. **(a)** Distorted bodies with the effective von Mises stress **(b)** mesh on a coarse level for the time steps t_k with $k = 1$, $k = 70$, $k = 100$, $k = 130$ and $k = 200$

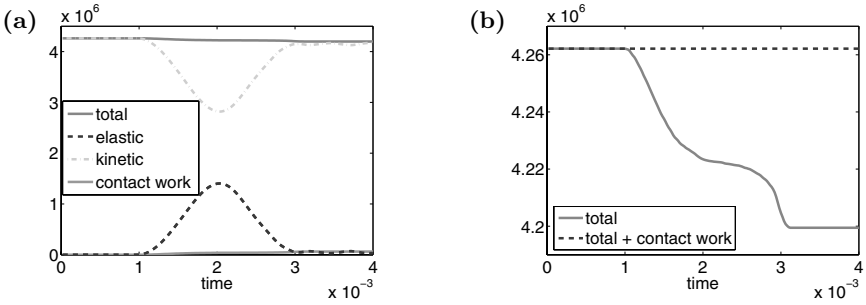


Fig. 10. **(a)** Total Energy $\mathcal{K} + \mathcal{E}^i$, elastic energy $\mathcal{E}^i$, kinetic energy $\mathcal{K}$ and contact work $\mathcal{E}^c$, **(b)** total energy $\mathcal{K} + \mathcal{E}^i$ with the contact work $\mathcal{E}^c$ for the example in Fig. 9

case and refer to [56] for details and the treatment of quadrilaterals in case of the Lamé equation. The error estimator is based on $H(\mathrm{div})$-conforming approximations for the stress using the Arnold–Winther element and on equilibrated fluxes, see [4, 47]. In comparison to other a posteriori error estimators the constant in the upper bound is up to higher order terms equal to one.

2.5.1 Definition of the Error Estimator

First we recall the basic idea of equilibrated fluxes and refer to [1] for details. Using the solution $(\mathbf{u}_h, \boldsymbol{\lambda}_h)$ of (11), we can find for each edge e a function $\mathbf{g}_e \in [P_1(e)]^2$ such that

$$a_T(\mathbf{u}_h, \mathbf{v}) = (\mathbf{f}, \mathbf{v})_T + \sum_{e \subset \partial T} \int_e (\mathbf{n}_e \cdot \mathbf{n}_T)(\mathbf{g}_e \cdot \mathbf{v}) \, \mathrm{d}s \,, \quad \mathbf{v} \in [P_1(T)]^2 \,, \tag{18}$$

$$\mathbf{g}_e|_{\Gamma_C^s} := -\boldsymbol{\lambda}_h|_e \,, \qquad \mathbf{g}_e|_{\Gamma_C^m} := \mathcal{P}_h \boldsymbol{\lambda}_h|_e \,.$$

Here, $\mathbf{n}_e$ denotes a fixed normal vector for each edge e and $a_T(\cdot,\cdot)$ is the local contribution of $a(\cdot,\cdot)$ to the element T and $\mathcal{P}_h$ is a suitable mortar projection.

Using biorthogonal basis functions for $\mathbf{g}_e$ leads to a local system at each node that is in general singular. We have the existence of a solution due to the variational consistent formulation (11), and a unique solution is found by solving a minimization problem, i.e., $\mathbf{g}_e$ is as close as possible to the average of the discrete flux $\{\boldsymbol{\sigma}(\mathbf{u}_h)\cdot\mathbf{n}_e\}$. Now using the Arnold–Winther element, see [4], we define a suitable approximation of the stress. The element is locally defined by the 24-dimensional space

$$S_T := \left\{\boldsymbol{\tau}_h \in [P_3(T)]^{2\times 2},\ (\boldsymbol{\tau}_h)_{12} = (\boldsymbol{\tau}_h)_{21},\ \operatorname{div}\boldsymbol{\tau}_h \in [P_1(T)]^2\right\},$$

and a global $H(\mathrm{div})$-conforming finite element space can be obtained using

- the nodal values at each node p,
- the zero and first order moments of $(\boldsymbol{\tau}_h\cdot\boldsymbol{n}_e)$ on each edge e,
- the mean value on each element T

as degrees of freedom. We define an approximation $\boldsymbol{\tau}_h$ of $\boldsymbol{\sigma}(\mathbf{u})$ by setting

$$\boldsymbol{\tau}_h(p) = \frac{1}{N_T^p}\sum_{T\in\mathcal{T}_p}\boldsymbol{\sigma}(\mathbf{u}_h)|_T(p)\,, \tag{19}$$

$$\int_e(\boldsymbol{\tau}_h\cdot\mathbf{n}_e)\cdot\mathbf{q}\,\mathrm{d}s = \int_e\mathbf{g}_e\cdot\mathbf{q}\,\mathrm{d}s\,,\qquad \mathbf{q}\in[P_1(e)]^2\,, \tag{20}$$

$$\int_T\boldsymbol{\tau}_h:\nabla\mathbf{v}\,\mathrm{d}s = a_T(\mathbf{u}_h,\mathbf{v})\,,\qquad \mathbf{v}\in[P_1(T)]^2\,. \tag{21}$$

Here, $\mathcal{T}_p$ denotes the set of all triangles having p as a node and N_T^p the number of those triangles. We satisfy (19) for all interior nodes and modify (19) according to the boundary conditions at the remaining nodes. The error estimator is then defined by

$$\eta^2 := \sum_T\eta_T^2\,,\qquad \eta_T^2 := \|\mathcal{C}^{-1/2}(\boldsymbol{\tau}_h-\boldsymbol{\sigma}(\mathbf{u}_h))\|_{0,T}^2\,.$$

It is easy to verify that conditions (20) and (21) together with (18) yield $\operatorname{div}\boldsymbol{\tau}_h = -\Pi_1\mathbf{f}$, where Π_1 is the L^2-projection onto piecewise affine functions. To estimate the error in terms of the error estimator, the error of $\boldsymbol{\lambda}_h$ has to be taken into account. Using the results of Lemmas 2, we obtain an upper bound with constant one for the energy norm $\|\cdot\|_a := \sqrt{a(\cdot,\cdot)}$.

Theorem 2. *Under suitable regularity assumptions, the error estimator provides a sharp bound for all $\varepsilon > 0$,*

$$\|\mathbf{u}-\mathbf{u}_h\|_a \le \eta + \mathcal{O}(h^{3/2})\,.$$

Here, the $h^{3/2}$-term comes from the estimation of $b(\mathbf{u}-\mathbf{u_h},\boldsymbol{\lambda}-\boldsymbol{\lambda}_h)$. We note that for the upper bound, the property $\operatorname{div}\boldsymbol{\tau}_h = -\Pi_1\mathbf{f}$ is crucial. Condition (19) does not play a role in the upper bound, but is important to establish a lower bound.

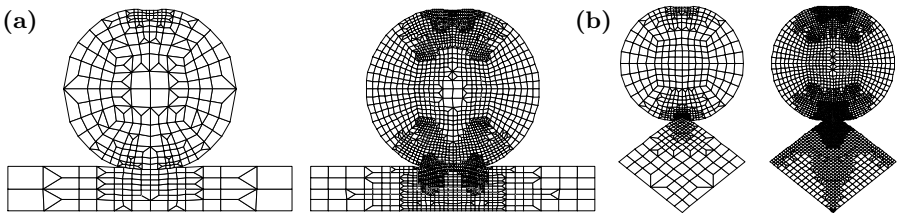


Fig. 11. Deformed adaptive meshes after 3 and 6 refinement steps, **(a)** first example, **(b)** second example

Theorem 3. *Under suitable regularity assumptions, there exists a constant c independent of h such that for all T*

$$c\eta_T^2 \leq \|\mathbf{u} - \mathbf{u}_h\|_{a,\omega_T}^2 + \sum_{e \subset \Gamma_C^s, e \in \omega_e} h_e \inf_{\mu_h \in M_h} \|\boldsymbol{\lambda} - \boldsymbol{\mu}_h\|_{0,e}^2 ,$$

where ω_T is the union of all elements $\tilde{T}$ such that $\partial T \cap \partial \tilde{T} \neq \emptyset$ and ω_e is the union of all edges e such that $\bar{T} \cap \bar{e} \neq \emptyset$.

2.5.2 Numerical Example

We apply the error estimator to Hertz type contact problems, see Sect. 2.2.4, and use it to obtain adaptive mesh refinements. Here, we use quadrilateral meshes on a circle with radius 0.48 that is pushed onto a plane with height 0.25 and width 1.6 in the first example and onto a diamond with a diagonal length of 1.0 in horizontal and 0.8 in vertical direction, see Fig. 11. For the circle, we use, Young modulus $E^s = 7000$ and Poisson ratio $\nu^s = 0.3$, for the plane and the diamond, we use, $E^m = 6000$ and $\nu^m = 0.3$. The boundary forces on the top of the circle are given by $\mathbf{p}^s = (0, p_y)^\top$, $p_y(x) = c\big(|x|^2 - 0.125^2\big)^2$, where c is chosen such that $\int p_y(x)\,\mathrm{d}x = -140$, the plane is fixed on the lower side and the diamond is fixed on the two lower sides. We note that in the second problem setting, the corner in the contact zone results in a lower regularity of the solution than in the first setting. We use a mean-value strategy with factor 0.9 for the mesh-refinement. In Fig. 11 the adaptively refined meshes for the two examples are depicted. The estimated errors using adaptive and uniform mesh refinement are compared in Fig. 12. Here the increase of the efficiency using adaptive meshes for low-regularity problems can clearly be seen.

In the third, example we use our error estimator to a dynamic frictional contact problem to get adaptive refined meshes. The meshes at different time steps are depicted in Fig. 13.

3 Experiments in Contact Dynamics

In this section we present the results of the engineering part of our joint project. An analysis of impact models is first performed in Sect. 3.1, where

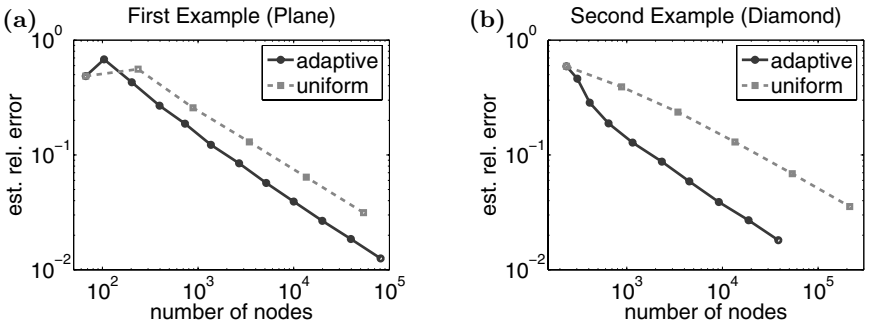


Fig. 12. Error reduction **(a)**for the first example and **(b)** for the second example using adaptive and uniform refinement

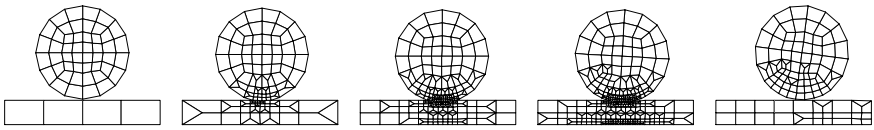


Fig. 13. Adaptive grid at time step t_k, $k = 0$, $k = 20$, $k = 40$, $k = 80$, $k = 120$

a hybrid impact model is developed for interpreting the experimental results. We then describe the experimental setup and methods in Sect. 3.2. In Sect. 3.3 we analyze the experimental results and discuss effects of the strip flexibility and the disc initial velocity on the rebound behavior of the disc.

3.1 Impact Models

In order to describe the impact of elastic compact-slender bodies, impact models are needed and corresponding properties are to be measured. In this section we first present briefly an instantaneous impact model by Walton [69] with three free impact parameters to be chosen. Then a finite element model is established for getting more insight into the impact process. We finally extend the oblique impact model by Maw et al. [53] and Johnson [42] to a hybrid impact model, which takes into account structural vibrations of the slender body.

3.1.1 Instantaneous Impact Model

Walton's model is one of representative instantaneous impact models often used in impact dynamics. In Walton's model [69], impact is regarded as an instantaneous event and three phenomenological impact parameters are defined for connecting the pre-impact and post-impact kinematics. These three parameters can accommodate the compliant and frictional nature of the contact

interactions and the coupling between translational and rotational relative motion of the impacting bodies.

Let us suppose that a sphere with mass m, radius r and moment of inertia I approaches a target wall with a pre-impact translational velocity $\mathbf{v}_i$ at its center of mass and rotational velocity ω_i at an incidence angle θ_i. The velocity of the sphere at the contact point before impact is then $\mathbf{v}_{ci}$. The new values of velocities and the angle after the interaction with the wall are denoted by $\mathbf{v}_r$, ω_r, θ_r and $\mathbf{v}_{cr}$, respectively. The impact kinematics is illustrated in Fig. 14.

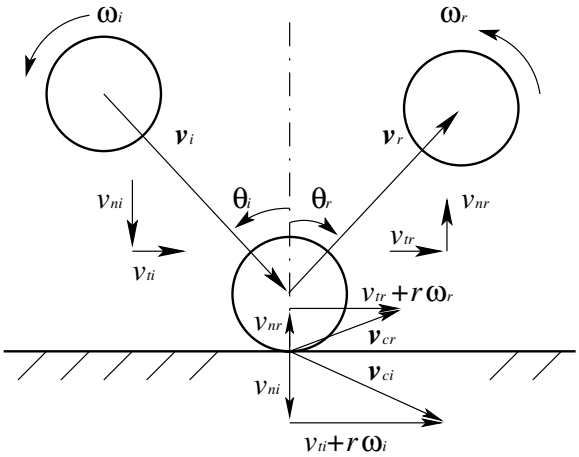


Fig. 14. Kinematics of a sphere striking a plane surface

The coefficients of normal and tangential restitution are defined respectively as

$$e_n = \frac{v_{nr}}{v_{ni}} \,, \tag{22}$$

$$e_t = -\frac{v_{tr} + r\omega_r}{v_{ti} + r\omega_i} \,. \tag{23}$$

Walton also assumes that a given impact is either *slipping* or *sticking* throughout the contact duration. In the *slipping* mode, the tangential impulse P_t is related to the normal impulse P_n by

$$P_t = \mu P_n \tag{24}$$

where μ is the Coulomb coefficient of friction, usually referred to as the impulse ratio. In the *sticking* mode, the coefficient of tangential restitution is assumed to be a constant $e_t = e_{t0}$. By performing some simple mathematical manipulation, the equations linking the pre-impact and post-impact kinematics read as [69, 46]

$$e_t = \begin{cases} -1 - \mu(1 + 1/K^2)(1 + e_n)\cot\gamma_i & \text{for } \gamma_i \geq \gamma_0 \,, \\ e_{t0} & \text{for } \gamma_i < \gamma_0 \end{cases} \tag{25}$$

where $K = \sqrt{I/(mr^2)}$, $\cot\gamma_i = v_{ni}/(v_{ti} + r\omega_i)$ and γ_0 is a critical angle of incidence below which (24) does no longer hold. Therefore, three parameters e_n, μ and e_{t0} are the impact properties to be determined by experiments.

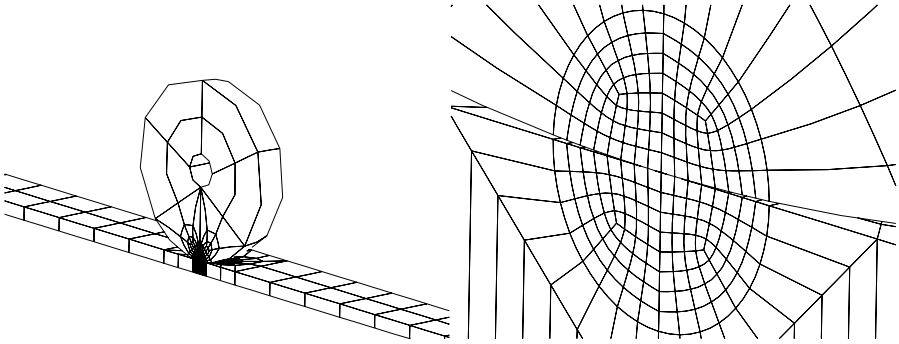
Above all, Walton's model is shown to be able to describe both energy loss in normal direction and compliance in tangential direction for the impact process. Hence, it is reasonable to study it in an experimental investigation of the impact of a rotating disc striking an elastic strip.

3.1.2 Nonlinear Finite Element Model

For getting further information upon the impact process, a finite element (FE) model with continuum elements is established. The nonlinearity here arises from the need of fulfilling the contact boundary of the non-penetration condition in normal direction and the Coulomb friction condition in tangential direction [76]. The key issues to be carefully considered are the modeling of local contact deformations and that of the resulting global wave motion in the strip. For capturing accurate information on the local deformation, two points should be noticed: (i) the refined mesh size for the contact region; (ii) the parameters of the contact algorithms, that is, the Lagrange method or the penalty method. Their appropriate selection can usually be done by testing the convergence of the results of forces or stresses in the contact region. Due to the higher accuracy in describing deformation, continuum elements of second order are used in the contact region, where contact deformation is predominant. In order to include the sufficient significant modes for representing wave motions of the slender body, besides suitable spatial discretization, the time step size must also be chosen correctly [61]. The impact problems investigated in this study show that the contact duration is down to 0.2 ms according to measurements. Hence a time step size of 10^{-6}s and an overall small element size of 16 mm for the strip of 600–1000 mm length are required. Figure 3.1.2 shows the mesh of the three-dimensional FE model in ANSYS. Because of geometrical and loading symmetries, only half of the model is considered and discretized. A graded meshing scheme is adopted for gaining as high computational efficiency as possible without a significant loss of accuracy. The major part of the strip is modeled by linear elements for their better behavior in dynamic analyses. The disc and strip are connected by surface-to-surface contact elements using the augmented Lagrange formulation. The choice of the tolerance is determined through some experimentation by checking the independency of the corresponding results.

3.1.3 Hybrid Impact Model

Instantaneous impact models describe the oblique impact by phenomenological parameters that are not fundamental material properties. Finite element



(a) Global FE mesh **(b)** Local FE mesh

Fig. 15. Finite element model of the disc to strip impact

modeling, though based on continuum theories, cannot demonstrate the insight into the impact process without a high computational effort. Following the idea of combining the local contact model with the global vibration model, proposed by Sears [60] and extended by Hu and Eberhard [17, 32], we develop a hybrid impact model by extending the impact theory by Maw et al. [53] and Johnson [42].

Their oblique impact theory is restricted to compact-body impact, implying that the coefficient of normal restitution is one [54]. For impact between compact and slender bodies, this is no longer valid due to the fact that a certain part of compact body's kinetic energy is to be transferred to the slender elastic body in the form of structural vibrations. However, this theory allows the quasi-static loading based on Hertz's contact theory and the fact that the normal force is not affected by the tangential force. This ensures a possible extension of this theory to the case of compact-slender-body impact via replacing the normal load history derived from compact-body impact by a new normal load history derived from compact-slender-body impact.

Therefore, our extension consists of two steps: (i) pre-calculatation of the load history using a combination of Hertz's theory and wave theory and (ii) solution of the local oblique impact process. In the following, we present this approach by an example of sphere-beam impact.

At first, the load history is to be resolved without considering frictional effects. The motion of the sphere with radius r can be described by

$$m\ddot{w}_1 = -F_n(t) \tag{26}$$

where m and w_1 are the mass and displacement of the mass, respectively, and F_n is the impact force given as

$$F_n(t) = \begin{cases} F_c & \text{if } F_c \geq 0 \,, \\ 0 & \text{if } F_c < 0 \end{cases} \tag{27}$$

where F_c is obtained from the Hertz's contact law. The governing equation for an isotropic beam subject to lateral loading is given as

$$E_2 I_2 \frac{\partial^4 w_2}{\partial x^4} + \rho_2 A \frac{\partial^2 w_2}{\partial x^2} = p(x,t) \tag{28}$$

where $E_2 I_2$ is the bending stiffness defined as usual, ρ is the mass density, A is the area of cross-section, w_2 is the transverse deflection and p is the lateral load per unit length. The Hertz contact law is written as

$$F_c(\alpha) = k_H \alpha^{3/2} \tag{29}$$

where in the case of a sphere to be colliding with a flat surface the contact stiffness k_H is given by [23]

$$k_H = \frac{4}{3}\sqrt{r}E^* \tag{30}$$

with

$$\frac{1}{E^*} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \tag{31}$$

where ν_1, E_1 and ν_2, E_2 are the Poisson ratios and the Young's moduli of the sphere and the beam, respectively and α is the contact indentation defined as

$$\alpha(t) = w_1(t) - w_2(t)\ . \tag{32}$$

The above equations of motion can be numerically solved by a standard finite element procedure including a nonlinear spring to obtain the discretized load history $F_{ni}(t_i)$, $i = 1, 2, \ldots, N$, where N is the number of time increments, and the contact duration is T_c.

Secondly, the force history in the oblique impact theory by Maw et al. [53] and Johnson [42] is to be replaced by the pre-calculated force history $F_{ni}(t_i)$. Thus the instantaneous radius of contact area becomes

$$b = \Big(\frac{3rF_{ni}(t_i)}{4E^*}\Big)^{1/3}\ . \tag{33}$$

The non-dimensional form of tangential velocity v_{tc} at a contact point during the impact remains the same, $\psi = \kappa v_{tc}/\mu v_{ni}$, where

$$\kappa = \frac{\frac{1-\nu_1}{G_1} + \frac{1-\nu_2}{G_2}}{\frac{1-\nu_1/2}{G_1} + \frac{1-\nu_2/2}{G_2}} \tag{34}$$

where $G = E/(2(1+\nu))$. Nonetheless, its behavior during the impact depends on a second parameter χ which is modified from $\chi = \kappa(1 + 1/K^2)/2$ to

$$\chi = \frac{F_m T_c}{3.6970 m v_{ni}} \frac{\kappa}{2}\Big(1 + \frac{1}{K^2}\Big) \tag{35}$$

where F_m is the maximum normal force. Hence it reflects the influences of structural flexibility and initial conditions, which will be discussed later in detail with measurements.

3.2 Experimental Setup and Methods

For studying the impact properties of a rotating disc striking an elastic strip, experimental measurements of the kinematics of the disc before and after impact have to be performed. Hence, an experimental setup was constructed, as shown in Fig. 16. In principle it consists of two main parts: (i) the impact unit, which includes a disc and a throwing machine with its control system, a strip properly fixed; (ii) the measurement unit, composed of a high-speed digital camera system and an image processing unit. Before impact, the disc is held by the throwing machine[3], accelerated to have the desired motion. After being released the disc follows a ballistic trajectory before and after impact. Meanwhile the camera system is triggered and a sequence of images is taken, transferred to a PC and saved for measurements using an image processing package. Every aspect of the setup will be described in more detail in the following.

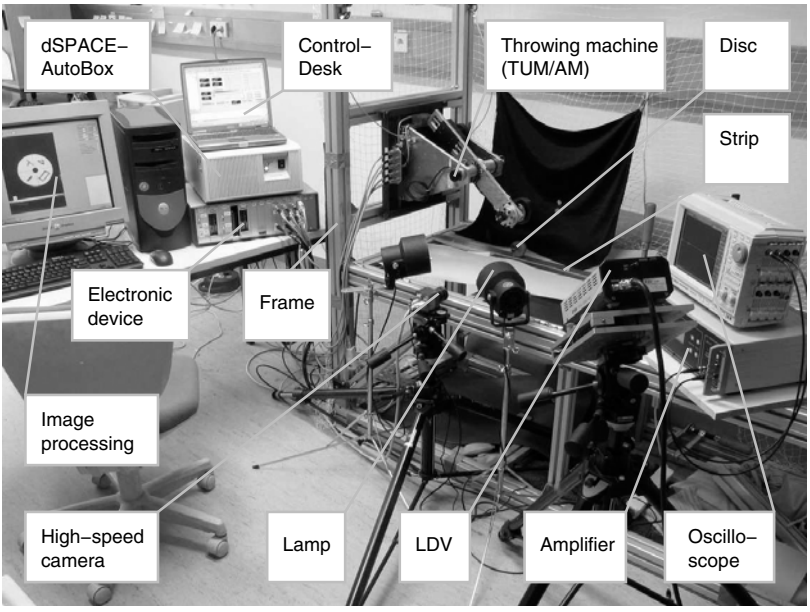


Fig. 16. Experimental setup

[3] The throwing machine is borrowed from the Institute for Applied Mechanics of TU München (Prof. Ulbrich, Prof. Pfeiffer). It was originally produced by Dr. Beitelschmidt/TUM for his dissertation [6] and is kindly provided tu us during the project time. The power electronics, the control logic and the data processing part for the throwing machine were, however, developed on our own.

3.2.1 Electronic Control

In the impact tests reported in literature only very few of them were concerned with a striker with initial rotation or spin, which however occurs very widely in reality. In the current study a throwing machine is employed for accelerating the disc to fulfill the desired pre-impact conditions including initial rotation. In order to use this machine, its driving electronics must have been designed and realized.

The throwing machine is composed of (i) an arm driven by a motor for generating translational motions for the disc, (ii) a holding mechanism with an electromagnet and (iii) a second motor for generating the rotational motions for the disc. At the shafts of the two motors two encoders are mounted which measure the angles by which the shafts rotate. In order to supply the power needed by the two motors and the electromagnet and to transfer the measured and controlling signals between the throwing machine and a control device, the dSPACE-AutoBox, an electronic device was designed and realized, see Fig. 17.

In order to let the disc strike the strip with desired velocities, a control strategy was employed, which includes three steps: (i) trajectory planning, (ii) drive control and (iii) process control. Based on the kinematics in Fig. 18, the pre-impact quantities can be derived through simple calculations as below

$$v_{ti} = R\dot{\phi}_1 \cos\phi_1 \ , \tag{36}$$

$$v_{ni} = -\sqrt{(R\dot{\phi}_1 \sin\phi_1)^2 + 2g(h - r + R\cos\phi_1)} \ , \tag{37}$$

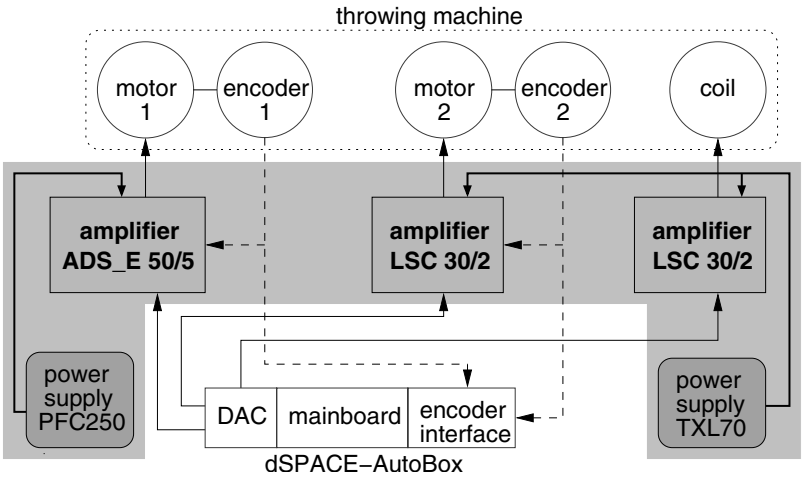
$$\omega_i = \dot{\phi}_1 + \dot{\phi}_2 \ , \tag{38}$$

$$x_i = R\dot{\phi}_1 \cos\phi_1 t_f + R\sin\phi_1 \ . \tag{39}$$

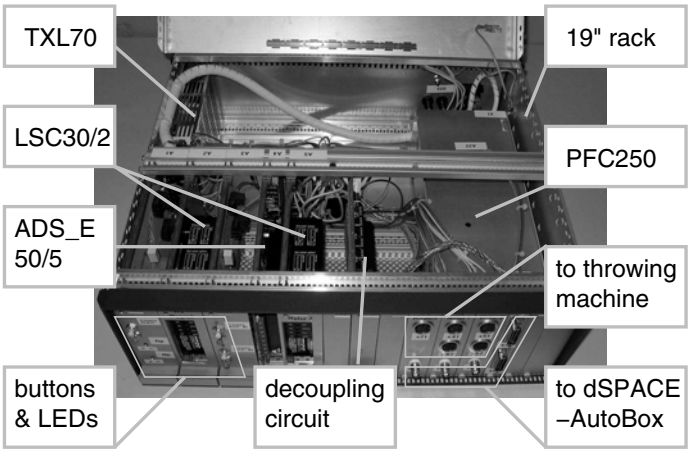
Before the release, the disc has to be accelerated to have the desired rotational and translational velocities. This was realized based on the dSPACE-AutoBox, a real-time computer instrumented by a PC. A PID controller, as shown in Fig. 19, was used for achieving appropriate accuracy and response speed for each motor. A criterion was then added to the control system, which ensures that the electromagnet is switched off when the difference between ϕ_1 and its reference value is less than a small prescribed tolerance. Thus the disc can be released at the prescribed angle ϕ_1. The MATLAB-SIMULINK package was chosen to conduct these control tasks.

Having the steps (i) and (ii) realized, the release process is then organized by a process control with the procedure

a. Initialization, in which computation of the trajectory planning is performed;
b. Calibration, for the two encoders measuring ϕ_1 and ϕ_2;
c. Acceleration, in which two motors are speeded up to the desired values $\dot{\phi}_1$ and $\dot{\phi}_2$;
d. Release, being executed when the criterion on the release angle is satisfied;



(a) Design



(b) Realization

Fig. 17. Design and realization of the electronic controller device and power electronics

e. Finalization, to set each instrument to its initial state.

This step is carried out on the host PC for the dSPACE-AutoBox based on the instrumentation software ControlDesk.

3.2.2 Specimens

The aim of this experimental study requires such impact conditions that the oblique impact event must be clearly captured with apparent structural vibrations being excited. According to Christoforou and Yigit [14], the impact

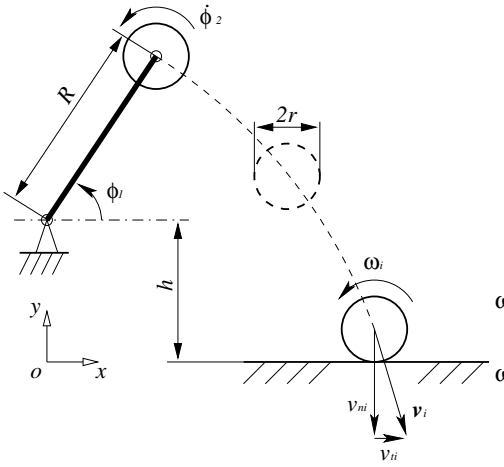


Fig. 18. Trajectory planning

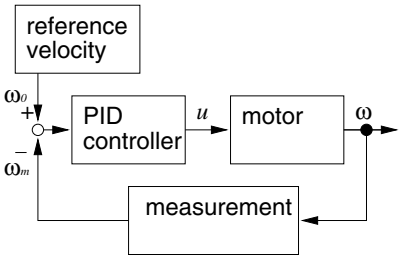


Fig. 19. Drive control

conditions in which local contact deformation dominates is selected, since in the situation where global vibrations dominate, the contact duration is considerably longer than the selected case, the configuration can no longer be assumed to remain unchanged and the camera system does not suffice consequently; and in the case where global and local effects are coupled, multiple impact and small rebound height are likely to happen in very short time, so that capturing a clear impact event would possibly be beyond the capability of the camera system.

Thereafter, the specimens should possess the properties: (i) the disc is to be made of a material of low mass density but relatively high stiffness; (ii) the strip is to have the material of relatively high mass density and moderate thickness. We choose PVC material for the disc and steel for the strip. The disc was produced in such a way that it is a symmetrical slice cut from a sphere. And the two key dimensions of the strip are its width and height. Their geometrical and material properties are listed in Table 4.

Table 4. Geometrical and material properties of specimens

Specimen	Material	Geometry [mm]	Young's modulus [GPa]	Poisson ratio	Density [kg/m^3]
Disc					
	PVC	ϕ54×20	2.982	0.38	1391
Strip					
	Steel	60×8	210	0.3	7800
	Steel	40×4	210	0.3	7800

3.2.3 Fixture of the Strip

The strip struck by the disc is clamped at the two ends by clamps and hard blocks on a very rigid foundation shown in Fig. 20. From wave theory, see e.g. [25], the impact between the disc and the strip will be influenced by the boundary only if during contact the waves have sufficient time to propagate through the strip from the impact position to the boundary and return to the contact region. According to Sondergaard et al. [65], the clamp boundary is to influence the impact response when the impact happens on a position within its critical distance. To avoid the boundary effects, the fixture of the strip should ensure that the impact happens out of the critical distance of the clamp boundary.

Flexural waves are excited due to the impact and travel through the strip, which can be described by the Bernoulli-Euler theory of beams. According to this theory the wave velocity is then

$$c = \sqrt{\sqrt{\frac{E_2 I_2}{\rho_2 A}}\omega} \tag{40}$$

where A is the cross-section area of the strip and ω is the wave frequency. Moreover, the contact time T_c can be reasonably estimated using Hertz's contact theory by

$$T_c = 4.53\Big(\frac{m}{E^*\pi\sqrt{Rv_{ni}}}\Big)^{2/5} \tag{41}$$

where v_0 is the initial normal velocity of the disc. And, finally, we assume that the vibrational mode with the period of $2T_c$ is predominantly excited, implying that $\omega = \pi/T_c$. Then, the critical distance is

$$d_c = \frac{1}{2}cT_c = \frac{1}{2}\sqrt{4.53\pi\sqrt{\frac{E_2 I_2}{\rho_2 A}}\left(\frac{m}{E^*\pi\sqrt{Rv_{ni}}}\right)^{2/5}}\,. \tag{42}$$

The critical distances for the PVC disc striking different strips are listed in Table 5. This estimation is consistent with the predictions from finite element calculations, in which a cantilever beam is struck at different positions by a PVC disc, see Fig. 21.

Table 5. Critical distances for the PVC disc striking steel strips

v_{ni} [m/s]	0.1	1	2	3
Thick strip[mm]	65.7	52.2	48.7	46.8
Thin strip[mm]	46.5	36.9	34.5	33.1

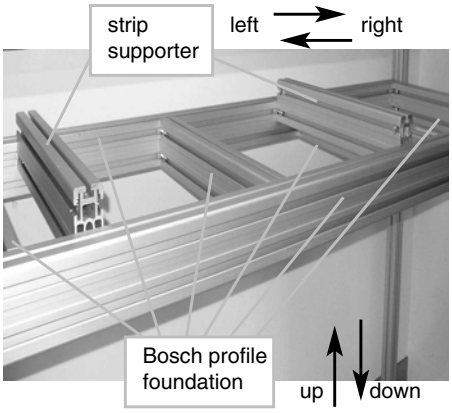


Fig. 20. Support for the strip

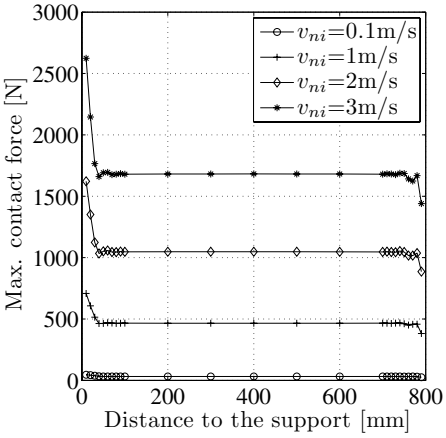


Fig. 21. Simulations on critical distances

3.2.4 Housing of the Impact Event

A frame was also built up using structural elements of Bosch profiles, as shown in Fig. 22. On one hand, this frame serves as a rigid platform for supporting the throwing machine and the foundation of the strip at their working positions. Therefore, it becomes the reference for placing other instruments and performing measurements. Many additional structural elements for reinforcement and connections are also used as auxiliaries. On the other hand, with the help of a safeguard net, the frame can also protect surroundings from accidents caused by unexpected instrumentations and operations. Likewise, the impact event can be less affected by the environment. In addition, two doors are also mounted for an easy and safe operation of the experiment.

3.2.5 High-Speed Digital Camera System and Measurement Devices

The measurement unit is comprised of the image recording part and the image processing part. The image recording part is a high-speed digital camera system, including

- a high-speed digital CMOS camera (Photonfocus MV-D1024-160), which can record at the rate of 150 frames per second with full resolution 1024×1024; based on the *Region Of Interest* technique, it can reach a much higher frame rate by reducing the resolution, e.g., 7930 frames per second with the resolution 128×128;
- a frame grabber card (microEnable III), which can transfer images from the camera to a PC in real time;

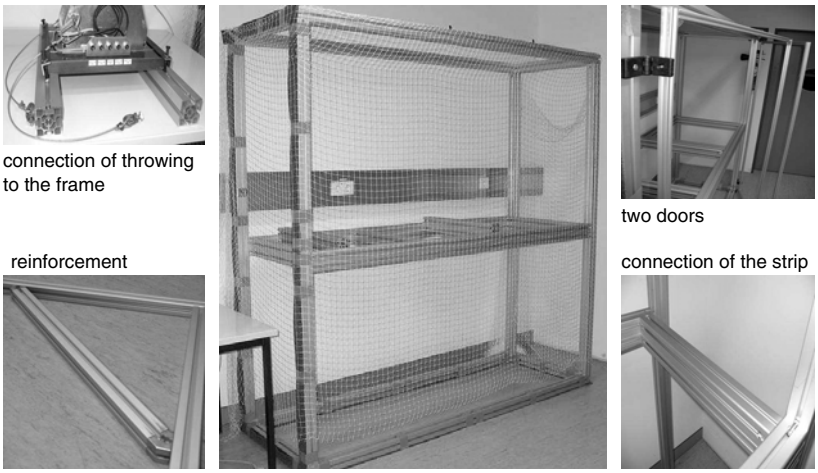
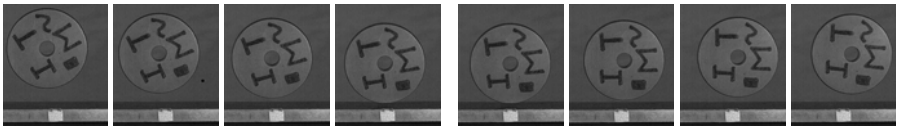


Fig. 22. Frame with safeguard net

- a host PC (P4, CPU 2.8 GHz and RAM 1 GByte), on which working mode selecting, parameter setting and sequence saving are performed based on the control software for the camera and the frame grabber;
- a triggering device made on our own, which can trigger the camera by the release signal from the throwing machine;
- two 1000 W halogen tungsten lamps (KOBOLD-LICHT), which are positioned symmetrically on both sides of the camera.

Compared with the traditional photography with film, this system has the advantages of checking the sequence of images immediately after the acquisition, no cost for consumables and extra work for image digitalization or transferring. The main drawback is the lower resolution, which, however, in the current experiments is acceptable and can effectively be compensated by the image processing in the next stage.

In operation, the camera is first positioned properly such that its focal plane is parallel to that in which the disc moves and its horizontal and vertical axes are parallel to those of the reference frame, respectively. The focal length and aperture size are correctly adjusted. The key parameters, such as exposure time, frame rate, region of interest and image buffer size, are carefully set. When the lamps are switched on, the camera acquisition is triggered by the release signal from the throwing machine. The sequence of images is continuously put into the host PC's buffer of memory till the prescribed number is reached. Then the sequence will be assessed by replaying. Once it is accepted it is saved to the hard disc of the PC. Then, an amount of consecutive images are selected for measurements. The number is usually of the order of 8 to 12 images with approximately as many images before as after impact. Figure 23 shows such a sequence of images.



(a) Four images before impact **(b)** Four images after impact

Fig. 23. Recorded sequence of images

The image recording part is then followed by the image processing part. The measurement of the disc's positions is carried out using an image processing package *Impact-Tracking*, developed on our own based on the basic image processing libraries from Stemmer Imaging GmbH. To determine the translational position of the disc's mass center and its rotational position pattern recognition techniques [67] are adopted, by which sub-pixel resolution of positions is obtained. Once the positions are determined, we perform a second degree polynomial fit of their respective trajectories, before and after impact independently. These fits are then used to find the precise 'impact time', namely, the time at which the two trajectories intersect. The velocities of the disc before and after the impact are then computed by differentiating the fits at the 'impact time'.

In addition a Laser-Doppler-Vibrometer and strain gauges are also mounted and used for validating the impact models on the microscopic and fast-time level.

3.3 Analysis of Results

In order to study the impact properties in cases of a disc elastically striking a strip, we perform a series of oblique impact experiments with specimens at low velocities. The impact parameters and other relative variables are extracted and the effects of strip flexibility and disc initial motions are discussed.

3.3.1 Effects of Strip Flexibility

For investigating the effects of strip flexibility, two steel strips with different dimensions are selected to be obliquely struck by a PVC disc. The disc is released to follow the approximately same trajectory with, however, varied initial rotational velocities. The normal velocity v_{ni} is 1.17 m/s.

Figure 24 shows the coefficient of normal restitution e_n versus the incidence angle $180 \arctan(v_{tci}/v_{nci})/\pi$. In this plot we observe that e_n in general behaves as being more or less a constant with values of 0.099 ± 0.011 and 0.644 ± 0.035 for the impact of the thin and the thick strip, respectively. In the case of thin-strip impact, a smaller scattering is obtained, due to much smaller rebound normal velocities of the disc v_{nr}. Furthermore, it is also observed that the thin-strip impact possesses a much lower value e_n in contrast to the thick-strip impact. This is consistent with the predictions of the hybrid

impact model, where tangential forces do not influence the normal forces and those from Christoforou and Yigit [14] that the impact of a beam with lower mass per unit length $\rho_2 A$ causes higher kinetic energy loss.

Figure 25 shows the coefficient of tangential restitution e_t versus the incidence angle. It is seen that for either the thin or the thick strip, e_t has both positive and negative values. The positive values distribute within a range of small incidence angles, which implies that *sticking* is dominating and tangential compliance plays a key role in this region, and negative tangential rebound velocities are therefore obtained. With the increase of incidence angle, e_t becomes negative and *slipping* comes to dominate. We also notice that e_t for the thin-strip impact qualitatively reaches zero at about 15 ° while it is about 30 ° for the thick-strip impact, which indicates that the thick-strip impact possesses a bigger range of tangential compliance. This is due to the fact that the lower flexibility of the thick strip leads to a larger local normal contact deformation, which correspondingly allows a larger tangential contact deformation and hence a bigger range of tangential compliance.

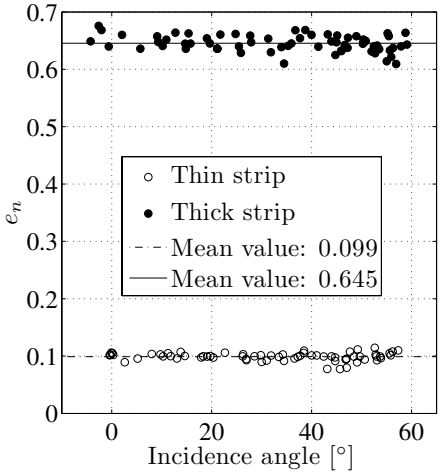


Fig. 24. e_n for varied flexibilities

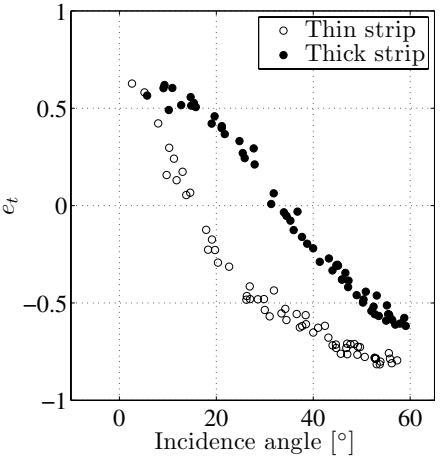


Fig. 25. e_t for varied flexibilities

From Fig. 26 we can observe the variation of the impulse ratio μ with the incidence angle. Similar as for e_t, μ also has two distinct regions. In the range of small incidence angles for *sticking* μ rises up when the incidence angle increases, whilst μ remains to be more or less a constant in the range for *slipping*. Likewise, the flexibility again influences the transition incidence angle from *sticking* to *slipping* due to the same reason. It can also be seen that μ is more scattered in the range for *slipping*, which may be caused by the lack of repeatability in friction conditions on the contact surface.

Figure 27 presents the measurements of the change of rotational velocity versus incidence angle. Two regions for *sticking* and *slipping* are also observed. The change of rotational velocity becomes larger when the incidence angle increases in the *sticking* region, and remains to be more or less a constant in the *slipping* region. The change of rotational velocity in the thick-strip impact is much larger when it is in the region of *slipping* because of the larger normal contact deformation and the higher value of impulse ratio.

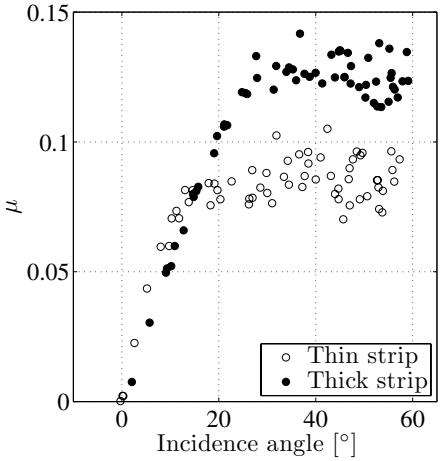


Fig. 26. μ for varied flexibilities

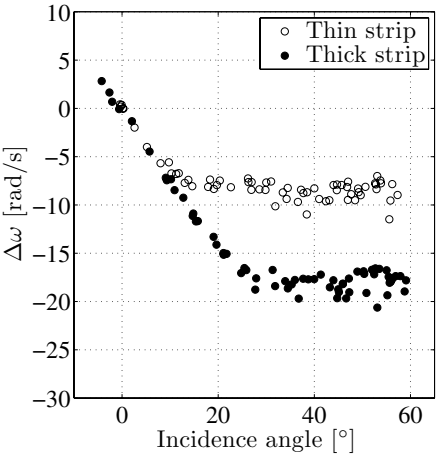


Fig. 27. $\Delta\omega$ for varied flexibilities

According to the instantaneous impact model, parameters μ and e_{t0} are identified to be $\mu = 0.073$ and $e_{t0} = 0.60$ for the thin-strip impact and $\mu = 0.113$ and $e_{t0} = 0.56$ for the thick-strip impact, as illustrated in Fig. 28. In this figure we can observe qualitatively that for small values of $-(1 + e_n)v_{ni}/v_{tci}$, e_t increases until it reaches a maximum positive value e_{t0}. We also notice that the measured data are in generally good agreement with the fitted curves. This can be interpreted by the fact that in the impact model e_n is included in the tangential description. Thus, effects of flexibility are taken into account by e_n. Consequently, the instantaneous impact model can provide an adequate description, but only for the impact dynamics of the disc.

Measurement data are also compared with predictions by the hybrid impact model. Calculations for this model only use the fundamental material properties of the disc and the strip, as well as constant friction coefficients. The results are shown in Fig. 29. In this plot the non-dimensional tangential rebound velocity ψ_2 versus the non-dimensional tangential incidence velocity ψ_1 is demonstrated. The curve for an impact between the disc and a half-space, according to Maw et al. [53] and Johnson [42] is also presented for reference. It can be seen from all of the three curves that a *sticking and micro-slip* region exists when $\psi_2 \leq 0$, in which partial sticking and slipping occur in the impact

process; a *gross slipping* region is apparently featured by $\psi_2 > 0$, in which only slipping occurs. The computed results are consistent with the measurements. Hence the hybrid model explained the fact that increasing flexibility reduces the size of the *sticking and micro-slip* region.

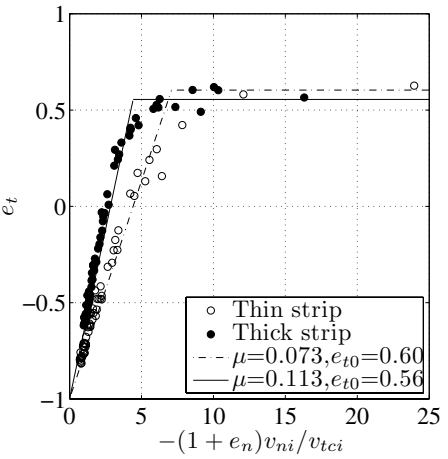


Fig. 28. e_t versus $-(1+e_n)\ v_{ni}/v_{tci}$

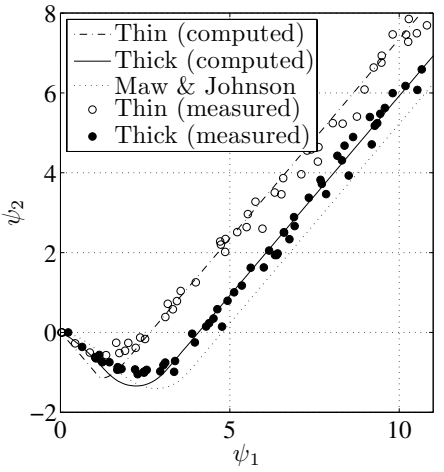


Fig. 29. Variation of ψ_2 with ψ_1

3.3.2 Effects of Disc Initial Normal Velocity

Measurements are also conducted for the thin strip using two different values of the normal incidence velocity of the disc v_{ni}, 1.17 m/s and 2.03 m/s. The disc is released with varied initial rotational velocities for obtaining oblique impact behaviors.

Figure 30 presents the coefficient of normal restitution e_n versus the incidence angle. In this plot we observe that the higher-velocity impact exhibits a slightly lower e_n against the lower-velocity impact. This is consistent with the predictions of (41), where due to the fifth power, the dependence on v_{ni} is not very strong for elastic impact problems.

Figure 31 shows the coefficient of tangential restitution e_t versus the incidence angle. It is seen that e_t for the lower-velocity impact qualitatively reaches zero at an incidence angle about 15 ° while it is about 20 ° for the higher-velocity impact, which indicates that the higher-velocity impact possesses a larger range of tangential compliance. This is due to the fact that the higher velocity of the disc leads to a larger local normal contact deformation, which correspondingly allows a larger tangential contact deformation and hence a bigger range of tangential compliance. Because of the aforementioned reason, this effect of change in v_{ni} is weak.

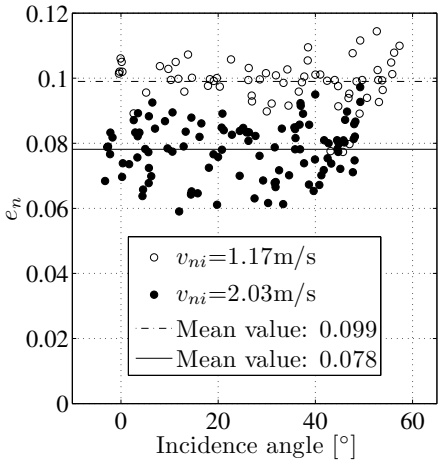


Fig. 30. e_n for varied v_{ni}

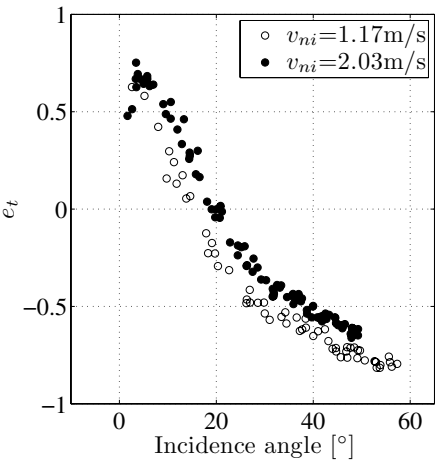


Fig. 31. e_t for varied v_{ni}

In Fig. 32 which demonstrates the variation of the impulse ratio μ with the incidence angle, we can also observe the slightly bigger transition incidence angle about 20 ° from *sticking* to *slipping* for the higher-velocity impact. In addition, it is noticed that the impulse ratio μ for the higher-velocity impact exhibits relatively higher values, which may be attributed to the coarsened surface conditions of the disc after a number of experiments.

Figure 33 shows the measurements of change of rotational velocity versus incidence angle. Two regions for *sticking* and *slipping* are again observed. But the change of rotational velocity in the higher-velocity impact is considerably larger when it is in the region of *slipping*. By referring to Fig. 31 and 32 it can be explained by both the larger normal contact force and the larger impulse ratio resulted from the higher-velocity impact.

For the instantaneous impact model, the parameters μ and e_{t0} are identified to be $\mu = 0.103$, $e_{t0} = 0.63$ for the higher-velocity impact and $\mu = 0.073$, $e_{t0} = 0.60$ for the lower-velocity impact, as shown in Fig. 34. In the plot we can see that a generally good agreement between the measurements and fitted curves is obtained for either the lower-velocity impact or the higher-velocity impact. And the effects of the disc initial normal velocity can be taken into account because of the same reason about e_n. Therefore, the instantaneous impact model can also provide adequate predictions for the disc in this case.

Calculations based on the hybrid model are also performed. The results are compared with measurements as shown in Fig. 35. In this plot we observe the measurements of higher-velocity and lower-velocity in general overlap with each other and the calculations for these impacts are coincident. Comparisons between measurements and calculations show that the measurements are in a good agreement with the calculations in the *gross slipping* region; in the *sticking and micro-slip* region, however, measured values of ψ_2 are lower than

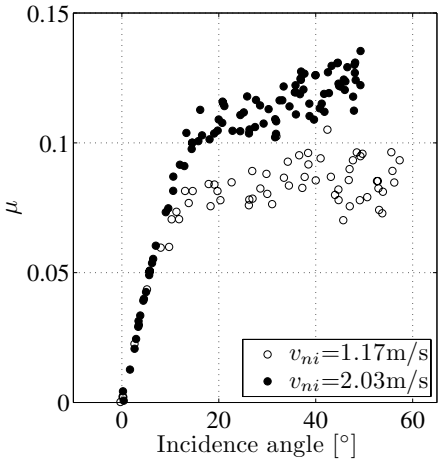


Fig. 32. μ for varied v_{ni}

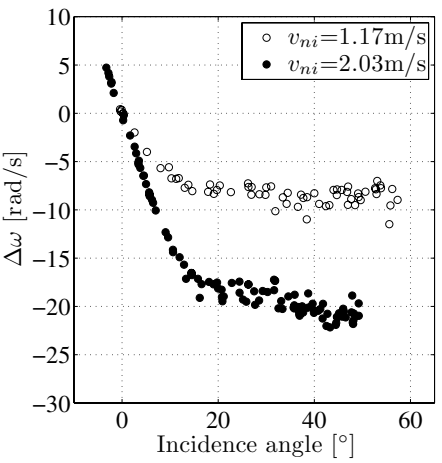


Fig. 33. $\Delta\omega$ for varied v_{ni}

computed. This reveals an energy dissipation which may partially be due to the insufficient surface finish of the impact pair and partially caused by possible tilting motions when the rotational velocity is low.

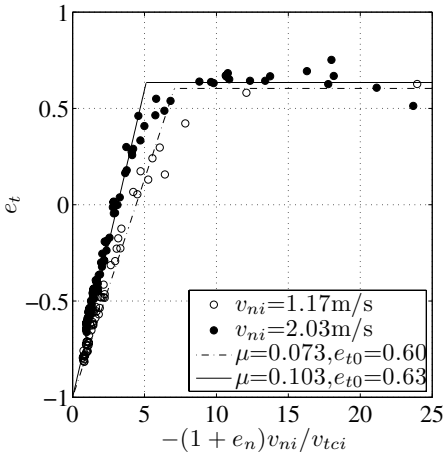


Fig. 34. e_t versus $-(1+e_n)\ v_{ni}/v_{tci}$

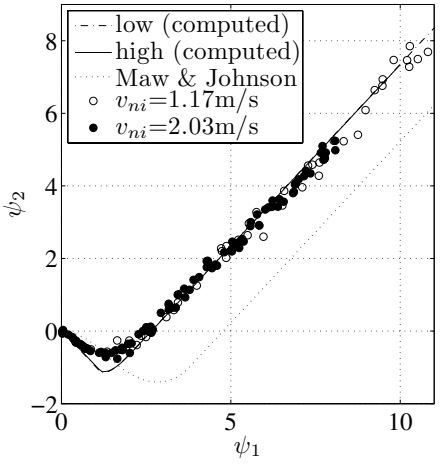


Fig. 35. Variation of ψ_2 with ψ_1

4 Conclusions

Since the start of the project in January 2004, it helped to accomplish many important contributions to mortar discretization methods for two-body frictional contact problems. New optimal a priori error estimates have been established as well as methods for the numerically efficient treatment of such nonlinear problems. The first goal was to show optimal a priori error estimates for the mortar discretization based on dual Lagrange multipliers. The results can be seen in [34, 35, 36, 38]. The second focus was on combining efficient iterative solvers based on multigrid techniques with the idea of semismooth Newton methods leading to primal-dual active set strategies. This approach leads to an efficient iterative solver for frictional contact problems which can be adjusted to nonlinear material laws with large deformations and to the dynamic case. Energy conserving time stepping schemes, see, e.g. [12, 50, 63], have been adapted to our primal-dual active set strategy. For the results, we refer to [35, 37, 39]. In the framework of this project, the convergence of a mortar multigrid method for the linear problem with a $\mathcal{V}$-cycle was proved, see [74]. Additionally, a locking-free formulation for the nearly incompressible case and the corresponding optimal a priori error estimates were achieved, see [11, 48]. Furthermore, an a posteriori error estimate, which could be extended to contact problems as described in Sect. 2.5, was developed, see [56]. An overview of the application of mortar techniques to contact problems is given in [40].

Although during the research work in this project questions concerning frictional contact problems both from the theoretical as well as from the numerical point of view were answered, there are still many open tasks in this active field. Especially, there is still a need for stable and consistent time integration schemes preserving the energy.

An experimental setup was designed to perform measurements upon the impact of a rotating disc striking an elastic strip. Electronic control was realized to release the disc with prescribed translational and rotational velocities based on a control strategy. The impact event is captured by a measurement unit consisting of a high-speed digital camera system and an image processing package. Impact variables, such as coefficients of normal and tangential restitution, impulse ratio, rotational velocity change, incidence and rebound angles, were measured. The instantaneous impact model was shown to be capable of capturing the measurement data. An analysis of the experimental results is performed. It is shown that higher flexibility leads to more energy transfer to structural vibrations and a smaller transition incidence angle from sticking to slipping, while higher initial normal velocity results in more energy transfer to structural vibrations, too, but a larger transition incidence angle. For interpreting the experimental results, a hybrid impact model was developed, which can provide predictions that are in good agreement with measurements. The model can also give some insight into elastic effects of strip flexibility and initial motions.

Up to now the fundamental results for the mathematical, the numerical and the experimental parts of the project were achieved. This provides a good starting point to compare the results of the numerical simulation with the experimental data during the final period of the project.

Acknowledgments

The first and the third author (ITM) want to thank Prof. F. Pfeiffer and Prof. H. Ulbrich from the Technical University of Munich, Germany very much for providing us with the throwing machine which was developed at their Institute of Applied Mechanics. Many experimental data can be found in the PhD thesis of M. Beitelschmidt [6] for rigid body impact. After an exchange of the electronics and control and after reproducing their data acquisition and processing parts the machine worked beautifully for our purpose. The authors also wish to acknowledge the excellent advice and assistance from Dr.-Ing. A. Eiber.

References

1. M. Ainsworth and J. Oden. *A posteriori error estimation in finite element analysis*. John Wiley & Sons, New York, 2000.
2. M. Al-Mousawi. On experimental studies of longitudinal and flexural wave propagations: An annotated bibliography. *Applied Mechanics Review*, 39:835–864, 1986.
3. 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.
4. D. Arnold and R. Winther. Mixed finite element methods for elasticity. *Numerische Mathematik*, 92:401–419, 2002.
5. P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuß, H. Rentz–Reichert, and C. Wieners. UG – a flexible software toolbox for solving partial differential equations. *Computing and Visualization in Science*, 1:27–40, 1997.
6. M. Beitelschmidt. *Reibstöße in Mehrkörpersystemen*. Fortschritt-Berichte VDI, Reihe 11, Nr. 275. VDI Verlag, Düsseldorf, 1999.
7. F. Ben Belgacem. Numerical simulation of some variational inequalities arisen from unilateral contact problems by the finite element methods. *SIAM Journal on Numerical Analysis*, 37(4):1198–1216, 2000.
8. F. Ben Belgacem, P. Hild, and P. Laborde. Extension of the mortar finite element method to a variational inequality modeling unilateral contact. *Mathematical Models & Methods in Applied Sciences*, 9:287–303, 1999.
9. F. Ben Belgacem and Y. Renard. Hybrid finite element methods for the Signorini problem. *Mathematics of Computation*, 72(243):1117–1145, 2003.
10. P. Boieri, F. Gastaldi, and D. Kinderlehrer. Existence, Uniqueness, and Regularity Results for the Two-Body Contact Problem. *Applied Mathematics and Optimization*, 15:251–277, 1987.

11. K. Chavan, B. Lamichhane, and B. Wohlmuth. Locking-free finite element methods for linear and nonlinear elasticity in 2D and 3D. Technical Report 13, Universität Stuttgart, SFB 404, 2005.
12. V. Chawla and T. Laursen. Energy consistent algorithms for frictional contact problems. *International Journal for Numerical Methods in Engineering*, 42:799–827, 1998.
13. P. Christensen and J. Pang. Frictional contact algorithms based on semismooth Newton methods. In *Reformulation: nonsmooth, piecewise smooth, semismooth and smoothing methods*, volume 22 of *Applied Optimization*, pages 81–116. Kluwer, 1999.
14. A. Christoforou and A. Yigit. Effect of flexibility on low velocity impact response. *Journal of Sound and Vibration*, 217(3):563–578, 1998.
15. P. Coorevits, P. Hild, K. Lhalouani, and T. Sassi. Mixed finite element methods for unilateral problems: convergence analysis and numerical studies. *Mathematics of Computation*, 71(237):1–25, 2001.
16. P. Eberhard. *Kontaktuntersuchungen durch hybride Mehrkörpersystem/Finite Elemente Simulationen*. Shaker Verlag, Aachen, 2000.
17. P. Eberhard and B. Hu. *Advanced Contact Dynamics*. Southeast University Press, Nanjing, 2003. (in Chinese).
18. P. Eberhard and W. Schiehlen. Computational dynamics of multibody systems: History, formalisms, and applications. *Journal of Computational and Nonlinear Dynamics*, 1(1):3–12, 2006.
19. R. Falk. Error estimates for the approximation of a class of variational inequalities. *Mathematics of Computation*, 28:963–971, 1974.
20. B. Flemisch, M. Puso, and B. Wohlmuth. A new dual mortar method for curved interfaces: 2D elasticity. *International Journal for Numerical Methods in Engineering*, 63(6):813–832, 2005.
21. S. Foerster, M. Louge, H. Chang, and K. Allia. Measurements of the collision properties of small spheres. *Physics of Fluids*, 6(3):1108–1115, 1994.
22. R. Glowinski. *Numerical methods for nonlinear variational problems*. Springer-Verlag, New York, 1984.
23. W. Goldsmith. *Impact: The Theory and Physical Behaviour of Colliding Solids*. Edward Arnold, London, 1960.
24. D. Gorham and A. Kharaz. The measurement of particle rebound characteristics. *Powder Technology*, 112(3):193–202, 2000.
25. K. Graff. *Wave Motion in Elastic Solids*. Oxford University Press, Oxford, 1975.
26. J. Haslinger and I. Hlaváček. Contact between two elastic bodies – I. continuous problems. *Aplikace Mathematiky*, 25:324–347, 1980.
27. J. Haslinger, I. Hlaváček, and J. Nečas. Numerical methods for unilateral problems in solid mechanics. In P. Ciarlet and J.-L. Lions, editors, *Handbook of Numerical Analysis*, volume IV, pages 313–485. North-Holland, 1996.
28. H. Hertz. Über die Berührung fester elastischer Körper. *Journal für die reine und angewandte Mathematik*, 92:156–171, 1882.
29. P. Hild. Numerical implementation of two nonconforming finite element methods for unilateral contact. *Computer Methods in Applied Mechanics and Engineering*, 184(1):99–123, 2000.
30. P. Hild and P. Laborde. Quadratic finite element methods for unilateral contact problems. *Applied Numerical Mathematics*, 41:410–421, 2002.

31. M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2003.
32. B. Hu and P. Eberhard. Experimental and theoretical investigation of a rigid body striking an elastic rod. Institute Report IB-32, Institute B of Mechanics, University of Stuttgart, Stuttgart, Germany, 1999.
33. H. Hu. On some variational principles in the theory of elasticity and the theory of plasticity. *Scientia Sinica*, 4:33–54, 1955.
34. S. Hüeber, M. Mair, and B. Wohlmuth. A priori error estimates and an inexact primal-dual active set strategy for linear and quadratic finite elements applied to multibody contact problems. *Applied Numerical Mathematics*, 54:555–576, 2005.
35. S. Hüeber, A. Matei, and B. Wohlmuth. Efficient algorithms for problems with friction. Technical Report 007, Universität Stuttgart SFB 404, 2005. To appear in SIAM Journal on Scientific Computing.
36. S. Hüeber, A. Matei, and B. Wohlmuth. A mixed variational formulation and an optimal a priori error estimate for a frictional contact problem in elasto-piezoelectricity. *Bulletin Mathématique de la Société des Sciences Mathématiques de Roumanie*, 48(96)(2):209–232, 2005.
37. S. Hüeber, G. Stadler, and B. Wohlmuth. A primal-dual active set algorithm for three-dimensional contact problems with Coulomb friction. Technical report, Pré-publicações do Departamento de Matemática da Universidade de Coimbra 06-16, 2006.
38. S. Hüeber and B. Wohlmuth. An optimal a priori error estimate for non-linear multibody contact problems. *SIAM Journal on Numerical Analysis*, 43(1):157–173, 2005.
39. S. Hüeber and B. Wohlmuth. A primal-dual active set strategy for non–linear multibody contact problems. *Computer Methods in Applied Mechanics and Engineering*, 194:3147–3166, 2005.
40. S. Hüeber and B. Wohlmuth. Mortar methods for contact problems. In P. Wriggers and U. Nackenhorst, editors, *Analysis and Simulation of Contact Problems*, volume 27 of *Lecture Notes in Applied and Computational Mechanics*, pages 39–47. Springer-Verlag, Berlin, 2006.
41. Y. Jang and P. Eberhard. An experimental and numerical study of deformable bodies contact. In *Proceedings of Applied Mathematics and Mechanics*, 2006. Submitted for publication.
42. K. Johnson. *Contact Mechanics*. Cambridge University Press, Cambridge, 1985.
43. E. Kasper and R. Taylor. A mixed-enhanced strain method. Part I: geometrically linear problems. *Computers and Structures*, 75:237–250, 2000.
44. A. Kharaz, D. Gorham, and A. Salman. An experimental study of the elastic rebound of spheres. *Powder Technology*, 120(3):281–291, 2001.
45. N. Kikuchi and J. Oden. *Contact problems in elasticity: A study of variational inequalities and finite element methods*. SIAM Studies in Applied Mathematics 8, Philadelphia, 1988.
46. L. Labous, A. Rosato, and R. Dave. Measurements of collisional properties of spheres using high-speed video analysis. *Physical Review E*, 56:5717–5725, 1997.
47. P. Ladevèze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM Journal on Numerical Analysis*, 20(3):485–509, 1983.

48. B. Lamichhane, B. Reddy, and B. Wohlmuth. Convergence in the incompressible limit of finite element approximations based on the Hu–Washizu formulation. Technical Report 05, University of Stuttgart, SFB 404, 2004.
49. T. Laursen. *Computational Contact and Impact Mechanics*. Springer-Verlag, Berlin, 2002.
50. T. Laursen and V. Chawla. Design of energy conserving algorithms for frictionless dynamic contact problems. *International Journal for Numerical Methods in Engineering*, 40:836–886, 1997.
51. K. Lhalouani and T. Sassi. Nonconfirming mixed variational inequalities and domain decomposition for unilateral problems. *East-West Journal of Numerical Mathematics*, 7:23–30, 1999.
52. A. Lorenz, C. Tuozzolo, and M. Louge. Measurements of impact properties of small, nearly spherical particles. *Experimental Mechanics*, 37(3):292–298, 1997.
53. N. Maw, J. Barber, and J. Fawcett. The oblique impact of elastic spheres. *Wear*, 38(1):101–114, 1976.
54. N. Maw, J. Barber, and J. Fawcett. The role of elastic tangential compliance in oblique impact. *Journal of Lubrication Technology*, 103:74–80, 1981.
55. R. Mindlin and H. Deresiewicz. Elastic spheres in contact under varying oblique forces. *Journal of Applied Mechanics*, 20:327–344, 1953.
56. S. Nicaise, K. Witowski, and B. Wohlmuth. An a posteriori error estimator for the lamé equation based on $H(\mathrm{div})$-conforming stress approximations. Technical Report 005, University of Stuttgart, SFB 404, 2006.
57. F. Pfeiffer and C. Glocker. *Multibody Dynamics with Unilateral Contacts*. John Wiley & Sons, New York, 1996.
58. W. Schiehlen. Multibody system dynamics: Roots and perspectives. *Multibody System Dynamics*, 1(2):149–188, 1997.
59. W. Schiehlen, R. Seifried, and P. Eberhard. Elastoplastic phenomena in multibody impact dynamics. *Computer Methods in Applied Mechanics and Engineering*, 2006 (accepted for publication).
60. J. Sears. On the longitudinal impact of metal rods with rounded ends. *Transactions of the Cambridge Philosophical Society*, 21:49–106, 1912.
61. R. Seifried, B. Hu, and P. Eberhard. Numerical and experimental investigation of radial impacts on a half-circular plate. *Multibody System Dynamics*, 9(3):265–281, 2003.
62. A. Shabana. Flexible multibody dynamics: Review of past and recent developments. *Multibody System Dynamics*, 1(2):189–222, 1997.
63. J. Simo and N. Tarnow. The discrete energy-momentum method. Conserving algorithms for nonlinear elastodynamics. *Zeitschrift für Angewandte Mathematik und Physik*, 43(5):757–792, 1992.
64. J. Solberg and P. Papadopoulos. An analysis of dual formulations for the finite element solution of two-body contact problems. *Computer Methods in Applied Mechanics and Engineering*, 194:2734–1780, 2005.
65. R. Sondergaard, K. Chaney, and C. Brennen. Measurements of solid spheres bouncing off flat plates. *Journal of Applied Mechanics*, 57:694–699, 1990.
66. G. Stadler. Semismooth Newton and augmented Lagrangian methods for a simplified friction problem. *SIAM Journal on Optimization*, 15(1):39–62, 2004.
67. Stemmer Imaging GmbH. *Common Vision Blox–Tool: ShapeFinder*. Stemmer Imaging GmbH, 2003.
68. W. Stronge. *Impact Mechanics*. Cambridge University Press, 2000.

69. O. Walton. Numerical simulation of inelastic, frictional particle-particle interactions. In M. Roco, editor, *Particulate Two Phase Flow*, pages 884–911. Butterworth-Heinemann, Boston, 1993.
70. K. Washizu. *Variational methods in elasticity and plasticity*. Pergamon Press, Oxford, 1982.
71. K. Willner. *Kontinuums- und Kontaktmechanik*. Springer-Verlag, Berlin, 2003.
72. B. Wohlmuth. A mortar finite element method using dual spaces for the Lagrange multiplier. *SIAM Journal on Numerical Analysis*, 38:989–1012, 2000.
73. B. Wohlmuth. *Discretization Methods and Iterative Solvers Based on Domain Decomposition*. Springer, 2001.
74. B. Wohlmuth. A V-cycle multigrid approach for mortar finite elements. *SIAM Journal on Numerical Analysis*, 42:2476–2495, 2005.
75. B. Wohlmuth and R. Krause. Monotone methods on non-matching grids for non linear contact problems. *SIAM Journal on Scientific Computing*, 25:324–347, 2003.
76. P. Wriggers. *Computational Contact Mechanics*. John Wiley & Sons, Chichester, 2002.

Regularity of Elastic Fields in Composites*

Dorothee Knees[1] and Anna-Margarete Sändig[2]

[1] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany
knees@wias-berlin.de

[2] Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
saendig@ians.uni-stuttgart.de

Summary. It is well known that high stress concentrations can occur in elastic composites in particular due to the interaction of geometrical singularities like corners, edges and cracks and structural singularities like jumping material parameters. In the project C5 *Stress concentrations in heterogeneous materials* of the SFB 404 it was mathematically analyzed where and which kind of stress singularities in coupled linear and nonlinear elastic structures occur. In the linear case asymptotic expansions near the geometrical and structural peculiarities are derived, formulae for generalized stress intensity factors included. In the nonlinear case such expansions are unknown in general and regularity results are proved for elastic materials with power-law constitutive equations with the help of the difference quotient technique combined with a quasi-monotone covering condition for the subdomains and the energy densities. Furthermore, some applications of the regularity results to shape and structure optimization and the Griffith fracture criterion in linear and nonlinear elastic structures are discussed. Numerical examples illustrate the results.

Keywords: Regularity, elasticity, composites, stress singularities, Ramberg-Osgood model

1 Introduction

Composites play an important role in everyday life, examples are fiber-reinforced composites in car industry, piezo-electric stack actuators or semiconductor devices. From experience and experiments it is well known that very high mechanical stresses can occur in the composite in the vicinity of re-entrant corners, edges, cracks and near interior surfaces, where the different materials of the composite come together. These stress concentrations have a strong influence on the strength and physical lifetime of the structure. Their knowledge is fundamental for fracture and failure criteria.

* Research Project C5 "Stress Singularities in Heterogeneous Materials"

The project C5 *Stress concentrations in heterogeneous materials* was devoted to the mathematical analysis of stress singularities in linear and nonlinear elastic coupled nonsmooth structures. In the first years of this project we have focused on linear problems whereas in the last period we have concentrated on some nonlinear boundary-transmission problems.

In this article we give an overview on the regularity results, shortly for linear and more detailed for nonlinear elastic composites. We consider bodies which are composed of several elastic substructures with different material properties. The whole body as well as the interfaces, which separate the substructures, may have corners or edges. Throughout the whole paper we assume small strains and consider constitutive laws which lead to linear elliptic systems of partial differential equations or to quasilinear elliptic systems of p-structure for the displacement fields. These PDEs have piecewise constant coefficients due to the heterostructure of the composite.

The Linear Case

In this case the substructures consist of linear elastic materials and the singular behavior of the displacement and stress fields can be completely characterized by means of an asymptotic expansion of the solution near the mentioned geometrical and structural peculiarities. In two dimensions this expansion reads in the neighborhood of a corner point or an interior cross point S for a displacement field u (polar coordinates with respect to S are used):

$$\eta^S u(r,\varphi) = u_{\mathrm{reg}}(r,\varphi) + \eta^S \sum_{0<\Re\alpha<1} c_\alpha r^\alpha v_\alpha(\ln r,\varphi)\ . \tag{1}$$

The singular exponents $\alpha \in \mathbb{C}$ are eigenvalues of a corresponding nonlinear eigenvalue problem and the functions v_α consist of (generalized) eigenfunctions and powers of $\ln r$. The constants c_α are generalized stress intensity factors and depend on the given external loading. η^S is a cut-off function with respect to S and u_{reg} is a regular function. The regularity of the solution u is determined by the singular exponent α with the smallest positive real part. This exponent and the corresponding function v_α can be explicitly calculated for fixed geometries and material parameters and do not depend on the given external forces.

Expansions like (1) are well established for homogeneous materials and, more general, for linear elliptic systems with smooth coefficients, see for example [12, 20, 36]. In Sect. 2 we will demonstrate that the Mellin technique as an appropriate mathematical tool guarantees such expansions for solutions of general elliptic boundary-transmission problems in composites, too [43, 44]. Furthermore, we will formulate explicit formulae for the constants c_α in this case and give a numerical example. Very small real parts of the exponents α can appear in general. In [27, 29, 45], we derived a criterion for linear elliptic systems with piecewise constant coefficients, that guarantees $\mathrm{Re}\ \alpha \geq \frac{1}{2}$ in the

two dimensional case and similar results for higher dimensions. This condition, the quasi-monotonicity condition, was originally defined for the Laplace operator with piecewise constant coefficients in a completely different context [16]. Its relevance for the regularity of weak solutions was discovered in [48] for the two dimensional Laplacian and in [45] for isotropic bi-materials. We discuss this condition in Sect. 2.4.

The Nonlinear Case

Much less is known about the regularity of displacement and stress fields of nonlinear elastic materials. For some classes of semilinear and quasilinear systems of partial differential equations (e.g. stationary Navier-Stokes equations, semiconductor equations) it can be shown [1, 6, 49] that the regularity of the solutions is dominated by assigned linearized problems. Furthermore, a comparison principle and barrier functions are used for a special class of scalar nonlinear equations on two-dimensional domains (see e.g. [8, 15, 55]) to get similar results.

Nonlinear elastic field equations of power-law type do not fit in this framework in general and it is an open question whether the behavior of the elastic fields can be completely characterized by leading terms in an asymptotic expansion similar to the linear case (1). First investigations into this direction were done at the end of the sixties in [22, 50] for homogeneous materials of Ramberg-Osgood type (power-law models). In order to describe the elastic fields near a crack tip an ansatz (HRR-field) of the form

$$u = r^{\alpha} v_{\alpha}(\varphi) \tag{2}$$

was inserted into the corresponding field equations and led to a fully nonlinear eigenvalue problem for the determination of the eigenpairs (α, v_{α}).

The field equations of the Ramberg-Osgood model are closely related to general systems of quasilinear elliptic partial differential equations of p-structure, see e.g. [18] for a definition. In [18, 19, 52], Ebmeyer, Frehse and Savaré obtained independently global regularity results for weak solutions of such systems on nonsmooth domains with a difference quotient technique. The difference quotient method is also applicable to the Ramberg-Osgood model [31, 32] and we cite and discuss the corresponding results in Sect. 4.2.

At the beginning of the last research period of the SFB 404 only very few regularity results for transmission problems of p-growth were reported in the literature; e.g. two subdomains were considered with either smooth interface and different p_i [39] or with nonsmooth interface and $p_1 = p_2 = 2$ [52]. In Sect. 3 we will present recently derived regularity results for transmission problems in composites where on the substructures Ω_i we have different quasilinear elliptic systems of p_i-structure. We do not restrict the number of subdomains and the growth properties of the differential operators may vary from subdomain to subdomain.

For obtaining the regularity results, the main idea is to combine the difference quotient technique with the concept of a quasi-monotone distribution of material parameters known from the linear problems. This leads to the new concept of a quasi-monotone covering condition for the subdomains and the energy densities which determine the differential operators on the subdomains. The very special case of the linear Laplace equation with piecewise constant coefficients is included and our general quasi-monotone covering condition coincides in this case up to an additional geometric condition with the original definition of quasi-monotonicity from [16].

In the last section of this paper we discuss applications of the regularity results for linear and nonlinear elastic problems. These are shape and structure optimization problems in nonsmooth domains, sensitivity analysis for compound elastic structures and the Griffith fracture criterion for a nonlinear elastic model of power-law type. Relying on the proved regularity results formulae for shape derivatives and the energy release rate are derived which are suitable for computations.

2 Linear Elastic Composites

2.1 Weak Formulation

We start with the weak formulation of the elastic field equations in a composite. Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded domain consisting of pairwise disjoint subdomains $\Omega_i \subset \Omega$, $1 \le i \le M$, with $\overline{\Omega} = \cup_{i=1}^{M} \overline{\Omega_i}$. We assume for simplicity that Ω_i are Lipschitz domains of polygonal or polyhedral type. We distinguish between exterior boundary pieces

$$\Gamma_i = \operatorname{int}(\partial\Omega_i \cap \partial\Omega) \tag{3}$$

and interior boundary pieces, the common interfaces of Ω_i and Ω_j,

$$\Gamma_{ij} = \operatorname{int}(\partial\Omega_i \cap \partial\Omega_j)\,. \tag{4}$$

In each domain Ω_i we consider the equilibrium equations for two classes of fields: potential fields (antiplanar case, (5)) and linear, anisotropic elastic fields, that means (6):

$$-\operatorname{div}(\mu_i \nabla u_i) = f_i \quad i = 1, \dots, M\,, \tag{5}$$

$$-\operatorname{div}\sigma^i(u^i) = F_i \quad i = 1, \dots, M\,. \tag{6}$$

Here, μ_i are given positive constants (shear moduli), u_i in (5) are the scalar potentials, $u^i = (u^i_1, \dots, u^i_d)^\top$ in (6) the displacement fields and $\sigma^i = (\sigma^i_{kj})_{kj} \in \mathbb{R}^{d\times d}$ are the stress tensors. For small strains, Hooke's law yields

$$\sigma^i_{kj}(u^i) = \sum_{m,n=1}^{d} C^i_{kjmn}\varepsilon_{mn}(u^i), \quad i = 1, \dots, M,\ k, j = 1, \dots, d\,,$$

where $\varepsilon_{mn}(u^i) = \frac{1}{2}(\frac{\partial u^i_m}{\partial x_n} + \frac{\partial u^i_n}{\partial x_m})$ are the components of the linearized strain tensor $\varepsilon(u^i) \in \mathbb{R}^{d\times d}$. It is assumed that the material tensors C^i are symmetric and positive definite

$$(C^i\xi) : \xi = \sum_{k,j,m,n=1}^{d} C^i_{kjmn}\xi_{kj}\xi_{mn} \geq M_i \sum_{k,j=1}^{d} |\xi_{kj}|^2 \tag{7}$$

for every $\xi = (\xi_{kj})_{kj} \in \mathbb{R}^{d\times d}_{\text{sym}}$. We consider Dirichlet and Neumann boundary conditions on $\partial\Omega = \overline{\Gamma_D} \cup \overline{\Gamma_N}$. These conditions read for (6)

$$\sigma\mathbf{n} = g_i \quad \text{on } \Gamma_N \ , \tag{8}$$

$$u = 0 \quad \text{on } \Gamma_D \ , \tag{9}$$

where $\sigma = \sigma^i$ on $\Gamma_N \cap \Gamma_i$, $u = u^i$ on $\Gamma_D \cap \Gamma_i$ and $\Gamma_D \cap \Gamma_N = \emptyset$. Furthermore, we assume that the subdomains Ω_i are bonded, which is expressed by the following transmission conditions on the skeleton $\Gamma = \cup_{i,j=1}^{M}\Gamma_{ij}$ ($\mathbf{n}_i$ denotes the exterior unit normal vector on Γ_{ij} with respect to Ω_i):

$$u^i = u^j \qquad \text{on } \Gamma_{ij} \ , \tag{10}$$

$$\sigma^i\mathbf{n}_i = \sigma^j\mathbf{n}_i \quad \text{on } \Gamma_{ij} \ . \tag{11}$$

We are now in a position to formulate the boundary-transmission problems in a weak sense. Let

$$V = \{u \in W^{1,2}(\Omega) : u\big|_{\Gamma_D} = 0\} \ , \tag{12}$$

where the restriction to Γ_D is to be understood in the trace-sense. We assume that $f \in V'$, $f_i = f$ on Ω_i, $g \in W^{-\frac{1}{2},2}(\Gamma_N)$, $g = g_i$ on $\Gamma_N \cap \Gamma_i$. The weak formulation reads: Find $u \in V$ such that for every $v \in V$

$$a(u, v) = \langle f, v\rangle_\Omega + \langle g, v\rangle_{\Gamma_N} \ , \tag{13}$$

where for the Poisson equations (5) we have

$$a(u, v) = \int_\Omega \mu(x)\nabla u\nabla v \, \mathrm{d}x = \sum_{i=1}^{M} \int_{\Omega_i} \mu_i \nabla u_i \nabla v_i \, \mathrm{d}x \tag{14}$$

and for the equations of linear elasticity (6) we have

$$a(u, v) = \int_\Omega \sigma(u) : \varepsilon(u) \, \mathrm{d}x = \sum_{i=1}^{M} \int_{\Omega_i} \sigma^i(u^i) : \varepsilon(u^i) \, \mathrm{d}x \ . \tag{15}$$

Both problems, (13)+(14) and (13)+(15), respectively, are elliptic boundary-transmission problems (see [43, 51]) due to the assumptions on the material parameters. Therefore it holds

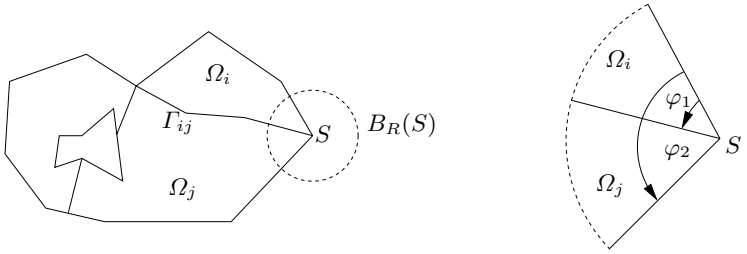


Fig. 1. Polygonal domain

1. If the surface measure of Γ_D does not vanish, then there exists a unique solution $u \in V$ of (13).
2. If $\Gamma_D = \emptyset$ and if $\langle f, r\rangle_\Omega + \langle g, r\rangle_{\Gamma_N} = 0$ for every $r \in \mathcal{R} \subset W^{1,2}(\Omega)$, where $\mathcal{R}$ is the set of rigid body displacements, then there exists a solution $u \in W^{1,2}(\Omega)$ which is unique up to elements from $\mathcal{R}$.

If Ω is a homogeneous body with a smooth boundary, then the regularity of a weak solution of problem (13) is determined by the regularity of the data f, g. But, if Ω is a composite and if the boundary of Ω has corners and edges, then such a regularity is not longer valid even if f and g are smooth. The geometric (corners, edges) and structural singularities (discontinuous material parameters) have an essential influence on the regularity of the solutions. Weak solutions of boundary-transmission problems of the form (13) can be decomposed into singular terms describing the behavior near corners, edges and jumps of the material parameters and into a more regular term. This decomposition can be described and calculated by Mellin techniques, similar to pure boundary-value problems. Special ansatzes frequently used in mechanics [23, 57, 58, 59] lead also to such results. In the following section we outline the use of the ansatzes and the Mellin technique for two dimensional transmission problems and we mention results for the 3D-case.

2.2 2D-Corner Singularities

Let $\mathcal{S}$ be the set of the corner points of $\partial\Omega_i$ and of the points of $\overline{\Gamma_D} \cap \overline{\Gamma_N}$, see Fig. 1. For a fixed point $S \in \mathcal{S}$ we introduce polar coordinates (r, φ) with respect to S and search a corner singularity of the form

$$u_{\text{sing}}(r, \varphi) = r^\alpha \Phi_\alpha(\varphi) \ , \tag{16}$$

The corner singularity (16) has to be a solution of

$$a(u_{\text{sing}}, v) = 0 \tag{17}$$

for $v \in V$ with $\operatorname{supp} v \subset \Omega_S = \overline{\Omega} \cap B_R(S)$, where $B_R(S)$ is an open ball centered at S with appropriate small radius and $a(\cdot, \cdot)$ is the bilinear form

from (14) or (15). Choosing the special test function $v = \eta(r)\Psi(\varphi)$, where $\Psi \in V_\varphi = \{\Psi \in W^{1,2}(0,\varphi_2) : \Psi|_{\Gamma_D} = 0\}$ and $\eta \in \mathcal{C}_0^\infty((-R,R))$, we obtain from (17) a bilinear form which depends on $S \in \mathcal{S}$ and the parameter α:

$$a_S(\alpha; \Phi_\alpha, \Psi) = 0 \,, \tag{18}$$

see for example [37, 41, 45] for explicit formulae. Relation (18) defines a quadratic eigenvalue problem and we refer to Sect. 2.4 for some examples.

By this procedure one can generate weak solutions having a singular behavior of the form (16) near a point $S \in \mathcal{S}$. Now the question is whether terms of the form (16) characterize the singular behavior of weak solutions completely. This problem can be answered using an appropriate integral transform, the Mellin transform [37]:

$$\mathcal{M}(u(\cdot,\varphi))[\alpha] = \hat{u}(\alpha,\varphi) = \int_0^\infty r^{-\alpha-1} u(r,\varphi)\,\mathrm{d}r \,. \tag{19}$$

This transform maps $r\partial_r$ into the complex parameter α. The Mellin transform applied to the differential operators in (5), (6), (8)–(11), generates a Fredholm operator pencil $\mathcal{A}_S(\alpha)$ with

$$\mathcal{A}_S(\alpha) : V_\varphi \to V_\varphi' \,. \tag{20}$$

For explicit formulas for $\mathcal{A}_S(\alpha)$ see the examples in [37, 45]. The operator pencil $\mathcal{A}_S(\alpha)$ and the bilinear form $a_S(\alpha;\cdot,\cdot)$ from (18) are related as follows

$$\langle \mathcal{A}_S(\alpha)\Phi, \Psi\rangle = a_S(\alpha;\Phi,\Psi) \quad \text{ for every } \Phi,\Psi \in V_\varphi \,. \tag{21}$$

The corresponding quadratic eigenvalue problem $\mathcal{A}_S(\alpha)\Phi_\alpha = 0$ has a finite number of eigenvalues in any strip $c_1 \le \mathrm{Re}\ \alpha \le c_2$, see [37, 45]. The following regularity theorem is proved with the Mellin technique and connects the eigenvalue problems (18) to the global regularity of weak solutions.

Theorem 1 (Regularity theorem). *[20, 36, 37, 45] Let the volume force densities f of (5) and F of (6) be elements of $W^{l,2}(\Omega)$ and let the Neumann datum g be in $W^{l+\frac{1}{2},2}(\Gamma_N)$ with $l \in \mathbb{N}_0$. Assume that $\mathcal{A}_S(\alpha)$ is invertible on the line* $\mathrm{Re}\ \alpha = l+1$. *Then the weak solution $u \in V$ of the boundary-transmission problem admits the following decomposition:*

$$u = u_{reg} + \sum_{\substack{S\in\mathcal{S},\\ \gamma\in\Lambda_{S,l}}} \eta_S c_\gamma^S v_\gamma^S(r,\varphi) \,. \tag{22}$$

Here, $u_{reg}\big|_{\Omega_i} \in W^{l+2,2}(\Omega_i)$ and

$$\Lambda_{S,l} = \{\gamma = (\alpha,\mu,\kappa) : \alpha \text{ is an eigenvalue of } \mathcal{A}_S(\alpha) \text{ in the strip } 0 < \mathrm{Re}\ \alpha < l+1;\ \mu = 1,\dots,I_\alpha^S;\ \kappa = 0,\dots,M_{\alpha,\mu}^S\} \,. \tag{23}$$

I_α^S denotes the geometrical multiplicity of α, $\{\Phi^S_{\alpha,\mu,\kappa}, \mu = 1,\dots,I^S_\alpha; \kappa = 0,\dots,M^S_{\alpha,\mu}\}$ is a canonical system of Jordan chains of $\mathcal{A}_S(\alpha)$ with respect to eigenvalue α, $M^S_{\alpha,\mu}+1$ are the lengths of the Jordan chains, η_S are cut-off functions which equal to 1 near S and the singular functions v^S_γ are of the form

$$v^S_\gamma(r,\varphi) = r^\alpha \sum_{q=0}^{\kappa} \frac{(\ln r)^q}{q!} \Phi^S_{\alpha,\mu,\kappa-q}(\varphi)\ . \tag{24}$$

The constants c^S_γ are also called generalized stress intensity factors and depend on the data.

Coefficient Formulae

The coefficients c^S_γ in (22) express the intensity of the singular functions $v^S_\gamma(r,\varphi)$. In particular, they can vanish and then u is regular. Damage and crack criteria rely on these coefficients. The coefficients depend on the exterior forces, the elastic material parameters and on the geometry of the domain and can be calculated via so-called coefficient formulae. In the case of homogeneous materials these formulae are well known [37, 41]. In this section we derive analogous formulae of Mazya/Plamenevski type for potential fields in composites. Corresponding coefficient formulae for linear elastic fields in bonded structures are described in [5].

Lemma 1. *Let $\Omega, \Omega_1, \Omega_2 \subset \mathbb{R}^2$ be bounded polygons, $\overline{\Omega} = \cup_{i=1}^2 \overline{\Omega_i}$ and $S \in \partial\Omega \cap \partial\Omega_1 \cap \partial\Omega_2$ a corner point (see Fig. 1). For a weak solution u of the boundary-transmission problem* (13) *with bilinear form* (14) *and $M = 2$ the expansion* (22) *reads near the point S ($u_i = u\big|_{\Omega_i}$):*

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \eta_S c_0 r^{\alpha_0} M(\alpha_0) \begin{pmatrix} \sin(\alpha_0\varphi) \\ \cos(\alpha_0\varphi) \end{pmatrix} + \textit{remainder}\ . \tag{25}$$

Here, $M(\alpha_0)$ is a matrix depending on the kind of the boundary conditions near the corner S:

$$M(\alpha_0) = M_{\text{D-D}} = \begin{pmatrix} -\sin(\alpha_0(\varphi_2-\varphi_1)) & 0 \\ \cos(\alpha_0\varphi_2)\sin(\alpha_0\varphi_1) & -\sin(\alpha_0\varphi_2)\sin(\alpha_0\varphi_1) \end{pmatrix}\ ,$$

$$M(\alpha_0) = M_{\text{N-N}} = \begin{pmatrix} 0 & \cos(\alpha_0(\varphi_2-\varphi_1)) \\ \sin(\alpha_0\varphi_2)\cos(\alpha_0\varphi_1) & \cos(\alpha_0\varphi_1)\cos(\alpha_0\varphi_2) \end{pmatrix}\ ,$$

$$M(\alpha_0) = M_{\text{N-D}} = \begin{pmatrix} \cos(\alpha_0(\varphi_2-\varphi_1)) & 0 \\ \sin(\alpha_0\varphi_1)\sin(\alpha_0\varphi_2) & \sin(\alpha_0\varphi_1)\cos(\alpha_0\varphi_2) \end{pmatrix}\ .$$

The exponent α_0 is a zero of the equation

$$\mu_2\cos(\alpha(\varphi_2-\varphi_1))\sin(\alpha\varphi_1) + \mu_1\sin(\alpha(\varphi_2-\varphi_1))\cos(\alpha\varphi_1) = 0 \text{ (D-D)}\ ,$$
$$\mu_2\cos(\alpha\varphi_1)\sin(\alpha(\varphi_2-\varphi_1)) + \mu_1\sin(\alpha\varphi_1)\cos(\alpha(\varphi_2-\varphi_1)) = 0 \text{ (N-N)}\ ,$$
$$\mu_1\cos(\alpha\varphi_1)\cos(\alpha(\varphi_2-\varphi_1)) - \mu_2\sin(\alpha\varphi_1)\sin(\alpha(\varphi_2-\varphi_1)) = 0 \text{ (N-D)}\ .$$

The corresponding coefficient c_0 is given by

$$c_0 = \frac{1}{\alpha_0 K(\alpha_0)} \left(\int_\Omega \mu(f s_- + u \triangle s_-)\,\mathrm{d}x + \int_{\Gamma_N} \mu(g s_- - u \frac{\partial s_-}{\partial n}) \mathrm{d}\sigma \right) , \quad (26)$$

where $s_- = \eta_S r^{-\alpha_0} M(\alpha_0) \left(\sin(\alpha_0\varphi), \cos(\alpha_0\varphi)\right)^\top$ *and*

$$K(\alpha_0) = K_{\text{D-D}} = \mu_1\varphi_1 \sin^2 \alpha_0(\varphi_2 - \varphi_1) + \mu_2(\varphi_2 - \varphi_1)\sin^2 \alpha_0\varphi_1 ,$$
$$K(\alpha_0) = K_{\text{N-N}} = \mu_1\varphi_1 \cos^2 \alpha_0(\varphi_2 - \varphi_1) + \mu_2(\varphi_2 - \varphi_1)\cos^2 \alpha_0\varphi_1 ,$$
$$K(\alpha_0) = K_{\text{N-D}} = \mu_1\varphi_1 \cos^2 \alpha_0(\varphi_2 - \varphi_1) + \mu_2(\varphi_2 - \varphi_1)\sin^2 \alpha_0\varphi_1 .$$

For the meaning of the angles φ_1, φ_2 see Fig. 1.

Sketch of the proof. The formula (25) was derived via the Mellin technique in [45]. It remains to show (26): For a fixed cut-off function $\eta_S = \eta_S(r)$ we consider a family of balls $B_\delta(S) = \{x \in \mathbb{R}^2 : |x - S| = r < \delta\}$ such that $\eta \equiv 1$ on B_δ. We apply Green's formula on $\Omega_\delta = \Omega \backslash \overline{B_\delta(S)}$ and obtain

$$\begin{aligned}
&\int_{\Omega_\delta} \mu(\Delta u s_- - u \Delta s_-)\,\mathrm{d}x \\
&= \sum_{i=1}^{2} \int_{\Omega_\delta \cap \Omega_i \cap \operatorname{supp}\eta_S} \mu_i(\Delta u_i s_{-,i} - u_i \Delta s_{-,i})\,\mathrm{d}x \\
&= \mu_1\delta \int_0^{\varphi_1} \left(u_1 \frac{\partial s_{-,1}}{\partial r} - \frac{\partial u_1}{\partial r} s_{-,1} \right) \mathrm{d}\varphi + \mu_2\delta \int_{\varphi_1}^{\varphi_2} \left(u_2 \frac{\partial s_{-,2}}{\partial r} - \frac{\partial u_2}{\partial r} s_{-,2} \right) \mathrm{d}\varphi \\
&\quad + \int_{\Gamma_N \cap \partial\Omega_\delta} \mu \left(g s_- - u \frac{\partial s_-}{\partial \mathbf{n}} \right) \mathrm{d}\sigma .
\end{aligned}$$

Inserting the expansion (25) of u and considering the limit $\delta \to 0$ we get the coefficient formula (26).

Computation of Stress Intensity Factors for Interface Cracks

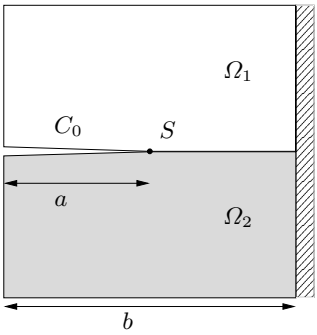
As an example we consider the linear isotropic elasticity problem in a one-sided clamped laminated structure with an interface crack, see Fig. 2. There are no volume forces and tensions of $\pm 2000\,\mathrm{MPa}$. The Young moduli are $E_1 = 200000\,\mathrm{MPa}$ and $E_2 = 400000\,\mathrm{MPa}$, the Poisson ratios are chosen as $\nu_1 = \nu_2 = 0.3$. The asymptotic expansion (22) near the crack tip S reads [5, 60]

$$u = \eta_S \left(c_1 r^{\frac{1}{2}+i\varepsilon} v_1(\varphi) + c_2 r^{\frac{1}{2}-i\varepsilon} v_2(\varphi) \right) + u_{reg} ,$$

where

$$\varepsilon = \frac{1}{2\pi} \ln \frac{1+\beta}{1-\beta}, \ \beta = \frac{\mu_2(1-2\nu_1) - \mu_1(1-2\nu_2)}{2(\mu_2(1-\nu_1) + \mu_1(1-\nu_2))} .$$

The coefficients c_i as well as the functions v_i are complex valued. The stress intensity factor $c_1 = K_1 + iK_2$ is computed via coefficient formulas similar to formula (26) and the real part and imaginary part are plotted in Fig. 2, [4]. Coefficient formulas for linear isotropic elasticity are derived in [5].



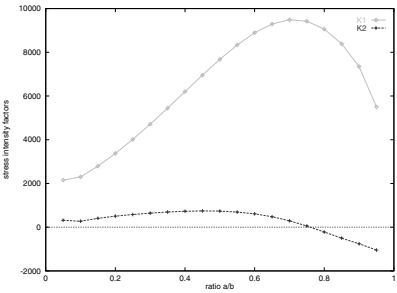


Fig. 2. **a)** interface crack, **b)** real part (K_1) and imaginary part (K_2) of c_1

2.3 3D Vertex and Edge Singularities

Let us assume that $\Omega_i \subset \mathbb{R}^3$ are polyhedrons and that the set $\Gamma_D \cap \Gamma_N$ consists of straight edges. We analyze now the behavior of the potential fields and elastic fields near the set $\mathcal{S}$:

$$\mathcal{S} = \{\text{edges of } \partial\Omega_i\} \cup \{\text{vertices of } \partial\Omega_i\} \ .$$

The *vertex singularities* can be described analogously to the 2D corner singularities taking an ansatz in spherical coordinates:

$$u_{\text{sing}}(r, \varphi, \theta) = r^{\alpha} \Phi_{\alpha}(\varphi, \theta) \ ,$$

where $r = |x - v|$ and v is a vertex as in Fig. 3. We proceed as in the 2D case and get (17), (18), (8) correspondingly and finally an operator pencil

$$\mathcal{A}_v(\alpha) : V(\Omega \cap S_R(v)) \to V(\Omega \cap S_R(v))' \ ,$$

where $S_R(v) = \{x \in \mathbb{R}^3 : |x - v| = R\}$ for sufficiently small fixed R and $V(\Omega \cap S_R(v)) = \{u \in W^{1,2}(\Omega \cap S_R(v)) : u|_{\Gamma_D} = 0\}$. There is a finite number of eigenvalues of the Fredholm operator bundle $\mathcal{A}_v(\alpha)$ in any strip $c_1 \leq \operatorname{Re} \alpha_v \leq c_2$. We call an eigenvalue non-defective, if its algebraic and geometric multiplicities coincide. Assuming this for simplicity we get together with the eigensolutions $\Phi_{\alpha_v}(\varphi, \theta)$ singular vertex functions for weak solutions of the form

$$u_{\text{vertex}} = \sum_{-\frac{1}{2} < \operatorname{Re} \alpha_v < \frac{1}{2}} \eta_v(r) c_{\alpha_v} r^{\alpha_v} \Phi_{\alpha_v}(\varphi, \theta) \ , \tag{27}$$

where η_v is a cut-off function and c_{α_v} are constants which depend on the right hand sides.

The *edge singularities* are generated by 2D operator pencils. For a fixed edge $\mathbf{e}$, we introduce an orthogonal system of coordinates (y_1, y_2, y_3), where

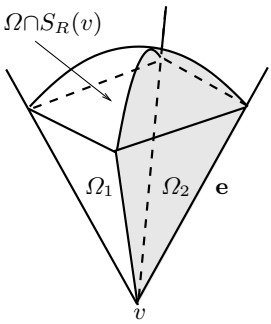


Fig. 3. The vertex neighborhood

the y_3-axis is directed along the edge $\mathbf{e}$. We denote by $K_\mathbf{e}$ the straight plane angular cone of opening $\varphi_\mathbf{e}$ in the $\{y_1, y_2\}$-plane such that Ω coincides with $K_\mathbf{e} \times \mathbb{R}$ in a neighborhood of $\mathbf{e}$. Removing the derivative ∂_{y_3} in the operators L defined by (5) and (6) we get 2D-operators $L_\mathbf{e}$ in $K_\mathbf{e}$. Introducing polar coordinates $(\rho, \varphi_\mathbf{e})$ and applying the Mellin technique we get edge singularities of the weak solution of the following form provided the eigenvalues are non-defective:

$$u_{\text{edge}} = \sum_{0<\operatorname{Re}\alpha_\mathbf{e}<1} c_{\alpha_\mathbf{e}}(y_3)\rho^{\alpha_\mathbf{e}}\Phi_{\alpha_\mathbf{e}}(\varphi_\mathbf{e}) \,. \tag{28}$$

The exponents $\alpha_\mathbf{e}$ are the eigenvalues and $\Phi_{\alpha_\mathbf{e}}(\varphi_\mathbf{e})$ the corresponding eigensolutions of the eigenvalue problem as formulated in (18).

The asymptotic expansion of weak solutions of (13), (14) and (15) reads in a vicinity of a vertex v, provided the eigenvalues are nondefective [2, 13, 37, 45]:

$$\begin{aligned}\eta_v u &= \eta_v\big(u_{\text{vertex}} + \sum_{\text{edges}} u_{\text{edge}}\big) + u_{reg} \\ &= \eta_v \sum_{-\frac{1}{2}<\operatorname{Re}\alpha_v<\frac{1}{2}} c_{\alpha_v} r^{\alpha_v}\Phi_{\alpha_v}(\varphi,\theta) + \eta_v \sum_{\text{edges}} \sum_{0<\operatorname{Re}\alpha_\mathbf{e}<1} \tilde{c}_{\alpha_\mathbf{e}}(y_3, r)\rho^{\alpha_\mathbf{e}}\Phi_{\alpha_\mathbf{e}}(\varphi_\mathbf{e}) \\ &\quad + u_{\text{reg}} \,,\end{aligned} \tag{29}$$

where $u_{\text{reg}}|_{\Omega_i} \in H^{2-\varepsilon}(\Omega_i)$. Here, we accumulated the interaction of vertex and edge singularities in the coefficient $\tilde{c}_{\alpha_\mathbf{e}}(y_3, r)$.

2.4 Examples

Vertex Exponents

We start with the Dirichlet problem for linear elastic fields in a composed Fichera domain with the Young moduli $E_1 = 1, E_2 = 10$ and Poisson ratios

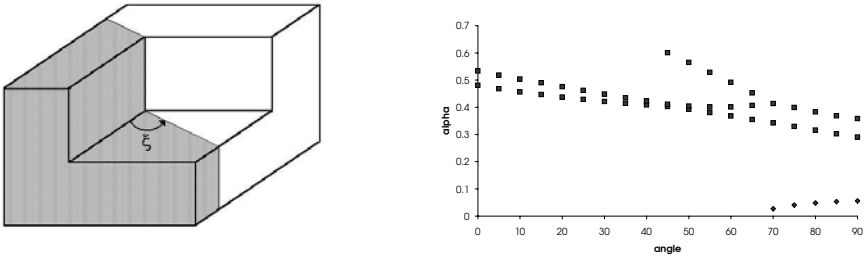


Fig. 4. **a)** Fichera corner, **b)** vertex exponent

$\nu_1 = \nu_2 = 0.3$, see Fig. 4. The plotted squares in Fig. 4 represent the real parts whereas the diamonds show the imaginary parts of the vertex exponents. Figure 4 is from [45], the computations where done by D. Leguillon.

Kellogg's Example

The following two dimensional example by R. B. Kellogg shows that the singular exponents in expansion (22) for solutions of the Laplace equation can have arbitrary small positive real parts.

Let be $\Omega = (-1,1)^2 \subset \mathbb{R}^2$, $\Omega_i = \{x = r\binom{\cos\varphi}{\sin\varphi} \in \Omega : (i-1)\pi/2 < \varphi < i\pi/2\}$ for $1 \le i \le 4$ and we consider the Laplace equation (5) on this domain. In the neighborhood of S, the solution u has the structure (22), the singular exponents α in (24) are real and there are no logarithmic terms. If we choose $\mu_1 = \mu_3 = 1$, $\mu_2 = \mu_4 = h > 0$, then $\alpha > 0$ is a singular exponent of (24) if and only if $\cos(\alpha\pi) = 1 - 8h/(1+h)^2$ [11, 24]. It follows from this relation that $\alpha_{\min} = \min\{\alpha > 0,\ \alpha \text{ is singular exponent}\}$ tends to 0 for $h \to 0$ or $h \to \infty$, see also Fig. 5. Therefore, the regularity of weak solutions can get arbitrarily low, i.e. for general situations it can only be guaranteed that $u\big|_{\Omega_i} \in W^{1+\epsilon,2}(\Omega_i)$, $\epsilon > 0$ small.

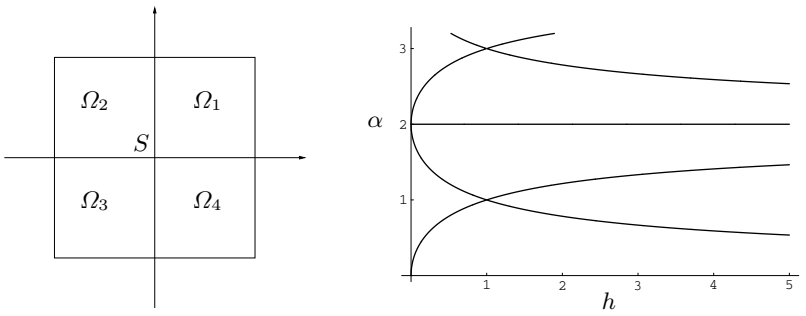


Fig. 5. Domain and eigenvalues α for Kellogg's example

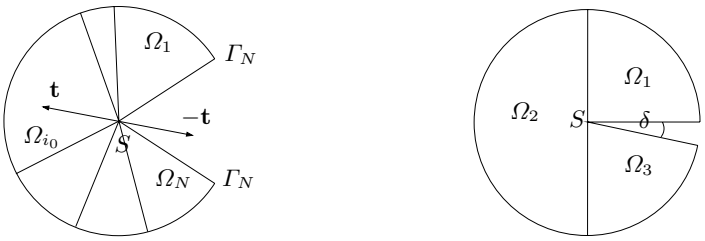


Fig. 6. Example for the quasi-monotonicity condition

Quasi-monotonicity

Different authors investigated in detail the dependence of the singular exponents in (24) and (29) on the geometry, material parameters and the number of subdomains and we refer to [27, 42, 45, 48] for the Laplace equation and to [27, 45] for linear elasticity. M. Petzoldt observed and proved for two dimensional polygons that a *quasi-monotone* distribution of the parameters μ_i in the Laplace equation (5) leads to general positive lower bounds for Re $\alpha_{\min}$ and thus guarantees a higher minimum regularity of weak solutions [48]. In [26, 27, 29] this condition was slightly modified and extended to composites of linear elastic materials in 2D and 3D polyhedral domains and the real parts of the singular exponents were estimated. We give here the definition of quasi-monotonicity from [29] for a cross point S which is situated on the Neumann boundary of a two dimensional polygon and formulate the corresponding regularity result.

Let $R > 0$, $N \in \mathbb{N}$. For $\Phi_0 < \Phi_1 < \ldots < \Phi_N \leq \Phi_0 + 2\pi$ we introduce

$$\Omega_i = \{x \in \mathbb{R}^2 : 0 < |x| < R,\, \Phi_{i-1} < \varphi < \Phi_i\},\ 1 \leq i \leq N$$

and $\Omega = \{x \in \mathbb{R}^2 : 0 < |x| < R,\, \Phi_0 < \varphi < \Phi_N\}$. Let furthermore C^i be the elasticity tensor corresponding to the subdomain Ω_i. It is assumed that the boundaries $\Gamma_i = \{x \in \mathbb{R}^2 : |x| \leq R,\, \varphi = \Phi_i\}$, $i \in \{0, N\}$, are parts of the Neumann boundary. Then the quasi-monotonicity condition is satisfied if there exists an index $i_0 \in \{0, \ldots, N\}$ such that (Fig. 6)

QM1 $C^1 \leq \ldots \leq C^{i_0-1} \leq C^{i_0} \geq C^{i_0+1} \geq \ldots \geq C^N$,
QM2 There exists $\mathbf{t} \in \mathbb{R}^2 \backslash \{0\}$ such that $\mathbf{t} \in \Omega_{i_0}$ and $-\mathbf{t} \notin \overline{\Omega}$.

By $C^i \geq C^j$ we mean that $(C^i A) : A \geq (C^j A) : A$ for every $A \in \mathbb{R}^{2\times 2}_{\text{sym}}$. If the materials are isotropic, then $C^i \varepsilon = \lambda_i \text{tr}\,(\varepsilon) I + 2\mu_i \varepsilon$ for $\varepsilon \in \mathbb{R}^{2\times 2}$ with Lamé constants λ_i, μ_i and **QM1** is equivalent to

$$\mu_1 \leq \ldots \leq \mu_{i_0} \geq \ldots \geq \mu_N\ , \tag{30}$$

$$\lambda_1 + \mu_1 \leq \ldots \leq \lambda_{i_0} + \mu_{i_0} \geq \ldots \geq \lambda_N + \mu_N\ . \tag{31}$$

Theorem 2. *[26, 27, 29, 45] Let $\Omega \subset \mathbb{R}^2$ be as described above and let $u \in W^{1,2}(\Omega)$ satisfy* (13) *with bilinear form* (15) *for every $v \in W^{1,2}(\Omega)$ with $v\big|_{\partial\Omega\cap\partial B_R(0)} = 0$. Assume furthermore that $f \in L^2(\Omega)$ and $g_i = 0$ (for simplicity). Let finally the quasi-monotonicity condition* **QM1**, **QM2** *be satisfied. Then it holds for the exponents α in the asymptotic expansion* (22) *with respect to S:* $\operatorname{Re}\alpha \geq \frac{1}{2}$. *Thus $\eta_S u\big|_{\Omega_i} \in W^{\frac{3}{2}-\epsilon,2}(\Omega_i)$ for every $\epsilon > 0$ and a cut-off function η_S. Moreover, if the materials are isotropic for every i, then* $\operatorname{Re}\alpha > \frac{1}{2}$ *and $\eta_S u\big|_{\Omega_i} \in W^{\frac{3}{2}+\epsilon,2}(\Omega_i)$ for an appropriate $\epsilon > 0$.*

The isotropic case is proved in [26, 27, 45] with a homotopy method. The proof of the general case relies on a difference quotient technique and we go into details in Sect. 3, see also [29, 31]. Analogous results are derived for more general linear elliptic systems in two and three dimensions [29, 31, 45].

As a special application of Theorem 2 we consider an elastic, isotropic bimaterial with a crack perpendicular to the interface. This example is studied in [35]. Let $\Phi_0 = 0$, $\Phi_1 = \frac{\pi}{2}$, $\Phi_2 = \frac{3\pi}{2}$ and $\Phi_4 = 2\pi - \delta$ for small $\delta > 0$. We assume that vanishing Neumann conditions are imposed near the cross point $S = (0,0)$, see Fig. 6. Let Ω_1 and Ω_3 be occupied by zirconia oxide ZrO_2 and Ω_2 by aluminum oxide Al_2O_3. The corresponding material parameters are $\mu_1 = \mu_3 = 0.73$ $[10^5\,\mathrm{N/mm^2}]$, $\lambda_1 = \lambda_3 = 1.096$ $[10^5\,\mathrm{N/mm^2}]$ and $\mu_2 = 1.46$ $[10^5\,\mathrm{N/mm^2}]$, $\lambda_2 = 2.19$ $[10^5\,\mathrm{N/mm^2}]$. The quasi-monotonicity condition **QM1**, **QM2** is satisfied for $\delta > 0$ and thus $\operatorname{Re}\alpha_{\min} \geq \frac{1}{2}$ due to Theorem 2. The numerical calculations in [35] confirm this with $\alpha_{\min} = 0.57$ for $\delta = 0$. If on the other hand the materials are interchanged, i.e. Ω_1 and Ω_3 are occupied by Al_2O_3 and Ω_2 by ZrO_2, then the quasi-monotonicity condition is violated and the calculations from [35] give $\alpha_{\min} = 0.42$ for $\delta = 0$. Following the discussion in [35] it seems that the quasi-monotonicity condition does not only lead to higher regularity results but also describes a class of composites which can sustain higher loads before breaking.

3 Nonlinear Elliptic Systems of *p*-Structure

In this section we consider boundary-transmission problems for quasilinear elliptic equations and systems of p-structure, where the p-Laplace equation is a typical example. We admit that the growth properties of the differential operators vary from subdomain to subdomain. To the authors' knowledge it is an unsolved problem whether the behavior of weak solutions of such nonlinear transmission problems can completely be characterized by asymptotic expansions similar to (22). By a difference quotient method, Savaré [52] and Ebmeyer and Frehse [18, 19] obtained global regularity results for quasilinear elliptic boundary-value problems with smooth coefficients on Lipschitz domains. They assumed that the domains satisfy an additional geometrical condition near those points, where the boundary conditions change. Their results describe a minimum regularity in Sobolev-Slobodeckij spaces for weak solutions on this class of domains.

The main idea for obtaining global regularity results also for transmission problems is to combine the difference quotient technique with the quasi-monotonicity condition, which originally was introduced for linear elliptic transmission problems (Sect. 2.4). We explain here this concept in detail.

3.1 Systems of *p*-Structure

It is assumed that the differential operators under consideration can be derived from convex minimization problems. Let $\overline{\Omega} = \cup_{i=1}^{M} \overline{\Omega_i} \subset \mathbb{R}^d$ and assume that M functions $W_i : \mathbb{R}^{m\times d} \to \mathbb{R}$ are given for some $m \geq 1$. The conditions on the functions W_i are specified later. The boundary-transmission problem reads for $u : \Omega \to \mathbb{R}^m$, $u\big|_{\Omega_i} = u_i$ and given functions f, g:

$$\operatorname{div}\left(DW_i(\nabla u_i)\right) + f_i = 0 \quad \text{in } \Omega_i \,, \tag{32}$$

$$u_i - u_j = 0 \quad \text{on } \Gamma_{ij} \,, \tag{33}$$

$$DW_i(\nabla u_i)\mathbf{n}_i + DW_j(\nabla u_j)\mathbf{n}_j = 0 \quad \text{on } \Gamma_{ij} \,, \tag{34}$$

$$u = 0 \quad \text{on } \Gamma_D \,, \tag{35}$$

$$DW_i(\nabla u_i)\mathbf{n}_i = g \quad \text{on } \Gamma_N \cap \Gamma_i \,. \tag{36}$$

Here, we use the notation $DW_i(A) = \left(\frac{\partial W_i(A)}{\partial A_{kl}}\right)_{kl} \in \mathbb{R}^{m\times d}$ and $A : B = \sum_{k=1}^{m}\sum_{l=1}^{d} A_{kl}B_{kl}$, $|A| = \sqrt{A : A}$ for $A, B \in \mathbb{R}^{m\times d}$. It is assumed that the differential operators (32) are of p-structure. This means that the energy densities W_i satisfy **H1**–**H4** here below for some $p_i \in (1, \infty)$:

H1 $W_i \in \mathcal{C}^1(\mathbb{R}^{m\times d}, \mathbb{R}) \cap \mathcal{C}^2(\mathbb{R}^{m\times d}\backslash\{0\}, \mathbb{R})$.

H2 There exist $c_0^i \in \mathbb{R}$, $c_1^i, c_2^i > 0$ such that for every $A \in \mathbb{R}^{m\times d}$

$$c_0^i + c_1^i \,|A|^{p_i} \leq W_i(A) \leq c_2^i(1 + |A|^{p_i}) \,.$$

H3 There exist $c^i > 0$ such that for every $A \in \mathbb{R}^{m\times d}$:

$$|DW_i(A)| \leq c^i(1 + |A|^{p_i - 1}) \,, \quad \left|D^2W_i(A)\right| \leq c^i \,|A|^{p_i - 2} \,.$$

H4 There exist $c_i > 0$ such that for every $A, B \in \mathbb{R}^{m\times d}$, $A \neq 0$:

$$D^2W_i(A)[B, B] = \sum_{k,j=1}^{m}\sum_{r,s=1}^{d} \frac{\partial^2 W_i(A)}{\partial A_{ks}\partial A_{jr}} B_{ks}B_{jr} \geq c_i \,|A|^{p_i - 2} \,|B|^2 \,.$$

Condition **H4** implies that the functions W_i are strictly convex and that the corresponding differential operators (32) are elliptic. The p-Laplace equation is included here with $W_i(\nabla u) = \frac{1}{p_i} |\nabla u|^{p_i}$ for $u : \Omega \to \mathbb{R}$.

Appropriate function spaces for a weak formulation of (32)–(36) were first introduced and studied by W.B. Liu [39]. Let $\mathbf{p} = (p_1, \ldots, p_M)$ with $p_i \in (1, \infty)$ corresponding to W_i and $p_{\min} = \min\{p_i,\, 1 \leq i \leq M\}$. Then

$$W^{1,\mathbf{p}}(\Omega) = \{u \in W^{1,p_{\min}}(\Omega) : u\big|_{\Omega_i} \in W^{1,p_i}(\Omega_i)\} .$$

Since $W^{1,\mathbf{p}}(\Omega) \subset W^{1,p_{\min}}(\Omega)$, traces are well defined for elements of $W^{1,\mathbf{p}}(\Omega)$ and thus, the following definition is meaningful:

$$V^{\mathbf{p}}(\Omega) = \{u \in W^{1,\mathbf{p}}(\Omega) : u\big|_{\Gamma_D} = 0\} .$$

Assume now for simplicity that the Neumann datum g in (36) vanishes and that $\Gamma_D \neq \emptyset$. The weak formulation to (32)–(36) reads for given $f \in L^{\mathbf{q}}(\Omega)$ with $\mathbf{q} = (q_1, \ldots, q_M)$ and $q_i^{-1} + p_i^{-1} = 1$: Find $u \in V^{\mathbf{p}}(\Omega)$ such that for every $v \in V^{\mathbf{p}}(\Omega)$

$$\sum_{i=1}^{M} \int_{\Omega_i} DW_i(\nabla u) : \nabla v \, \mathrm{d}x = \int_{\Omega} f v \, \mathrm{d}x . \tag{37}$$

It follows from the main theorem on monotone operators that (37) has a unique weak solution $u \in V^{\mathbf{p}}(\Omega)$.

3.2 The Quasi-monotone Covering Condition and Regularity

Kellogg's example in Sect. 2.4 shows that even in the linear case one cannot expect to obtain general minimum regularity results without any further assumptions on the geometry of the subdomains or the distribution of the coefficients. Furthermore, when proving the regularity results with a difference quotient technique, one has to ensure that functions of the form $\eta^2(x)(u(x + he_j) - u(x))$, where η is a cut-off function, $\{e_1, \ldots, e_d\}$ a basis of $\mathbb{R}^d$, $h > 0$ and $u \in W^{1,\mathbf{p}}(\Omega)$, are admissible test functions. In particular, the translated function $\eta^2 u(\cdot + he_j)$ should be an element of $W^{1,\mathbf{p}}(\Omega)$ as well. This cannot be guaranteed for an arbitrary geometry and an arbitrary distribution of the parameters p_i.

Our main assumption on the boundary-transmission problem is that the subdomains Ω_i together with the energy densities W_i satisfy the quasi-monotone covering condition. We formulate here this condition for an interior cross point S and refer to [29, 31] for the general case.

Definition 1. *Let $S \in \Omega \subset \mathbb{R}^d$ be an interior cross point of the subdomains Ω_i, $1 \leq i \leq N$ and $R > 0$ such that $B_R(S) \Subset \Omega$. Let W_i, $1 \leq i \leq N$, be the energy densities corresponding to the subdomains Ω_i.*

The pairs $\{(\Omega_i, W_i), 1 \leq i \leq N\}$ satisfy the quasi-monotonicity condition on $B_R(S)$ if there exist numbers $k_i \in \mathbb{R}$ and an open cone $\mathcal{K} \subset \mathbb{R}^d$ with vertex in 0 such that for every $h \in \mathcal{K}$, $1 \leq i, j \leq N$ and $A \in \mathbb{R}^{m \times d}$ it holds

$$\text{if } \Omega_i + h \cap \Omega_j \neq \emptyset, \text{ then } W_j(A) + k_j \geq W_i(A) + k_i . \tag{38}$$

See Fig. 7 (left) for an example. For interior cross points on two dimensional domains which are composed of polygonal subdomains, this definition can be

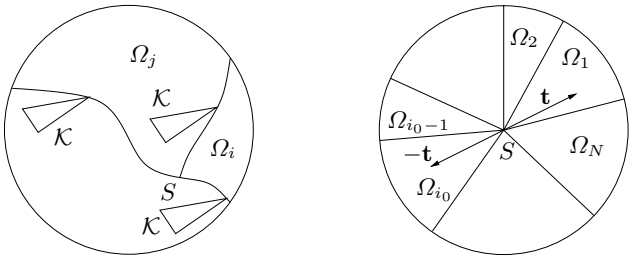


Fig. 7. Examples for the quasi-monotonicity condition

reformulated in a more illustrative way. Let $\Omega = B_R(0) \subset \mathbb{R}^2$ and $\Omega_i = \{x \in \mathbb{R}^2 : \Phi_{i-1} < \varphi < \Phi_i, |x| < R\}$ for $\Phi_0 < \ldots < \Phi_N = \Phi_0 + 2\pi$. The quasi-monotonicity condition is satisfied if there exist an index $i_0 \in \{2, \ldots, N\}$ and numbers $k_i \in \mathbb{R}$ such that it holds (see Fig. 7, right):

QM3 For every $A \in \mathbb{R}^{m\times 2}$ we have

$$W_1(A) + k_1 \geq W_2(A) + k_2 \geq \ldots \geq W_{i_0}(A) + k_{i_0} \leq \\ \ldots \leq W_N(A) + k_N \leq W_1(A) + k_1 \,.$$

QM4 There exists $\mathbf{t} \in \mathbb{R}^2 \backslash \{0\}$ such that $\mathbf{t} \in \Omega_1$ and $-\mathbf{t} \in \Omega_{i_0}$.

Example 1. Let $W_i(A) = \frac{1}{p_i} |A|^{p_i}$ for $A \in \mathbb{R}^d$. Then **QM3** is equivalent to

$$p_1 \geq p_2 \geq \ldots \geq p_{i_0} \leq \ldots \leq p_N \leq p_1 \,.$$

The energy densities $W_i(A) = \frac{\mu_i}{2} |A|^2$, $A \in \mathbb{R}^2$, in Kellogg's example (Sect. 2.4) do not satisfy the quasi-monotonicity condition with respect to $S = (0, 0)$.

Example 2. Let $\Omega, \Omega_1, \Omega_2 \subset \mathbb{R}^d$ be bounded Lipschitz domains with $\Omega_1 \cap \Omega_2 = \emptyset$, $\overline{\Omega} = \overline{\Omega_1} \cup \overline{\Omega_2}$ and $\Omega_2 \Subset \Omega$. Let furthermore $\partial\Omega = \Gamma_D$ or $\partial\Omega = \Gamma_N$ and assume that W_1, W_2 satisfy **H1**–**H4** with $p_1 \neq p_2$. Then the quasi-monotone covering condition is satisfied for the pairs $\{(\Omega_i, W_i),\ i = 1, 2\}$, see Fig. 8. We refer to [29, 31] for further examples.

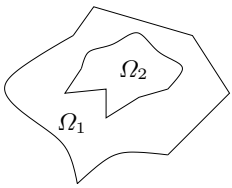


Fig. 8. Nested Lipschitz domains

Theorem 3. *[29, 30, 31] Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain with $\overline{\Omega} = \cup_{i=1}^M \overline{\Omega_i}$. We assume that the functions $W_i : \mathbb{R}^{m\times d} \to \mathbb{R}$, $1 \le i \le M$, satisfy* **H1**–**H4** *for some $p_i \in (1,\infty)$ and the quasi-monotone covering condition. Let finally $f \in L^{\mathbf{q}}(\Omega)$ with $p_i^{-1} + q_i^{-1} = 1$.*

Then the weak solution $u \in V^{\mathbf{p}}(\Omega)$ of (37) and the stress field $\sigma = DW(\nabla u) \in L^{\mathbf{q}}(\Omega)$ have the following regularity: For every $\delta > 0$

$$u\big|_{\Omega_i} \in W^{1+\frac{1}{p_i}-\delta,p_i}(\Omega_i),\ \sigma\big|_{\Omega_i} \in W^{\frac{1}{2}-\delta,r(q_i)}(\Omega_i) \quad \text{if } p_i \in [2,\infty)\,, \tag{39}$$

$$u\big|_{\Omega_i} \in W^{\frac{3}{2}-\delta,r(p_i)}(\Omega_i),\ \ \sigma\big|_{\Omega_i} \in W^{\frac{1}{q_i}-\delta,q_i}(\Omega_i) \quad \text{if } p_i \in (1,2]\,. \tag{40}$$

Here, $r(s) = \frac{2ds}{2d-2+s}$ and for $s \in (1,2]$ it is $s \le r(s) \le 2$.

Non vanishing Dirichlet and Neumann conditions are treated in [29, 31]. The regularity theorem corresponds well with the results of references [18, 19, 52] for pure boundary-value problems. Moreover, the results of Theorem 2 (linear case) are recovered by Theorem 3. Let us remark that Theorem 3 is applicable in the situation described in Example 2.

Remarks on the proof. The proof of Theorem 3 is carried out with a difference quotient technique, where the domain Ω is covered by a finite number of balls and Theorem 3 is proved for each ball separately. We give here a sketch for an interior cross point S for which the quasi-monotonicity condition of Definition 1 is satisfied. For simplicity we assume that $p_i \ge 2$ for every i. For the full proof we refer to [29, 31].

The goal is to show the following estimate for $h \in \mathcal{K}$, where $\mathcal{K}$ is the cone of Definition 1:

$$\sum_{i=1}^N \int_{\Omega_i \cap B_R(S)} |\nabla u(x+h) - \nabla u(x)|^{p_i}\ \mathrm{d}x \le c\,|h|\ , \tag{41}$$

and the constant c is independent of $h \in \mathcal{K}$. Inequality (41) implies that $u\big|_{\Omega_i\cap B_R(S)}$ is an element of the Nikolskii space $\mathcal{N}^{1+\frac{1}{p_i},p_i}(\Omega_i \cap B_R(S))$ [56]. The embedding theorems for Nikolskii and Sobolev-Slobodeckij spaces lead to $u\big|_{\Omega_i\cap B_R(S)} \in W^{1+\frac{1}{p_i}-\delta,p_i}(\Omega_i \cap B_R(S))$ for every $\delta > 0$ [56]. We prove now estimate (41). **H1**–**H4** imply the following convexity inequality (see e.g. [29])

$$\begin{aligned}\sum_{i=1}^N \int_{\Omega_i} c_i\eta^2\,|\nabla u(x+h) - \nabla u(x)|^{p_i}\ \mathrm{d}x \\ \le \sum_{i=1}^N \int_{\Omega_i} \eta^2 DW(\nabla u) : \nabla(u(x+h) - u(x))\,\mathrm{d}x \\ + \sum_{i=1}^N \int_{\Omega_i} \eta^2\,(W_i(\nabla u(x+h) - W_i(\nabla u(x)))\ \mathrm{d}x\,,\end{aligned} \tag{42}$$

where η is a smooth cut-off function with $\eta = 1$ on $B_R(S)$. The quasi-monotonicity condition guarantees that $v(x) = \eta^2(u(x+h) - u(x))$, $h \in \mathcal{K}$, is an an element of $V^{\mathbf{P}}(\Omega)$ and thus is an admissible test function for the weak formulation. Using (37), the first term on the right hand side in (42) can be controlled after some technical calculations by $c\,|h|$. The second term in (42) can be rewritten as follows with $\triangle_h w(x) = w(x+h) - w(x)$ and the numbers k_i from Definition 1:

$$\begin{aligned}\sum_{i=1}^{N} \int_{\Omega_i} \eta^2 \left(W_i(\nabla u(x+h)) + k_i - W_i(\nabla u(x)) - k_i\right) \mathrm{d}x \\ = \sum_{i=1}^{N} \int_{\Omega_i} \triangle_h(\eta^2(W_i(\nabla u) + k_i))\,\mathrm{d}x \\ - \sum_{i=1}^{N} \int_{\Omega_i} (\triangle_h \eta^2)(W_i(\nabla u(x+h)) + k_i)\,\mathrm{d}x\,. \qquad (43)\end{aligned}$$

The quasi-monotonicity condition implies that the first term on the right hand side in (43) is ≤ 0. Since η is smooth, the second term can be estimated by $c\,|h|$. Thus (41) holds and the proof is finished.

Inequalities like convexity inequality (42) are essential for obtaining regularity results via a difference quotient technique. On the basis of inequalities like (42) we proved a global regularity result for a shear thinning fluid of power-law type [28] and extended a local regularity result by Carstensen and Müller [9] for stress fields of not strictly convex energies to a global one [30].

4 Application of the Regularity Results

The derived regularity results and coefficient formulae can be applied in different fields. We discuss here two of them in detail, namely sensitivity analysis for linear elastic fields and the derivation of formulas in fracture mechanics for a nonlinear elastic model of power-law type.

4.1 Sensitivity Analysis

The goal of shape and structure optimization in elasticity is to determine an elastic body or composite which is optimal with respect to objective and constraint functionals. For example, if one wants to avoid plastification the values of the von Mises yield functional should be small enough, or if one wants to avoid crack growth, the energy release rate (or the stress intensity factors) should not exceed their critical values. The influence of the shape or the structure of the domain on the stress behavior has been studied by many authors [21, 53] and the corresponding sensitivity analysis is well developed

for problems in smooth domains. Here, we focus on the sensitivity analysis for linear elastic fields in two-dimensional non smooth domains and study a class of functionals with respect to shape perturbation.

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain. We introduce a family of mappings $\{\Phi_\varepsilon \in [\mathcal{C}^3(\overline{\Omega})]^2,\ \varepsilon \in [0, \varepsilon_0]\}$ which admit Taylor expansions

$$\Phi_\varepsilon(x) = x + \varepsilon\Phi(x) + \varepsilon^2\Phi_R(\varepsilon, x)$$

with $\Phi, \Phi_R \in [\mathcal{C}^3(\overline{\Omega})]^2$. The function $\Phi_R(\varepsilon, x)$ is bounded with respect to ε for every $x \in \Omega$. The perturbations $(\Omega_\varepsilon, \Gamma_\varepsilon, \Gamma_\varepsilon^D, \Gamma_\varepsilon^N)$ of the reference configuration $(\Omega, \Gamma, \Gamma^D, \Gamma^N)$ are defined by

$$\Omega_\varepsilon = \Phi_\varepsilon(\Omega)\ , \quad \Gamma_\varepsilon = \Phi_\varepsilon(\Gamma)\ , \quad \Gamma_\varepsilon^D = \Phi_\varepsilon(\Gamma^D)\ , \quad \Gamma_\varepsilon^N = \Phi_\varepsilon(\Gamma^N)\ .$$

Since $\Phi_\varepsilon \in [\mathcal{C}^3(\overline{\Omega})]^2$ the number of singular points in Ω_ε is independent of ε.

Let be u_ε a solution of (6) in Ω_ε with mixed boundary conditions on $\Gamma_\varepsilon^D, \Gamma_\varepsilon^N$ and corresponding interface conditions on Γ_ε. We consider functionals associated with the elastic fields u_ε and $\sigma(u_\varepsilon)$

$$J(\Omega_\varepsilon) = \int_{\Omega_\varepsilon} F(u_\varepsilon, \sigma(u_\varepsilon))dx_\varepsilon\ , \tag{44}$$

where the function F satisfies for a positive constant c the growth conditions

$$F(p, q) \le a(p)(c + |q|^2),\ \partial_q F(p, q) \le a(p)(c + |q|) \tag{45}$$

for some $a \in \mathcal{C}(\mathbb{R}^2)$ and all $p \in \mathbb{R}^2,\ q \in \mathbb{R}^4$.

Our goal is to derive formulae for the sensitivity of the functional J with respect to the perturbation mapping Φ_ε, i.e. we want to calculate the shape derivative

$$dJ(\Omega, \Phi) = \lim_{\varepsilon \to 0} \frac{J(\Phi_\varepsilon(\Omega)) - J(\Omega)}{\varepsilon} \tag{46}$$

and to express $dJ(\Omega, \Phi)$ as an integral over $\partial\Omega$.

Sensitivity Formulae with Material and Shape Derivatives

We give different formulae for $dJ(\Omega, \Phi)$ with material and shape derivatives. The *material derivatives* of u_ε and σ_ε are defined as

$$\dot{u} := \left.\frac{d(u_\varepsilon \circ \Phi_\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}\ , \tag{47}$$

whereas the *shape derivative* is given as

$$u' := \left.\frac{du_\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} = \dot{u} - Du_0\Phi\ . \tag{48}$$

It is proved in [7] that $\dot{u}$ and u' are well defined. We assume that the transformed force densities $f_\varepsilon \circ \Phi_\varepsilon$ and $g_\varepsilon \circ \Phi_\varepsilon$ depend smoothly on ε

$$\begin{aligned} f_\varepsilon \circ \Phi_\varepsilon &= f_0 + \varepsilon f_1 + \varepsilon^2 f_R(\varepsilon) \,, \\ g_\varepsilon \circ \Phi_\varepsilon &= g_0 + \varepsilon g_1 + \varepsilon^2 g_R(\varepsilon) \,. \end{aligned}$$

Furthermore,

$$\begin{aligned} \mathbf{n}_\varepsilon \circ \Phi_\varepsilon &= \quad \mathbf{n}_0 + \varepsilon \dot{\mathbf{n}} + \varepsilon^2 \mathbf{n}_R(\varepsilon) \,, \\ (u_\varepsilon \circ \Phi_\varepsilon)(x) &= u_0(x) + \varepsilon \dot{u}(x) + O(\varepsilon^2) \,. \end{aligned}$$

Theorem 4. *[7] Let F in (44) be continuously differentiable with respect to all its arguments and satisfy the growth conditions (45). Furthermore, let $\alpha* = \min\{\mathrm{Re}\ \alpha_j \in (0,1)\}$, where α_j is defined by (23). Then*

$$dJ(\Omega, \Phi) = \int_\Omega (\partial_u F(u_0, \sigma_0) \cdot \dot{u} + \partial_\sigma F(u_0, \sigma_0) : \dot{\sigma} + F(u_0, \sigma_0) \mathrm{div}\, \Phi)\ \mathrm{d}x \,.$$

If additionally $\alpha* \geq \frac{1}{2}$, then

$$dJ(\Omega, \Phi) = \int_\Omega (\partial_u F(u_0, \sigma_0) \cdot u' + \partial_\sigma F(u_0, \sigma_0) : \sigma')\, \mathrm{d}x + \int_{\partial\Omega} F(u_0, \sigma_0) \Phi \cdot \mathbf{n}_0 \mathrm{d}s_x \,.$$

If $\alpha* > \frac{1}{2}$, then

$$\begin{aligned} dJ(\Omega, \Phi) = &\int_{\partial\Omega} F(u_0, \sigma_0)(\Phi \cdot \mathbf{n}_0) \mathrm{d}s_x \\ &+ \int_{\Gamma^N} w \cdot ((\Phi \cdot \mathbf{n}_0)(f + \kappa g) - \mathrm{div}_\Gamma((\Phi \cdot \mathbf{n}_0)\sigma_T(u_0)))\, \mathrm{d}s_x \\ &- \int_{\Gamma^D} (\Phi \cdot \mathbf{n}_0)(C\partial_\sigma F(u_0, \sigma_0) - \sigma(w))\mathbf{n}_0 \cdot \partial_n u_0 \mathrm{d}s_x \,. \end{aligned} \tag{49}$$

$\sigma_T(u_0)$ is the tangential component of the stress tensor on $\partial\Omega$, κ is the curvature of $\partial\Omega$, the tangential divergence operator div_Γ is defined by

$$\mathrm{div}_\Gamma\, v = \mathrm{div}\, v - Dv\mathbf{n}_0 \cdot \mathbf{n}_0 \,,$$

and w is the so-called adjoint displacement field, see [7].

Remark 1. If $\alpha* = \frac{1}{2}$ and a homogeneous material is given, then we have to add stress intensity factors to (49), see [7]. For straight propagation of cracks in linear, isotropic, elastic materials with the energy functional

$$J(\Omega_\varepsilon) = \frac{1}{2} \int_{\Omega_\varepsilon} \sigma(u_\varepsilon) : \varepsilon(u_\varepsilon) dx_\varepsilon \tag{50}$$

this yields to the well-known Irwin formula

$$dJ(\Omega, \Phi) = \gamma \sum_{i=1}^{2} K_i(u_0)^2 \,.$$

Here, K_i are the classical stress intensity factors and γ is a material constant.

4.2 Griffith's Fracture Criterion for a Power-Law Model

A special case in sensitivity analysis is the derivation of the formulas for the energy release rate of bodies with pre-existing cracks. We describe here recently derived results for a power-law model.

The Ramberg-Osgood Model and Regularity

The Ramberg-Osgood model is applied to describe materials with low proportionality limit and with strain hardening behavior [10, 47]. The field equations for the displacement field $u : \Omega \to \mathbb{R}^d$ and stress field $\sigma : \Omega \to \mathbb{R}^{d\times d}$ read as follows:

$$\operatorname{div}\sigma + f = 0 \quad \text{in } \Omega\,, \tag{51}$$

$$\varepsilon(u) - A\sigma - \alpha\left|\sigma^D\right|^{q-2}\sigma^D = 0 \quad \text{in } \Omega \tag{52}$$

together with boundary conditions on $\partial\Omega$. Here, $\sigma^D = \sigma - \frac{1}{d}\operatorname{tr}\sigma I$ denotes the deviator of σ, $q \geq 2$ is the strain hardening parameter, A the tensor of elastic compliances ($A^{-1} = C$ with C from (7)) and $\alpha > 0$ a further material constant. The field equations (51)–(52) are closely related to quasilinear elliptic systems of p-structure (Sect. 3). The following global regularity results are derived in [31, 32] with a difference quotient technique for weak solutions on admissible domains ($q \geq 2$, $p^{-1} + q^{-1} = 1$):

$$u \in W^{\frac{3}{2}-\delta,\frac{2dp}{2d-2+p}}(\Omega)\,, \quad \sigma,\, \operatorname{div} u \in W^{\frac{1}{2}-\delta,2}(\Omega) \cap W^{\frac{1}{q}-\delta,2}(\Omega) \tag{53}$$

for every $\delta > 0$. We call a domain admissible if either **A1** or **A2** here below is satisfied:

A1 $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain and $\overline{\Gamma_D} \cap \overline{\Gamma_N} = \emptyset$.

A2 $\Omega \subset \mathbb{R}^d$ is a Lipschitz polyhedron where at most d faces intersect near points $S \in \overline{\Gamma_D} \cap \overline{\Gamma_N}$. Furthermore, the interior opening angle between Γ_D and Γ_N is less than π, see Fig. 9 and [17].

A slightly more general definition of admissible domains is given in [32]. Local regularity results are proved by Bensoussan and Frehse in [3].

As in the case of quasilinear elliptic systems of p-structure, it is also an unsolved problem for the Ramberg-Osgood model, whether the behavior of weak solutions near re-entrant corners, edges or crack tips can be completely characterized by asymptotic expansions. A comparison between singularities obtained with ansatzes of the form $u(r,\varphi) = r^\alpha v(\varphi)$ and between the regularity results (53) shows good agreement:

Let $\Omega_{2\pi} = \{x \in \mathbb{R}^2 : -\pi < \varphi < \pi,\ |x| < R\}$ for $R > 0$, and assume that $\Gamma_N \supset \{x : \varphi = \pm\pi\}$, i.e. $\Omega_{2\pi} \subset \mathbb{R}^2$ is a domain with a crack on the negative x_1-axis. First investigations on crack tip singularities for Ramberg-Osgood materials were done by Hutchinson [22] and Rice and Rosengren [50]. Based

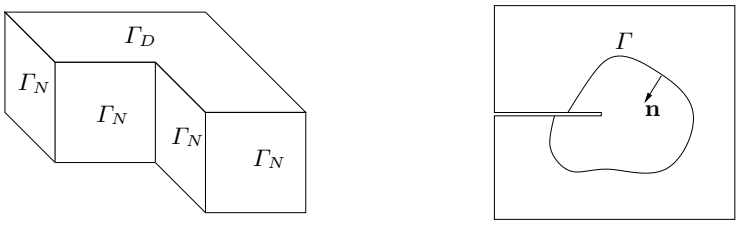


Fig. 9. Examples for an admissible domain and a cracked domain

on the assumption that the displacement and stress fields have an asymptotic structure like in the linear case (22), they derived a strongly nonlinear eigenvalue problem from which they calculated the leading terms in the asymptotic expansion. In particular, they obtained

$$u(r,\varphi) = r^{\frac{1}{q}} v(\varphi) + u_{\mathrm{reg}}(r,\varphi)\ , \quad \sigma(r,\varphi) = r^{-\frac{1}{q}} \tau(\varphi) + \sigma_{\mathrm{reg}}(r,\varphi) \tag{54}$$

near the crack tip. The singular terms are called HRR-fields. Relation (54) fits well with the regularity results (53) since the function $v(x) = |x|^\gamma$ is an element of $W^{\frac{3}{2}-\delta,\frac{4p}{2+p}}(\Omega)$, which is the space from (53) for $d = 2$, if and only if $\gamma \geq \frac{1}{q}$. Furthermore, numerical investigations (see e.g. [61]) show the dependence of the exponents in (54) on the opening angle of the domain and indicate that the singular behavior of weak solutions is completely characterized by asymptotic expansions as in the linear case. But to our knowledge there is no rigorous proof of this conjecture and therefore, we do not use relation (54) for the derivation of formulas for the energy release rate.

Griffith's Fracture Criterion and Energy Release Rate

Griffith's fracture criterion is an energetic criterion and reads as follows for a domain Ω_0 with pre-existing crack C_0 and loading F [38]:

The crack C_0 is stationary with respect to the applied loading F if the total potential energy Π of the body in the actual configuration is minimal compared to every admissible neighboring configuration.

We consider here the simplest case and assume that plane strain conditions hold, that the crack is part of a straight line and that the crack can grow straight on, only. Admissible neighboring configurations are characterized as follows: For $\delta \in \mathbb{R}$ let $S_\delta = \{x \in \mathbb{R}^2 : x_1 \leq \delta, x_2 = 0\}$ and let $\tilde{\Omega} \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary such that $\delta \binom{1}{0} \in \tilde{\Omega}$ for $|\delta| < \delta_0$. For $|\delta| < \delta_0$ we define $\Omega_\delta = \tilde{\Omega} \backslash S_\delta$, $C_\delta = \tilde{\Omega} \cap S_\delta$ and call Ω_0 actual configuration with crack C_0. The domains Ω_δ, $\delta > 0$, are admissible neighboring configurations with cracks C_δ.

The total potential energy Π has the following form for a displacement field u and external forces F:

$$\Pi(\Omega_\delta) = I_{\mathrm{el}}(\Omega_\delta, u) - W(\Omega_\delta, u, F) + D(\Omega_\delta)\ . \tag{55}$$

Here, I_{el} denotes the elastic strain energy, W the work of the external forces and D is a dissipative energy which we assume to be proportional to the crack length: $D(\Omega_\delta) = D(\Omega_0) + 2\gamma\delta$. The constant $\gamma > 0$ is the specific surface energy or fracture toughness and depends on the material. The elastic strain energy of the Ramberg-Osgood model reads for a displacement field u_δ, a corresponding stress field σ_δ with $\varepsilon(u_\delta) = A\sigma_\delta + \alpha \left|\sigma_\delta^D\right|^{q-2} \sigma_\delta^D = D_A W_c(\sigma_\delta)$ and complementary energy density $W_c(\sigma_\delta) = \frac{1}{2} A\sigma_\delta : \sigma_\delta + \frac{\alpha}{q} \left|\sigma_\delta^D\right|^q$:

$$I_{\mathrm{el}}(\Omega_\delta, u_\delta) = \int_{\Omega_\delta} \sigma_\delta : \varepsilon(u_\delta) - W_c(\sigma_\delta)\,\mathrm{d}x\ . \tag{56}$$

Let $f \in L^q(\tilde{\Omega})$, $g \in \left(W^{\frac{1}{q},q}(\Gamma_N)\right)'$ and $F = (f, g)$. Then $W(\Omega_\delta, u_\delta, F) = \int_{\Omega_\delta} f u_\delta\,\mathrm{d}x + \langle u_\delta, g\rangle_{\Gamma_N}$. The quantity $E(\Omega_\delta, u_\delta, F) = I_{\mathrm{el}}(\Omega_\delta, u_\delta) - W(\Omega_\delta, u_\delta, F)$ describes the potential deformation energy. Let $(u_\delta, \sigma_\delta)$ be a weak solution of (51)–(52). Griffith's fracture criterion takes now the form: If

$$E(\Omega_0, u_0, F) - E(\Omega_\delta, u_\delta, F) < D(\Omega_\delta) - D(\Omega_0) = 2\gamma\delta \tag{57}$$

for every small $\delta > 0$, then the crack C_0 is stationary. This motivates the following definition, which is a special case of (46) with $J = -E$:

Definition 2 (Energy release rate). *For $\delta \geq 0$ let u_δ, σ_δ be a weak solution of* (51)–(52). *The energy release rate, shortly ERR, for the domain Ω_0 with crack C_0 and exterior forces F is defined as*

$$ERR(\Omega_0, F) = \lim_{\delta \searrow 0} \frac{1}{\delta}(E(\Omega_0, u_0, F) - E(\Omega_\delta, u_\delta, F))\ .$$

With this definition the fracture criterion reads:

If $ERR(\Omega_0, F) < 2\gamma$, then the crack is stationary, otherwise it will grow.

The question is whether the energy release rate is well defined and whether there exist formulas for calculating this quantity. In the case of linear elastic materials such formulas (Griffith's formula, J-integral, formulas based on the stress intensity factors) are rigorously proved in [14, 25, 40]. For nonlinear elastic models these formulas were derived in the literature under the assumption that the elastic fields u_0, σ_0 are smooth enough or that they can be characterized by certain asymptotic expansions near the crack tip. However, such regularity results are not known in general. Using the regularity results for the Ramberg-Osgood model from Sect. 4.2 we proved the following theorem:

Theorem 5. *[33, 31] Let $\theta \in \mathcal{C}_0^\infty(\tilde{\Omega})$ with $\theta = 1$ in a neighborhood of the crack tip. Let furthermore Γ be a non-intersecting, Lipschitz-continuous path around the crack tip with normal vector $\mathbf{n} = (n_1, n_2)^\top$ pointing towards the*

crack tip. Let finally $f \in \mathcal{C}^1(\overline{\tilde{\Omega}})$ *with* $\frac{\partial}{\partial x_1} f = 0$ *in a neighborhood of the crack tip and* Γ *(see Fig. 9). Then the energy release rate is well defined for the Ramberg-Osgood model and the* Griffith formula *holds:*

$$ERR(\Omega_0, F) = \int_{\Omega_0} \sigma_0 : (\partial_1 u_0 \otimes \nabla\theta)_{sym} \,\mathrm{d}x + \int_{\Omega_0} u_0 \cdot \partial_1(\theta f) \,\mathrm{d}x - \int_{\Omega_0} \big(\sigma_0 : \varepsilon(u_0) - W_c(\sigma_0)\big)\partial_1\theta \,\mathrm{d}x \,. \quad (58)$$

Furthermore, after integration by parts,

$$ERR(\Omega_0, F) = \int_\Gamma (\sigma_0 \mathbf{n}) \cdot \partial_1 u_0 \mathrm{d}s - \int_\Gamma \big(\sigma_0 : \varepsilon(u_0) - W_c(\sigma_0)\big) n_1 \mathrm{d}s + \int_\Gamma u \cdot f n_1 \mathrm{d}s \,. \quad (59)$$

This path integral is called J-integral. *The integrands of* (58) *and* (59) *are* L^1*-functions and* (58)–(59) *are independent of* θ *and of the path* Γ.

The formulas for the energy release rate have the same structure as in the linear case. Moreover, the proof of Theorem 5 runs parallel to the linear case and is based on the mapping $T_\delta(x) = x - \delta \binom{\theta(x)}{0}$, which is a diffeomorphism from the domain Ω_δ to Ω_0. The J-integral is meaningful due to the regularity results in (53). In a recent paper we extended Theorem 5 to geometrically nonlinear elastic models with polyconvex energy densities [34], results for dynamical crack propagation (linear case) are proved in [46].

Numerical Examples for the Energy Release Rate

The following example is studied in [54]. Let $\Omega_0 = (-5,5)^2 \backslash S_0$ be a compound of two materials with an interface crack and energy densities corresponding to modified p_i-Laplace operators, i.e. $W_i(A) = p_i^{-1}(\kappa_i + |A|^2)^{\frac{p_i}{2}}$ for $A \in \mathbb{R}^2$, $i = 1, 2$. This example can be interpreted as an anti-plane case of the Ramberg-Osgood model. The same notation as in the previous section is used here, see also Fig. 2. The field equations for $u : \Omega_0 \to \mathbb{R}$ read

$$\operatorname{div} DW_i(\nabla u_i) = 0 \quad \text{in } \Omega_i \,, \qquad DW_i(\nabla u_i)\mathbf{n}_i = 0 \quad \text{on } \Gamma_N \cup C_0$$

together with the Dirichlet conditions

$$u(x) = \begin{cases} -2x_1 + x_2 + 15 & \text{if } x_2 = -5,\ x_1 \in (-5,5) \,, \\ 0 & \text{if } x_1 = 5 \,, \\ 2x_1 + x_2 - 15 & \text{if } x_2 = 5,\ x_1 \in (-5,5) \,. \end{cases}$$

The energy release rate can be expressed by the Griffith formula

$$ERR(\Omega_0) = \sum_{i=1}^{2} \int_{\Omega_i} \partial_{x_1} u_i DW_i(\nabla u_i) \cdot \nabla\theta - W_i(\nabla u_i)\partial_{x_1}\theta \,\mathrm{d}x \,,$$

where θ is a cut-off function centered at the crack tip. Figure 10 shows the energy release rate for $\kappa_1 = \kappa_2 = 10^{-7}$, different lengths of the crack C_0 and varying parameters μ_i and p_i.

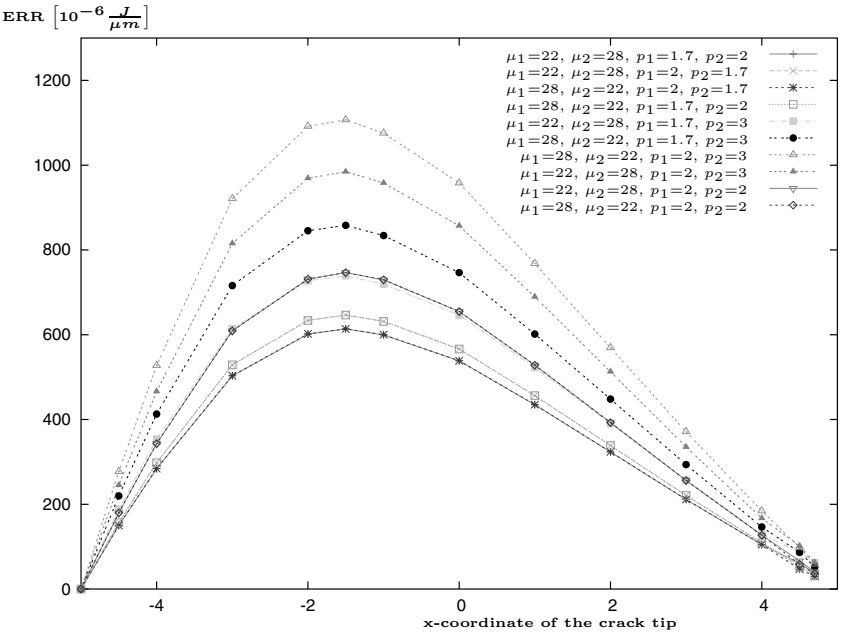


Fig. 10. Energy release rate versus crack length, μ_1, μ_2, p_1, p_2

5 Conclusions

High mechanical stresses can occur in linear and nonlinear elastic composites in the vicinity of re-entrant corners, edges, cracks and near interior surfaces, where the different materials of the composite come together. These stress concentrations have a strong influence on the strength and physical life of the structure. Their knowledge is fundamental for fracture and failure criteria.

In the linear case the substructures consist of linear elastic materials and the singular behavior of the displacement and stress fields can be completely characterized by means of an asymptotic expansion of the solution near the mentioned geometrical and structural peculiarities. Detailed formulas are derived. For some classes of semilinear and quasilinear systems of partial differential equations (e.g. stationary Navier-Stokes equations, semiconductor equations) it can be shown that the regularity of the solutions is dominated by assigned linearized problems.

Nonlinear elastic field equations of power-law type do not fit in this framework in general and it is an open question whether the behavior of the elastic

fields can be completely characterized by leading terms in an asymptotic expansion similar to the linear case. In order to obtain global regularity results for nonlinear elastic field equations of power-law type a combination of the difference quotient technique with the concept of a quasi-monotone distribution of material parameters was used. This leads to the new concept of a quasi-monotone covering condition for the subdomains and the energy densities which determine the differential operators on the subdomains.

The regularity results for linear and nonlinear elastic problems can be applied in shape and structure optimization problems in nonsmooth domains, sensitivity analysis for compound elastic structures and the Griffith fracture criterion for a nonlinear elastic model of power-law type. Relying on the proved regularity results formulae for shape derivatives and the energy release rate are derived. Numerical experiments show their relevance for computations.

References

1. F. Ali Mehmeti, M. Bochniak, S. Nicaise, and A.-M. Sändig. Quasilinear elliptic systems of second order in domains with corners and edges: Nemytskij operator, local existence and asymptotic behaviour. *Zeitschrift für Analysis und ihre Anwendungen*, 21:57–90, 2002.
2. T. Apel, A.-M. Sändig, and S. Solov'ev. Computation of 3d vertex singularities for linear elasticity: Error estimates for a finite element method on graded meshes. *Mathematical Modelling and Numerical Analysis*, 36:1034–1070, 2002.
3. A. Bensoussan and J. Frehse. Asymptotic behaviour of Norton-Hoff's law in plasticity theory and H^1 regularity. In J.-L. Lions and C. Baiocchi, editors, *Boundary Value Problems for Partial Differential equations and Applications*, number 29 in Research Notes in Applied Mathematics, pages 3–26. Masson, Paris, 1993.
4. M. Bochniak. Analytische und numerische Behandlung von Spannungssingularitäten in elastischen Strukturen. Dissertation, Universität Stuttgart, 2000.
5. M. Bochniak and A.-M. Sändig. Computation of generalized stress intensity factors for bonded elastic structures. *Mathematical Modeling and Numerical Analysis*, 33:853–878, 1999.
6. M. Bochniak and A.-M. Sändig. Local solvability and regularity results for a class of semilinear elliptic problems in nonsmooth domains. *Mathematica Bohemica*, 124:245–254, 1999.
7. M. Bochniak and A.-M. Sändig. Sensitivity analysis of 2d elastic structures in presence of stress singularities. *Archives of Mechanics*, 51:275–291, 1999.
8. M. Borsuk and M. Dobrowolski. On the behavior of solutions of the Dirichlet problem for a class of degenerate elliptic equations in the neighborhood of conical boundary points. *Nonlinear Boundary Value Problems*, 9:29–34, 1999.
9. C. Carstensen and S. Müller. Local stress regularity in scalar nonconvex variational problems. *SIAM Journal on Mathematical Analysis*, 34(2):495–509, 2002.
10. J.-L. Chaboche and J. Lemaitre. *Mechanics of solid materials*. Cambridge University Press, 1990.
11. M. Costabel, M. Dauge, and S. Nicaise. Singularities of Maxwell interface problems. *Mathematical Modeling and Numerical Analysis*, 33:627–649, 1999.

12. M. Dauge. *Elliptic Boundary Value Problems on Corner Domains. Smoothness and Asymptotic Expansion*, volume 1341 of *Lecture Notes in Mathematics*. Springer Verlag, 1988.
13. M. Dauge. Singularities of corner problems and problems of corner singularities. *ESAIM Proceedings*, 6:19–40, 1998.
14. P. Destuynder and M. Djaoua. Sur une interpretation mathématique de l'intégrale de Rice en théorie de la rupture fragile. *Mathematical Methods in the Applied Sciences*, 3:70–87, 1981.
15. M. Dobrowolski. On quasilinear elliptic equations in domains with conical boundary points. *Journal für die reine und angewandte Mathematik*, 394:186–195, 1989.
16. M. Dryja, M. V. Sarkis, and O. B. Widlund. Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions. *Numerische Mathematik*, 72:313–348, 1996.
17. C. Ebmeyer. Steady flow of fluids with shear-dependent viscosity under mixed boundary value conditions in polyhedral domains. *Mathematical Models and Methods in Applied Sciences*, 10(5):629–650, 2000.
18. C. Ebmeyer. Nonlinear elliptic problems with p-structure under mixed boundary value conditions in polyhedral domains. *Advances in Differential Equations*, 6(7):873–895, 2001.
19. C. Ebmeyer and J. Frehse. Mixed boundary value problems for nonlinear elliptic equations in multidimensional non-smooth domains. *Mathematische Nachrichten*, 203:47–74, 1999.
20. P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman Press, Bath, Avon, 1985.
21. E. J. Haug, K. K. Choi, and V. Komkov. *Design Sensitivity analysis of structural systems*. Academic Press Inc., 1986.
22. J. W. Hutchinson. Singular behaviour at the end of a tensile crack in a hardening material. *Journal of the Mechanics and Physics of Solids*, 16:13–31, 1968.
23. G. R. Irwin. Analysis of stresses and strains near the end of a crack traversing a plate. *Journal of Applied Mechanics*, 24:361–364, 1957.
24. R. B. Kellogg. On the Poisson equation with intersecting interfaces. *Applicable Analysis*, 4:101–129, 1975.
25. A. M. Khludnev and J. Sokolowski. Griffith formulae for elasticity systems with unilateral conditions in domains with cracks. *European Journal of Mechanics and Solids*, 19:105–119, 2000.
26. D. Knees. Regularitätsaussagen für zweidimensionale elastische Felder in Kompositen. Diplomarbeit, Universität Stuttgart, 2001.
27. D. Knees. Regularity results for transmission problems for the Laplace and Lamé operators on polygonal or polyhedral domains. Bericht SFB404 2002/10, Universität Stuttgart, 2002.
28. D. Knees. On the regularity of weak solutions of a shear thinning fluid of power-law type. Bericht SFB 404 2003/23, Universität Stuttgart, 2003.
29. D. Knees. On the regularity of weak solutions of nonlinear transmission problems on polyhedral domains. *Zeitschrift für Analysis und ihre Anwendungen*, 23(3):509–546, 2004.
30. D. Knees. Global stress regularity for convex and some nonconvex variational problems on nonsmooth domains. Bericht SFB 404 2005/04, Universität Stuttgart, 2005. (submitted).

31. D. Knees. *Regularity results for quasilinear elliptic systems of power-law growth in nonsmooth domains: boundary, transmission and crack problems*. PhD thesis, Universität Stuttgart, 2005. http://elib.uni-stuttgart.de/opus/volltexte/2005/2191/.
32. D. Knees. Global regularity of the elastic fields of a power-law model on Lipschitz domains. *Mathematical Methods in the Applied Sciences*, Published Online: 15 Feb 2006, DOI: 10.1002/mma.727, 2006.
33. D. Knees. Griffith-formula and J-integral for a crack in a power-law hardening material. *Mathematical Models and Methods in Applied Sciences*, 2006. accepted.
34. D. Knees and A. Mielke. Energy release rate for cracks in finite-strain elasticity. WIAS Preprint No. 1100, Weierstrass Institute for Applied Analysis and Stochastics, Berlin, 2006. (submitted).
35. Z. Knésl, L. Náhlík, and J. Radon. Influence of interface on fatigue threshold values in elastic bimaterials. *Computational Materials Science*, 28:620–627, 2003.
36. V. A. Kondrat'ev. Boundary value problems for elliptic equations in domains with conical or angular points. *Transactions of the Moscow Mathematical Society*, 10:227–313, 1967.
37. V. A. Kozlov, V. G. Maz'ya, and J. Rossmann. *Elliptic Boundary Value Problems in Domains with Point Singularities*. American Mathematical Society, 1997.
38. H. Liebowitz, editor. *Fracture*, volume 1-6. Academic Press, New York, 1968.
39. W.-B. Liu. Degenerate quasilinear elliptic equations arising from bimaterial problems in elastic-plastic mechanics. *Nonlinear Analysis, Theory, Methods & Applications*, 35:517–529, 1999.
40. V. G. Maz'ya and S. A. Nazarov. Asymptotics of energy integrals under small perturbations of the boundary close to angular and conic points. *Trudy Moskovskogo Matematicheskogo Obshchestva*, 50:79–129, 1987.
41. V. G. Maz'ya and B. A. Plamenevsky. On the coefficients in the asymptotics of solutions of elliptic boundary value problems in domains with conical points. *Mathematische Nachrichten*, 76:29–60, 1977.
42. D. Mercier. Minimal regularity of the solutions of some transmission problems. *Mathematical Methods in the Applied Sciences*, 26:321–348, 2003.
43. S. Nicaise and A.-M. Sändig. General interface problems, part I. *Mathematical Methods in the Applied Sciences*, 17:395–430, 1994.
44. S. Nicaise and A.-M. Sändig. General interface problems, part II. *Mathematical Methods in the Applied Sciences*, 17:431–450, 1994.
45. S. Nicaise and A.-M. Sändig. Transmission problems for the Laplace and elasticity operators: Regularity and boundary integral formulation. *Mathematical Methods in the Applied Sciences*, 9:855–898, 1999.
46. S. Nicaise and A.-M. Sändig. Dynamical crack propagation in a 2d elastic body. The out-of plane state. Preprint 2005/001, Institut für Angewandte Analysis und Numerische Simulation, Universität Stuttgart, 2005. (submitted).
47. W. R. Osgood and W. Ramberg. Description of stress-strain curves by three parameters. NACA Technical Note 902, National Bureau of Standards, Washington, 1943.
48. M. Petzoldt. Regularity results for Laplace interface problems in two dimensions. *Zeitschrift für Analysis und ihre Anwendungen*, 20(2):431–455, 2001.

49. L. Recke. Applications of the implicit function theorem to quasilinear elliptic boundary value problems with non–smooth data. *Comm. Part. Diff. Equat.*, 20:1457–1479, 1995.
50. J. R. Rice and G. F. Rosengren. Plane strain deformation near a crack tip in a power-law hardening material. *Journal of the Mechanics and Physics of Solids*, 16:1–12, 1968.
51. A.-M. Sändig. The Shapiro-Lopatinskij condition for boundary and transmission problems for the Lamé system. Bericht 98/14, SFB 404, Universität Stuttgart, 1998.
52. G. Savaré. Regularity results for elliptic equations in Lipschitz domains. *Journal of Functional Analysis*, 152:176–201, 1998.
53. J. Sokolowski and J. P. Zolesio. *Introduction to Shape Optimization.* Springer Verlag, 1992.
54. M. Thomas. Mode-III-Interface-Bruch in einem Verbund aus nichtlinearen Materialien. Diplomarbeit, Universität Stuttgart, 2006.
55. P. Tolksdorf. On the Dirichlet problem for quasilinear equations in domains with conical boundary points. *Communications in Partial Differential Equations*, 8:773–817, 1983.
56. H. Triebel. *Spaces of Besov-Hardy-Sobolev type.* Teubner-Texte zur Mathematik. Teubner, 1978.
57. H. M. Westergaard. Bearing pressures and cracks. *Journal of Applied Mechanics*, 6:A49–A53, 1939.
58. K. Wieghardt. Über das Spalten und Zerreißen elastischer Körper. Auf Grund eines Ansatzes von A. Sommerfeld. *Zeitschrift für Mathematik und Physik*, 55:60–103, 1907.
59. M. Williams. Stress singularities resulting from various boundary conditions in angular corners of plates in extension. *Journal of Applied Mechanics*, 74:526–528, 1952.
60. M. Williams. The stresses around a fault or crack in dissimilar media. *Bull.Seismol.Soc.Amer.*, 49:199–204, 1959.
61. H. Yuan and G. Lin. Analysis of elastoplastic sharp notches. *International Journal of Fracture*, 67:187–216, 1994.

Multilevel FEM for Heterogeneous Structures: From Homogenization to Multigrid Solvers*

Christian Miehe[1] and Claus G. Bayreuther[2]

[1] Institute of Applied Mechanics, Pfaffenwaldring 7, 70569 Stuttgart, Germany
cm@mechbau.uni-stuttgart.de
[2] P+Z Engineering GmbH, Anton-Ditt-Bogen 3, 80939 München, Germany
c.bayreuther@puz.de

Summary. Heterogeneous structures like composites often need a fine-scale resolution of micro-effects which influence the macroscopic overall response. This is of particular relevance in the fully nonlinear range of large strains and inelastic material response of the constituents. Suitable solution methods introduce a multifield scenario of hierarchically superimposed states on different length scales. For big differences of micro- and macro-scales, the argument of *scale separation* induces the application of homogenization methods. Such types of *physical multiscale approaches* can be treated by nested multilevel finite element analyses that discretize both the fine-scale micro-structure as well as the macroscopic boundary-value problem. In contrast, small scale differences require *full resolution* of the heterogeneous structure. Effective solution methods for the resulting large-scale problems with strongly oscillating properties are suitably designed geometric multigrid techniques, which may be considered as *numerical multiscale approaches*. In both scenarios, a key ingredient is the suitable formulation of *scale bridging algorithms* that govern the transfer between different scales. The paper outlines new mesh-bridging techniques in a deformation-driven context for fully nonlinear response, which exploit in a nontrivial manner weak constraints on the average deformation in typical finite element patches. The framework is based on an incremental variational structure of finite inelasticity. The proposed new formulations provide variational-based *homogenization algorithms* for physical multiscale scenarios and problem-dependent *optimal finite element grid transfers* for numerical multiscale scenarios of heterogeneous materials.

Keywords: Heterogeneous materials, homogenization, multigrid methods, micro-structures, energy minimization

1 Introduction

The paper investigates formulations for the numerical analysis of inelastic composites *with* or *without* scale separation. For problems with scale sep-

* Research Project C6 "Multiscale Models and Multigrid Methods"

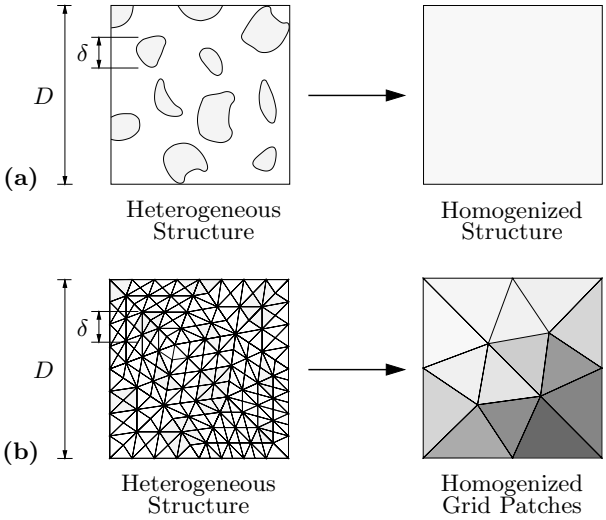


Fig. 1. Homogenization and multi-grid applied to heterogeneous structures. **(a)** Large scale differences $\delta \ll D$ allow scale separation and definition of a homogenized macro-response. **(b)** Small scale differences $\delta \leq D$ need full resolution and efficient multi-grid FE solvers with homogenization-based grid transfer on FE patches

aration, we outline distinct theoretical and computational formulations of strain-driven homogenization for representative composite aggregates with an emphasis on the numerical implementation of boundary constraints of the micro-structure. For problems without scale separation, we develop a specific multigrid solution procedure for composites based on homogenization-based transfer operators. It is shown that the developments on the *phisical multiscale approaches* for problems with scale separation inspire the developments on *numerical multiscale approaches* for problems without scale separation. Both formulations are outlined for the geometrically and phisically nonlinear range in a compact format based on recently developed incremental variational formulations.

1.1 Homogenization: Physical Multiscale with Scale Separation

A heterogeneous material with coherent material constituents is schematically depicted in Fig. 1a. While the macroscopic problem is defined on a scale with characteristic length D, the micro-structural constituents are of size δ. The argument of scale separation can be applied if the ratio of both scales is very small

$$R := \delta / D \ll 1 \, . \tag{1}$$

Typical technical applications of this category cover micro-heterogeneous materials such as metal-matrix composites. For such a scenario, the macroscopic

overall response can be defined in a *physically averaged or homogenized sense* as indicated in Fig. 1a by so-called micro-to-macro transitions, which are important ingredients of continuum micromechanics. Comprehensive reviews of overall properties of heterogeneous materials are provided by the articles [22, 46, 21, 41, 33, 37], see also the mathematical literature on asymptotic expansion methods [39, 7] and the references therein. Starting with the works of Suquet [41], in recent years several numerical methods have been developed which discretize fine-scale fields on representative micro-structures. Numerical formulations for the analysis of representative cells of composites have been published in [16, 31, 40, 24, 42, 23, 26, 27, 30] based on the discretization of composite micro-structures in terms of finite element methods or fast Fourier transforms

In this work, we develop computational aspects of homogenization-based micro-to-macro transitions which supplement the works cited above. This concerns in a first step the introduction of a variational-based approach to the definition of the overall properties of micro-structures in the fully nonlinear range of inelastic materials. Here, we apply a principle of *minimum averaged incremental strain energy* proposed in [26] and [30]. It extends the classical concept in the homogenization theory of finite elasticity, used for example in [32] and [37] and denoted as the average free energy variational principle, to the range of *standard dissipative materials*. Nonlinear standard materials cover a broad spectrum of constitutive descriptions for the constituents of a micro-structure, such as models of finite elasticity, viscoelasticity, plasticity or damage mechanics. In a discretized finite element setting of a composite micro-structure $\mathcal{B}$ as shown in Fig. 1a, we minimize the averaged incremental energy

$$\bar{W} = \inf_{\mathbf{d}} \frac{1}{|\mathcal{V}|} \int_{\mathcal{B}} W(\mathbf{1} + \mathbf{B}\mathbf{d}; \mathbf{X}) \, \mathrm{d}V \ , \tag{2}$$

with respect to the nodal displacement $\mathbf{d} = \bar{\mathbf{d}} + \tilde{\mathbf{d}}$, which decomposes into a homogeneous part $\bar{\mathbf{d}}$ governed by the prescribed macro-deformation $\bar{\mathbf{F}}$ and a fluctuation part $\tilde{\mathbf{d}}$. $\bar{W}$ serves as a potential for the overall stresses of the micro-structure. Based on (2), we construct distinct families of algorithms and matrix representations of *overall stresses and tangent moduli* for micro-structures of heterogeneous materials undergoing inelastic large-strain deformations. The basic underlying approach is the partitioning of the displacement degrees $\mathbf{d}$ into those of the interior domain and those associated with the boundary of the micro-structure as visualized in Fig. 2. It is then shown that the overall stresses and tangent moduli of a typical micro-structure may exclusively be defined in terms of discrete forces and stiffness properties associated with the displacement degrees on the boundary. We focus on deformation-driven micro-structures where the overall macroscopic deformation $\bar{\mathbf{F}}$ is controlled as indicated in Fig. 2. In this context, three classical types of boundary conditions are investigated. Linear deformation and uniform tractions on the boundary provide upper and lower bounds to the micro-structural stiffness.

Furthermore, we consider periodic deformations on the boundary, yielding exact results for unit-cells of periodic materials. The alternative boundary constraints are implemented based on a *penalty extension* of the incremental energy functional (2), providing a *algorithmic class of computational homogenization* in the nonlinear range. The algorithm proposed can be applied to a broad spectrum of model problems. It provides a basis for the computational scale bridging of scale-separated hierarchical micro-structures in multiscale composites.

1.2 Multigrid: Numerical Multiscale Without Scale Separation

The second part of this work, considers heterogeneous materials where the size of the constituents δ is of the order of the macroscopic length scale D

$$R := \delta / D \approx 1 \, , \tag{3}$$

as depicted in Fig. 1b. In this case, the argument of scale separation is not applicable. As a consequence, an analysis must *resolve the full heterogeneous structure*. The resulting finite element discretization can only be handled by application of *fast iterative solvers* for the inherent large system of algebraic equations. Efficient methods for solving large-scale problems with low numerical effort are *multigrid solvers*. The basic multigrid idea is to damp high-frequency errors on the fine grid and to approximate low-frequency errors on coarse grids. In this sense, multigid methods may be understood to be driven by a *numerical multiscale modeling of different grids*. Extensive reviews of basic multigrid concepts are provided by [2, 10, 17, 45], among others. Two different classes of transfer strategies are possible. The *geometric transfer* induces a coarsening process of a sequence of hierarchical grids, see [1, 12] and references therein. An *algebraic transfer* yields an artificial hierarchy of coarse-grid problems. Associated transfer operators are constructed based on the structure of the fine-grid matrices, see [38, 9, 43, 44, 18, 19] and references therein. With respect to analyses of heterogeneous media, many multigrid approaches in the literature focus on diffusion or convection-diffusion problems with oscillating coefficients or anisotropies. A combination of homogenization and multigrid methods for an efficient solution of heterogeneous structures has been considered in [13, 14, 34, 6, 4]. For elliptic partial differential equations with periodically oscillating coefficients, [34] proves robust V-cycle convergence of a multigrid algorithm, where the construction of transfer operators is coupled with a homogenization technique. However, all mentioned approaches are restricted to materials with periodic elastic micro-structures in the small strain context.

Conceptually following the work of [3], we extend in [5, 29] the above mentioned results by constructing a new class of *homogenization-based geometric transfer operators* for inelastic composites with arbitrary shape of

its micro-structure. The new operators are applicable to irregular finite element discretizations of heterogeneous structures. We assume the heterogeneous structure to be resolved by a fine-grid finite element discretization as conceptually shown in Fig. 1b. In order to get geometric transfers in a multigid hierarchy which reflect properties of the heterogeneous material, we couple the two-scale computational homogenization method outlined in the first part of this work with a typical two-grid solution algorithm. The key idea is to construct suitable transfer operators based on a *numerical homogenization of finite element patches* in the multigrid hierarchy. We apply the incremental minimization condition (2) to *finite element patches* $\mathcal{B}^f$ indicated in Fig. 1b

$$W^c = \inf_{\mathbf{d}^f} \frac{1}{|\mathcal{B}^f|} \int_{\mathcal{B}^f} W(\mathbf{1} + \mathbf{B}\mathbf{d}^f; \mathbf{X}) \, \mathrm{d}V \ , \tag{4}$$

defining the *fine-scale nodal displacement* $\mathbf{d}^f = \bar{\mathbf{d}}^f + \tilde{\mathbf{d}}^f$ which decomposes into a homogeneous part $\bar{\mathbf{d}}^f$ and the fluctuation part $\tilde{\mathbf{d}}^f$. For a linearized problem associated with a typical Newton-multigrid scheme, W^c provides a potential for the overall stiffness of the finite element patch. These overall moduli need to be stored on coarse levels if multigrid cycles with more than two levels are applied. Furthermore, assuming the above problem to be driven by a coarse-scale deformation $\mathbf{F}^c$ that depends linearly on the *coarse-scale displacement* $\mathbf{d}^c$ of the patch, we derive from the minimization principle (4) a distinct class of homogenization-based prolongation operators. These transfer operators differ with respect to the boundary conditions applied to the finite element patch $\mathcal{B}^f$ as depicted in Fig. 7. Thus, we get in a procedure conceptually identical to the computational homogenization of scale-separated problems different transfer operators for homogeneous deformation, homogeneous stresses and periodic deformation on the boundary of the patch. These constraints are incorporated by a *penalty extension* of the incremental energy functional (4), providing for the linearized problem a closed-form matrix representation of the desired homogenization-based transfer operators. Comparison with standard transfers from the literature show excellent performances in applications of highly anisotropic composites with finite elastic and plastic material response.

2 Homogenization of Heterogeneous Structures

2.1 Nonlinear Constitutive Response of the Micro-Structure

We consider a general nonlinear constitutive response of the micro-structure in the finite strain context. Here, the first Piola stress $\mathbf{P}$ is assumed to be related to $\mathbf{F}$ by some constitutive material model that governs the local response of the constituents of the micro-structure. In a finite time increment $[t_n, t_{n+1}]$ we assume the potential equation

$$\mathbf{P} = \partial_{\mathbf{F}} W(\mathbf{F}; \mathbf{X}) \quad \text{in } \mathcal{B} \ , \tag{5}$$

for the current stresses at time t_{n+1}, governed by the objective incremental stress potential W. As shown in [26] and [30], such a potential can be constructed for generalized standard dissipative materials by the *constitutive minimization problem*

$$W(\mathbf{F};\mathbf{X}) = \inf_{\boldsymbol{\mathcal{I}}\in\mathcal{G}} \int_{t_n}^{t_{n+1}} [\,\dot{\psi} + \phi\,]\,\mathrm{d}t \quad \text{with} \quad \boldsymbol{\mathcal{I}}(t_n) = \boldsymbol{\mathcal{I}}_n\ , \tag{6}$$

with respect to a generic vector of internal (history) variables $\boldsymbol{\mathcal{I}} \in \mathcal{G}$, which may geometrically be constraint to a manifold $\mathcal{G}$. Here, $\psi(\mathbf{F},\boldsymbol{\mathcal{I}};\mathbf{X})$ is a constitutive *free energy storage function* and $\phi(\dot{\boldsymbol{\mathcal{I}}},\boldsymbol{\mathcal{I}};\mathbf{X})$ a *dissipation function* of the material of the micro-structure. Nonlinear standard materials based on these two scalar functions cover a broad spectrum of constitutive models in finite elasticity, viscoelasticity, plasticity or damage mechanics, see for example [8, 47, 15, 20, 35]. Details of the formulation and discretization of the constitutive minimization principle (6) for models of finite viscoplasticity are outlined in [30].

2.2 Deformation-Driven Homogenization of Micro-Structures

Minimum Averaged Incremental Energy

For the above outlined constitutive response of the micro-structure with constituents of standard materials it is possible to define the micro-to-macro transition as an incremental energy minimization problem associated with the time interval $[t_n, t_{n+1}]$ under consideration. Following [26, 27, 30], we define the macro-stresses by

$$\bar{\mathbf{P}} = \partial_{\bar{\mathbf{F}}}\bar{W}(\bar{\mathbf{F}})\ , \tag{7}$$

in terms of a macro-stress potential $\bar{W}$ that depends on the macroscopic deformation $\bar{\mathbf{F}}$. This potential is defined in terms of the *incremental minimization problem of homogenization*

$$\bar{W}(\bar{\mathbf{F}}) = \inf_{\boldsymbol{\varphi}} \bar{W}(\boldsymbol{\varphi}) \quad \text{with} \quad \bar{W}(\boldsymbol{\varphi}) = \frac{1}{|\mathcal{V}|}\int_{\mathcal{B}} W(\nabla\boldsymbol{\varphi};\mathbf{X})\,\mathrm{d}V\ . \tag{8}$$

It minimizes the average incremental energy of the micro-structure subjected to linear deformations, uniform tractions or periodic displacements and antiperiodic tractions on the boundary of the micro-structure. These boundary conditions can be enforced in a deformation-driven scenario via a penalty functional as outlined below. Note that $|\mathcal{V}|\bar{W}(\boldsymbol{\varphi})$ is the total incremental energy of the micro-structure. For a given macroscopic deformation $\bar{\mathbf{F}}$ the minimization problem determines the deformation field $\boldsymbol{\varphi}$ of the micro-structure up to possible rigid body motions, which can be suppressed by some artificial constraints. The minimization problem (8) extends the well-known average variational principle of homogenization in nonlinear elasticity, outlined for

Table 1. Energy Minimization Problems of Heterogeneous Micro-Structures

1. *Local Consitutive Minimization.* In a time step $[t_n, t_{n+1}]$ define an incremental micro-stress potential at $\mathbf{X} \in \mathcal{B}$ of the micro-structure $\mathcal{B}$ by minimization

$$W(\mathbf{F};\mathbf{X}) = \inf_{\boldsymbol{\mathcal{I}} \in \mathcal{G}} \int_{t_n}^{t_{n+1}} [\,\dot{\psi} + \phi\,]\,\mathrm{d}t \quad \text{with} \quad \boldsymbol{\mathcal{I}}(t_n) = \boldsymbol{\mathcal{I}}_n$$

with respect to internal variables $\boldsymbol{\mathcal{I}} \in \mathcal{G}$ for standard dissipative materials with energy storage function $\psi(\mathbf{F}, \boldsymbol{\mathcal{I}}; \mathbf{X})$ and dissipation function $\phi(\dot{\boldsymbol{\mathcal{I}}}, \boldsymbol{\mathcal{I}}; \mathbf{X})$.

2. *Minimization Problems of Homogenization.* Define incremental macro-stress potentials by minimization of the average incremental energy of micro-structure

$$\tilde{W}(\boldsymbol{\varphi}) = \frac{1}{|\mathcal{V}|} \int_{\mathcal{B}} W(\nabla\boldsymbol{\varphi}; \mathbf{X})\,\mathrm{d}V$$

with respect to the deformation $\boldsymbol{\varphi}$ for deformation–driven constraints on the boundary of the micro–structure in terms of a *penalty formulation*

$$\begin{aligned}
\bar{W}_D^\epsilon(\bar{\mathbf{F}};\bar{\mathbf{X}}) &= \inf_{\boldsymbol{\varphi}}\{\ \tilde{W}(\boldsymbol{\varphi}) - \frac{\epsilon}{2|\mathcal{V}|}\int_{\partial\mathcal{V}} |\,\boldsymbol{\varphi} - \bar{\mathbf{F}}\mathbf{X}\,|^2\,\mathrm{d}A\ \} \\
\bar{W}_P^\epsilon(\bar{\mathbf{F}};\bar{\mathbf{X}}) &= \inf_{\boldsymbol{\varphi}}\{\ \tilde{W}(\boldsymbol{\varphi}) - \frac{\epsilon}{2|\mathcal{V}|}\int_{\partial\mathcal{V}^+} |\,[\![\boldsymbol{\varphi}]\!] - \bar{\mathbf{F}}\,[\![\mathbf{X}]\!]\,|^2\,\mathrm{d}A\ \} \\
\bar{W}_S^\epsilon(\bar{\mathbf{F}};\bar{\mathbf{X}}) &= \inf_{\boldsymbol{\varphi}}\{\ \tilde{W}(\boldsymbol{\varphi}) - \frac{\epsilon}{2}|\frac{1}{|\mathcal{V}|}\int_{\partial\mathcal{V}} \boldsymbol{\varphi}\otimes\mathbf{N}\mathrm{d}A - \bar{\mathbf{F}}\,|^2\ \}
\end{aligned}$$

example in [32, 36, 37], to the incremental setting of generalized standard materials. Details of the formulation and numerical implementation based on the explicit consideration of a displacement fluctuation field are outlined in [26, 27, 30].

Formulation Based on Penalty Functionals

We now consider a family of penalty functionals which link the deformation $\boldsymbol{\varphi}$ of the micro-structure to a prescribed macro-deformation $\bar{\mathbf{F}}$ via the three characteristic classes of boundary constraints considered above. To this end, we recast the incremental minimization principle (8) into the *penalty functionals*

$$\bar{W}_I^\epsilon(\bar{\mathbf{F}}) = \inf_{\boldsymbol{\varphi}} \left[\ \frac{1}{|\mathcal{V}|}\int_{\mathcal{B}} W(\mathbf{F};\mathbf{X})\,\mathrm{d}V - p_I(\boldsymbol{\varphi};\bar{\mathbf{F}})\ \right], \tag{9}$$

for three alternative formats $I = D, P, S$, where the penalty terms p_I associated with the surface ∂V of the micro-structure $\mathcal{V}$ are given by

$$p_I(\boldsymbol{\varphi};\bar{\mathbf{F}}) := \begin{cases} \dfrac{\epsilon}{2|\mathcal{V}|}\displaystyle\int_{\partial\mathcal{V}} |\,\boldsymbol{\varphi} - \bar{\mathbf{F}}\mathbf{X}\,|^2\,\mathrm{d}A\,, \\ \dfrac{\epsilon}{2|\mathcal{V}|}\displaystyle\int_{\partial\mathcal{V}^+} |\,[\![\boldsymbol{\varphi}]\!] - \bar{\mathbf{F}}\,[\![\mathbf{X}]\!]\,|^2\,\mathrm{d}A\,, \\ \dfrac{\epsilon}{2}|\dfrac{1}{|\mathcal{V}|}\displaystyle\int_{\partial\mathcal{V}} \boldsymbol{\varphi}\otimes\mathbf{N}\mathrm{d}A - \bar{\mathbf{F}}\,|^2\,, \end{cases} \tag{10}$$

in terms of the penalty parameter ϵ. The Euler-Lagrange equations of the variational problem (9) first determine the static equilibrium in the interior domain $\mathcal{B}$ of the micro-structure by

$$\mathrm{DIV}[\mathbf{P}] = \mathbf{0} \;\text{ in }\; \mathcal{B} \;\text{ with }\; \mathbf{P} := \partial_{\mathbf{F}} W(\nabla\boldsymbol{\varphi};\mathbf{X}) \tag{11}$$

in terms of the nominal stress tensor $\mathbf{P}$ at point $\mathbf{X}\in\mathcal{B}$ of the micro-structure. The remaining Euler-Lagrange equations give the equations

$$\begin{cases} \mathbf{P}\mathbf{N} = \epsilon\,\big[\,\boldsymbol{\varphi} - \bar{\mathbf{F}}\mathbf{X}\,\big] \;\text{ on }\; \partial\mathcal{V}\,, \\ \pm\,[\,\mathbf{P}\mathbf{N}\,]^{\pm} = \epsilon\,\big[\,[\![\boldsymbol{\varphi}]\!] - \bar{\mathbf{F}}\,[\![\mathbf{X}]\!]\,\big] \;\text{ on }\; \partial\mathcal{V}^{\pm}\,, \\ \mathbf{P}\mathbf{N} = \epsilon\left(\dfrac{1}{|\mathcal{V}|}\displaystyle\int_{\partial\mathcal{V}} \boldsymbol{\varphi}\otimes\mathbf{N}\,dA - \bar{\mathbf{F}}\right)\mathbf{N} \;\text{ on }\; \partial\mathcal{V}\,, \end{cases} \tag{12}$$

on the boundary $\partial\mathcal{V}$ of the micro-structure. These three equations provide *approximate representations* of the exact boundary. In particular, equation $(12)_1$ provides an approximate formulation of the *linear displacement condition* (D), where the traction vector of the boundary $\partial\mathcal{V}$ is determined by a constitutive function in terms of the penalty parameter ϵ. Clearly, the parameter penalizes deviations from the exact enforcement of the displacement boundary condition. Similarly, $(12)_2$ determines approximately the *non-trivial periodic condition* (P), where the traction vector of the boundary $\partial\mathcal{V}^+$ is obtained by a constitutive function in terms of ϵ. Finally, $(12)_3$ approximately determines *uniform tractions* (S) on the boundary, where the overall Piola stress $\bar{\mathbf{P}}$ is directly given by a constitutive function in terms of the penalty parameter ϵ.

The homogenized stresses are derived from the functional (12) by a straightforward exploitation of the definition (7) as

$$\bar{\mathbf{P}}^{\epsilon}_I := \partial_{\bar{\mathbf{F}}}\,\bar{W}^{\epsilon}_I = \begin{cases} \dfrac{1}{|\mathcal{V}|}\displaystyle\int_{\partial\mathcal{V}} \epsilon\,\big[\,\boldsymbol{\varphi} - \bar{\mathbf{F}}\mathbf{X}\,\big]\otimes\mathbf{X}\mathrm{d}A\,, \\ \dfrac{1}{|\mathcal{V}|}\displaystyle\int_{\partial\mathcal{V}^+} \epsilon\,\big[\,[\![\boldsymbol{\varphi}]\!] - \bar{\mathbf{F}}\,[\![\mathbf{X}]\!]\,\big]\otimes[\![\mathbf{X}]\!]\,\mathrm{d}A\,, \\ \epsilon\left(\dfrac{1}{|\mathcal{V}|}\displaystyle\int_{\partial\mathcal{V}} \boldsymbol{\varphi}\otimes\mathbf{N}\mathrm{d}A - \bar{\mathbf{F}}\right)\,. \end{cases} \tag{13}$$

The overall stresses are exclusively governed by the penalty surface terms (10). These results again highlight the constitutive equations for the traction

vector for approximately linear deformation and periodic boundary conditions (D) and (P), governed by the penalty parameter. In the case of approximate uniform traction conditions (S) the overall stresses $\bar{\mathbf{P}}$ are directly given by a constitutive equation in terms of the macro-deformation $\bar{\mathbf{F}}$. The variational formulations are summarized in Table 1.

2.3 FE Formulation of Deformation-Driven Homogenization

FE Discretization of the Micro-Structure

Now we consider the discretization of the micro-structure $\mathcal{B}$ by a standard displacement-type finite element method. Considering two- and three-dimensional micro-structures with $d = 2$ or $d = 3$, respectively, we assume a spatial discretization of the domain $\mathcal{B} \in \mathcal{R}^d$ in terms of a mesh with N nodal points as visualized in Fig. 2. Assume furthermore the existence of an incremental stress potential W defined in (6) at discrete time t_{n+1}. This induces the incremental minimization problem (8) of homogenization. Its spatial finite element discretization yields the finite-dimensional minimization problem

$$\bar{W} = \inf_{\mathbf{d}} \tilde{W}(\mathbf{d}) \quad \text{with} \quad \tilde{W}(\mathbf{d}) = \frac{1}{|\mathcal{V}|} \int_{\mathcal{B}} W(\mathbf{1} + \mathbf{B}\mathbf{d}; \mathbf{X}) \, \mathrm{d}V \,, \tag{14}$$

of the discretized averaged incremental energy with respect to the discrete *global nodal displacement vector* $\mathbf{d} \in \mathcal{R}^{d \cdot N}$, subject to discrete counterparts of the boundary constraints introduced above. Here, the deformation gradient $\nabla \boldsymbol{\varphi} = \mathbf{1} + \mathbf{B}\mathbf{d}$ in a typical element domain $\mathcal{B}^e \subset \mathcal{B}$ is discretized by an element-specific interpolation matrix $\mathbf{B}(\mathbf{X})$. The derivative of the discretized incremental energy functional $\tilde{E}(\mathbf{d}) := |\mathcal{V}|\tilde{W}(\mathbf{d})$ of the micro-structure yields the *internal nodal force vector*

$$\mathbf{f}(\mathbf{d}) := |\mathcal{V}| \partial_{\mathbf{d}} \tilde{W}(\mathbf{d}) = \int_{\mathcal{B}} \mathbf{B}^T \partial_{\mathbf{F}} W(\mathbf{1} + \mathbf{B}\mathbf{d}; \mathbf{X}) \, \mathrm{d}V \in \mathcal{R}^{d \cdot N} \,, \tag{15}$$

at time t_{n+1}. For nonlinear material response, $\mathbf{f}$ is a nonlinear function of $\mathbf{d}$. The sensitivity of the internal forces with respect to a change in the nodal positions defines the *tangent matrix*

$$\mathbf{K}(\mathbf{d}) := |\mathcal{V}| \partial^2_{\mathbf{d}\mathbf{d}} \tilde{W}(\mathbf{d}) = \int_{\mathcal{B}} \mathbf{B}^T \partial^2_{\mathbf{F}\mathbf{F}} W(\mathbf{1} + \mathbf{B}\mathbf{d}; \mathbf{X}) \mathbf{B} \, \mathrm{d}V \in \mathcal{R}^{(d \cdot N) \times (d \cdot N)} \,, \tag{16}$$

of the discretized nonlinear structural model. Note that the internal force vector $\mathbf{f}$ and the tangent matrix $\mathbf{K}$ are *standard ingredients* of finite element structural models.

Partition of Internal and Boundary Degrees

With regard to the discrete formulation of the boundary constraints, the nodes of the mesh are partitioned into those on the surface $\partial\mathcal{V}$ of the representative

volume and those in the interior of $\mathcal{V}$. Assume that $M < N$ nodes of the mesh lie on the boundary $\partial\mathcal{V}$ as indicated in Fig. 2. Then we consider the partitions of the displacement and internal nodal force vectors

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_a \\ \mathbf{d}_b \end{bmatrix} := \begin{bmatrix} \mathbb{P}_a \mathbf{d} \\ \mathbb{P}_b \mathbf{d} \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_a \\ \mathbf{f}_b \end{bmatrix} := \begin{bmatrix} \mathbb{P}_a \mathbf{f} \\ \mathbb{P}_b \mathbf{f} \end{bmatrix} . \tag{17}$$

Here, $\mathbb{P}_a \in \mathcal{R}^{d\cdot(N-M)\times d\cdot N}$ and $\mathbb{P}_b \in \mathcal{R}^{d\cdot M\times d\cdot N}$ are projection matrices which define the interior contribution and the contributions on the boundary, respectively. In line with (17), the tangent matrix defined in (16) is partitioned

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{aa} & \mathbf{K}_{ab} \\ \mathbf{K}_{ba} & \mathbf{K}_{bb} \end{bmatrix} := \begin{bmatrix} \mathbb{P}_a \mathbf{K} \mathbb{P}_a^T & \mathbb{P}_a \mathbf{K} \mathbb{P}_b^T \\ \mathbb{P}_b \mathbf{K} \mathbb{P}_a^T & \mathbb{P}_b \mathbf{K} \mathbb{P}_b^T \end{bmatrix} \tag{18}$$

into contributions associated with interior nodes and nodes on the surface of the representative volume.

Discrete Formulation of the Boundary Constraints

The *linear deformation condition* (D) defined in $(10)_1$ induces at each node $q = 1 \dots M$ of the boundary the condition $\mathbf{d}_q = (\bar{\mathbf{F}} - \mathbf{1})\mathbf{X}_q$ on the nodal displacement in terms of the prescribed macroscopic deformation gradient $\bar{\mathbf{F}}$. Introducing a matrix representation $[\bar{\mathbf{F}}] \in \mathcal{R}^{d^2}$, for example $[\bar{\mathbf{F}}] := [\bar{F}_{11}, \bar{F}_{22}, \bar{F}_{12}, \bar{F}_{21}]^T$, we may represent the nodal displacement constraints in the form

$$\mathbf{d}_q = \mathbb{D}_q^T([\bar{\mathbf{F}}] - [\mathbf{1}]) \quad \text{for} \quad q = 1 \dots M , \tag{19}$$

where $\mathbb{D}_q \in \mathcal{R}^{d^2\times d}$ is a matrix that depends on the coordinates of a nodal point q in the reference configuration, for example

$$\underline{\mathbb{D}}_q := \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \\ X_2 & 0 \\ 0 & X_1 \end{bmatrix}_q , \tag{20}$$

in the case of $d = 2$. Defining the global coordinate matrix $\mathbb{D} := [\mathbb{D}_1\ \mathbb{D}_2 \dots \mathbb{D}_M] \in \mathcal{R}^{d^2\times N_b}$ associated with all $N_b := d\cdot M$ nodes on the surface and taking into account the partition (17) of the global displacement vector, we can represent *all* linear displacement constraints in the compact form

$$\mathbf{d}_b - \mathbb{D}^T([\bar{\mathbf{F}}] - [\mathbf{1}]) = \mathbf{0} . \tag{21}$$

The implementation of the *non-trivial periodic boundary conditions* (P) defined in $(10)_2$ needs a discretization of constraints associated with "pairs" of node positions $\mathbf{X}_q^+$ and $\mathbf{X}_q^-$ associated with the parts $\partial\mathcal{B}^+$ and $\partial\mathcal{B}^-$ of the surface of the micro-structure, see Fig. 2c. This needs a mesh with P corresponding nodes on the surface. For the 2-d micro-structure visualized in

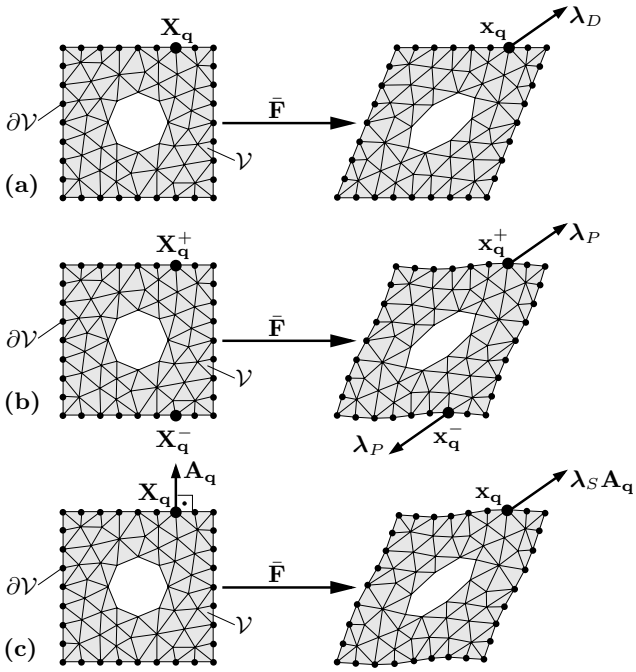


Fig. 2. Partitioning of nodes and deformation-driven boundary constraints $\boldsymbol{\mathcal{A}}_I\mathbf{d}_b - \boldsymbol{\mathcal{B}}_I([\bar{\mathbf{F}}] - [\mathbf{1}]) = \mathbf{0}$ for $I = D, P, S$. **(a)** Prescribed deformation $I = D$ with nodal forces on boundary. **(b)** Periodic deformation $I = P$ with forces at corresponding node pairs on boundary. **(c)** Uniform traction $I = S$ with average first Piola stress $\boldsymbol{\lambda}_S$

Fig. 2c, we have $P = M/2 + 2$ associated pairs of nodes on the surface. For each node pair $q = 1 \dots P$ we assume the periodicity condition $(10)_2$, i.e. $[\![\mathbf{d}_q]\!] = (\bar{\mathbf{F}} - \mathbf{1})[\![\mathbf{X}_q]\!]$ with jump notation $[\![\mathbf{d}_q]\!] := \mathbf{d}_q^+ - \mathbf{d}_q^-$. Using the matrix notation (20), we may recast these constraints into the form

$$\mathbf{d}_q^+ - \mathbf{d}_q^- = (\mathbb{D}_q^{+T} - \mathbb{D}_q^{-T})([\bar{\mathbf{F}}] - [\mathbf{1}]) \quad \text{for} \quad q = 1 \dots P \,. \tag{22}$$

Then, taking into account the partition (17) of the global displacement vector, we can represent *all* constraints in the discretized form

$$\mathbb{P}\mathbf{d}_b - \mathbb{P}\mathbb{D}^T([\bar{\mathbf{F}}] - [\mathbf{1}]) = \mathbf{0} \,. \tag{23}$$

Here, $\mathbb{P} \in \mathcal{R}^{P \times N_b}$ is a link topology matrix whose entries consist of $\{0, 1, -1\}$ only, i.e.

$$\mathbb{P}_{ij} = \begin{cases} +1 \text{ for} i = j \,, \\ -1 \text{ for coupled degrees } i \neq j \text{ on surface} \,, \\ \;\; 0 \text{ otherwise} \quad . \end{cases} \tag{24}$$

It is effectively stored as a standard topology array in finite element analysis.

The *uniform traction boundary condition* (S) is incorporated in the deformation-driven context by the weak constraint $(10)_3$ for the deformation on the boundary of the representative volume. The implementation needs a discretization of the surface integral. To this end we consider the limit $\mathbf{N}\mathrm{d}A \to \mathbf{A}_q$ of the continuous area vector to a discrete *outward-normal area vector* at node q as visualized in Fig. 2. We then may approximate the integral $(10)_3$ by the discrete sum $\frac{1}{|\mathcal{V}|}\sum_{q=1}^{M} \mathbf{d}_q \otimes \mathbf{A}_q = \bar{\mathbf{F}} - \mathbf{1}$, which can be recast into the matrix representation

$$\frac{1}{|\mathcal{V}|}\sum_{q=1}^{M} \mathbb{S}_q \mathbf{d}_q = [\bar{\mathbf{F}}] - [\mathbf{1}] \,, \tag{25}$$

where $\mathbb{S}_q \in \mathcal{R}^{d^2 \times d}$ is a matrix depending on the area vector $\mathbf{A}_q$ of the nodal point q in the reference configuration. The representation, in analogy to (20), takes the form

$$\mathbb{S}_q := \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \\ A_2 & 0 \\ 0 & A_1 \end{bmatrix}_q \,, \tag{26}$$

for $d = 2$. Defining a global coordinate matrix $\mathbb{S} := [\mathbb{S}_1\ \mathbb{S}_2 \dots \mathbb{S}_M] \in \mathcal{R}^{d^2 \times N_b}$ associated with *all* M nodal points on the surface of the micro-structure, we can represent *all* constraints in the discretized form

$$\mathbb{S}\mathbf{d}_b - |\mathcal{V}|([\bar{\mathbf{F}}] - [\mathbf{1}]) = \mathbf{0} \,. \tag{27}$$

For a compact representation of the subsequent developments, we represent all three conditions (21), (23) and (26) in the form

$$\boldsymbol{\mathcal{A}}_I \mathbf{d}_b - \boldsymbol{\mathcal{B}}_I([\bar{\mathbf{F}}] - [\mathbf{1}]) = \mathbf{0} \,, \tag{28}$$

for the three alternative formats $I = D, P, S$ with the coefficient matrices $\boldsymbol{\mathcal{A}}_I$ and $\boldsymbol{\mathcal{B}}_I$ defined by

$$\boldsymbol{\mathcal{A}}_I := \begin{cases} \mathbb{I} \in \mathcal{R}^{N_b \times N_b} \,, \\ \mathbb{P} \in \mathcal{R}^{P \times N_b} \,, \\ \mathbb{S} \in \mathcal{R}^{d^2 \times N_b} \,, \end{cases} \quad \text{and} \quad \boldsymbol{\mathcal{B}}_I := \begin{cases} \mathbb{D}^T \in \mathcal{R}^{N_b \times d^2} \,, \\ \mathbb{P}\mathbb{D}^T \in \mathcal{R}^{P \times d^2} \,, \\ |\mathcal{V}|\mathbb{I} \in \mathcal{R}^{d^2 \times d^2} \,, \end{cases} \tag{29}$$

where $\mathbb{I}$ denotes the identity matrix.

Discrete Penalty Functionals

With the above compact representation of the discrete boundary constraints, the discrete counterpart of the continuous penalty functional (9) reads

$$\bar{W}_I^{\epsilon}(\bar{\mathbf{F}}; \bar{\mathbf{X}}) = \inf_{\mathbf{d}} \left\{ \tilde{W}(\mathbf{d}) \, - \, \frac{\epsilon}{2} | \, \boldsymbol{\mathcal{A}}_I \mathbf{d}_b - \boldsymbol{\mathcal{B}}_I([\bar{\mathbf{F}}] - [\mathbf{1}]) \, |^2 \right\} \,, \tag{30}$$

Table 2. Algorithmic Micro-Macro-Transition in Penalty Form

1. Set initial value $\mathbf{d} = \boldsymbol{0}$
2. Assemble nodal force vector and tangent of FE-discretized micro-structure
   $\mathbf{f} = \int_{\mathcal{B}} \mathbf{B}^T \partial_{\mathbf{F}} W(\mathbf{1} + \mathbf{Bd}) \mathrm{d}V$ and $\mathbf{K} = \int_{\mathcal{B}} \mathbf{B}^T \partial^2_{\mathbf{FF}} W(\mathbf{1} + \mathbf{Bd}) \mathbf{B} \mathrm{d}V$
   in partitions associated with nodal positions $\mathbf{X}_a \in \mathcal{B}$ and $\mathbf{X}_b \in \partial\mathcal{B}$
   $$\mathbf{f} = \begin{bmatrix} \mathbf{f}_a \\ \mathbf{f}_b \end{bmatrix} \quad \text{and} \quad \mathbf{K} = \begin{bmatrix} \mathbf{K}_{aa} & \mathbf{K}_{ab} \\ \mathbf{K}_{ba} & \mathbf{K}_{bb} \end{bmatrix}$$
3. Compute residual $\mathbf{r} = [\mathbf{r}_a, \mathbf{r}_b]^T$ with
   $$\mathbf{r}_a = \mathbf{f}_a$$
   $$\mathbf{r}_b = \mathbf{f}_b - \epsilon |\mathcal{V}| \, \boldsymbol{\mathcal{A}}_I^T \big[\boldsymbol{\mathcal{A}}_I \mathbf{d}_b - \boldsymbol{\mathcal{B}}_I ([\bar{\mathbf{F}}] - [\mathbf{1}]) \big]$$
4. Compute condensed matrices
   $$\tilde{\mathbf{r}}_b = \mathbf{r}_b - \mathbf{K}_{ba} \mathbf{K}_{aa}^{-1} \mathbf{r}_a \qquad \tilde{\mathbf{K}}_{bb} = \mathbf{K}_{bb} - \epsilon |\mathcal{V}| \, \boldsymbol{\mathcal{A}}_I^T \boldsymbol{\mathcal{A}}_I - \mathbf{K}_{ba} \mathbf{K}_{aa}^{-1} \mathbf{K}_{ab}$$
5. Compute increments
   $$\Delta \mathbf{d}_b = -\tilde{\mathbf{K}}_{bb}^{-1} \tilde{\mathbf{r}}_b$$
   $$\Delta \mathbf{d}_a = -\mathbf{K}_{aa}^{-1} (\mathbf{K}_{ab} \, \Delta \mathbf{d}_b + \mathbf{r}_a)$$
   and update nodal displacements $\mathbf{d} \Leftarrow \mathbf{d} + \Delta \mathbf{d}$
6. For $|\mathbf{r}| >$ tol go to 2
7. Compute effective stresses and overall tangent moduli
   $$[\bar{\mathbf{P}}_I^\epsilon] = \epsilon \, \boldsymbol{\mathcal{B}}_I^T \big[\boldsymbol{\mathcal{A}}_I \mathbf{d}_b - \boldsymbol{\mathcal{B}}_I ([\bar{\mathbf{F}}] - [\mathbf{1}]) \big]$$
   $$[\bar{\mathbb{C}}_I^\epsilon] = -\epsilon^2 |\mathcal{V}| \boldsymbol{\mathcal{B}}_I^T \boldsymbol{\mathcal{A}}_I \tilde{\mathbf{K}}_{bb}^{-1} \boldsymbol{\mathcal{A}}_I^T \boldsymbol{\mathcal{B}}_I - \epsilon \boldsymbol{\mathcal{B}}_I^T \boldsymbol{\mathcal{B}}_I$$

for the three alternative formats $I = D, P, S$. The first variation of this functional for prescribed macro-deformation $\bar{\mathbf{F}}$ gives the discrete form of the Euler-Lagrange equations

$$\begin{cases} \mathbf{0} = \mathbf{f}_a \, , \\ \mathbf{0} = \mathbf{f}_b - \epsilon |\mathcal{V}| \, \boldsymbol{\mathcal{A}}_I^T \big[\boldsymbol{\mathcal{A}}_I \mathbf{d}_b - \boldsymbol{\mathcal{B}}_I ([\bar{\mathbf{F}}] - [\mathbf{1}]) \big] \, . \end{cases} \tag{31}$$

For the solution of this nonlinear problem by a Newton algorithm, we define the residual $\mathbf{r} = [\mathbf{r}_a, \mathbf{r}_b]^T$ with

$$\begin{cases} \mathbf{r}_a := \mathbf{f}_a \, , \\ \mathbf{r}_b := \mathbf{f}_b - \epsilon |\mathcal{V}| \, \boldsymbol{\mathcal{A}}_I^T \big[\boldsymbol{\mathcal{A}}_I \mathbf{d}_b - \boldsymbol{\mathcal{B}}_I ([\bar{\mathbf{F}}] - [\mathbf{1}]) \big] \, . \end{cases} \tag{32}$$

The iterative solution of (31) is based on a sequence of linearizations at the position $\{\mathbf{d}_a, \mathbf{d}_b\}$

$$\begin{cases} \mathbf{0} = \mathbf{r}_a + \mathbf{K}_{aa} \Delta \mathbf{d}_a + \mathbf{K}_{ab} \Delta \mathbf{d}_b \, , \\ \mathbf{0} = \mathbf{r}_b + \mathbf{K}_{ba} \Delta \mathbf{d}_a + (\mathbf{K}_{bb} - \epsilon |\mathcal{V}| \, \boldsymbol{\mathcal{A}}_I^T \boldsymbol{\mathcal{A}}_I) \Delta \mathbf{d}_b \end{cases} \tag{33}$$

along with the update of the nodal displacements

$$\mathbf{d} \Leftarrow \mathbf{d} + \Delta\mathbf{d} \; . \tag{34}$$

The successive steps for the solution of the linear system (33) is outlined in Table 2. The iteration is terminated for the case $|\mathbf{r}| < \text{tol}$.

When an equilibrium state has been obtained, we compute the homogenized stresses from the functional (30) by a straightforward exploitation of the definition (7) as

$$[\bar{\mathbf{P}}_I^\epsilon] := \partial_{\bar{\mathbf{F}}} \bar{W}_I^\epsilon = \epsilon \, \boldsymbol{\mathcal{B}}_I^T \big[\boldsymbol{\mathcal{A}}_I \mathbf{d}_b - \boldsymbol{\mathcal{B}}_I([\bar{\mathbf{F}}] - [\mathbf{1}]) \big] \; . \tag{35}$$

Observe that the macro-stresses are exclusively governed by the penalty surface terms (30). The macroscopic tangent moduli are obtained by a further derivative of the macro-stresses $\bar{\mathbf{P}}$ with respect to the macroscopic deformation gradient $\bar{\mathbf{F}}$. To this end, we need the sensitivity of the nodal displacements with respect to the macroscopic deformation. Starting from (32), we derive the coupled sensitivities

$$\begin{cases} \Delta\mathbf{d}_b = -\epsilon |\mathcal{V}| \tilde{\mathbf{K}}_{bb}^{-1} \boldsymbol{\mathcal{A}}_I^T \boldsymbol{\mathcal{B}}_I [\Delta\bar{\mathbf{F}}] \; , \\ \Delta\mathbf{d}_a = -\mathbf{K}_{aa}^{-1} \mathbf{K}_{ab} \Delta\mathbf{d}_b \end{cases} \tag{36}$$

with the condensed matrix $\tilde{\mathbf{K}}_{bb} = \mathbf{K}_{bb} - \epsilon |\mathcal{V}| \, \boldsymbol{\mathcal{A}}_I^T \boldsymbol{\mathcal{A}}_I - \mathbf{K}_{ba} \mathbf{K}_{aa}^{-1} \mathbf{K}_{ab}$. With these sensitivities at hand, the straightforward derivative of (35) gives the effective macro-moduli

$$[\bar{\mathbb{C}}_I^\epsilon] = \partial_{[\bar{\mathbf{F}}]} [\bar{\mathbf{P}}_I^\epsilon] = -\epsilon^2 |\mathcal{V}| \, \boldsymbol{\mathcal{B}}_I^T \boldsymbol{\mathcal{A}}_I \tilde{\mathbf{K}}_{bb}^{-1} \boldsymbol{\mathcal{A}}_I^T \boldsymbol{\mathcal{B}}_I - \epsilon \boldsymbol{\mathcal{B}}_I^T \boldsymbol{\mathcal{B}}_I \; . \tag{37}$$

Table 2 summarizes the algorithmic steps.

2.4 Numerical Examples of Nonlinear Homogenization

We now apply the algorithms for computational homogenization summarized in Table 2 to representative model problems. Here, we focus on two-phase composites with elastic and inelastic constituents at finite deformations.The elastic response is based on an isotropic Neo-Hookean-type free energy function

$$\psi(\mathbf{F}^e) = \frac{\mu}{2} (\text{tr}[\mathbf{F}^{eT} \mathbf{F}^e] - 3) + \frac{\mu}{\beta} ((\det[\mathbf{F}^e])^{-\beta} - 1) \; , \tag{38}$$

with $\mathbf{F}^e := \mathbf{F} \mathbf{F}^{p-1}$ and $\beta := \kappa/\mu - 2/3$, where κ represents the bulk modulus and μ the shear modulus. The plastic response is assumed to be governed by the isotropic dissipation function of the von Mises type

$$\phi(\mathbf{L}^p) = \sup_{(\boldsymbol{\Sigma},\beta) \in \mathbb{E}} \{ \boldsymbol{\Sigma} : \mathbf{L}^p + \beta \dot{\alpha} \} \quad \text{with} \quad \mathbb{E} := \{ (\boldsymbol{\Sigma}, \beta) \, | \, ||\boldsymbol{\Sigma}|| - \sqrt{\frac{2}{3}} (y_0 + \beta) \le 0 \} \; , \tag{39}$$

with $\mathbf{L}^p := \dot{\mathbf{F}}^p \mathbf{F}^{p-1}$. y_0 is the yield stress and $\beta = h\alpha$ describes linear hardening with modulus h. These functions enter the constitutive variational formulation (6). Associated stress update algorithms are outlined in [25, 28].

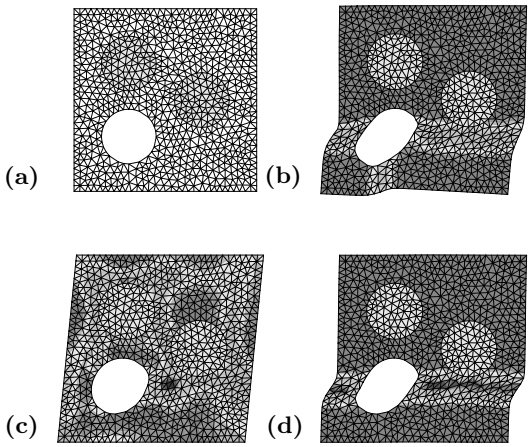


Fig. 3. Simple shearing of a composite micro-structure for different micro-to-macro transitions: **(a)** initial configuration, deformed configuration with distribution of equivalent plastic strains for **(b)** uniform tractions $I = S$, **(c)** linear displacements $I = D$, **(d)** periodic displacements $I = P$ on the boundary of the micro–structure. Micro–structure stiffness **(d)** is bounded by **(b)** and **(c)**

Finite Elastic-Plastic Deformations of a Composite

The first example is a composite undergoing finite elastic-plastic strains. We investigate the cube-shaped representative micro-structure depicted in Fig. 3a for a simple shear mode. The side length of the micro-structure is $l = 1.0$, the diameter of the hole and the inclusions are $d = 0.3$. In a local coordinate system located at the center of the micro-structure, the hole is at the position $(-0.2, -0.2)$. Two inclusions are located at $(-0.2, 0.2)$ and $(-0.2, 0)$. The elastic material parameters in (38) are $\kappa_{\mathcal{M}} = 17.5$, $\mu_{\mathcal{M}} = 8.0$ for the matrix and $\kappa_{\mathcal{I}} = 100\kappa_{\mathcal{M}}$, $\mu_{\mathcal{I}} = 100\mu_{\mathcal{M}}$ for the inclusion. The response of the stiff inclusions is assumed to be elastic ($y_0 \to \infty$) while the matrix material is assumed to be elastic-plastic with parameters $y_0 = 0.45$ and $h = 0.1$ (39). The micro-structure has been discretized with 1501 Mini-type mixed triangles as conceptually outlined in [11]. It is driven by a simple shear mode in 20 equal time steps up to the final value $\bar{F}_{12} = 0.1$. Figure 3 depicts the deformed micro-structures for the different micro-macro transitions $I = D, P, S$ associated with uniform tractions, linear and periodic boundary deformations. Observe that for the unit-cell-type periodic case $I = P$ in Fig. 3d a horizontal shear band develops starting from the hole. In the case of stress boundary conditions $I = S$ in Fig. 3b an additional vertical shear band develops that is also aligned to the hole. The plot in Fig. 3b documents the fact that the RVE-type stress boundary condition provides maximum flexibility, yielding the softest response of the micro-structure. These bounding properties are also documented in the diagram in Fig. 4 depicting a shear component of the macroscopic Kirchhoff stress $\bar{\boldsymbol{\tau}} := \bar{\mathbf{P}}\bar{\mathbf{F}}^T$ that develops during the deformation

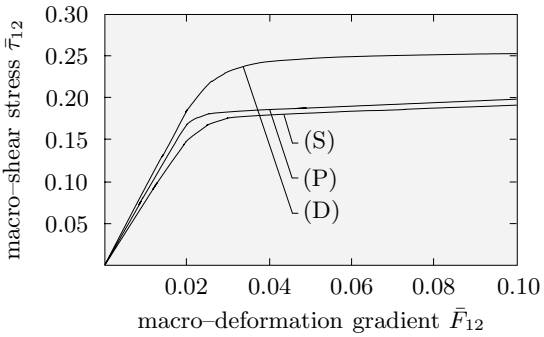


Fig. 4. Simple shearing of a composite micro-structure. Macroscopic stress-strain curves for analyses with uniform tractions $I = S$, linear displacement $I = D$ and periodic deformation $I = P$ on the boundary of the micro-structure

process. The RVE-type linear deformation and uniform traction conditions $I = D, S$ provide upper and lower bounds of the macro-stress response. The unit-cell-type solution $I = P$ with periodic deformation and antiperiodic traction lies between them.

Two-Scale Scenario of an Elastic-Plastic Composite

As an elementary example of a two-scale micro-to-macro transition in nonlinear elasto-plasticity, we consider a disc with a hole as a macro-structure, under an equibiaxial extension as depicted in Fig. 5. The macroscopic constitutive response at each integration point is governed by the homogenization procedure outlined above. We consider a composite micro-structure that consists of two stiff inclusions and one hole embedded in a soft matrix. Figures 5a and b show the deformed macro-structure with the distribution of macroscopic equivalent plastic strains for the case a) of linear displacements $I = D$ or b) for the case of uniform stresses $I = S$ on the boundaries of the micro-structures. Furthermore, Figs. 5a and b depict the deformed micro-structures associated with integration points of four finite elements of the macro-mesh close to the hole. The results obtained in Figs. 5a and b provide upper and lower bounds to the stiffness of the macro-structure.

3 Multigrid Solver for Heterogeneous Structures

3.1 A Basic Multigrid Algorithm for Nonlinear Response

We first review a basic multigrid algorithm, which is conceptually well-known in the literature. Furthermore, we address the basic algorithmic ingredients for the iterative solving of large sparse systems of equations. The individual notations and definitions introduced are required in the subsequent sections.

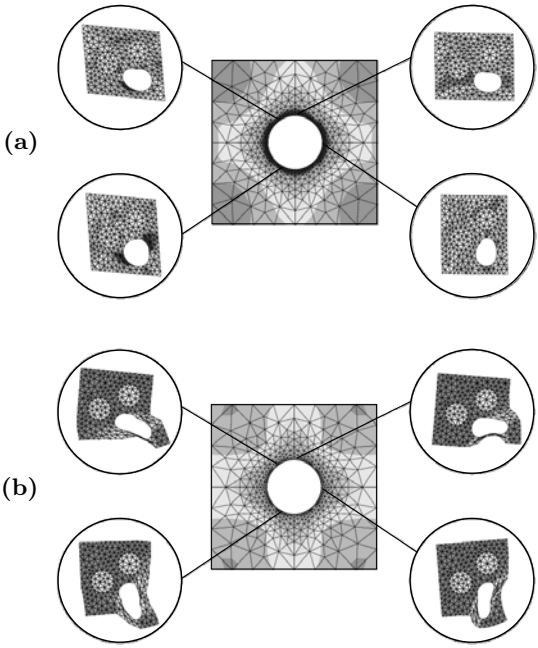


Fig. 5. Two-scale bridging for a micro-heterogeneous composite with alternative constraints. Discretized macro-structure of a square disc with a hole under uniaxial tension. Equivalent plastic strains for **(a)** linear displacements $I = D$ and **(b)** uniform stresses $I = S$ on the boundaries of the microstructures

Newton-Multigrid Algorithm

We focus on a general nonlinear response of heterogeneous structures such as nonlinear elasticity or plasticity at large strains. Displacement-type finite element discretizations of fine-scale boundary-value problems of composites as depicted in Figs. 9 and 12 typically yield *large nonlinear algebraic systems* of the form

$$\mathbf{R}(\mathbf{D}) = \mathbf{0}\,, \tag{40}$$

where $\mathbf{D} \in \mathcal{R}^{n_1}$ and $\mathbf{R} \in \mathcal{R}^{n_1}$ denote the fine-scale displacement and residual vectors, respectively. n_1 is the dimension of the discrete fine-scale solution space. The solution of (40) may be obtained by a combination of an iterative solution concept of the incrementally nonlinear problem with an efficient iterative solver for the *linear system* obtained based on the linearization

$$\mathrm{Lin}\mathbf{R}(\Delta\mathbf{D};\mathbf{D}_0) := \mathbf{R}(\mathbf{D}_0) + \mathbf{K}(\mathbf{D}_0)\Delta\mathbf{D} = \mathbf{0} \text{ with } \mathbf{K}(\mathbf{D}_0) := \partial_{\mathbf{D}}\mathbf{R}(\mathbf{D}_0) \tag{41}$$

of (40) at $\mathbf{D}_0$ in a typical increment. $\mathbf{K} \in \mathcal{R}^{n_1 \times n_1}$ is the tangent matrix at $\mathbf{D}_0$. The numerical efficiency of direct or classical iterative solvers deteriorates for large systems (41) of linear equations. However, an effective treatment of

Table 3. Newton-Multigrid Algorithm for Nonlinear Fine-Scale Problem.

Newton-Loop for Nonlinear Fine-Scale Problem

1. Initialize solution vector $\mathbf{D} = \mathbf{D}_0$ and iteration counter $m = 0$
2. Assemble fine-scale residual $\mathbf{R}_m(\mathbf{D})$ and tangent matrix $\mathbf{K}_m(\mathbf{D})$
3. Check convergence of Newton iteration: If $|\mathbf{R}_m| < \theta|\mathbf{R}_0|$ stop
4. For $m > 1$ set $\kappa_{nili} := |\mathbf{R}_m|/|\mathbf{R}_{m-1}|$ and $tol := \max(|\mathbf{R}_m|\kappa_{nili}^2, \theta|\mathbf{R}_0|)$

Multigrid Cycle for Solving Linear System $\mathbf{K}_m \Delta\mathbf{D} = \mathbf{R}_m$

5. Initialize vectors $\mathbf{r}_0 = \mathbf{R}_m$, $\Delta\mathbf{D} = \boldsymbol{0}$ and iteration counter $i = 0$
6. Determine $\Delta\mathbf{D}$ by multigrid cycle Iter$(\Delta\mathbf{D}, \mathbf{R}_m, \mathbf{K}_m)$ in Table 4
7. If $i \leq 1$ or $\kappa_{nili} > 1$ go to 11.
8. If $i > 1$ and $|\mathbf{r}_0|\kappa_{lin} < tol$ with $\kappa_{lin} := (|\mathbf{r}_i|/|\mathbf{r}_0|)^{\frac{i+1}{i}}$ go to 11.
9. Determine linear residual $\mathbf{r}_{i+1} = \mathbf{R}_m - \mathbf{K}_m \Delta\mathbf{D}$
10. If $|\mathbf{r}_{i+1}| > tol$ set $i = i + 1$ and go to 6.

11. Update solution vector $\mathbf{D} \Leftarrow \mathbf{D} + \Delta\mathbf{D}$, set $m \Leftarrow m + 1$ and go to 2.

this class of problems is accomplished by means of multigrid methods. The algorithm in Table 3 depicts a Newton-type solver for the nonlinear problem (40) with an *embedded iterative multigrid cycle* for the linearized problem within each step of the Newton iteration. The dynamic formulation of the solution tolerance *tol* of the inner multigrid iteration in Table 3 related to the convergence of the outer Newton iteration represents an important speed-up with respect to direct solvers. Such a scheme provides a general tool for a wide spectrum of problems in nonlinear structural mechanics of heterogeneous media. The critical point with regard to the efficiency of such a method is a specific design of the multigrid cycle that takes into account the heterogeneity of the linearized structural problem.

Multigrid of the Linearized Problem

During the iterative solution of large linear systems (41), the current *iteration defect* can be decomposed into *long-wave* and *short-wave parts*. The fundamental idea of a multigrid solver is the reduction of the defect on different meshes. To this end, one smoothes the fine-scale iteration defect on a hierarchy of coarse meshes based on classical iterative schemes such as Jacobi or Gauss-Seidel methods. These methods smooth out the highly oscillating defects after few steps. Long-wave defects on fine grids appear as short-wave defects on coarse grids and allow on coarser grids a fast smoothing by the standard iterative methods. In order to apply multigrid methods, we establish a hierarchy of conforming *grid levels* $l = 1, 2, \ldots, L$ such as depicted in

Table 4. Iteration Iter($\mathbf{d}^1, \mathbf{f}^1, \mathbf{K}^1$) of Linearized Fine-Scale Problem

$\mathbf{r} = \mathbf{f}^1$, $\mathbf{d}^1 = \mathbf{0}$	...	Initializations
If $\|\mathbf{r}\| \geq \theta \|\mathbf{f}^1\|$ do		
call MG($\Delta\mathbf{d}^1, \mathbf{f}^1, \mathbf{K}^1$)	...	Multigrid iteration
$\mathbf{d}^1 \Leftarrow \mathbf{d}^1 + \Delta\mathbf{d}^1$	...	Update solution vector
$\mathbf{r} \Leftarrow \mathbf{r} - \mathbf{K}^1\Delta\mathbf{d}^1$	...	Residual vector
Enddo		

Figs. 8 and 11. With each level l we associate the linear system

$$\mathbf{K}^l\mathbf{d}^l = \mathbf{f}^l \ , \tag{42}$$

where level $l = 1$ refers to the *fine-grid linearized problem* (41) and the level $l = L$ to the matrix system on the *coarsest grid level*. Here, $\mathbf{K}^l \in \mathcal{R}^{n_l \times n_l}$ is the grid matrix, $\mathbf{d}^l \in \mathcal{R}^{n_l}$ the grid solution vector and $\mathbf{f}^l \in \mathcal{R}^{n_l}$ the grid right-hand side. n_l indicates the number of unknowns of a typical grid l. In a typical iteration, the *defect* of the grid l is defined by

$$\mathbf{r}^l := \mathbf{f}^l - \mathbf{K}^l\mathbf{d}^l \tag{43}$$

and the update of the solution vector reads

$$\mathbf{d}^l \Leftarrow \mathbf{d}^l + \Delta\mathbf{d}^l \ , \tag{44}$$

where the increment $\Delta\mathbf{d}^l$ follows from a multigrid cycle as pointed out below. We intend to solve the fine-grid problem (42) on level $l = 1$ based on the iterative solution algorithm Iter($\mathbf{d}^1, \mathbf{f}^1, \mathbf{K}^1$) outlined in Table 4. The classical way is to apply the multigrid cycles specified in Table 5 until a given tolerance is obtained, for example $\theta = 10^{-8}$. We employ a multiplicative method based on the multigrid V-cycle MG($\mathbf{d}^l, \mathbf{r}^l, \mathbf{K}^l$) in Table 5, which represents a recursive two-grid scheme. This classical two-grid method consists of two crucial steps, an *iterative smoothing procedure* and a *coarse-grid correction*.

The Standard Two-Grid Algorithm

In order to reduce the fine-scale oscillating part of the defect, the two-grid cycle starts with *pre-smoothing* the incremental solution vector $\Delta\mathbf{d}^l$ of the fine grid by means of ν_1 relaxation steps

$$\Delta\mathbf{d}^l \Leftarrow \Delta\mathbf{d}^l + (\mathbf{W}^l_{pre})^{-1}\mathbf{r}^l \ , \tag{45}$$

with the iteration matrix $\mathbf{W}^l_{pre}$. The pre-smoothing of $\Delta\mathbf{d}^l$ starts with $\Delta\mathbf{d}^l = \mathbf{0}$ and is performed by the Block-Backward Gauss-Seidel algorithm

Table 5. Multigrid Algorithm $\mathrm{MG}(\mathbf{d}^l, \mathbf{r}^l, \mathbf{K}^l)$

If $l < L$ then	
1. Initialization	$\mathbf{d}^l = \boldsymbol{0}$
2. Pre-Smoothing	Call $\mathrm{S}_{\mathrm{pre}}(\mathbf{d}^l, \mathbf{r}^l, \mathbf{K}^l)$
3. Coarse-Grid Matrix	$\mathbf{K}^{l+1} = \mathbf{R}^{l+1}{}_l \mathbf{K}^l \mathbf{P}^l{}_{l+1}$
4. Restricted Defect	$\mathbf{r}^{l+1} = \mathbf{R}^{l+1}{}_l [\mathbf{r}^l - \mathbf{K}^l \mathbf{d}^l]$
5. Multigrid Iteration	Call $\mathrm{MG}(\mathbf{d}^{l+1}, \mathbf{r}^{l+1}, \mathbf{K}^{l+1})$
6. Prolongation, Update	$\mathbf{d}^l \Leftarrow \mathbf{d}^l + \tau \mathbf{P}^l{}_{l+1} \mathbf{d}^{l+1}$
7. Post-Smoothing	Call $\mathrm{S}_{\mathrm{post}}(\mathbf{d}^l, \mathbf{r}^l, \mathbf{K}^l)$
Else	
Solve $\mathbf{K}^L \mathbf{d}^L = \mathbf{r}^L$ by direct solver, e.g. Gauss algorithm	
Endif	

Table 6. Pre-Smoothing Algorithm $\mathrm{S}_{\mathrm{pre}}(\mathbf{d}^l, \mathbf{r}^l, \mathbf{K}^l)$

Block-Backward Gauss-Seidel Iterations

$$\mathbf{d}_i^{l(m+1)} = (\mathbf{K}_{ii}^l)^{-1} [\mathbf{r}_i^l - \sum_{j=1}^{i-1} \mathbf{K}_{ij}^l \mathbf{d}_j^{l(m)} - \sum_{j=i+1}^{n_l} \mathbf{K}_{ij}^l \mathbf{d}_j^{l(m+1)}]$$

with $m = 1, .., \nu_1$ and $i = n_l, .., 1$

Table 7. Post-Smoothing Algorithm $\mathrm{S}_{\mathrm{post}}(\mathbf{d}^l, \mathbf{r}^l, \mathbf{K}^l)$

Block-Forward Gauss-Seidel Iterations

$$\mathbf{d}_i^{l(m+1)} = (\mathbf{K}_{ii}^l)^{-1} [\mathbf{r}_i^l - \sum_{j=1}^{i-1} \mathbf{K}_{ij}^l \mathbf{d}_j^{l(m+1)} - \sum_{j=i+1}^{n_l} \mathbf{K}_{ij}^l \mathbf{d}_j^{l(m)}]$$

with $m = 1, .., \nu_2$ and $i = 1, .., n_l$

$\mathrm{S}_{\mathrm{pre}}(\mathbf{d}^l, \mathbf{r}^l, \mathbf{K}^l)$ outlined in Table 6. The remaining coarse-scale part is then approximated by a correction on a coarse-grid, where it appears as a short-wave defect. This coarse-grid correction step requires the construction of grid transfer operators. The *coarse-to-fine-grid operator* $\mathbf{P}^l{}_{l+1} \in \mathcal{R}^{n_l \times n_{l+1}}$ and the *fine-to-coarse-grid operator* $\mathbf{R}^{l+1}{}_l \in \mathcal{R}^{n_{l+1} \times n_l}$ define the mappings between the fine- and coarse-grid ansatz spaces, i.e.

$$\mathbf{P}^l{}_{l+1} : \mathcal{S}^{l+1} \to \mathcal{S}^l \quad \text{and} \quad \mathbf{R}^{l+1}{}_l : \mathcal{S}^l \to \mathcal{S}^{l+1} \,, \tag{46}$$

respectively. With these transfer operators at hand, the coarse-grid-matrix $\mathbf{K}^{l+1} \in \mathcal{R}^{n_{l+1} \times n_{l+1}}$ and the restricted defect $\mathbf{r}^{l+1} \in \mathcal{R}^{n_{l+1}}$ are defined by

$$\mathbf{K}^{l+1} = \mathbf{R}^{l+1}{}_l \mathbf{K}^l \mathbf{P}^l{}_{l+1} \quad \text{and} \quad \mathbf{r}^{l+1} = \mathbf{R}^{l+1}{}_l [\mathbf{f}^l - \mathbf{K}^l \Delta \mathbf{d}^l] \,. \tag{47}$$

For symmetric matrices $\mathbf{K}^l$ and $\mathbf{K}^{l+1}$ we define

$$\mathbf{R}^{l+1}{}_l = (\mathbf{P}^l{}_{l+1})^T \ . \tag{48}$$

We then perform the coarse-grid correction by solving first the problem

$$\mathbf{K}^{l+1} \Delta \mathbf{d}^{l+1} = \mathbf{r}^{l+1} \ , \tag{49}$$

for the incremental solution vector $\Delta \mathbf{d}^{l+1}$ on the coarse grid. This is performed by a recursive call of the multigrid algorithm $\mathrm{MG}(\mathbf{d}^{l+1}, \mathbf{r}^{l+1}, \mathbf{K}^{l+1})$ in Table 5, where a direct solver is used to solve the coarsest system for $l = L$. Next, we obtain by prolongation the correction update of the fine-grid solution vector

$$\Delta \mathbf{d}^l \Leftarrow \Delta \mathbf{d}^l + \mathbf{P}^l{}_{l+1} \Delta \mathbf{d}^{l+1} \ . \tag{50}$$

Finally, the fine-grid solution vector is post-smoothed by ν_2 relaxation steps via a classical iterative method

$$\Delta \mathbf{d}^l \Leftarrow \Delta \mathbf{d}^l + (\mathbf{W}^l_{\mathrm{post}})^{-1} \mathbf{r}^l \ , \tag{51}$$

with iteration matrix $\mathbf{W}^l_{\mathrm{post}}$. Here, we apply the standard Block-Forward Gauss-Seidel algorithm $\mathrm{S}_{\mathrm{post}}(\mathbf{d}^l, \mathbf{r}^l, \mathbf{K}^{l+1})$ outlined in Table 7. This algorithm is repeated until convergence is achieved within a given tolerance θ in Table 4.

3.2 Numerical Homogenization of Heterogeneous Substructures

Of fundamental importance for an effective iteration performance of the recursive two-grid algorithm $\mathrm{MG}(\mathbf{d}^{l+1}, \mathbf{r}^{l+1}, \mathbf{K}^{l+1})$ in Table 5 is the design of the transfer operators $\mathbf{P}^f{}_c = (\mathbf{R}^c{}_f)^T$ which map variables between fine and coarse grids. These operators need to take into account the properties of the heterogeneous structure. In what follows, we construct a new class of transfer operators for heterogeneous composites which are based on a homogenization of sub-structures as indicated in Fig. 1b. The basic idea for the construction is visualized in Fig. 6. We interprete the transfer between a fine-grid f and a coarse-grid c as a scale transition between a heterogeneous micro-scale and a homogeneous macro-scale. With this viewpoint at hand, we base the new class of transfer operators based on the numerical homogenization procedures outlined in Sect. 2. With the incremental constitutive variational formulation (6) at hand, we demand *energetic consistency* between a typical *heterogeneous fine-grid patch* $\mathcal{B}^f \subset \mathcal{R}^3$ and the associated *homogenized coarse-grid patch* $\mathcal{B}^c \subset \mathcal{R}^3$ as depicted in Fig. 6. This defines for the *linearized setting* associated with a typical step of the Newton iteration outlined in Table 3 the homogenized moduli $\mathbb{C}^c$ of the coarse-grid patch in terms of the moduli $\mathbb{C}^f$ of the fine-grid patch for alternative boundary conditions $I = D, P, S$ as considered in Sect. 2. Furthermore, we compute for a given macroscopic deformation $\mathbf{F}^c$ of the coarse-grid patch the fine scale nodal displacements $\mathbf{d}^f$ of the linearized problem, which subsequently induce the desired form of the prolongation operator $\mathbf{P}^f{}_c$.

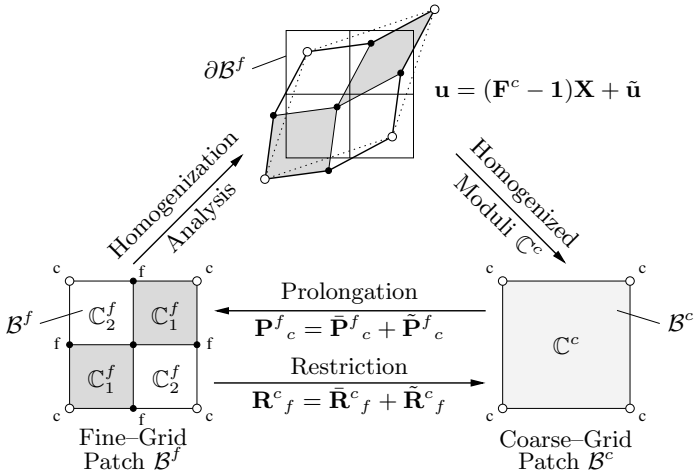


Fig. 6. Construction of transfer operators based on homogenization analysis. A numerical homogenization analysis of fine-grid finite element patches $\mathcal{B}^f$ provides the homogenized moduli $\mathbb{C}^c$ and the fluctuation part $\tilde{\mathbf{P}}^f{}_c = (\tilde{\mathbf{R}}^f{}_c)^T$ of the transfer operators

Homogenization of Linearized Patch Response

The numerical homogenization of a typical fine-grid patch $\mathcal{B}^f$ shown in Fig. 6 is based on the incremental setting of the discrete micro-to-macro variational principle

$$W^c = \inf_{\mathbf{d}^f} \tilde{W}^c(\mathbf{d}^f) \quad \text{with} \quad \tilde{W}^c(\mathbf{d}^f) = \frac{1}{|\mathcal{B}^f|} \int_{\mathcal{B}^f} W^f(\mathbf{1} + \mathbf{B}^f \mathbf{d}^f; \mathbf{X}^f)\, \mathrm{d}V \ , \quad (52)$$

in complete analogy to the formulation (14). For the linearized incremental problem associated with a typical step of the Newton algorithm outlined in Table 3, we assume a *quadratic form* of the stress potential at local points $\mathbf{X}^f \in \mathcal{B}^f$ of the fine-grid patch

$$W^f(\mathbf{F}^f; \mathbf{X}^f) = \frac{1}{2}(\mathbf{F}^f - \mathbf{1}) : \mathbb{C}^f : (\mathbf{F}^f - \mathbf{1}) \ . \quad (53)$$

Here, the fourth-order stiffness tensor $\mathbb{C}^f$ is obtained by successive homogenization based micro-to-macro transitions from fine-grid patches to coarse-grid patches. A typical transition is visualized in Fig. 6. Within the hierarchy of the multigrid meshes $l = 1, 2, \ldots, L$, we define the moduli by

$$\mathbb{C}^f := \begin{cases} \partial^2_{\mathbf{FF}} W(\mathbf{F}; \mathbf{X}^f) & \text{at the finest scale } l = 1 \ , \\ \mathbb{C}^c(\mathbf{X}^f) & \text{at all other scales} \ . \end{cases} \quad (54)$$

On the finest grid level, the tensor is given by the second derivative of the incremental stress potential W defined in (6), which characterizes the local

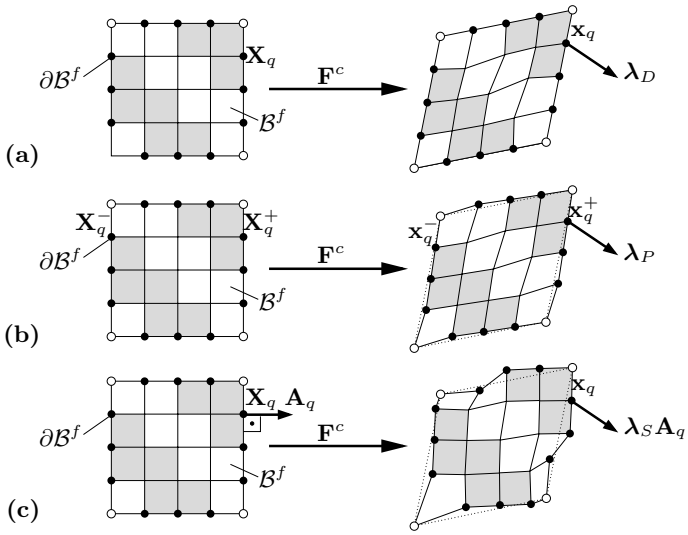


Fig. 7. Deformation-driven boundary constraints $\boldsymbol{\mathcal{A}}_I^f \mathbf{d}_b^f - \boldsymbol{\mathcal{B}}_I^f([\mathbf{F}^c] - [\mathbf{1}]) = \mathbf{0}$ with $I = D, P, S$ of finite element patches $\mathcal{B}^f$ in a multi-grid mesh hierarchy. **(a)** Prescribed deformation $I = D$, **(b)** periodic deformation $I = P$ and **(c)** uniform traction $I = S$. The driving patch deformation $\mathbf{F}^c$ is related to the coarse-grid nodal displacement $\mathbf{d}^c$

material response under consideration. On coarse grid levels, the moduli are assumed to be the homogenized moduli $\mathbb{C}^c$ of the associated grid-patch. With the quadratic potential (53) at hand, the internal nodal force vector of the fine-scale patch

$$\mathbf{f}^f := |\mathcal{B}^f| \partial_{\mathbf{d}^f} \tilde{W}^c(\mathbf{d}^f) = \int_{\mathcal{B}^f} \mathbf{B}^{fT} \mathbb{C}^f \mathbf{B}^f \mathbf{d}^f \, \mathrm{d}V \tag{55}$$

is a *linear function* of the fine-scale nodal displacements $\mathbf{d}^f$. The finite element tangent matrix of the patch takes the constant representation

$$\mathbf{K}^f := |\mathcal{B}^f| \partial^2_{\mathbf{d}^f \mathbf{d}^f} \tilde{W}^c(\mathbf{d}^f) = \int_{\mathcal{B}^f} \mathbf{B}^{fT} \mathbb{C}^f \mathbf{B}^f \, \mathrm{d}V \, . \tag{56}$$

Boundary Constraints of FE Grid Patches

With regard to the discrete formulation of the boundary constraints, the nodes of the fine-grid patch are partitioned into those on the surface $\partial\mathcal{B}^f$ and those in the interior $\mathcal{B}^f$ of the patch according to (17) and (18)

$$\mathbf{d}^f = \begin{bmatrix} \mathbf{d}_a^f \\ \mathbf{d}_b^f \end{bmatrix} \quad \text{and} \quad \mathbf{K}^f = \begin{bmatrix} \mathbf{K}_{aa}^f & \mathbf{K}_{ab}^f \\ \mathbf{K}_{ba}^f & \mathbf{K}_{bb}^f \end{bmatrix} \, . \tag{57}$$

We subject the minimization problem (52) to different boundary constraints on the fine-scale displacement $\mathbf{d}^f$ on the boundary $\partial\mathcal{B}^f$ of the fine-grid patch $\mathcal{B}^f$. Furthermore, we assume the fine-grid to be driven by a *homogeneous coarse-grid deformation* $\mathbf{F}^c$ as depicted in Fig. 7. In analogy to (28), we consider the constraints

$$\boldsymbol{\mathcal{A}}_I^f \mathbf{d}_b^f - \boldsymbol{\mathcal{B}}_I^f([\mathbf{F}^c] - [\mathbf{1}]) = \mathbf{0} \,, \tag{58}$$

on the fine-scale displacements $\mathbf{d}_b^f$ on the boundary of the patch. $I = D, P, S$ characterizes *linear deformation*, *non-trivial periodicity* and *uniform traction* constraints as visualized in Fig. 7. The associated coefficient matrices $\boldsymbol{\mathcal{A}}_I^f$ and $\boldsymbol{\mathcal{B}}_I^f$ are defined in (29).

Solution Based on Penalty Functionals

The solution of the constraint minimization problem can be performed based on penalty functionals. The discrete penalty extension of the minimization problem (52) reads

$$W_I^{c\,\epsilon}(\mathbf{F}^c) = \inf_{\mathbf{d}^f} \left\{ \tilde{W}^c(\mathbf{d}^f) \; - \; \frac{\epsilon}{2} | \, \boldsymbol{\mathcal{A}}_I^f \mathbf{d}_b^f - \boldsymbol{\mathcal{B}}_I^f([\mathbf{F}^c] - [\mathbf{1}]) \, |^2 \right\} \,, \tag{59}$$

for $I = D, P, S$. Its solution with respect to the fine-grid displacements $\mathbf{d}^f$ and the homogenized moduli $\mathbb{C}^c$ is obtained by the algorithm summarized in Table 2. For the quadratic assumption (53), we get from Table 2 the closed-form expression for the fine-scale displacements

$$\begin{cases} \mathbf{d}_a^{f\,\epsilon} = \quad \epsilon |\mathcal{B}^f| \mathbf{K}_{aa}^{f\,-1} \mathbf{K}_{ab}^f \tilde{\mathbf{K}}_{bb}^{f\,-1} \boldsymbol{\mathcal{A}}_I^{f\,T} \boldsymbol{\mathcal{B}}_I^f([\mathbf{F}^c] - [\mathbf{1}]) \,, \\ \mathbf{d}_b^{f\,\epsilon} = -\epsilon |\mathcal{B}^f| \tilde{\mathbf{K}}_{bb}^{f\,-1} \boldsymbol{\mathcal{A}}_I^{f\,T} \boldsymbol{\mathcal{B}}_I^f([\mathbf{F}^c] - [\mathbf{1}]) \,, \end{cases} \tag{60}$$

in terms of the condensed matrix $\tilde{\mathbf{K}}_{bb}^f := \mathbf{K}_{bb}^f - \epsilon |\mathcal{B}^f| \, \boldsymbol{\mathcal{A}}_I^{f\,T} \boldsymbol{\mathcal{A}}_I^f - \mathbf{K}_{ba}^f \mathbf{K}_{aa}^{f\,-1} \mathbf{K}_{ab}^f$. The homogenized moduli of the fine-scale patch derived from Table 2 appear in the form

$$[\mathbb{C}^c]_I^\epsilon := \partial_{\mathbf{F}^c} W_I^{c\,\epsilon}(\mathbf{F}^c) = -\epsilon^2 |\mathcal{B}^f| \boldsymbol{\mathcal{B}}_I^{f\,T} \boldsymbol{\mathcal{A}}_I^f \tilde{\mathbf{K}}_{bb}^{f\,-1} \boldsymbol{\mathcal{A}}_I^{f\,T} \boldsymbol{\mathcal{B}}_I^f - \epsilon \boldsymbol{\mathcal{B}}_I^{f\,T} \boldsymbol{\mathcal{B}}_I^f \,. \tag{61}$$

Sensitivity of Fine-Scale Nodal Displacements

Note that the fine-scale displacements (60) are linear functions of the coarse-grid deformation $\mathbf{F}^c$. These displacements are *fictitious quantities* associated with the linearized problem (53) and do not reflect the real nodal displacements of the patch. However, their *sensitivity* with respect to coarse-grid displacements can be used for the subsequent construction of homogenization-based transfer operators. To this end, we need to relate the coarse-grid deformation $\mathbf{F}^c$ to the coarse-grid displacements $\mathbf{d}^c$. Here, we assume the formulation

$$[\mathbf{F}^c] - [\mathbf{1}] = \bar{\mathbf{B}}^c \mathbf{d}^c \quad \text{with} \quad \bar{\mathbf{B}}^c := \frac{1}{|\mathcal{B}^c|} \int_{\mathcal{B}^c} \mathbf{B}^c(\mathbf{X}^c)\, \mathrm{d}V \,, \tag{62}$$

in terms of the *coarse-grid strain interpolation matrix* $\bar{\mathbf{B}}^c$. It is obtained by an average of the $\mathbf{B}$-matrices of the coarse-grid space associated with the patch $\mathcal{B}^c$. Insertion of the assumption (62) into the closed-form expression (60) yields the linear relationship

$$\mathbf{d}^f = [\mathbf{P}^f{}_c]^\epsilon_I \, \mathbf{d}^c \,, \tag{63}$$

between the fine-scale displacements $\mathbf{d}^f$ and the coarse-scale nodal displacements $\mathbf{d}^c$ of the patch. Thus, we obtain closed-form expressions for the sensitivity of the fine-grid-displacements with respect to the coarse-scale displacements from the penalty functional (59)

$$[\mathbf{P}^f{}_c]^\epsilon_I := \begin{cases} \epsilon |\mathcal{B}^f| \mathbf{K}^{f\,-1}_{aa} \mathbf{K}^f_{ab} \tilde{\mathbf{K}}^{f\,-1}_{bb} \boldsymbol{\mathcal{A}}^{f\,T}_I \boldsymbol{\mathcal{B}}^f_I \, \bar{\mathbf{B}}^c & \text{in } \mathcal{B}^f \,, \\ -\epsilon |\mathcal{B}^f| \tilde{\mathbf{K}}^{f\,-1}_{bb} \boldsymbol{\mathcal{A}}^{f\,T}_I \boldsymbol{\mathcal{B}}^f_I \, \bar{\mathbf{B}}^c & \text{on } \partial\mathcal{B}^f \,. \end{cases} \tag{64}$$

This sensitivity is viewed as *homogenization-based prolongation* for grid-patches in a multigrid mesh hierarchy associated with the concept visualized in Fig. 6.

3.3 Construction of Homogenization-Based Transfer Operators

With the above closed-form sensitivities for the fine-scale nodal displacements at hand, we are now able to construct effective prolongation operators for heterogeneous multigrid meshes. The prolongation (64) with the homogenization-based patch boundary constraints $I = D, P, S$ cannot be directly used in the multigrid algorithm in Table 5 because they neglect the interaction of neighboring patches. Because of the small-scale difference between two typical grids in the multigrid hierarchy, we cannot apply in a straightforward manner the argument of scale separation which underlines the homogenization theory.

Decomposition of Transfer Operators

As a consequence, we decompose the fine-scale displacements into *long-wave* and *short-wave contributions* by the additive split

$$\underbrace{\mathbf{d}^f}_{\text{total}} = \underbrace{\bar{\mathbf{d}}^f}_{\text{long-wave}} + \underbrace{\tilde{\mathbf{d}}^f}_{\text{short-wave}} \,. \tag{65}$$

The long-wave contributions are associated with the homogeneous contributions of a typical grid patch. The short-wave part represents fluctuations due to the heterogeneous structure of the patch. Assuming both contributions to be obtained by prolongation of the coarse-grid displacements $\mathbf{d}^c$

$$\bar{\mathbf{d}}^f = \bar{\mathbf{P}}^f{}_c \, \mathbf{d}^c \quad \text{and} \quad \tilde{\mathbf{d}}^f = \tilde{\mathbf{P}}^f{}_c \, \mathbf{d}^c \,, \tag{66}$$

we define for heterogeneous structures the *additive decomposition* of the prolongation operator

$$\mathbf{P}^f{}_c = \bar{\mathbf{P}}^f{}_c + \tilde{\mathbf{P}}^f{}_c \tag{67}$$

into homogeneous long-wave part $\bar{\mathbf{P}}^f{}_c$ and short-wave fluctuation part $\tilde{\mathbf{P}}^f{}_c$.

Long-Wave Contribution to Transfer Operator

The long-wave part $\bar{\mathbf{P}}^f{}_c$ should take into account the consistency of interacting grid-patches. Hence, it cannot be identified with the homogeneous part of the homogenization-based prolongation (64). As a consequence, we relate the homogeneous part to the interpolation of the coarse-grid displacements. Thus the long-wave contribution to the prolongation operator is identified with the *geometric transfer operator*

$$\bar{\mathbf{P}}^f{}_c = \bar{P}^{ia}{}_{IA} := N_I{}^c(X_i{}^f)\,\delta_{aA} \quad \text{with} \quad I = 1\dots n \;;\; i = 1\dots n_f \,. \tag{68}$$

Short-Wave Contribution to Transfer Operator

The fluctuation part $\tilde{\mathbf{P}}^f{}_c$ of the prolongation operator reflects the heterogeneity of the finite element patch under consideration. We define this contribution to be identical with the fluctuation part of the homogenization-based formulation (64). Hence, we have to subtract from (64) the homogeneous contribution. The homogeneous part of the fine-scale nodal displacements associated with the coarse-grid deformation $\mathbf{F}^c$ assumed in (62) reads

$$[\bar{\mathbf{d}}^f]_{\text{homo}} := \mathbb{D}^T([\mathbf{F}^c] - [\mathbf{1}]) = \mathbb{D}^T\bar{\mathbf{B}}^c\mathbf{d}^c \,, \tag{69}$$

with the nodal-coordinate matrix $\mathbb{D}^T$ of the finite element patch defined in analogy to (20). The associated short-wave contribution to the prolongation then simply takes the form

$$[\bar{\mathbf{P}}^f{}_c]_{\text{homo}} = \mathbb{D}^T\bar{\mathbf{B}}^c \,. \tag{70}$$

With this identification at hand, we define the short-wave contribution to the prolongation by the homogenization-based fluctuation part

$$[\tilde{\mathbf{P}}^f{}_c]_I^{\lambda,\epsilon} := [\mathbf{P}^f{}_c]_I^{\lambda,\epsilon} - [\bar{\mathbf{P}}^f{}_c]_{\text{homo}} \,, \tag{71}$$

for the three alternative boundary constraints $I = D, P, S$. Insertion of (64) gives the closed-form representations

$$[\tilde{\mathbf{P}}^f{}_c]_I^{\epsilon} := \begin{cases} [\;\; \epsilon|\mathcal{B}^f|\mathbf{K}_{aa}^{f\,-1}\mathbf{K}_{ab}^f\tilde{\mathbf{K}}_{bb}^{f\,-1}\boldsymbol{\mathcal{A}}_I^{fT}\boldsymbol{\mathcal{B}}_I^f - \mathbb{D}_a^T\,]\,\bar{\mathbf{B}}^c & \text{in } \mathcal{B}^f \,, \\ [\,-\epsilon|\mathcal{B}^f|\tilde{\mathbf{K}}_{bb}^{f\,-1}\boldsymbol{\mathcal{A}}_I^{fT}\boldsymbol{\mathcal{B}}_I^f - \mathbb{D}_b^T\,]\,\bar{\mathbf{B}}^c & \text{on } \partial\mathcal{B}^f \,. \end{cases} \tag{72}$$

According to (48), we set $[\tilde{\mathbf{R}}^f{}_c]_I^{\lambda,\epsilon} := ([\tilde{\mathbf{P}}^f{}_c]_I^{\lambda,\epsilon})^T$. This represent the desired homogenization-based transfer operator.

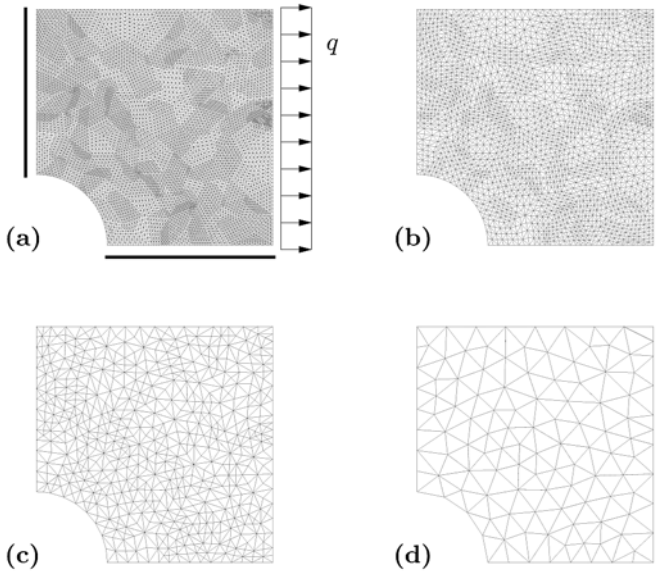


Fig. 8. Heterogeneous disc with a hole. Multi-grid method with four hierarchical grid levels. **(a)** Fine-grid with geometry and boundary conditions, **(b)**–**(d)** coarse-grid levels

3.4 Multigrid Model Computations of Nonlinear Composites

In order to demonstrate the efficiency of the new transfer operators with homogenization-based fluctuation part (72), we investigate representative model problems and compare the results with computations based on standard transfers. The abbreviations *dis*, *per* and *stress* stand for the homogenization-based transfer operators $I = D, P, S$ associated with linear displacement, periodic displacements and homogeneous stresses on the boundary, respectively. *geom* indicates the geometrical, *schur* the local Schur-complement-based, *smooth* the recovery-based and *alg* the algebraic transfer. *inject* is the abbreviation for injection. The efficiency of these different transfers applied to heterogeneous structures is reported by numerical convergence analyses and computing times. We again focus on two-phase composites with elastic and inelastic constituents at finite deformations. The basic constitutive functions are identical to those used in Sect. 2.4 as outlined in (38) and (39).

Elastic Heterogeneous Specimen with Centered Hole

The first example in Fig. 8 treats a heterogeneous specimen with centered hole which is discretized by an unstructured finite element mesh hierarchy. Due to the symmetry, we discretize only a quarter of the specimen of dimension 10x10 by using constant strain triangles. As shown in Fig. 9, we define

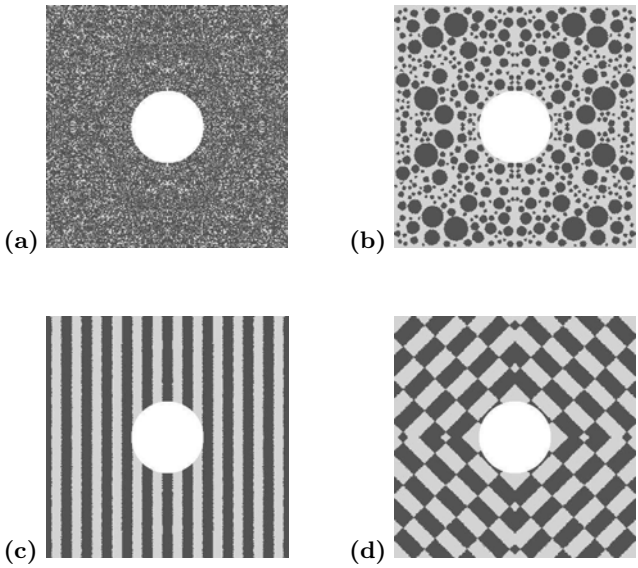


Fig. 9. Heterogeneous disc with a hole. Structures with **(a)** randomly distributed heterogeneity, **(b)** randomly distributed fibers, **(c)** laminates and **(d)** checkerboard patterns

different topologies of micro-heterogeneity in order to investigate the influence of the micro-structure on the multigrid convergence behavior. The material model for the constituents is governed by the elastic free energy function (38). For the stiff elastic inclusions we choose the material parameters $\kappa_{stiff} = 175$ and $\mu_{stiff} = 80$. The soft elastic matrix gets the values $\kappa_{soft} = \kappa_{stiff}/100$ and $\mu_{soft} = \mu_{stiff}/100$. We apply a tension load $q = 1$ along the right boundary in the plain-strain context as depicted in Fig. 8a. The left and the lower boundary are kept fixed in the horizontal and vertical direction, respectively. We consider a four-grid method with meshes of 22300, 5184, 1150 and 222 degrees as depicted in Figs. 8a-d. Figure 8d shows the coarsest mesh of level $l = 4$, which is hierarchically refined down to the level $l = 1$ in Fig. 8a. This level is assumed to resolve the heterogeneous structure. Four typically heterogeneous micro-structures are visualized in Figs. 9a-d. We consider (i) randomly distributed heterogeneity, (ii) randomly distributed fibers, (iii) laminates and (iv) checkerboard patterns. The distribution of the heterogeneity is obtained by specific density functions evaluated at the mid-points of the elements. In the case (i), we assign to each element on the finest mesh a number in the range 0.0-1.0 by means of a random number generator. For numbers greater than 0.5, the element is provided with stiff otherwise with soft material parameters. For micro-structure (ii) the diameter of the fibers is chosen randomly in the range 0.1-1.0. Furthermore, we randomly generate the locations of the

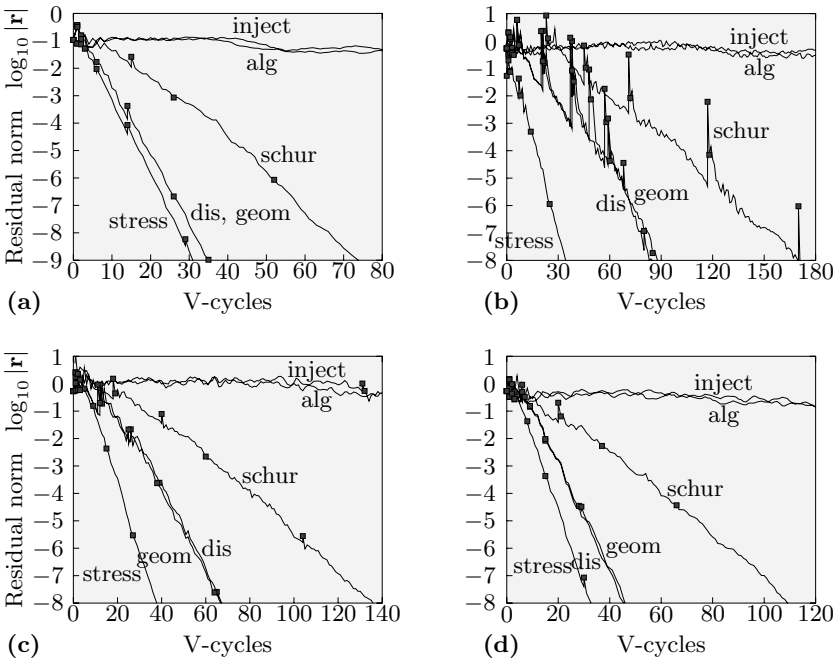


Fig. 10. Heterogeneous disc with a hole. Convergence study of tension test for finite elastic deformations of micro-structures with **(a)** randomly distributed heterogeneity, **(b)** randomly distributed fibers, **(c)** laminates and **(d)** checkerboard patterns. The black squares represent the end of a Newton iteration step

Table 8. Heterogeneous disc with a hole. Ratio of computation time relative to transfer method *stress* for elastic micro-structures at finite deformations

	Multi-grid–iteration	Transfer–matrices	Fine–grid matrix	Coarse–grid matrices	Comput. time
	Micro-structure with randomly distributed heterogeneity				
dis	0,94	0,19	0,75	1,00	0,86
geom	0,96	0,19	0,74	1,00	0,87
schur	2,28	0,70	0,87	1,17	1,85
	Micro-structure with randomly distributed fibers				
dis	1,64	0,19	1,16	0,82	1,39
geom	1,57	0,19	1,21	0,90	1,36
schur	3,54	0,70	1,26	0,97	2,75
	Micro-structure with laminates				
dis	1,50	0,19	0,94	0,71	1,27
geom	1,50	0,19	0,97	0,71	1,27
schur	3,44	0,70	0,88	0,85	2,71
	Micro-structure with checkerboard patterns				
dis	1,17	0,19	1,11	0,72	1,04
geom	1,17	0,19	1,00	0,72	1,03
schur	3,10	0,70	1,32	1,24	2,56

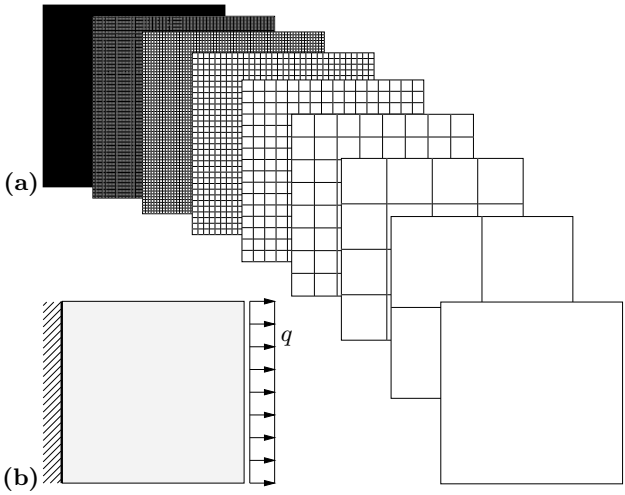


Fig. 11. Extension of heterogeneous strip. Multi-grid method with nine hierarchical grid-levels. **(a)** Regular discretizations of the grid-levels with $Q1P0$-elements. **(b)** Boundary condition with respect to the finest discretization

fibers in such a way, that intersections do not occur. Finite elements, which are covered by these circular inclusions are provided with stiff, the remaining elements with soft material parameters. In the cases of the aligned (iii) and the checkerboard patterns (iv), we cover the domain by a specific sinus function, which is evaluated at the element mid-points. Elements with negative value of the density function are provided with stiff, others with soft material parameters. Finite elements which are covered by the stiffer material appear shaded in Figs. 9a-d. The system of linear equations on the finest grid level is solved by a multigrid preconditioned CG-accelerated iteration. The tolerance of the residual is chosen to be $\theta = 10^{-8}$. A four-grid method with $V(3,3)$-cycles is considered. Pre-smoothing is performed by Block-Backward Gauss-Seidel and post-smoothing by Block-Forward Gauss-Seidel three times on each grid level. In Figs. 10a-d, the convergence behavior of the four-grid method with $V(3,3)$-cycles and CG-acceleration is documented. The new homogenization-based transfer operator *stress* shows a significantly better convergence behavior than the other transfers. The curves in Fig. 10a associated with the homogenization-based and the geometrical transfer are close to each other.

The efficiency of the homogenization-based transfer depends significantly on anisotropy effects. A strongly inhomogeneous distribution of stiffness in the relevant finite element patch affects the multigrid convergence behavior. In contrast, the influence of randomly distributed micro-heterogeneity is smoothed out by the homogenization, yielding an isotropic response. In the diagrams presented in Figs. 10a-d, the local Schur complement, the algebraic transfer and the injection yield worse convergence behavior when compared to others. This tendency is also observed by the required computing times in

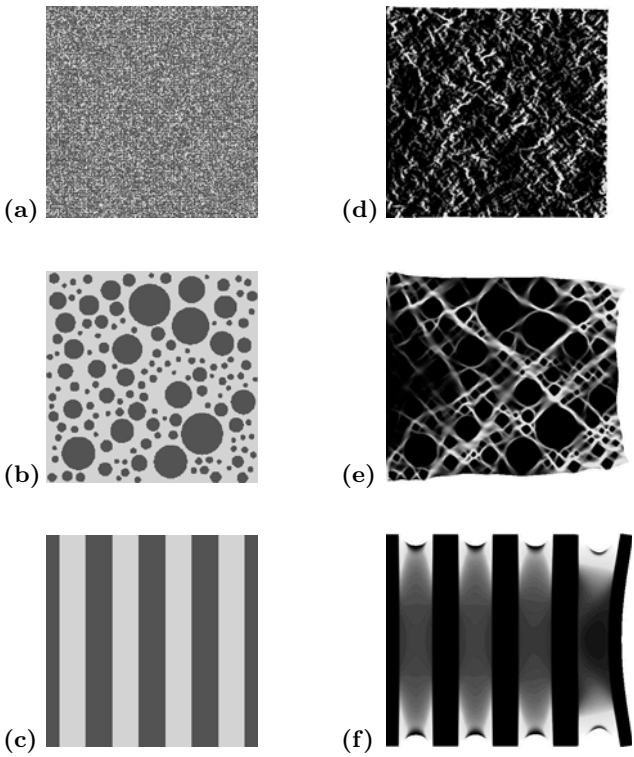


Fig. 12. Extension of heterogeneous strip. Heterogeneous structures with **(a)** randomly distributed heterogeneity, **(b)** randomly distributed fibers and **(c)** laminates. **(d)**–**(f)** Associated equivalent plastic strains with maximum values in light grey

Table 8. Observe that the homogenization-based transfer *stress* and the Schur complement transfer *schur* are expensive with regard to the computation time due to supplementary fill-ins in the coarse-grid matrices. Despite the excellent convergence behavior of *stress* and *schur*, they yield higher computation times for randomly distributed micro-heterogeneity than for the transfers *dis* and *geom*. In Fig. 10, black squares represent the end of Newton iterations and the assembly of a new fine-grid matrix. The arrangement of the symbols reflects the quadratic convergence. The numerical examples demonstrate in the case of anisotropy effects, that the new stress-induced transfer *stress* is distinctly superior to alternative transfer types. Recovery-based transfer operators yield a bad conditioning of coarse-grid matrices and a failure of the multigrid method. Regarding the convergence behavior, the transfer types *schur*, *smooth*, *alg* and *inject* are not acceptable for simulations of heterogeneous composites.

Inelastic Heterogeneous Strip in Tension

The second example in Fig. 11 deals with inelastic composites. The geometry and boundary conditions are depicted in Fig. 11b. The specimen has the dimensions 1x1, is clamped at the left boundary and is subjected to a tensile load q along the right boundary in a plain-strain context. The domain is regularly discretized by mixed $Q1P0$-elements. Four neighboring elements arranged at the fine-grid level represent one element on the next higher grid level. We consider a hierarchical nine-grid method. Figure 11a depicts the nine hierarchical grid levels starting with the finest up to the coarsest mesh. We use meshes with 131584, 33024, 8320, 2112, 544, 144, 40, 12 and 4 degrees. The finest mesh resolves the heterogeneity element-wise. The heterogeneous structures considered and the developed equivalent plastic strains in the deformed state are visualized in Figs. 12a-c and d-f, respectively. In Figs. 12a-c, stiff material properties are shaded. The distribution of the heterogeneity is realized similar to the previous example by density functions. The new transfer operators are analyzed for a von Mises-type plasticity model. The material parameters are the bulk modulus $\kappa_1 = 175.00$ and the shear modulus $\mu_1 = 80.00$ for the stiff elastic inclusions, $\kappa_2 = 17.50$ and $\mu_2 = 8.00$ for the elastic matrix, the flow stress $y_0 = 0.45$, the saturation stress $y_\infty = 0.45$, the linear hardening parameter $h = 1.00$ and the saturation parameter $\omega = 0.01$ for the inelastic matrix material. The resulting linear equation system of the finite element discretization is solved on the finest mesh via a nine-grid method with $V(3,3)$-cycles and its CG-accelerated version. The tolerance of the residual is chosen to be $\theta = 10^{-5}$. Again, pre-smoothing is performed by Block-Backward Gauss-Seidel and post-smoothing by Block-Forward Gauss-Seidel three times on each grid level. The load q is increased from 0 until the final loading state by constant load increments of 0.1 for all investigated models. The final loading state for the structures visualized in Figs. 12a and b is set to $q = 5$, for the structure depicted in Fig. 12c to $q = 3$ and for the structure in Fig. 12d to $q = 1$. The convergence behavior associated with different transfer operators is documented along with the required number of nonlinear Newton iterations in Fig. 13. Black squares representing the end of a loading step.

The graphs demonstrate that the homogenization-based stress induced transfer *stress* can be regarded as optimal. Regarding efficiency, it directly follows the transfer *dis* with linear patch boundary displacements. The bilinear displacement-type elements used are not able to resolve a homogeneous deformation. As the transfer *stress* implies an increase of fill-ins in the coarse-grid matrices, it may so happen, that the transfer *dis* is at least equivalent to *stress* regarding the numerical efficiency. It should be pointed out that the consistency between the Garlekin-operator and operator built from effective moduli is not ensured for bilinear ansatz functions. Hence, the robustness for the *stress* transfer type is not generally guaranteed. Furthermore, the local Schur-complement-based transfer *schur*, the algebraic transfer *alg* as well as the injection *inject* do not converge or show an extremely bad convergence be-

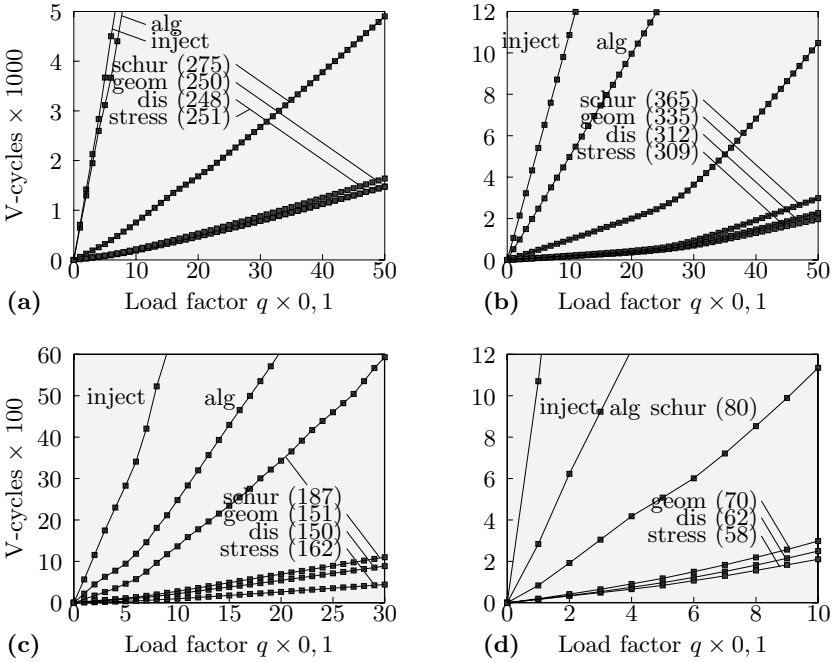


Fig. 13. Extension of heterogeneous strip. Convergence study of inelastic microstructure with **(a)** randomly distributed heterogeneity, **(b)** randomly distributed fibers, **(c)** laminates and **(d)** checkerboard pattern. Numbers in brackets report the Newton-iteration steps. Black squares represent the end of a loading step

Table 9. Extension of heterogeneous strip. Ratio of computation time relative to transfer method *stress* for inelastic micro-structures

	Multi-grid–iteration	Transfer-matrices	Fine–grid matrix	Coarse–grid matrices	Comput. time
	Micro-structure with randomly distributed heterogeneity				
dis	0,90	0,08	0,95	0,92	0,91
geom	1,06	0,07	1,05	1,00	1,04
schur	2,64	1,06	1,07	1,49	2,25
	Micro-structure with randomly distributed fibers				
dis	1,03	0,08	1,07	0,81	0,95
geom	1,22	0,55	1,09	0,78	1,07
schur	4,19	1,05	1,35	1,44	3,19
	Micro-structure with laminates				
dis	1,34	0,61	0,89	0,66	1,03
geom	1,47	0,61	0,94	0,67	1,10
schur	6,50	0,92	1,12	1,22	4,02
	Micro-structure with checkerboard patterns				
dis	0,71	0,07	0,84	0,55	0,64
geom	1,65	0,06	1,39	0,95	1,32
schur	4,28	0,57	1,05	1,10	2,77

havior. Alternative recovery based multigrid methods *smooth* are not suitable for heterogeneous materials. The numerical investigations demonstrate that homogenization-based transfer operators with linear patch boundary displacements are distinctly superior to alternative transfer types when bilinear ansatz functions are involved. In summary, the homogenization-based transfer operators distinctly yield a better convergence behavior than standard transfer methods. Furthermore, the transfers *schur*, *smooth*, *alg* and *inject* turn out to be not suitable for simulations of heterogeneous materials. The numerical investigations show that the new homogenization based transfer operators, coupled with periodic and stress boundary constraints, are very efficient for three-noded finite elements in the case of anisotropy due to the heterogeneity. Otherwise, transfer operators coupled with linear displacement boundary constraints are recommended.

4 Conclusion

We outlined new computational procedures for analyses of heterogeneous composites with and without scale separation in the fully nonlinear range. For problems with scale separation, we developed a computational setting of strain-driven homogenization for representative aggregates with respect to different boundary constraints of the micro-structure. The algorithm summarized in Table 2 based on a penalty form for the averaged incremental energy is suitable for a large spectrum of practical applications. For problems without scale separation, we developed a specific multigrid solution procedure for composites based on the new homogenization-based geometric transfer operator (72). It provides an optimal transfer for heterogeneous highly anisotropic grid patches and reduces in the case of homogeneous materials to the standard geometrical transfer. It was shown that the formulations of computational homogenization for problems with scale separation inspire the developments on multigrid methods for problems without scale separation. Both settings were outlined in a compact format based on incremental variational formulations. The proposed new formulations provide variational-based homogenization algorithms for physical multiscale scenarios and problem-dependent optimal finite element grid transfers for numerical multiscale scenarios of heterogeneous materials. Representative model simulations demonstrated their excellent performance.

Acknowledgment. We thank Manuel Birkle for his support and many helpful discussions.

References

1. R. E. Alcouffe, A. Brandt, J. E. Dendy, and J. W. Painter. The multigrid method for the diffusion equation with strongly discontinous coefficients. *SIAM Journal of Scientific Statistical Computation*, 2:430–454, 1981.
2. P. Bastian, W. Hackbusch, and G. Wittum. Additive and multiplicative multigrid – a comparison. *Computing*, 60:345–364, 1998.
3. C. G. Bayreuther. *Mehrskalenmodelle in der Festkörpermechanik und Kopplung von Mehrgittermethoden mit Homogenisierungsverfahren.* PhD thesis, Institut für Mechanik (Bauwesen), Report No. I-16, Universität Stuttgart, 2005.
4. C. G. Bayreuther and C. Miehe. Coupling of homogenization techniques with multigrid solvers for unstructured meshes. In W. Wendland and M. Effendiev, editors, *Lecture Notes in Applied and Computational Mechanics*, pages 67–72. Springer-Verlag, 2003.
5. C. G. Bayreuther and C. Miehe. Construction of multigrid transfer operators for heterogeneous materials with arbitrary microstructures based on homogenization techniques. *Computers & Structures*, 2006. submitted to.
6. C. G. Bayreuther, C. Miehe, and J. Schröder. Aspects of homogenization techniques and multigrid solving. In A. Sändig, W. Schiehlen, and W. Wendland, editors, *Multifield Problems – state of the art*, pages 88–95. Springer-Verlag, 2000.
7. A. Bensoussan, J. L. Lions, and G. Papanicolaou. Asymptotic analysis for periodic structures. In J. Lions, G. Papanicolaou, and R. Rockafellar, editors, *Studies in mathematics and its applications, Vol. 5.* North-Holland Publishing Company, 1978.
8. M. A. Biot. *Mechanics of incremental deformations.* John Wiley & Sons, Inc., New York, 1965.
9. D. Braess. Towards algebraic multigrid for elliptic problems of second order. *Computing*, 55:379–393, 1995.
10. A. Brandt. Guide to multigrid development. In W. Hackbusch and U. Trottenberg, editors, *Multigrid Methods, Vol. 960*, Lecture notes in mathematics. Springer-Verlag, 1982.
11. M. Crouzeix and P. A. Raviart. Conforming and nonconforming finite element methods for solving the stationary stokes equations. *Revue Francaise D'Automatique, d'Informatique et de Recherche Operationelle / Mathematical Modelling and Numerical Analysis*, 7:33–76, 1973.
12. P. M. De, Zeeuw. Matrix-dependent prolongations and restrictions in a black box multigrid solver. *Journal of computational and applied mathematics*, 33:1–27, 1990.
13. J. Fish and V. Belsky. Multigrid method for periodic heterogeneous media. part 1: convergence studies for one-dimensional case. *Computer Methods in Applied Mechanics and Engineering*, 126:1–16, 1995.
14. J. Fish and V. Belsky. Multigrid method for periodic heterogeneous media. part 2: multiscale modeling and quality control in multidimensional case. *Computer Methods in Applied Mechanics and Engineering*, 126:17–38, 1995.
15. P. Germain. *Cours de mécanique des milieux continus.* Masson et Cie, Paris, 1973.

16. S. Gosh, K. Lee, and S. Moorthy. Two-scale analysis of heterogeneous elastic-plastic materials with asymptotic homogenization and voronoi cell finite element model. *Computer Methods in Applied Mechanics and Engineering*, 132:63–116, 1996.
17. W. Hackbusch. *Multigrid methods and applications*. Springer-Varlag, 1985.
18. W. Hackbusch and S. Sauter. Composite finite elements for problems containing small geometric details. part ii: implementation and numerical results. *Computing and Visualization in Science*, 1:15–25, 1997.
19. W. Hackbusch and S. Sauter. Composite finite elements for the approximation of pdes on domains with complicated micro-structures. *Numerische Mathematik*, 75:447–472, 1997.
20. B. Halphen and Q. S. Nguyen. Sur les matéraux standards généralisés. *Journal de Mécanique*, 40:39–63, 1975.
21. Z. Hashin. Analysis of composite materials – a survey. *ASME Journal of Applied Mechanics*, 50:481–505, 1983.
22. R. Hill. On constitutive macro-variables for heterogeneous solids at finite strain. *Proceedings of the Royal Society of London (Serie A)*, 326:131–147, 1972.
23. V. Kouznetsova, T. J. R. Brekelmans, and F. T. P. Baaijens. An approach to micro-macro modelling of heterogeneous materials. *Computational Mechanics*, 27:37–48, 2001.
24. J. Michel, H. Moulinec, and P. Suquet. A computational method based on augmented lagrangians and fast fourier transforms for composites with high contrast. *Computer Modeling in Engineering and Sciences*, 1(2):79–88, 2000.
25. C. Miehe. A formulation of finite elastoplasticity based on dual co- and contra-variant eigenvector triads normalized with respect to a plastic metric. *Computer Methods in Applied Mechanics and Engineering*, 159:223–260, 1998.
26. C. Miehe. Strain-driven homogenization of inelastic microstructures and composites based on an incremental variational formulation. *International Journal for Numerical Methods in Engineering*, 55:1285–1322, 2002.
27. C. Miehe. Computational micro-to-macro transitions for discretized micro-structures of heterogeneous materials at finite strains based on the minimization of averaged incremental energy. *Computer Methods in Applied Mechanics and Engineering*, 192:559–591, 2003.
28. C. Miehe, N. Apel, and M. Lambrecht. Anisotropic additive plasticity in the logarithmic strain space. modular kinematic formulation and implementation based on incremental minimization principles for standard materials. *Computer Methods in Applied Mechanics and Engineering*, 191:5383–5425, 2002.
29. C. Miehe and C. G. Bayreuther. Multiscale finite element analyses for heterogeneous structures: From homogenization to multigrid solvers. *International Journal for Numerical Methods in Engineering*, page submitted to, 2006.
30. C. Miehe, J. Schotte, and M. Lambrecht. Homogenization of inelastic solid materials at finite strains based on incremental minimization principles. *Journal of the Mechanics and Physics of Solids*, 50:2123–2167, 2002.
31. H. Moulinec and P. Suquet. A numerical method for computing the overall response of nonlinear composites with complex microstructure. *Computer Methods in Applied Mechanics and Engineering*, 157:69–94, 1998.
32. S. Müller. Homogenization of nonconvex integral functionals and cellular elastic materials. *Archive of Rational Mechanics and Analysis*, 99:189–212, 1987.

33. S. Nemat-Nasser and M. Hori. *Micromechanics: overall properties of heterogeneous materials*. North-Holland series in applied mathematics and mechanics, Vol. 36. Elsevier Science Publisher B.V., 2. edition, 1999.
34. N. Neuß. *Homogenisierung und Mehrgitter*. PhD thesis, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, 1995. Berichtnr. N96/7.
35. Q. S. Nguyen. *Stability and nonlinear solid mechanics*. John Wiley & Sons, LTD, Chichester, 2000.
36. P. Ponte Castañeda. The effective mechanical properties of nonlinear isotropic composites. *Journal of the Mechanics and Physics of Solids*, 39:45–71, 1991.
37. P. Ponte Castañeda and P. Suquet. Nonlinear composites. *Advances in Applied Mechanics*, 34:171–303, 1998.
38. J. W. Ruge and K. Stüben. Algebraic multigrid. In S. McCormick, editor, *Multigrid methods*. SIAM, 1987.
39. E. Sanchez-Palenzia. Nonhomogeneous media and vibration theory. *Lecture Notes in Physics*, 127:193–278, 1980.
40. R. J. M. Smit, W. A. M. Brekelmans, and H. E. H. Meijer. Prediction of the mechanical behavior of nonlinear heterogeneous systems by multi-level element modeling. *Computer Methods in Applied Mechanics and Engineering*, 155:181–192, 1998.
41. P. M. Suquet. Elements of homogenization for inelastic solid mechanics. In E. Sanchez-Palenzia and A. Zaoui, editors, *Lecture Notes in Physics: Homogenization Techniques for Composite Materials*, pages 193–278. Springer–Verlag, vol. 272 edition, 1987.
42. K. Terada and T. Kikuchi. A class of general algorithms for multi-scale analyses of heterogeneous media. *Computer Methods in Applied Mechanics and Engineering*, 190:5427–5464, 2001.
43. P. Vaněk, J. Mandel, and M. Brezina. Algebraic multigrid based on smoothed aggregation for second and fourth order problems. *Computing*, 56:179–196, 1996.
44. P. Vaněk, J. Mandel, and M. Brezina. Algebraic multigrid on unstructured meshes. Technical report, Center for Computational Mathematics, University of Colorado at Denver, 1996.
45. P. Wesseling. *http://www.mgnet.org/mgnet-books-wesseling.html An introduction to multigrid methods*. John Wiley & Sons, 1992.
46. J. R. Willis. Variational and related methods for the overall properties of composites. *Advances in Applied Mechanics*, 21:1–78, 1981.
47. H. Ziegler. Some extremum principles in irreversible thermodynamics with application to continuum mechanics. In I. N. Sneddon and R. Hill, editors, *Progress in solid mechanics*. North–Holland Publishing Company, Amsterdam, vol. iv edition, 1963.

A Mathematical Framework for Generalized Standard Materials in the Rate-Independent Case*

Alexander Mielke[1,2]

[1] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany
mielke@wias-berlin.de
[2] Department of Mathematics, Humboldt-Universität zu Berlin, Rudower Chaussee 25, 12489 Berlin-Adlershof, Germany

Summary. Generalized standard materials are described by an elastic energy density and a dissipation potential. The latter gives rise to the evolution equation (flow law) for the internal variables. The energeti cformulation provides a very weak, derivative-free form of this flow law. It is based on a global stability condition and an energy balance. Using time-incremental minimization problems, which allow for the usage of the rich theory in the direct method of the calculus of variations, it is possible to establish general, abstract existence results as well as convergence for numerical approximations. Applications to shape-memory materials and to magnetostrictive or piezoelectric materials are surveyed.

Keywords: Rate independence, energetic formulation, Gamma convergence, relaxation, shape-memory material, magnetostriction, piezoelectricity

1 Introduction

On the mechanical side the theory of generalized standard materials was developed in the early 1970s, see [29, 63, 68, 80]. The mathematics for these models was studied in parallel, but was mostly restricted to the case of convex potentials with applications in small-strain elastoplasticity, cf. [32, 64].

The theory of rate-independent hysteresis operators advanced much further, see [11, 36, 37, 79], mainly in the field of scalar-valued hysteresis operators. In parallel, the mathematical theory of solid mechanics had major breakthroughs in the treatment of finite-strain elastostatics [6, 13] and in the study of microstructures in modern materials [7, 66].

* Project C7 "Continuum Mechanical Models with Hysteresis Effects"

The theory presented here is located in a triangle that has its corners in the rich area of existing engineering models, in the theory of hysteresis models, and in the methods of calculus of variations that were derived for nonconvex material models. The major fact is that rate independence is still so close to statics that very similar methods can be employed. Nevertheless it allows us to study evolutionary effects on slow timescales.

In Sect. 2 we will present the theory of generalized standard materials and will show how these models are linked to the so-called energetic formulation. In Sect. 3 we summarize the existence theory for energetic solutions developed in a quite abstract setting, see [22, 43, 48].

In Sect. 4 we discuss the question of approximation of the energetic formulation. Based on abstract Γ-convergence ideas it is possible to derive convergence results for numerical approximations via finite-element methods, see [38, 54]. Moreover, homogenization results are established, see [62]. Finally, a relaxation result is presented that is due to [38, 56].

The final section is devoted to a list of several applications. The whole work was initiated through the need for a better understanding of the hysteretic evolution of microstructures in shape-memory alloys [57, 59]. In the Sects. 5.1 to 5.4 we report on more recent developments of the analysis of different models. Further applications occur in damage [20, 55], in delamination [35] and in brittle fracture [12, 15, 21]. The modeling of ferroelectric and magnetostricitive materials also fits into this framework, see Sects. 5.5 and 5.6. Moreover, the theory of elastoplasticity has been one of the major driving forces of the theory of rate-independent processes. The recent advances in this topic will be surveyed in another article of this volume, see [26].

2 Modeling Materials with Internal Variables

2.1 Generalized Standard Materials

This theory was developed in [29, 80] and plays, nowadays, a central rôle in the area of material modeling on the phenomenological level, see [23, 27, 45] for some recent references.

We consider an elastic body with reference domain $\Omega \subset \mathbb{R}^d$. The deformation $\varphi : \Omega \to \mathbb{R}^d$ gives rise to the strain tensor $F = \nabla\varphi$. We assume that the state at a material point $x \in \Omega$ is described by $F \in \mathbb{R}^{d\times d}$ and a further variable $z \in Z$ which is often called internal variable. Here z may denote plastic variables, damage, magnetization, polarization or some phase indicator. The admissible set Z is in general a submanifold (with boundary) of $\mathbb{R}^m$ for some $m \in \mathbb{N}$.

The material behavior is described by two constitutive functions, the stored-energy density $W = \widehat{W}(x, F, z)$ (also called elastic potential) and the dissipation potential $R = \widehat{R}(x, z, \dot{z})$. While W is the potential for the stress-strain relation, R is the potential for the dissipational forces versus the rate

$\dot z$, viz.,

$$T = \tfrac{\partial}{\partial F}\widehat{W}(x,F,z) \text{ and } f_{\rm diss} = -\tfrac{\partial}{\partial \dot z}\widehat{R}(x,z,\dot z)\ .$$

The time evolution of the material is now described by the quasistatic elastic equilibrium

$$-\mathrm{div}\left(\tfrac{\partial}{\partial F}\widehat{W}(x,\nabla\varphi,z)\right) = f_{\rm ext} \text{ plus boundary conditions}$$

and by the flow law for the internal variable which involves the thermodynamically conjugated driving force $X_Z = -\frac{\partial}{\partial z}\widehat{W}(x,F,z)$, viz.,

$$-(f_{\rm diss}+X_z) = 0 = \tfrac{\partial}{\partial \dot z}\widehat{R}(x,z,\dot z) + \tfrac{\partial}{\partial z}\widehat{W}(x,\nabla\varphi,z)\ .$$

Rate independence means that $\widehat{R}(x,z,\cdot)$ is homogeneous of degree 1, i.e., the dissipation is proportional to the rate $v=\dot z$. Then, $\frac{\partial}{\partial \dot z}\widehat{R}$ has to be understood as the multi-valued subdifferential of convex analysis

$$\partial_{\dot z}\widehat{R}(x,z,v) = \{\,\eta\in \mathrm{T}_z^*Z \mid \forall\, w\in \mathrm{T}_zZ\colon \widehat{R}(x,z,w)\ge \widehat{R}(x,z,v)+\langle \eta, w-v\rangle\,\}\ .$$

To provide a mathematical framework we introduce $\mathcal{F}$ as the set of admissible deformations, which typically is an affine subspace of some Sobolev space $\mathrm{W}^{1,p}(\Omega,\mathbb{R}^d)$ due to the Dirichlet boundary conditions. Moreover, we let $\mathcal{Z}=\mathrm{L}^1(\Omega,Z)$ for the function space of admissible internal states. For the state space $\mathcal{Q}=\mathcal{F}\times\mathcal{Z}$ we set $q=(\varphi,z)$ and

$$\begin{aligned}\mathcal{E}(t,q) &= \textstyle\int_\Omega \widehat{W}(x,\nabla\varphi,z)\,\mathrm{d}z - \int_\Omega f_{\rm ext}(t,x)\cdot\varphi(x)\,\mathrm{d}x\ ,\\ \mathcal{R}(z,\dot z) &= \textstyle\int_\Omega \widehat{R}(t,z(x),\dot z(x))\,\mathrm{d}x\ .\end{aligned}$$

Here $\mathcal{E}$ is the *energy storage functional* and $\mathcal{D}$ is the global dissipation function. The evolutionary problem takes the form

$$\begin{aligned}&\mathrm{D}_\varphi\mathcal{E}(t,\varphi(t),z(t)) = 0,\\ &0\in\partial_{\dot z}\mathcal{R}(z(t),\dot z(t)) + \mathrm{D}_z\mathcal{E}(t,\varphi(t),z(t)),\quad z(0)=z_0\ .\end{aligned}\tag{2.1}$$

2.2 The Energetic Formulation

In general the manifold $Z\subset\mathbb{R}^m$ might be complicated and the definition of $\dot z$ might be nontrivial. Moreover, in rate-independent systems it is to be expected that solutions develop jumps. Hence, it is desirable to find a weaker formulation avoiding derivatives. To this end we introduce the *dissipation distance* $D(x,\cdot,\cdot): Z\times Z\to[0,\infty]$ which is associated with the Finslerian dissipation metric $\widehat{R}(x,\cdot,\cdot): \mathrm{T}Z\to[0,\infty]$, viz.,

$$D(x,z_0,z_1) = \inf\{\,\textstyle\int_0^1 \widehat{R}(x,\widetilde z(s),\dot{\widetilde z}(s))\mathrm{d}s \mid \widetilde z\in \mathrm{C}^1([0,1],Z), \widetilde z(0)=z_0, \widetilde z(1)=z_1\,\}\ .$$

On $\mathcal{Z}$ this induces the distance $\mathcal{D}$ with $\mathcal{D}(z_0, z_1) = \int_\Omega D(x, z_0(x), z_1(x))\,\mathrm{d}x$, and we are able to define the dissipation along an arbitrary path $z : [0,T] \to \mathcal{Z}$ via

$$\mathrm{Diss}_{\mathcal{D}}(z,[s,t]) = \sup\{\sum_{j=1}^{N} \mathcal{D}(z(t_{j-1}), z(t_j)) \mid N \in \mathbb{N}, s \le t_0 < t_1 < \cdots t_N \le t\}\ .$$

For smooth paths this definition is compatible with the classical dissipation

$$\mathrm{Diss}_{\mathcal{D}}(z,[s,t]) = \int_s^t \mathcal{R}(z(\tau), \dot z(\tau))\,\mathrm{d}\tau = \int_s^t \int_\Omega \widehat{R}(x, z(\tau,x), \dot z(\tau,x))\,\mathrm{d}x\,\mathrm{d}\tau\ .$$

Our weak form of (2.1) is the *energetic formulation* involving the *stability condition* (S) and the *energy balance* (E). A process $q = (\varphi, z) : [0,T] \to \mathcal{F} \times \mathcal{Z} = \mathcal{Q}$ is called *energetic solution* for $(\mathcal{E}, \mathcal{D})$, if for all $t \in [0,T]$ we have

$$\begin{array}{ll} \textbf{(S)} & q(t) \in \mathcal{S}(t) \stackrel{\mathrm{def}}{=} \{\, q \in \mathcal{Q} \mid \mathcal{E}(t,q) < \infty,\ \forall\, \widetilde q \in \mathcal{Q}\colon \mathcal{E}(t,q) \le \mathcal{E}(t,\widetilde q) + \mathcal{D}(q,\widetilde q)\,\} \\ \textbf{(E)} & \mathcal{E}(t,q(t)) + \mathrm{Diss}_{\mathcal{D}}(q,[0,t]) = \mathcal{E}(0,q(0)) + \int_0^t \partial_s \mathcal{E}(s,q(s))\,\mathrm{d}s\ . \end{array}$$

Here $\partial_s \mathcal{E}(s,q(s)) = \frac{\partial}{\partial s}\mathcal{E}(s,q(s))$ is called the power of the external forces and we implicitly assume that $t \mapsto \partial_t \mathcal{E}(t,q(t))$ lies in $\mathrm{L}^1((0,T))$.

In the case that $\mathcal{Q}$ is a Banach space, that $\mathcal{E}$ and $\mathcal{R}$ are Gateaux differentiable and that the energetic solution q lies in $\mathrm{W}^{1,1}([0,T],\mathcal{Q})$ it is easy to see that (S) implies $\mathrm{D}_\varphi \mathcal{E}(t,\varphi(t),z(t)) = 0$ and $O \in \partial_{\dot z}\mathcal{R}(z(t),0) + \mathrm{D}_z\mathcal{E}(t,\varphi(t),z(t))$. Moreover, differentiating (E) with respect to the time t yields $\mathrm{D}_z\mathcal{E}(t,\varphi(t),z(t))[\dot z(t)] + \mathcal{R}(z(t),\dot z(t)) = 0$. This is exactly (2.1). In the case that $\mathcal{E}(t,\cdot)$ is strictly convex on the Banach space $\mathcal{Q}$ and that $\mathcal{R}$ does not depend on z, it is shown in [58] that (2.1) is in fact equivalent to (S) & (E). See also [19, 51] for more general results on this equivalence.

However, as we are mostly interested in nonconvex models we will mainly focus on the energetic formulation (S) & (E). Note that a significant simplification occurs due to the fact that (S) is a purely static condition.

2.3 Formulations that Minimize Locally

A major drawback of the energetic formulation is that (S) involves a *global* stability condition, while *local* stability would be more physical. However, the word "local" means that we need to specify a topology in which neighborhoods will be defined. A physically motivated way of doing this is to consider systems with small viscosity and to study the limit of vanishing viscosity,

$$\begin{aligned} 0 &= \varepsilon A_1 \dot\varphi + \mathrm{D}_\varphi \mathcal{E}(t,\varphi,z)\ , \\ 0 &\in \partial \mathcal{R}(z,\dot z) + \varepsilon A_2 \dot z + \mathrm{D}_z \mathcal{E}(t,\varphi,z)\ . \end{aligned}$$

A more mathematically motivated approach to the same problem is that of doing local minimization in the associated *time-incremental problem*

$$\boxed{(\mathrm{IP})^{\delta}_{\mathrm{loc}} \qquad q_k \in \mathrm{Argmin}\{\, \mathcal{E}(t_k,\widetilde{q}) + \mathcal{D}(q_{k-1},\widetilde{q}) \mid \widetilde{q} \in \mathcal{Q}, \|q_{k-1}-\widetilde{q}\| \le \delta \,\}\,,}$$

where $\|\cdot\|$ denotes a suitable norm.

It is shown in [18] that for the smooth finite-dimensional situation the associated solutions converge, after an arclength parameterization, to solutions of the following limit problem

$$0 \in \partial R_{\|\cdot\|}(z'(s)) + \mathrm{D}_z\mathcal{E}(t(s),z(s)) \quad \text{and} \quad 1 = t'(s) + \|z'(s)\|\,,$$

where $R_{\|\cdot\|}(v) = R(v)$ for $\|v\| \le 1$ and ∞ else. Generalizations of this idea to the infinite dimensional setting will be discussed in [52].

3 Analysis of the Energetic Formulation

3.1 The Basic Abstract Assumptions

Our state space $\mathcal{Q} = \mathcal{F} \times \mathcal{Z}$ is considered to be the product of two topological spaces $\mathcal{F}$ and $\mathcal{Z}$, both of which are assumed to be Hausdorff spaces. Throughout all topological notions like compactness, closedness and (semi-)continuity are meant in the sequential sense. For convergence we write $\overset{\mathcal{Q}}{\to}$, $\overset{\mathcal{F}}{\to}$ and $\overset{\mathcal{Z}}{\to}$, respectively.

We start with the assumptions on $\mathcal{D} : \mathcal{Z} \times \mathcal{Z} \to [0,\infty]$:

$$\forall\, z_1, z_2, z_3 \in \mathcal{Z} :\ \mathcal{D}(z_1,z_3) \le \mathcal{D}(z_1,z_2) + \mathcal{D}(z_2,z_3)\,. \tag{3.1}$$

$$\mathcal{D} : \mathcal{Z} \times \mathcal{Z} \to [0,\infty] \text{ is lower semi-continuous}\,. \tag{3.2}$$

$$\begin{aligned} &\text{For compact } \mathcal{K} \subset \mathcal{Z} \text{ and } (z_k)_{k\in\mathbb{N}} \subset \mathcal{K} \text{ we have:} \\ &\min\{\mathcal{D}(z_k,z), \mathcal{D}(z,z_k)\} \to 0 \quad \Longrightarrow \quad z_k \overset{\mathcal{Z}}{\to} z\,. \end{aligned} \tag{3.3}$$

For applications in continuum mechanics it is essential to allow $\mathcal{D}$ to attain the value $+\infty$ and to be non-symmetric, i.e., in general $\mathcal{D}(z_1,z_2) \neq \mathcal{D}(z_2,z_1)$.

An important abstract tool is a suitable generalization of Helly's selection principle, cf. [43]. If the functions $z_k : [0,T] \to \mathcal{K} \subset \mathcal{Z}$ with $\mathcal{K}$ compact satisfy $\mathrm{Diss}_{\mathcal{D}}(z_k,[0,T]) \le C < \infty$, then there exists a subsequence $(k_j)_{j\in\mathbb{N}}$ and a limit function $z : [0,T] \to \mathcal{K} \subset \mathcal{Z}$, such that for all $t \in [0,T]$ we have $z_{k_j}(t) \overset{\mathcal{Z}}{\to} z(t)$ and $\mathrm{Diss}_{\mathcal{D}}(z,[0,T]) \le \liminf_{k\to\infty} \mathrm{Diss}_{\mathcal{D}}(z_k,[0,T])$.

For the energy functional $\mathcal{E}$ the following assumptions proved to be useful:

$$\forall\, t \in [0,T]\ \forall\, E \in \mathbb{R} :\quad \{\, q \in \mathcal{Q} \mid \mathcal{E}(t,q) \le E \,\} \text{ is compact}\,; \tag{3.4}$$

$$\begin{aligned} &\exists\, c_0^E \in \mathbb{R}\ \exists\, c_1^E > 0\ \forall\,(t,q) \in [0,T]\times\mathcal{Q} \text{ with } \mathcal{E}(t,q) < \infty : \\ &\mathcal{E}(\cdot,q) \in \mathrm{C}^1([0,T],\mathbb{R}) \text{ and } |\partial_t\mathcal{E}(s,q)| \le c_1^E(\mathcal{E}(s,q)+c_0^E) \text{ on } [0,T]\,; \end{aligned} \tag{3.5}$$

$$\begin{aligned} &\forall\,\varepsilon > 0\ \forall\, E \in \mathbb{R}\ \exists\,\delta > 0\ \forall\, q \text{ with } \mathcal{E}(0,q) \le E : \\ &|t_1 - t_2| \le \delta \quad \Longrightarrow \quad |\partial_t\mathcal{E}(t_1,q) - \partial_t\mathcal{E}(t_2,q)| \le \varepsilon\,; \end{aligned} \tag{3.6}$$

$$\Big(q_k \in \mathcal{S}(t),\ \sup_{k\in\mathbb{N}} \mathcal{E}(t,q_k)<\infty,\ q_k \overset{\mathcal{Q}}{\to} q\Big) \implies \partial_t\mathcal{E}(t,q_k) \to \partial_t\mathcal{E}(t,q)\ . \tag{3.7}$$

The standard condition (3.4) implies lower semi-continuity and relative compactness of infimizing sequences. The other conditions concern the power of external forces $\partial_t\mathcal{E}$. Assumption (3.5) says that we are able to control the work of the external forces via the energy itself. The assumptions (3.6) and (3.7) concern continuity in t and q. They are easily checked in the Banach space setting if $\mathcal{E}$ has the form $\mathcal{E}(t,q) = \mathcal{E}_0(q) - \langle \ell(t), q\rangle$ with $\ell \in \mathrm{C}^1([0,T],\mathcal{Q}^*)$.

The final and crucial assumption controls the interplay of $\mathcal{E}$ and $\mathcal{D}$:

$$\forall\, t \in [0,T]:\ \mathcal{S}(t) \text{ is closed in } \mathcal{Q}\ . \tag{3.8}$$

In most applications of the present theory, the major work is needed for establishing (3.8). There are a few abstract results that establish (3.8). For instance, if $\mathcal{D}$ is continuous on $\mathcal{Z}$, then (3.8) can be easily derived by the use of (3.4).

The following lemma provides a more general condition. We refer to [43, 54, 56] for more discussion on ways to establish closedness of the stable set.

Lemma 1. *If for each sequence $(q_k)_{k\in\mathbb{N}}$ in $\mathcal{S}(t)$ with $q_k \overset{\mathcal{Q}}{\to} q$ and each $\widetilde{q} \in \widetilde{Q}$ there exists a* recovery sequence *$(\widetilde{q}_k)_{k\in\mathbb{N}}$ with $\widetilde{q}_k \overset{\mathcal{Q}}{\to} \widetilde{q}$ such that*

$$\limsup_{k\to\infty} \Big(\mathcal{E}(t,\widetilde{q}_k) + \mathcal{D}(q_k,\widetilde{q}_k) - \mathcal{E}(t,q_k)\Big) \le \mathcal{E}(t,\widetilde{q}) + \mathcal{D}(q,\widetilde{q}) - \mathcal{E}(t,q)$$

holds, then $\mathcal{S}(t)$ is closed.

Proof. We start from $q_k \in \mathcal{S}(t)$ with $q_k \to q$ and have to show $q \in \mathcal{S}(t)$. Let $\widetilde{q}$ be an arbitrary test function. Then, by the assumption of the lemma there exist $\widetilde{q}_k, k \in \mathbb{N}$, with $\widetilde{q}_k \overset{\mathcal{Q}}{\to} \widetilde{q}$. From $q_k \in \mathcal{S}(t)$ we know $0 \le \mathcal{E}(t,\widetilde{q}_k) + \mathcal{D}(q_k,\widetilde{q}_k) - \mathcal{E}(t,q_k)$ and hence the $\limsup_{k\to\infty}$ is nonnegative. We conclude $\mathcal{E}(t,\widetilde{q}) + \mathcal{D}(q,\widetilde{q}) - \mathcal{E}(t,q) \ge 0$ and obtain $q \in \mathcal{S}(t)$. □

3.2 The Existence Result

We approximate the time-continuous formulation (S) & (E) by the following time-incremental problem (IP). For a partition $\Pi = \{0 = t_0 < t_1 < \cdots < t_N = T\}$ and a given initial value $q_0 \in Q$ we consider

$(\mathrm{IP})_\Pi$	Find $q_1, q_2, \dots, q_n$ such that $q_k \in \mathrm{Argmin}\{\, \mathcal{E}(t_k,\widetilde{q}) + \mathcal{E}(q_{k-1},\widetilde{q}) \mid \widetilde{q} \in \mathcal{Q}\,\}$.

By assumption (3.2) and (3.4) it is immediate that $(\mathrm{IP})_\Pi$ is solvable and we are able to define the piecewise constant interpolant

$$q^\Pi : [0,T] \to \mathcal{Q} \text{ with } q^\Pi(t) = \begin{cases} q_{j-1} \text{ for } t \in [t_{j-1}, t_j)\,, \\ q_N \text{ for } t = T\ . \end{cases}$$

It is not difficult to see that the incremental solution satisfies $q^{\Pi}(t_j) \in \mathcal{S}(t_j)$ for $j = 1, \dots, N$ and

$$\mathcal{E}(t_j, q^{\Pi}(t_j)) + \mathrm{Diss}_{\mathcal{D}}(q^{\Pi}, [0, t_j]) \le \mathcal{E}(0, q^{\Pi}(0)) + \int_0^{t_j} \partial_s \mathcal{E}(s, q^{\Pi}(s)) \, \mathrm{d}s \ .$$

From this it is then possible to derive a priori estimates independent of Π for $\mathcal{E}(t, q^{\Pi}(t))$ and $\mathrm{Diss}_{\mathcal{D}}(q^{\Pi}, [0, T])$. Helly's selection principle for the z-component and the compactness of the sublevels of $\mathcal{E}$ allow us then to construct a converging subsequence and to pass to the limit. The final result reads as follows. We refer to [22, 43, 48] for the proof.

Theorem 1. *Let $\Pi_k = \{0 = t_0^k < t_1^k < \cdots < t_{N_k}^k = T\}, k \in \mathbb{N}$, be a sequence of partitions such that $\phi(\Pi_k) = \max\{ t_j^k - t_{j-1}^k \mid j = 1, \dots, N_k \}$ tends to 0. Let $q_0 \in \mathcal{S}(0)$ be an initial condition and $q^{\Pi_k} : [0, T] \to \mathcal{Q}$ be piecewise constant interpolants of the solution of (IP)$_{\Pi_k}$. Then there exists a subsequence $\overline{q}_n = q^{\Pi_{k_n}}$ and an energetic solution $q : [0, T] \to \mathcal{Q}$ of (S) & (E) with $q(0) = q_0$ such that for all $t \in [0, T]$ the following holds*

(i) $\overline{z}_n(t) \overset{Z}{\to} z(t)$,
(ii) $\mathcal{E}(t, \overline{q}_n(t)) \to \mathcal{E}(t, q(t))$,
(iii) $\mathrm{Diss}_{\mathcal{D}}(\overline{q}_n, [0, t]) \to \mathrm{Diss}_{\mathcal{D}}(q, [0, T])$,
(iv) $\exists$ *subsequence* $(N_l^t)_{l \in \mathbb{N}}$*:* $\overline{\varphi}_{N_l^t}(t) \overset{\mathcal{F}}{\to} \varphi(t)$ *for* $l \to \infty$.

Moreover, $\partial_t \mathcal{E}(\cdot, \overline{q}_n(\cdot)) \overset{*}{\rightharpoonup} \partial_t \mathcal{E}(\cdot, q(\cdot))$ *in* $\mathrm{L}^\infty((0, T))$.

The convergence of the φ-component occurs only on t-dependent subsequences $(N_l^t)_{l \in \mathbb{N}}$. Hence, in general, we cannot guarantee the measurability of the mapping $\varphi : [0, T] \to \mathcal{F}$. However, in [42] it is shown that measurability can also be obtained by applying suitable results for measurable selections of multi-valued mappings.

3.3 Results Based on Convexity

The abstract result of the previous section can be improved if additional properties are available. We now assume that $\mathcal{Q}$ is a Banach space, such that convexity methods can be used. In general, one should distinguish three different spaces X, Y and Z. The space Z is the one that provides coercivity of the dissipation distance, i.e.,

$$\forall\, q_0, q_1 \in \mathcal{Q} : \ \mathcal{D}(q_0, q_1) \ge \|q_1 - q_0\|_Z \ . \tag{3.9}$$

The space Y measures the uniform convexity of $\mathcal{J}_{t,q} : \widetilde{q} \mapsto \mathcal{E}(t, \widetilde{q}) + \mathcal{D}(q, \widetilde{q})$:

$$\forall\, q_0, q_1 \in \mathcal{Q} : \ \mathcal{J}_{t,q}\big(\tfrac{1}{2}(q_0 + q_1)\big) \le \tfrac{1}{2}(\mathcal{J}_{t,q}(q_0) + \mathcal{J}_{t,q}(q_1)) - \tfrac{\alpha}{2} \|q_0 - q_1\|_Y^2 \tag{3.10}$$

for some $\alpha > 0$. Finally, X relates to the coercivity of $\mathcal{E}$, i.e.,

$$\forall\, q \in \mathcal{Q}:\ \mathcal{E}(t,q) \geq g(\|q\|_X) \tag{3.11}$$

for some $g \in \mathrm{C}^0([0,\infty),\mathbb{R})$ with $g(t) \to \infty$ for $t \to \infty$.

The abstract results of Sect. 3.2 immediately imply that any solution of (S) & (E) satisfies

$$q = (\varphi, z) \in \mathrm{L}^\infty([0,T],X) \quad \text{and} \quad z \in \mathrm{BV}([0,T],Z)\ .$$

For a proof of the following result we refer to Theorem 3.4 in [48].

Proposition 1. *Assume that $\mathcal{E}$ and $\mathcal{D}$ satisfy the joint convexity condition* (3.10) *for some $\alpha > 0$ and that there exists $C_Y > 0$ such that*

$$\forall\, t \in [0,T]\ \forall\, q_0, q_1 \in \mathcal{Q}:\ |\partial_t\mathcal{E}(t,q_0) - \partial_t\mathcal{E}(t,q_1)| \leq C_Y \|q_0 - q_1\|_Y\ .$$

Then, every solution q of (S) & (E) satisfies

$$\forall\, t_1, t_2 \in [0,T]:\ \|q(t_1) - q(t_2)\|_Y \leq \frac{C_Y}{\alpha}|t_1 - t_2|\ .$$

As a typical example we consider the case $\mathcal{Q} = \mathcal{Z} = X$ with

$$\begin{aligned} &X = \mathrm{H}^1(\Omega)\ , \quad \mathcal{D}(z_0,z_1) = \textstyle\int_\Omega |z_0(x) - z_1(x)|\,\mathrm{d}x\ , \\ &\mathcal{E}(t,z) = \textstyle\int_\Omega W(\nabla z(x)) + \frac{\alpha}{2}|z(x)|^2 - f_{\mathrm{ext}}(t,x)z(x)\,\mathrm{d}x\ , \end{aligned}$$

with $\alpha > 0$, $f_{\mathrm{ext}} \in \mathrm{C}^1([0,T],\mathrm{L}^2(\Omega))$ and $W : \mathbb{R}^d \to [0,\infty)$, where W is convex and coercive, i.e., $W(A) \geq c|A|^2 - C$ for some $C, c > 0$ and all $A \in \mathbb{R}^d$. Then, we may choose $Z = \mathrm{L}^1(\Omega)$ and $Y = \mathrm{L}^2(\Omega)$.

In such situations it is possible to define $\dot{q}(t)$ almost everywhere, since jumps, which are allowed in the energetic formulation, can no longer occur. Hence, it is possible to study the local subdifferential formulation (2.1) instead. Using $q = (\varphi, z) \in X = Q$ and $\mathcal{R}(z,v) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\mathcal{D}(z, z+\varepsilon v)$ we write (2.1) in the compact form

$$X^* \ni 0 \in \partial_v\mathcal{R}(q(t),\dot{q}(t)) + \partial_q\mathcal{E}(t,q(t)) \text{ a.e. on } [0,T]\ . \tag{3.12}$$

This equation is called a doubly nonlinear equation and it relates to evolutionary quasi-variational inequalities (cf. [10]). We refer to [51, 58] for exact conditions which guarantee the equivalence between (S) & (E) and (3.12).

The latter work contains also a general existence result for Lipschitz continuous solutions to (3.12). Under quite severe additional assumptions it is even possible to prove uniqueness, see [10, 51, 58]. However, these assumptions are rarely met in material models except for very simple cases like linearized elastoplasticity with quadratic hardening, see [30, 32, 64]. Other uniqueness results are discussed in [60, 61] for piezoelectricity and in [4] for an isotropic model for shape-memory alloys, see also Sect. 5.

4 Approximation, Γ-Limits and Relaxation

In several circumstances it is desirable to consider sequences of functionals $(\mathcal{E}_k)_{k\in\mathbb{N}}$ and $(\mathcal{D}_k)_{k\in\mathbb{N}}$ which converge to limit functionals $\mathcal{E}_\infty$ and $\mathcal{D}_\infty$, respectively, in a suitable sense. The main question is which type of convergence guarantees that limits $q : [0,T] \to \mathcal{Q}$ of solutions $q_k : [0,T] \to \mathcal{Q}$ for $(\mathcal{E}_k, \mathcal{D}_k)$ are solutions for $(\mathcal{E}_\infty, \mathcal{D}_\infty)$.

Typical applications of this idea occur for

- numerical approximations with $\mathcal{E}_k(t,q) = \mathcal{E}_\infty(t,q)$ for $q \in \mathcal{Q}_k \subset \mathcal{Q}$ and ∞ otherwise, where each Q_k is a finite-dimensional subspace of $\mathcal{Q}$ such that $Q_k \subset \mathcal{Q}_{k+1}$ and $\bigcup_{k\in\mathbb{N}} Q_k$ is dense in $\mathcal{Q}$.
- problems with singular perturbations (like sharp interface models) or with penalization terms
- constant sequences $\mathcal{E}_k = \mathcal{E}_1, \mathcal{D}_k = \mathcal{D}_1$, where $\mathcal{E}_1(t,\cdot)$ and $\mathcal{D}_1(\cdot,\cdot)$ are not lower semi-continuous and differ from their Γ-limits $\mathcal{E}_\infty$ and $\mathcal{D}_\infty$.

The latter point relates to relaxations of rate-independent evolution which is an important topic in material modeling. It is a tool for deriving evolution equations for microstructures. We refer to [14, 47, 50, 59, 78] for discussions of this topic.

Here we present the theory originating from [38]. In [56] the abstract version was developed and in [54] it is applied to numerical approximations in several material models. The following version is a simplified version of the one developed in [56].

4.1 Γ-Convergence of Rate-Independent Systems

We let $\mathbb{N}_\infty := \mathbb{N} \cup \{\infty\}$ and state first the conditions on the dissipation distances $(\mathcal{D}_k)_{k\in\mathbb{N}_\infty}$. Each $\mathcal{D}_k, k \in \mathbb{N}_\infty$, is a pseudo distance on $\mathcal{Z}$, i.e.,

$$\forall\, z_j \in \mathcal{Z} :\ \mathcal{D}_k(z_1,z_1) = 0 \text{ and } \mathcal{D}_k(z_1,z_3) \le \mathcal{D}_k(z_1,z_2) + \mathcal{D}_k(z_2,z_3)\,. \tag{4.1}$$

To obtain solutions of incremental problems we impose that

$$\forall\, k \in \mathbb{N}_\infty :\ \mathcal{D}_k : \mathcal{Z}\times\mathcal{Z} \to [0,\infty] \text{ is lower semi-continuous}\,. \tag{4.2}$$

The limit distance $\mathcal{D}_\infty$ must be positive in the following sense

$$\begin{aligned} &\text{For compact } \mathcal{K} \subset \mathcal{Z} \text{ and } (z_k)_{k\in\mathbb{N}} \subset \mathcal{K} \text{ we have:}\\ &\min\{\mathcal{D}_\infty(z_k,z), \mathcal{D}_\infty(z,z_k)\} \to 0 \quad\Longrightarrow\quad z_k \overset{\mathcal{Z}}{\to} z\,. \end{aligned} \tag{4.3}$$

Finally, $\mathcal{D}_\infty$ must be bounded from above by the Γ-liminf of $(\mathcal{D}_k)_{k\in\mathbb{N}}$, i.e.,

$$\Big(z_k \overset{\mathcal{Z}}{\to} z \text{ and } \widetilde{z}_k \overset{\mathcal{Z}}{\to} \widetilde{z}\Big) \quad\Longrightarrow\quad \mathcal{D}_\infty(z,\widetilde{z}) \le \liminf_{k\to\infty} \mathcal{D}_k(z_k,\widetilde{z}_k)\,. \tag{4.4}$$

Next we state the conditions on the energy functionals. We start with the compactness of the sublevels:

$$\begin{aligned}&\forall\, t \in [0,T]\, \forall\, E \in \mathbb{R}:\\ &\text{(i)}\quad \forall\, k \in \mathbb{N}_\infty:\ \{\, q \in \mathcal{Q} \mid \mathcal{E}_k(t,q) \le E \,\} \text{ is compact},\\ &\text{(ii)}\ \textstyle\bigcup_{k\in\mathbb{N}} \{\, q \in \mathcal{Q} \mid \mathcal{E}_k(t,q) \le E \,\} \text{ is relatively compact}.\end{aligned} \tag{4.5}$$

The next three conditions provide suitable continuity properties of the powers $\partial_t \mathcal{E}_k(\cdot,\cdot)$ of the external forces.

$$\begin{aligned}&\exists\, c_0, c_1 > 0\, \forall\, k \in \mathbb{N}_\infty\, \forall\, (t,q) \in [0,T] \times \mathcal{Q} \text{ with } \mathcal{E}_k(t,q) < \infty:\\ &\mathcal{E}_k(\cdot,q) \in \mathrm{C}^1([0,T]) \text{ and } |\partial_t \mathcal{E}_k(s,q)| \le c_1(\mathcal{E}_k(s,q)+c_0) \text{ on } [0,T]\,;\end{aligned} \tag{4.6}$$

$$\begin{aligned}&\forall\, \varepsilon > 0\, \forall\, E > 0\, \exists\, \delta > 0\, \forall\, k \in \mathbb{N}_\infty\, \forall\, q \in \mathcal{Q} \text{ with } \mathcal{E}_k(0,q) \le E:\\ &|t_1 - t_2| \le \delta \quad\Longrightarrow\quad |\partial_t \mathcal{E}_k(t_1,q) - \partial_t \mathcal{E}_k(t_2,q)| \le \varepsilon\,;\end{aligned} \tag{4.7}$$

$$\Big(q_k \overset{\mathcal{Q}}{\to} q \text{ and } \sup_{k\in\mathbb{N}} \mathcal{E}_k(t,q_k) < \infty\Big) \quad\Longrightarrow\quad \partial_t \mathcal{E}_k(t,q_k) \to \partial_t \mathcal{E}(t,q)\,. \tag{4.8}$$

The final condition on $(\mathcal{E}_k)_{k\in\mathbb{N}_\infty}$ concerns the Γ-liminf, namely

$$q_k \overset{\mathcal{Q}}{\to} q \quad\Longrightarrow\quad \mathcal{E}_\infty(t,q) \le \liminf_{k\to\infty} \mathcal{E}_k(t,q_k)\,. \tag{4.9}$$

The crucial condition that connects the convergences of $\mathcal{D}_k$ to $\mathcal{D}_\infty$ and $\mathcal{E}_k$ to $\mathcal{E}_\infty$ involves the sets of stable states. For $k \in \mathbb{N}_\infty$ we have

$$\mathcal{S}_k(t) \overset{\text{def}}{=} \{\, q \in \mathcal{Q} \mid \mathcal{E}_k(t,q) < \infty \text{ and } \forall\, \widetilde{q} \in \mathcal{Q} : \mathcal{E}_k(t,q) \le \mathcal{E}_k(t,\widetilde{q}) + \mathcal{D}_k(q,\widetilde{q}) \,\}$$

and ask for the upper semi-continuity $\mathrm{Limsup}_{k\to\infty} \mathcal{S}_k(t) \subset \mathcal{S}_\infty(t)$, i.e.,

$$\Big(q_{k_\ell} \in \mathcal{S}_{k_\ell}(t) \text{ and } q_{k_\ell} \overset{\mathcal{Q}}{\to} q \text{ for } k_\ell \to \infty\Big) \quad\Longrightarrow\quad q \in \mathcal{S}_\infty(t)\,. \tag{4.10}$$

In typical applications in continuums mechanics it is hard to establish this condition. On the abstract level it is possible to provide sufficient conditions. For instance, we say that $\mathcal{E}_\infty$ is the *Γ-limit* of $(\mathcal{E}_k)_{k\in\mathbb{N}}$ if (4.9) holds and if for all $\widetilde{q} \in \mathcal{Q}$ there exists a recovery sequence $(\widetilde{q}_k)_{k\in\mathbb{N}}$ such that

$$\widetilde{q}_k \overset{\mathcal{Q}}{\to} \widetilde{q} \text{ and } \mathcal{E}_\infty(t,\widetilde{q}) \ge \limsup_{k\to\infty} \mathcal{E}_k(t,\widetilde{q}_k)\,. \tag{4.11}$$

A similar notion for the Γ-limit holds for $(\mathcal{D}_k)_{k\in\mathbb{N}}$. Of course, in the case of a constant sequence the Γ-limit is simply the lower semi-continuous hull.

It is shown in [56] that in general (4.10) does not hold if $\mathcal{E}_k$ Γ-converges to $\mathcal{E}_\infty$ and $\mathcal{D}_k$ Γ-converges to $\mathcal{D}_\infty$. Even more, the following Theorem 2 may be false. The next lemma gives a positive result.

Lemma 2. *If $\mathcal{E}_\infty = \Gamma\text{-}\lim_{k\to\infty}\mathcal{E}_k$, i.e., (4.9) and (4.11) hold, and if $\mathcal{D}_k$ converges continuously to $\mathcal{D}_\infty$, i.e.,*

$$\big(\, z_k \overset{\mathcal{Z}}{\to} z \text{ and } \widetilde{z}_k \overset{\mathcal{Z}}{\to} \widetilde{z} \,\big) \quad \Longrightarrow \quad \mathcal{D}_k(z_k,\widetilde{z}_k) \to \mathcal{D}_\infty(z,\widetilde{z}) \,, \tag{4.12}$$

then (4.10) holds.

Proof. Let $q_k = (\varphi_k, z_k) \in \mathcal{S}_k(t)$ be given such that $q_{k_\ell} \overset{\mathcal{Q}}{\to} q$. Moreover, let $\widetilde{q}$ be arbitrary. Then there exists a recovery sequence $\widetilde{q}_k = (\widetilde{\varphi}_k, \widetilde{z}_k)$ satisfying (4.11). Using (4.12) we conclude

$$\begin{aligned}
&\mathcal{E}_\infty(t,q) \le \liminf_{\ell\to\infty} \mathcal{E}_{k_\ell}(t,q_{k_\ell}) \\
&\le \liminf_{\ell\to\infty} \big(\mathcal{E}_{k_\ell}(t,\widetilde{q}_{k_\ell}) + \mathcal{D}_{k_\ell}(q_{k_\ell},\widetilde{q}_{k_\ell})\big) \;=\; \mathcal{E}_\infty(t,\widetilde{q}) + \mathcal{D}_\infty(q,\widetilde{q}) \,.
\end{aligned}$$

Here we use first (4.9), next $q_k \in \mathcal{S}_k(t)$ and last (4.11) and (4.12). Since $\widetilde{q} \in \mathcal{Q}$ was arbitrary, we have $q \in \mathcal{S}_\infty(t)$. □

The following result is concerned with the so-called *incremental problem* $(\mathrm{IP})_k$. For this choose a sequence $(\Pi_k)_{k\in\mathbb{N}}$ of partitions with $\Pi_k = \{0 = t_0^k < t_1^k < \ldots < t_{N_k}^k = T\}$ and fineness $\phi(\Pi_k) = \max\{\, t_j^k - t_{j-1}^k \mid j = 1,\ldots,N_k \,\}$:

$(\mathrm{IP})_k$ — Given $q_0^k \in \mathcal{Q}$, find iteratively
$q_j^k \in \mathrm{Argmin}\{\, \mathcal{E}_k(t_j^k,\widetilde{q}) + \mathcal{D}_k(q_{j-1}^k,\widetilde{q}) \mid \widetilde{q} \in \mathcal{Q} \,\}$.

Existence of solutions follows easily from (4.2) and (4.5). We define the constant interpolants $\overline{q}_k : [0,T] \to \mathcal{Q}$ via

$$\overline{q}_k(t) = q_{j-1}^k \text{ for } t \in [t_{j-1}^k, t_j^k) \text{ and } \overline{q}_k(T) = q_{N_k}^k \,.$$

Theorem 2. *Let the conditions (4.1) to (4.10) hold and let the partitions Π_k, $k \in \mathbb{N}$, satisfy $\phi(\Pi_k) \to 0$ for $k \to \infty$. Moreover, assume*

$$q_0^k \in \mathcal{S}_k(0),\; q_0^k \overset{\mathcal{Q}}{\to} q_0, \;\text{ and } \mathcal{E}_k(0,q_0^k) \to \mathcal{E}_\infty(0,q_0) \,.$$

Choose any sequence $(\overline{q}_k)_{k\in\mathbb{N}}$ of constant interpolants of solutions to $(IP)_k$. Then, there exists a solution $q : [0,T] \to \mathcal{Q}$ of (S) & (E) associated with $(\mathcal{E}_\infty, \mathcal{D}_\infty)$ and $q(0) = q_0$ and a subsequence $(\overline{q}_{k_\ell})_{\ell\in\mathbb{N}}$ such that for all $t \in [0,T]$ the following holds:

(i) $\mathcal{E}_{k_\ell}(t,\overline{q}_{k_\ell}(t)) \to \mathcal{E}_\infty(t,q(t))$,
(ii) $\mathrm{Diss}_{\mathcal{D}_{k_\ell}}(\overline{q}_{k_\ell},[0,t]) \to \mathrm{Diss}_{\mathcal{D}_\infty}(q,[0,t])$,
(iii) $z_{k_\ell}(t) \overset{\mathcal{Z}}{\to} z(t)$,
(iv) $\exists$ *subsequences* $(K_n^t)_{n\in\mathbb{N}}$ *of* $(k_\ell)_{\ell\in\mathbb{N}}$: $\varphi_{K_n^t}(t) \overset{\mathcal{F}}{\to} \varphi(t)$ *for* $n \to \infty$.

Moreover, we have $\partial_t\mathcal{E}_{k_\ell}(\cdot,\overline{q}_{k_\ell}(\cdot)) \overset{*}{\rightharpoonup} \partial_t\mathcal{E}_\infty(\cdot,q(\cdot))$ in $\mathrm{L}^\infty([0,T])$.

4.2 Relaxation in the Case of Missing Lower Semi-continuity

In applications it may occur that for the empirically derived functionals $\mathcal{E}$ and $\mathcal{D}$ it is not possible to choose a space $\mathcal{Q}$, such that the sublevels of $\mathcal{E}$ are compact. In particular, the time-incremental problems $(\mathrm{IP})_\Pi$ may not be solvable because of the missing lower semi-continuity, which has its mechanical counterpart in the formation of microstructure. In such situations it is desirable to find suitable *relaxations*, which allow for the calculation of suitable effective quantities associated with these microstructures. For rate-independent systems this question was first addressed in [59], where the separate relaxation $\mathcal{E}_\infty = \Gamma\text{-}\lim \mathcal{E}$ and $\mathcal{D}_\infty = \Gamma\text{-}\lim \mathcal{D}$ was postulated. Further developments are discussed in [14, 20, 47, 56, 78].

In [47, 56, 59] it is suggested to study the *approximate incremental problem*

$$
\begin{array}{|ll|}
\hline
 & \text{Given } q_0 \in \mathcal{Q}, \text{ find iteratively } q_1, q_2, \dots, q_N \text{ such that} \\
(\mathrm{AIP})_{\Pi,\alpha} & \mathcal{E}(t_j, q_j) + \mathcal{D}(q_{j-1}, q_j) \\
 & \quad \le (t_j - t_{j-1})\alpha + \inf_{\widetilde{q}\in\mathcal{Q}} \mathcal{E}(t_j, \widetilde{q}) + \mathcal{D}(q_{j-1}, \widetilde{q}). \\
\hline
\end{array}
$$

For $\alpha > 0$ this problem always has solutions and the question arises as to how the solutions behave for $\alpha \to 0$ and for smaller and smaller time steps.

Choose sequences $(\Pi_k)_{k\in\mathbb{N}}$ and $(\alpha_k)_{k\in\mathbb{N}}$ with $0 < \alpha_k \to 0$ and $\phi(\Pi_k) \to 0$. Then, for each $k \in \mathbb{N}$ a solution of $(\mathrm{AIP})_{\Pi_k,\alpha_k}$ exists and defines a piecewise constant interpolant $\overline{q}_k : [0,T] \to \mathcal{Q}$. In [56] it is shown under general abstract conditions that the interpolants contain a convergent subsequence in the sense above and that the limit $q : [0,T] \to \mathcal{Q}$ is an energetic solution for the Γ-limit potentials $\mathcal{E}_\infty$ and $\mathcal{D}_\infty$. One simple sufficient condition is that $\mathcal{D}$ is continuous, which implies $\mathcal{D}_\infty = \mathcal{D}$.

Another abstract relaxation result is derived in [50]. It is based on a kind of elliptic regularization of the subdifferential inclusion (3.12). We consider a sequence of functionals

$$
\mathcal{I}_{k,\delta}(q) = \int_0^T \mathrm{e}^{-t/\delta} (\mathcal{R}_k(\dot{q}(t)) + \tfrac{1}{\delta}\mathcal{E}_k(t, q(t)))\,\mathrm{d}t \ ,
$$

where each $\mathcal{R}_k : Q \to [0,\infty]$ is convex, lower semi-continuous and 1- homogeneous. If $\mathcal{R}_k$ and $\mathcal{E}_k$ are smooth, then the Euler-Lagrange equation reads

$$
\delta \mathrm{D}^2\mathcal{R}_k(\dot{q})[\ddot{q}] = \mathrm{D}\mathcal{R}_k(\dot{q}) + \mathrm{D}_q\mathcal{E}_k(t,q) \ ,
$$

which in the formal limit $\delta \to 0$ converges to (3.12).

Using the 1-homogeneity of $\mathcal{R}_k$ it is proved in [50] that minimizers $q : [0,T] \to \mathcal{Q}$ of $\mathcal{I}_{k,\delta}$ satisfy the δ-independent energy balance

$$
\mathcal{E}_k(t, q(t)) + \int_0^t \mathcal{R}_k(\mathrm{d}q) = \mathcal{E}_k(0, q(0)) + \int_0^t \partial_s\mathcal{E}_k(s, q(s))\,\mathrm{d}s \ .
$$

As in Sect. 3.2 this implies a priori bounds independent of $\delta > 0$ and of k, if (4.6) is used. Fixing $\delta > 0$ and letting $k \to \infty$, we obtain a Γ-limit $\mathcal{I}_{\infty,\delta}$ in the form

$$\mathcal{I}_{\infty,\delta}(q) = \int_0^T \mathrm{e}^{-t/\delta} \left(\mathcal{R}_\infty(\dot{q}(t)) + \tfrac{1}{\delta} \mathcal{E}_\infty(t, q(t)) \right) \mathrm{d}t \,,$$

if $\mathcal{E}_\infty = \Gamma\text{-}\lim_{k\to\infty} \mathcal{E}_k$ and $\mathcal{R}_k$ converges continuously to $\mathcal{R}_\infty$. Finally, under these assumptions it is shown that for minimizers $q_{k,\delta} : [0,T] \to \mathcal{Q}$ the accumulation points for $k \to \infty$ and $\delta \to 0$ are in fact solutions of the energetic formulation (S) & (E) associated with $\mathcal{E}_\infty$ and $\mathcal{D}_\infty : (q, \widetilde{q}) \mapsto \mathcal{R}_\infty(\widetilde{q} - q)$.

4.3 Numerical Space Discretization

We indicate one of the main applications of the Γ-convergence results. Consider a reflexive Banach space $\mathcal{Q}$ equipped with its weak topology. This space is approximated by a nested sequence $(Q_k)_{k\in\mathbb{N}}$ of finite-dimensional subspaces such that their union is dense, viz., $Q_k \subset Q_{k+1} \subset \mathcal{Q}$ and $\overline{\cup_{k\in\mathbb{N}} Q_k} = \mathcal{Q}$. Finally, assume that the functional $\mathcal{E}_\infty = [0,T] \times \mathcal{Q} \to \mathbb{R}_\infty$ is strongly continuous and $\mathcal{D}_\infty : \mathcal{Q} \times \mathcal{Q} \to [0,\infty]$ is weakly continuous in addition to the assumption (3.1) to (3.8). Now define the finite-dimensional (space) approximations via

$$\mathcal{E}_k(t,q) = \begin{cases} \mathcal{E}_\infty(t,q) \text{ for } q \in Q_k \,, \\ \infty \quad \text{otherwise} \,, \end{cases} \qquad \mathcal{D}_k(q,\widetilde{q}) = \begin{cases} \mathcal{D}_\infty(q,\widetilde{q}) \text{ for } q, \widetilde{q} \in Q_k \,, \\ \infty \quad \text{otherwise} \,. \end{cases}$$

Then it is easy to see that the assumptions (4.1) to (4.9) are satisfied. To establish the upper semi-continuity of the stable sets we proceed as follows. Starting from $q_k \in \mathcal{S}_k(t)$ with $q_k \rightharpoonup q$ we need to show $q \in \mathcal{S}_\infty(t)$. For $\widetilde{q} \in \mathcal{Q}$ we choose $\widetilde{q}_k \in Q_k$ with $\widetilde{q}_k \to q$ (strongly). Then we have

$$0 \le \mathcal{E}_k(t,\widetilde{q}_k) + \mathcal{D}_k(q_k,\widetilde{q}_k) - \mathcal{E}_k(t,q_k) = \mathcal{E}(t,\widetilde{q}_k) + \mathcal{D}(q_k,\widetilde{q}_k) - \mathcal{E}(t,q_k) \,.$$

Using strong continuity for $\mathcal{E}$, weak continuity for $\mathcal{D}$ and weak lower semi-continuity for $\mathcal{E}$ we take the limsup of the last expression and find

$$0 \le \limsup_{k\to\infty} \left(\mathcal{E}(t,\widetilde{q}_k) + \mathcal{D}(q_k,\widetilde{q}_k) - \mathcal{E}(t,q_k) \right) \le \mathcal{E}(t,\widetilde{q}) + \mathcal{D}(q,\widetilde{q}) - \mathcal{E}(t,q) \,,$$

which is the desired stability result, since $\widetilde{q}$ is arbitrary.

This theory is the basis for treating spatial discretizations of energetic formulations. In particular, Theorem 2 guarantees that each limit point of the *joint space-time discretization* provides a true solution of (S) & (E).

In [38] this numerical theory was developed for a model involving gradient Young measures to describe microstructures in shape-memory alloys. A more systematic treatment of different aspects of numerical space discretizations as well as penalizations or relaxations is given in [54]. Using more regularity and convexity assumptions, full convergence results, without choosing subsequences, are established in [4]. This is relates closely to the highly developed theory in linearized elastoplasticity, see [1, 30, 32, 77].

5 Applications to Material Models

The theory of rate-independent processes finds applications in many areas. This includes the theory of superconductivity [67, 75], dry friction on surfaces [2, 44, 74], delamination [35], damage [55] and brittle fracture [12, 15, 21]. The latter three areas involve applications where the internal variable is active only on submanifolds of the elastic body. Here we restrict ourselves to those situations where the internal variable z is distributed throughout the body.

The original driving force of this theory was the dynamic problem of linearized elastoplasticity, whereas, nowadays many other applications occur in phase transformations in shape-memory materials, in magnetostriction, in piezo- or ferroelectricity, and in damage. Finite-strain elastoplasticity is another very active area for rate-independent modeling. This will be surveyed in [26] of this volume.

5.1 Shape-Memory Materials

We consider an elastic body $\Omega \subset \mathbb{R}^d$ in its reference configuration, which we assume to have a Lipschitz boundary. The deformation $\varphi : \Omega \to \mathbb{R}^d$ describes the elastic behavior in the case of large strains (also called "finite strain"), whereas the displacement $u : \Omega \to \mathbb{R}^d$, $x \mapsto \varphi(x) - x$, is used in the case of small strains (also called "infinitesimal strain"). The elastic properties depend on the strain tensors $\nabla\varphi^\top\nabla\varphi$ and $\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^\top)$, respectively.

Since the shape-memory effect relies on the fact that these materials have several phases, the corresponding stored-energy density (also called stress potential) has a multi-well structure, which is usually given in the form

$$W(x, \nabla\varphi) = \min\{\, W_j(x, \nabla\varphi) \mid j = 1, \dots, N \,\} .$$

Here N is the number of (variants of) phases including the austenite and the martensites. Each $W_j(x, \cdot) : \mathbb{R}^{d\times d} \to [0, \infty]$ is assumed to behave nicely in terms of lower semi-continuity and coercivity.

However, $W(x, \cdot)$ is in general not rank-one convex, and hence formation of microstructures is to be expected. This is compatible with the physics, since the shape-memory effect relies heavily on the formation of martensitic laminates (also called "twinning"), see Fig. 1. We refer to [9, 70] for surveys on the mathematical modeling of *microsctructures* in shape-memory alloys.

To describe this mathematically it is advantageous to use gradient Young measures, see [7, 9, 39, 40, 65] for the static case and see [3, 28, 38, 47, 53, 70] for the evolution of microstructures. We will survey this work next. After that we will discuss several models which do not resolve the microstructure but keep certain volume fractions or effective properties, see [23, 24, 25, 42, 57, 59, 76]. Finally, we will indicate how these models may be generalized to include the temperature as an additional external parameter.

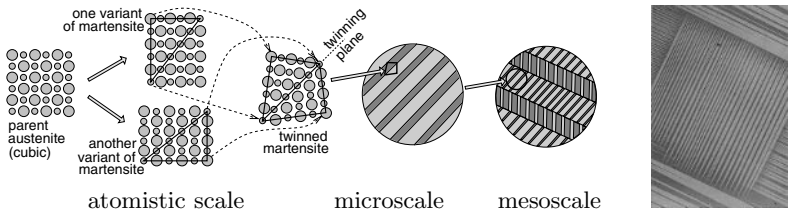


Fig. 1. Sketch of the multiscale structure of a sequential laminate in a shape-memory alloy (left to middle). Microscopic view of laminates by Chu and James (right)

5.2 Models Using Gradient Young Measures

A gradient Young measure is a function over the physical domain Ω which takes values in the set of probability measures on the set $\mathbb{R}^{d\times d}$ of deformation gradients, namely

$$\mathrm{Prob}(\mathbb{R}^{d\times d}) := \{\, \mu \in \mathcal{M}(\mathbb{R}^{d\times d}) \mid \mu \geq 0,\ \textstyle\int_{\mathbb{R}^{d\times d}} 1\, \mu(\mathrm{d}A) = 1 \,\} \,.$$

However, the addition "gradient" means that only those measures are considered, that can be generated via a sequence of gradients of deformations.

We say that a bounded sequence $(\varphi_k)_{k\in\mathbb{N}}$ in $\mathrm{W}^{1,p}(\Omega,\mathbb{R}^d)$ generates the gradient Young measure $\mu : \Omega \to \mathrm{Prob}(\mathbb{R}^{d\times d})$ and write $\nabla\varphi_k \overset{\mathrm{YM}}{\to} \mu$, if for all $\psi \in \mathrm{C}_\mathrm{c}^0(\Omega \times \mathbb{R}^{d\times d})$ we have, for $k \to \infty$,

$$\textstyle\int_\Omega \psi(x, \nabla\varphi_k(x))\,\mathrm{d}x \to \int_\Omega \int_{\mathbb{R}^{d\times d}} \psi(x,A)\mu(x,\mathrm{d}A)\,\mathrm{d}x \,.$$

Such μ have the additional property $\int_\Omega \int_{\mathbb{R}^{d\times d}} (1+|A|)^p \mu(x,\mathrm{d}A)\,\mathrm{d}x < \infty$, and we denote the set of all these measures by

$$\mathcal{G}_p(\Omega) = p\text{-integrable gradient Young measures} \,.$$

To model the hysteretic behavior in shape-memory materials with the energetic formulation discussed in Sect. 3 we need to introduce a phase indicator $z : \Omega \to Z_N$ where Z_N is usually taken as the Gibbs simplex

$$Z_N = \{\, z \in \mathbb{R}^N \mid z_i \geq 0,\ \textstyle\sum_{i=1}^N z_i = 1 \,\} \,.$$

The components z_i of $z \in Z_N$ measure the volume fraction of phase i in a representative volume element. For gradient Young measures we extract the phase fractions via a continuous mapping

$$\zeta : \Omega \times \mathbb{R}^{d\times d} \to Z_N$$

such that $\zeta(x,A) = e_j$ (unit vector in $\mathbb{R}^N$), if $W(x,A) = W_j(x,A) \leq W_k(x,A) - \delta$ for $k \neq j$. Here $\delta > 0$ is a suitable constant which is assumed to be much smaller than the depth of the wells.

Finally we introduce a dissipation distance $D : Z_N \times Z_N \to [0,\infty)$. It suffices to prescribe the values $\kappa_{j\to k} = D(e_j, e_k) > 0$, such that the triangle inequality holds, i.e., $\kappa_{j\to\ell} \le \kappa_{j\to k} + \kappa_{k\to\ell}$. Here $\kappa_{j\to k}$ denotes the energetic loss when the material jumps from a phase e_j into another phase e_k. Then, $D : Z_N \times Z_N \to [0,\infty)$ is defined via the optimal transport problem

$$D(z,\widetilde{z}) = \min\Big\{ \sum_{j,k=1}^N m_{jk}\kappa_{j\to k} \;\Big|\; m_{jk} \ge 0,\ \sum_{k=1}^N m_{jk}e_j = \widetilde{z},\ \sum_{j=1}^N m_{jk}e_k = z \Big\} .$$

It is shown in Proposition 4.7 in [59], that there exists a convex, 1-homogeneous $R : \mathbb{R}^N \to [0,\infty)$ such that $D(z,\widetilde{z}) = R(\widetilde{z} - z)$.

With these notations we now formulate the function spaces and the functionals. We assume that $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz boundary $\partial\Omega$ and that $\Gamma_{\rm Dir} \subset \partial\Omega$ is a set of positive surface measure on which we describe Dirichlet boundary data. We let

$$\mathcal{F} = \{\, \varphi \in \mathrm{W}^{1,p}(\Omega,\mathbb{R}^d) \mid \varphi|_{\Gamma_{\rm Dir}} = \Phi_{\rm Dir} \,\} \times \mathcal{G}_p(\Omega) \text{ and } \mathcal{Z} = \mathrm{L}^1(\Omega, Z_N) .$$

The state space is $\mathcal{Q} = \mathcal{F} \times \mathcal{Z}$ and a state consists of a triple $q = (\varphi,\mu,z)$. We further let $\mathcal{Q}_0 = \{\, (\varphi,\mu,z) \in \mathcal{Q} \mid \nabla\varphi = \mathrm{id}\bullet\mu,\ z = \zeta\bullet\mu \,\}$, where "$\bullet$" denotes the contraction over $A \in \mathbb{R}^{d\times d}$ but not over $x \in \Omega$, i.e., $(\mathrm{id}\bullet\mu)(x) = \int_{\mathbb{R}^{d\times d}} A\,\mu(x,\mathrm{d}A)$ and $(\zeta\bullet\mu)(x) = \int_{\mathrm{R}^{d\times d}} \zeta(x,A)\mu(x,\mathrm{d}A)$. With the prescribed external volume and surface loadings

$$\langle \ell(t), \varphi\rangle = \int_\Omega f_{\rm ext}(t,x)\cdot\varphi(x)\,\mathrm{d}x + \int_{\Gamma_{\rm Neu}} g_{\rm ext}(t,x)\cdot\varphi(x)\,\mathrm{d}a(x)$$

we define the energy-storage functional

$$\mathcal{E}(t,q) = \int_\Omega \int_{\mathbb{R}^{d\times d}} W(x,A)\mu(x,\mathrm{d}A) + \tfrac{\rho}{2}|\nabla^\alpha z|^2\,\mathrm{d}x - \langle \ell(t), q\rangle \text{ for } q \in \mathcal{Q}_0 \quad (5.1)$$

and $\mathcal{E}(t,q) = +\infty$ for $q = (\varphi,\mu,z) \in \mathcal{Q}\backslash\mathcal{Q}_0$. Here $\rho > 0$ and $\nabla^\alpha z$ with $\alpha > 0$ denotes a (fractional) derivative, for instance for $\alpha \in (0,1)$ we have

$$\int_\Omega |\nabla^\alpha z|^2\,\mathrm{d}x = \int_\Omega \int_\Omega \tfrac{|z(x)-z(\widetilde{x})|^2}{|x-\widetilde{x}|^{d+2\alpha}}\,\mathrm{d}x\,\mathrm{d}\widetilde{x} .$$

This regularizing term allows us to choose the strong topology in $\mathrm{L}^1(\Omega, Z_N)$ as this space is compactly embedded into $\mathrm{W}^{\alpha,2}(\Omega,\mathbb{R}^N)$ for $\alpha > 0$. Nevertheless, for $\alpha < 1/2$ the functions in $\mathrm{W}^{\alpha,2}(\Omega,\mathbb{R}^N)$ may have jumps along smooth interfaces such as the habit plane between austenite and martensite.

The dissipation distance is defined as

$$\mathcal{D}(z,\widetilde{z}) = \int_\Omega D(x,z(x),\widetilde{z}(x))\,\mathrm{d}x = \int_\Omega R(x,\widetilde{z}(x)-z(x))\,\mathrm{d}x . \quad (5.2)$$

Since $\mathcal{D}$ is (strongly) continuous on $\mathcal{Z}$ the crucial closedness condition (3.8) of the stable sets is easily obtained via Lemma 1 by taking $\widetilde{q}_k = \widetilde{q}$.

The following existence theorem is established in [38]. The earlier version in [53] was based on the much stronger assumption that $\mathcal{E}(t,(\cdot,\cdot,z))$ has a unique minimizer, but this condition is no longer needed because of the abstract developments in [22, 48].

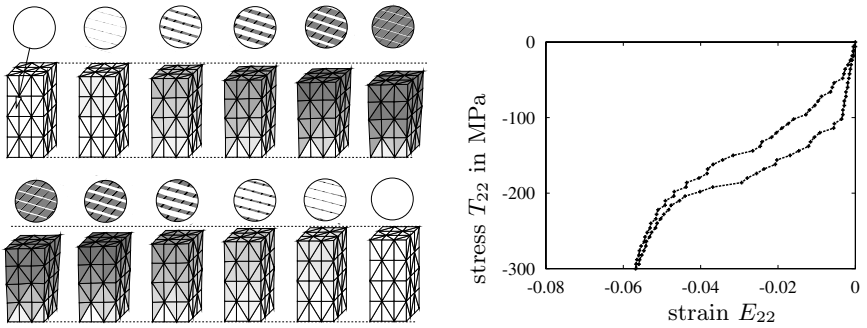


Fig. 2. Left: 12 steps of the compression cycle. The greyscale indicates the volume fraction of martensite in each element. Additionally, in the discs the microstructure on the indicated element is displayed (white=austenite, grey=martensite2, black=martensite3). Right: hysteresis in the overall stress-strain relation

Theorem 3. *Let $p \in (1,\infty)$, $\alpha \in (0,1)$, and $\rho > 0$. Assume that there exists $C > 0$ such that for $j = 1,\dots,N$, we have*

$$\forall\, A \in \mathbb{R}^{d\times d}:\ \tfrac{1}{C}|A|^P - C \le W_j(x,A) \le C|A|^P + C\,. \tag{5.3}$$

Further assume $\ell \in \mathrm{C}^1([0,T],\mathrm{W}^{1,p}(\Omega,\mathbb{R}^d)^*)$ and that $q_0 \in \mathcal{S}(0)$. Then, the energetic formulation (S) & (E) associated with $\mathcal{E}$ and $\mathcal{D}$ from (5.1) *and* (5.2)*, respectively, has a solution $q : [0,T] \to \mathcal{Q}_0$.*

The theory of Γ-convergence discussed in Sect. 4.1 can also be used to show that space-time discretizations of the energetic formulation contain subsequences which converge to energetic solutions. We use triangulations $\mathcal{T}_h$ of Ω. Moreover, we approximate gradient Young measures by sequential laminates of order $\kappa \in \mathbb{N}$, see [3, 8, 39, 71] and Sect. 5.4 in [47] for an introduction. Fixing $\kappa \in \mathbb{N}$ we define $\mathcal{Q}^h_\kappa$ as the space of functions $q = (\varphi,\mu,z) \in \mathcal{Q}$ for which $\nabla\varphi, \mu$ and z are constant on each simplex and μ is a laminate of order at most κ. Using the penalization parameter $\varepsilon > 0$ we define

$$\mathcal{E}_{h,\varepsilon}(t,q) = \mathcal{E}(t,q) + \tfrac{1}{\varepsilon}\textstyle\int_\Omega |z - \zeta\bullet\mu|^2\,\mathrm{d}x \text{ for } q \in \mathcal{Q}^n_\kappa \text{ with } \nabla\varphi = \mathrm{id}\bullet\mu$$

and $\mathcal{E}_{h,\varepsilon}(t,q) = +\infty$ otherwise on $\mathcal{Q}$.

In [38] a function $H : (0,\infty) \to (0,\infty)$ with $H(\varepsilon) \to 0$ for $\varepsilon \searrow 0$ is constructed such that the following holds: *If $(\Pi_k)_{k\in\mathbb{N}}$ is a sequence of partitions of $[0,T]$ with $\phi(\Pi_k) \to 0$, if $(\mathcal{T}_{h_k})$ is a sequence of triangulations of Ω and if $\varepsilon_k \to 0$ with $h_k \le H(\varepsilon_k)$, then the sequence of incremental solutions $\overline{q}_k$ associated with $(\mathcal{E}_{h_k,\varepsilon_k},\mathcal{D})$ on $\mathcal{Q}^h_\kappa$ contains a subsequence which converges to an energetic solution $q : [0,T] \to \mathcal{Q}$ for $(\mathcal{E},\mathcal{D})$.*

Figures 2 and 3 show results from a numerical simulation from [38] for a sample of 4×4×9 mm single-crystal alloy of CuAlNi. It has a cubic-to-orthorhombic phase transition with one austenite and 6 variants of a martensite (i.e., $N = 7$).

All phases are modeled by a Saint-Venant-Kirchhoff material $W_j(x,A) = \frac{1}{2}(A^\top A - C_j) : \mathbb{C}_j : (A^\top A - C_j) + d_j$, where $C_j \in \mathbb{R}^{3\times3}_{\rm sym}$, $\mathbb{C}_j \in {\rm Lin}(\mathbb{R}^{3\times3}_{\rm sym})$, and d_j are the experimentally measured values for each $j \in \{1,\dots,7\}$ at a fixed temperature of 312 K. The dissipation constants $D(e_j,e_\ell)$ are chosen to be 0.5 MPa for transformations between martensite and austenite or vice versa. Transformations between different variants of austenite are assumed to have much lower dissipation thresholds.

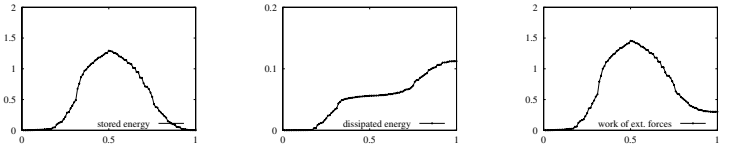


Fig. 3. Energies (in Joule) during the cycle. Left: stored energy $\mathcal{E}(t,q(t))$, middle: dissipated energy ${\rm Diss}_{\mathcal{D}}(q,[0,t])$ and right: work of external forces $\int_0^t \partial_s\mathcal{E}(s,q(s))\,{\rm d}s$

The discretization involves 180 tetrahedrons and second-order laminates. This leads to 20 degrees of freedom in each element, which lie in a nonlinear manifold with boundary (box constraints). The minimization technique for solving the highly nonconvex incremental problem is described in Sect. 6.3 in [38].

A microscopic model that does not allow for microstructure and is based on pure phases only is developed and analyzed in [42]. It uses an interfacial energy that is proportional to the surface of the interfaces. We denote the set of pure phases by $P_N = \{e_1,\dots,e_N\} \subset Z_N \in \mathbb{R}^N$ and let $\mathcal{Z}_{\rm pure} = {\rm L}^1(\Omega,P_N)$ equipped with the strong ${\rm L}^1$-topology. The space $\mathcal{F}$ of admissible deformations remains as above, whereas the energy functional $\mathcal{E} : [0,T]\times\mathcal{Q} \to \mathbb{R}_\infty$ takes the form

$$\mathcal{E}(t,\varphi,z) = \int_\Omega W_{j(x)}(x,\nabla\varphi(x))\,{\rm d}x + \rho\int_\Omega |{\rm D}z| - \langle \ell(t),\varphi\rangle\ ,$$

where $j(x) = k \ \Leftrightarrow\ z(x) = e_k$ and where $\int_\Omega |{\rm d}z|$ denotes the total variation

$$\int_\Omega |{\rm D}z| \stackrel{\rm def}{=} \sup\{\ \int_\Omega z\cdot{\rm div}\psi\,{\rm d}x \mid \psi\in{\rm C}^1(\Omega,\mathbb{R}^{n\times d}),\ |||\psi(x)||| \le 1 \text{ on } \Omega\ \}\ .$$

The norm $|||\cdot|||$ on $\mathbb{R}^{N\times d}$ can be adjusted to anisotropies in $\Omega\subset\mathbb{R}^d$ and to different weights for the interfaces between phases j and k, see [42] for the details and for generalizations.

Using the same dissipation distance $\mathcal{D}$ as above, an existence theory as in Theorem 3 can be derived, since ${\rm BV}(\Omega)$ embeds compactly into ${\rm L}^1(\Omega)$. The solution $q = (\varphi,z) : [0,T] \to \mathcal{Q}$ now satisfies

$$\begin{aligned}&\varphi \in {\rm L}^\infty([0,T],{\rm W}^{1,p}(\Omega,\mathbb{R}^d)) \text{ and}\\ &z \in {\rm BV}([0,T],{\rm L}^1(\Omega,P_N)) \cap {\rm L}^\infty_{\rm weak}([0,T],{\rm BV}(\Omega,\mathbb{R}^N))\ .\end{aligned}$$

5.3 Mesoscopic Models

Often it is not desirable or prohibitly costly to calculate the evolution of the microstructure during the hysteretic evolution process. If these details are not needed and if volume fractions or other effective quantities are sufficient, then simpler models may be used.

If we only care about volume fractions, then the *mixture function* can be used to describe the effective behavior of phase mixtures. Let $W_j(x,\cdot)$ be given as above for $j=1,\dots,N$. For $z\in Z_N$ and $A\in\mathbb{R}^{d\times d}$ we define

$$\mathbb{W}(x,A,z)=\inf\Big\{\int\limits_{(0,1)^d} W_{J(y)}(x,A+\nabla\psi(y))\,\mathrm{d}y \;\Big|\; J\in \mathrm{L}^1((0,1)^d,\{1,...,N\}),\ \int\limits_{(0,1)^d} e_{J(y)}\,\mathrm{d}y=z,\ \psi\in \mathrm{W}_0^{1,\infty}((0,1)^d,\mathbb{R}^d)\Big\}$$

where $(0,1)^d$ is a microscopic representative volume element, J a microscopic phase indicator, and $\nabla\psi$ microscopic fluctuation of the gradient. In [41] $\mathbb{W}$ is also called *cross-quasiconvexification* and in [25] the *free energy of mixing*.

Unfortunately, in general situations it is almost impossible to calculate $\mathbb{W}$ explicitly. Nevertheless $\mathbb{W}$ is cross-quasiconvex and hence, for each $x\in\Omega$ and $A\in\mathbb{R}^{d\times d}$, the function $\mathbb{W}(x,A,\cdot):Z_N\to[0,\infty)$ is convex and, for each $x\in\Omega$ and $z\in Z_N$, the function $\mathbb{W}(x,\cdot,z):\mathbb{R}^{d\times d}\to[0,\infty)$ is quasiconvex. Explicit formulas are only available in dimension $d=1$ or if each W_j is quadratic with an elastic tensor independent of j, viz.,

$$W_j(x,A)=\tfrac{1}{2}(\varepsilon(A)-\varepsilon_j(x)):\mathbb{C}(x):(\varepsilon(A)-\varepsilon_j(x))+d_j(x)$$

where $\varepsilon(A)=\frac{1}{2}(A+A^\top-2I)$. Then,

$$\mathbb{W}(x,A,z)=\textstyle\sum_{j=1}^N z_jW_j(x,A)+w_{\mathrm{mix}}(x,z)\ ,$$

where $w_{\mathrm{mix}}(x,e_k)=0$ and $w_{\mathrm{mix}}(x,\cdot):Z_N\to\mathbb{R}$ is convex. See [25, 46] for cases where w_{mix} can be calculated or estimated efficiently.

The advantage of the mixture theory is that we are not forced to work with quasiconvexity. We are able to use polyconvexity as well. Hence it is possible to use energy densities that take the value $+\infty$, as for instance in finite-strain elasticity where $W(x,A)=+\infty$ for $\det A\le 0$. Instead of cross-quasiconvexity we may use cross-polyconvexity, i.e., $\mathbb{W}(x,\cdot,\cdot):\mathbb{R}^{d\times d}\times Z_N\to[0,\infty]$ is called *cross-polyconvex*, if there exists a function $g(x,\cdot):\mathbb{R}^{m_d+N}\to[0,\infty]$ that is convex, lower semi-continuous and satisfies

$$\mathbb{W}(x,A,z)=g(x,\mathcal{M}(A),z)\ ,$$

where $\mathcal{M}(A)\in\mathbb{R}^{m_d}$ is the set of all minors.

We now define the state space $\mathcal{Q}=\mathcal{F}\times\mathcal{Z}$ for classical functions φ only, namely

$$\mathcal{F} = \{\, \varphi \in \mathrm{W}^{1,p}(\Omega,\mathbb{R}^d) \mid \varphi|_{\Gamma_{\mathrm{Dir}}} = \mathrm{id} \,\} \subset \mathrm{W}^{1,p}(\Omega,\mathbb{R}^d)$$

equipped with the weak topology. The stored-energy functional takes the form

$$\mathcal{E}(t,\varphi,z) = \textstyle\int_\Omega \mathbb{W}(x,\nabla\varphi(x),z(x)) + \frac{\rho}{2}|\nabla^\alpha z|^2 \,\mathrm{d}x - \langle \ell(t),\varphi\rangle \ . \tag{5.4}$$

For $\rho > 0$ and $\alpha > 0$ we take $\mathcal{Z} = \mathrm{L}^1(\Omega, Z_N)$ equipped with the strong topology. Under suitable coercivity and (poly)quasiconvexity assumptions on $\mathbb{W}(x,\cdot,z)$ it can then be shown that the sublevels of $\mathcal{E}(t,\cdot)$ are compact in $\mathcal{Q}$, which is our basic condition (3.4). In the case $\rho = 0$, this is more difficult, since $\mathcal{Z}$ then has to be equipped with the weak topology. Then, cross-(poly)quasiconvexity is necessary for weak lower semi-continuity of $\mathcal{E}$. However, for the case without regularization ($\rho = 0$) the best we can hope for is that solutions for the incremental problem $(\mathrm{IP})_\Pi$ exist. The passage to the limit of vanishing time incrementals strongly relies on the closedness condition (3.8) for the stable sets which, so far, cannot be established in cases without regularization.

The following result is a slight variant of the existence results in [22, 42, 48].

Theorem 4. *Let $p \in (1,\infty)$, $\alpha,\rho > 0$ and $\ell \in \mathrm{C}^1([0,T],\mathrm{W}^{1,p}(\Omega,\mathbb{R}^d)^*)$. Moreover, let $\mathcal{D}$ be given as in Sect. 5.2 and assume that $\mathcal{E}$ in* (5.4) *has compact sublevels in $\mathcal{Q} \subset \mathrm{W}^{1,p}(\Omega,\mathbb{R}^d)_{\mathrm{weak}} \times \mathrm{L}^1(\Omega,Z_N)_{\mathrm{strong}}$. Then, for each stable initial state $q_0 = (\varphi_0,z_0) \in \mathcal{Q}$ there exists an energetic solution $q = (\varphi,z) : [0,T] \to \mathcal{Q}$ for $(\mathcal{E},\mathcal{D})$ with $\varphi \in \mathrm{L}^\infty([0,T],\mathrm{W}^{1,p}(\Omega,\mathbb{R}^d))$ and $z \in \mathrm{BV}([0,T],\mathrm{L}^1(\Omega,Z_N)) \cap \mathrm{L}^\infty([0,T],\mathrm{H}^\alpha(\Omega,\mathbb{R}^N))$.*

The associated numerical convergence results are discussed in [54]. But all the above models have the disadvantage that the solutions are not unique. Hence, it is not possible to show that numerical solutions converge.

The next model goes back to [76] and was further developed in [4, 5]. This model is based on the linearized strain tensor $\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^\top)$ and the mesoscopic transformation strain $z \in Z = \{\, A \in \mathbb{R}^{d\times d} \mid A = A^\top, \mathrm{tr}\, A = 0 \,\}$. The dissipation is simply a multiple of the L^1-norm:

$$\mathcal{D}(z,\widetilde{z}) = \mathcal{R}(\widetilde{z}-z) = \textstyle\int_\Omega c_d|\widetilde{z}(x)-z(x)|\,\mathrm{d}x \ .$$

The energy functional takes the form

$$\mathcal{E}(t,u,z) = \textstyle\int_\Omega W(x,\varepsilon(u),z) + h(|z|) + \frac{\rho}{2}|\nabla^\alpha z|^2 \,\mathrm{d}x - \langle \ell(t),u\rangle \ ,$$

with $W(x,\varepsilon,z) = \frac{1}{2}(\varepsilon-z){:}\mathbb{C}{:}(\varepsilon-z)$. Again, the classical model has no regularization, i.e., $\rho = 0$. The hardening function $h : [0,\infty) \to [0,\infty]$ equals

$$h(r) = \begin{cases} c_1\sqrt{\delta^2+r^2} + c_2 r^2 & \text{for } r \in [0,r_*]\,, \\ \infty & \text{otherwise}\,, \end{cases}$$

in [5] and has $\delta = 0$ in [76]. In these cases it is easy to solve the incremental problems $(\mathrm{IP})_\Pi$ in the space $\mathcal{Q} = \mathrm{H}^1(\Omega,\mathbb{R}^d) \times \mathrm{H}^\alpha(\Omega,Z)$, for all $\alpha \geq 0$. However, for obtaining energetic solutions as stated in Theorem 4 it is necessary

to impose that ρ and α are strictly positive. Only then $\mathcal{D}$ is weakly continuous on $\mathcal{Z} = \mathrm{H}^\alpha(\Omega, Z)$.

A further variation is considered in [4], where h is replaced by a smooth, convex function taking finite values and growing at most quadratically, e.g.,

$$h(r) = c_1\sqrt{\delta^2 + r^2} + c_2 r^2 + \tfrac{c_3}{\delta}\tfrac{\max\{0, r-r_*\}^4}{r_*^2+r^2} \ .$$

Then, for $\alpha \geq d/6$ it can be shown that $\mathcal{E}(t,\cdot) : \mathrm{H}^1(\Omega,\mathbb{R}^d) \times \mathrm{H}^\alpha(\Omega, Z) \to \mathbb{R}$ is three times differentiable and uniformly convex. Hence, the theory of Sect. 7 in [58] is applicable. This allows us to conclude uniqueness of the solutions as well as strong convergence of the solutions of the incremental problems. In fact, the convergence rate is $(\phi(\Pi_k))^{1/2}$. In [4] also the convergence of spatial discretization will be discussed.

5.4 Temperature-Induced Phase Transformation

The original shape-memory effect is based on cooling and heating to switch between martensite occurring in several variants and the single austenite phase. So far the energetic formulation is only available for the isothermal case and thus is suited for stress-induced phase transformations only.

There is at least one nonisothermal case that can be treated via the energetic formulation as well, namely if the temperature field is prescribed a priori independently of the solution to be calculated. This means that the deformation and phase transformation process is so slow that all latent heat which is either consumed or generated via phase transformation can be transported via heat conduction into the environment.

Thus, our model is based on a temperature dependent stored-energy density $W(x, A, z, \theta)$ which is assumed to satisfy

$$\begin{aligned} &\exists\, c_3^W \in \mathbb{R}\ \exists\, c_4^W > 0\ \forall\, x \in \Omega\ \forall\, A \in \mathbb{R}^{d\times d}\ \forall\, z \in Z_N\ \forall\, \theta > 0\colon \\ &|\partial_\theta W(x, A, z, \theta)| \leq c_4^W (W(x, A, z, \theta) + c_3^W) \ . \end{aligned} \tag{5.5}$$

The given temperature profile θ should satisfy $(\log\theta) \in \mathrm{C}^1([0,T] \times \overline{\Omega})$, then the energy potential, which for simplicity is now without external forcing, takes the form

$$\mathcal{E}(t, \varphi, z) = \textstyle\int_\Omega W(x, \nabla\varphi(x), z(x), \theta(t,x))\,\mathrm{d}x$$

and the power associated to the temperature changes is

$$\partial_t\mathcal{E}(t, \varphi, z) = \textstyle\int_\Omega \partial_\theta W(x, \nabla\varphi(x), z(x), \theta(t,x))\partial_t\theta(t,x)\,\mathrm{d}x \ .$$

Using (5.5) it is easy to establish the condition (3.5) and, under suitable additional assumptions, the conditions (3.6) and (3.7) hold as well. In [49] we will provide the detailed assumptions for a full existence theory.

5.5 Poling Induced Piezoelectricity

Multifunctional materials derive their functionality from the combination of several properties such as elasticity, polarizability, and magnetizability. For such materials the polarization p or the magnetization m may be considered as the variable z used above. However, in addition we have to take into account the relevant version of the Maxwell equation.

In the quasi-static setting either the electric or the magnetic field vanishes such that we obtain two clearly distinguished cases, which are dual in a certain sense. Throughout we will restrict ourselves to the case of small strains, since otherwise the Maxwell equations have to be solved in the deformed configuraton, see the references at the beginning of Sect. 5.6.

The electric field E and the dielectric displacement D are defined on all of $\mathbb{R}^d$ whereas the polarization $P : \Omega \to \mathbb{R}^d$ is defined on the body only. These fields are related by the constitutive relation

$$D = \varepsilon_0 E + P \text{ in } \Omega \quad \text{and} \quad D = \varepsilon_0 E \text{ in } \mathbb{R}^d \backslash \Omega \,.$$

The reduced Maxwell equations are

$$\operatorname{div} D = 0 \text{ and } \operatorname{curl}(E - E_{\mathrm{ext}}(t,\cdot)) = 0 \text{ in } \mathbb{R}^d \,, \tag{5.6}$$

where $\operatorname{curl} \widetilde{E} = \nabla \widetilde{E} - (\nabla \widetilde{E})^\top$. We will implement these equations as part of the energetic formulation.

We consider the displacement $u : \Omega \to \mathbb{R}^d$ and the dielectric displacement D as variables in the space

$$\begin{aligned} \mathcal{F} &= \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega, \mathbb{R}^d) \times \mathrm{L}^2_{\mathrm{div}}(\mathbb{R}^d, \mathbb{R}^d) \\ \text{with } \mathrm{L}^2_{\mathrm{div}}(\mathbb{R}^d, \mathbb{R}^d) &= \{\, D \in \mathrm{L}^2(\mathbb{R}^d, \mathbb{R}^d) \mid \operatorname{div} D = 0 \,\} \,. \end{aligned}$$

The internal variable $p \in \mathcal{Z} = H^1(\Omega, \mathbb{R}^d)$ is the remanent polarization. For $q = (u, D, p) \in \mathcal{Q} = \mathcal{F} \times \mathcal{Z}$ and $t \in [0, T]$, the energy potential $\mathcal{E}$ is defined via

$$\begin{aligned} \mathcal{E}(t,q) = \textstyle\int_\Omega W(x, \varepsilon(u), p) - \frac{1}{\varepsilon_0} D \cdot P(x, \varepsilon(u), p) + \frac{\rho}{2} |\nabla p|^2 \,\mathrm{d}x \\ + \textstyle\int_{\mathbb{R}^d} \frac{1}{2\varepsilon_0} |D|^2 \,\mathrm{d}x - \langle \ell(t), (u, D) \rangle \end{aligned}$$

where the external forcing occurs via mechanical volume and surface loadings and via an external electric field

$$\langle \ell(t), (u, D) \rangle = \textstyle\int_\Omega f_{\mathrm{ext}}(t) \cdot u \,\mathrm{d}x + \int_{\Gamma_{\mathrm{Neu}}} g_{\mathrm{ext}}(t) \cdot u \,\mathrm{d}a + \int_{\mathbb{R}^d} E_{\mathrm{ext}}(t) \cdot D \,\mathrm{d}x \,.$$

The electric field is the dual variable to the dielectric displacement D, i.e.,

$$E = \tfrac{1}{\varepsilon_0}(D - P(x, \varepsilon(u), p)) \text{ in } \Omega \quad \text{and} \quad E = \tfrac{1}{\varepsilon_0} D \text{ in } \mathbb{R}^d \backslash \Omega \,. \tag{5.7}$$

The polarization is given as a constitutive function and poling induced piezoelectricity means that the piezoelectric tensor $\partial_\varepsilon P$ does not vanish.

Following [33, 69] the dissipation distance is the Legendre transform of the so-called switching function, namely

$$\mathcal{D}(p,\widetilde{p}) = \mathcal{R}(\widetilde{p}-p) = \int_\Omega R(x,\widetilde{p}(x)-p(x))\,\mathrm{d}x$$

with a Caratheodory function $R : \Omega \times \mathbb{R}^d \to [0,\infty)$ with $R(x,\cdot)$ being convex and 1-homogeneous. Under the assumption that $W(x,\cdot,\cdot,p) : \mathbb{R}^{d\times d}_{\mathrm{sym}} \times \mathbb{R}^d \to \mathbb{R}$ is convex and that W satisfies suitable upper and lower bounds, it is now straightforward to prove the existence of energetic solutions $(u,D,p) : [0,T] \to \mathcal{Q}$ with $(u,D) \in \mathrm{L}^\infty([0,T],\mathcal{F})$ and $p \in \mathrm{BV}([0,T],\mathrm{L}^1(\Omega,\mathbb{R}^d)) \cap \mathrm{L}^\infty([0,T],H^1(\Omega,\mathbb{R}^d))$.

To see the compatibility with the Maxwell equations (5.6) we note that the stability condition (S) implies that for all $t \in [0,T]$ we have

$$\mathrm{D}_D\mathcal{E}(t,u(t),D(t),p(t))[\widehat{D}] = 0 \text{ for all } \widehat{D} \in \mathrm{L}^2_{\mathrm{div}}(\mathbb{R}^d,\mathbb{R}^d)\,.$$

In Proposition 2.1 of [60] it is shown that the latter relation is equivalent to the Maxwell equations (5.6), if the definition (5.7) is used.

Moreover, in that work additional conditions are discussed which imply also uniqueness of solutions. For this the uniqueness theory of Sect. 7 in [58] is employed. However, the resulting conditions seem to be very restrictive.

5.6 Magnetostrictive Materials

We summarize the theory of [17] which is based on small-strain elasticity, see also [54]. For the much more complicated constitutive theory in the case of finite-strain elasticity we refer to [16, 31] and for some analysis for the static problem with second-order regularization of the deformation we refer to [73]. For small strain-models including microstructure via Young measures (like in Sect. 5.2) we refer to [71, 72].

In analogy to the case of polarizable materials we use the magnetization $m : \Omega \to \mathbb{R}^d$ as an internal variable. Usually the saturation assumption $|m(x)| = m_{\mathrm{sat}} > 0$ is added which we impose by letting $Z = \{\, m \in \mathbb{R}^d \mid |m| = m_{\mathrm{sat}} \,\}$. The magnetic induction $B : \mathbb{R}^d \to \mathbb{R}^d$ and the magnetic field $H : \mathbb{R}^d \to \mathbb{R}^d$ are related via the constitutive law

$$B = \mu_0(H+m) \text{ in } \Omega \quad \text{and} \quad B = \mu_0 H \text{ in } \mathbb{R}^d\backslash\Omega\,.$$

In this quasistatic setting the Maxwell equations reduce to

$$\mathrm{div} B = 0 \quad \text{and} \quad \mathrm{curl}\, H = 0 \text{ in } \mathbb{R}^d\,. \tag{5.8}$$

We choose $\mathcal{F} = \mathrm{H}^1_{\Gamma_{\mathrm{Dir}}}(\Omega,\mathbb{R}^d) \times \mathrm{L}^2_{\mathrm{div}}(\mathbb{R}^d,\mathbb{R}^d)$ equipped with the weak topology and $\mathcal{Z} = \mathrm{L}^1(\Omega,Z)$ with the strong topology. The energy potential reads

$$\mathcal{E}(t,u,B,m) = \int_\Omega W(x,\varepsilon(u),m) - B\cdot m + \tfrac{\rho}{2}|\nabla m|^2\,\mathrm{d}x$$

$$+ \int_{\mathbb{R}^d} \tfrac{1}{2\mu_0}|B|^2 \,\mathrm{d}x - \langle \ell(t), (u,B)\rangle$$

with an external forcing of the form

$$\langle \ell(t), (u,B)\rangle = \int_\Omega f_{\mathrm{ext}}(t)\cdot u \,\mathrm{d}x + \int_{\Gamma_{\mathrm{Neu}}} g_{\mathrm{ext}}(t)\cdot u \,\mathrm{d}a + \int_{\mathbb{R}^d} H_{\mathrm{ext}}(t)\cdot B \,\mathrm{d}x \,.$$

The parameter $\sqrt{\rho}$ relates to the exchange length, which determines the scalings for the width of domain walls. The dissipation distance may be chosen via an arbitrary distance $D(x,\cdot,\cdot)$ on $Z = m_{\mathrm{sat}}\mathbb{S}^{d-1}$, e.g.,

$$D(x,m,\widetilde{m}) = c_1 \arccos\big(\tfrac{m\cdot\widetilde{m}}{m_{\mathrm{sat}}^2}\big) + c_2|\widehat{e}\cdot(m-\widetilde{m})|$$

where $\widehat{e}$ is an "easy" axis and $c_2 = 0$ in the isotropic case. We define $\mathcal{D}(m,\widetilde{m}) = \int_\Omega D(x,m(x),\widetilde{m}(x))\,\mathrm{d}x$.

Using the standard coercivity assumptions on $W : \Omega\times\mathbb{R}^{d\times d}_{\mathrm{sym}}\times Z \to [0,\infty)$, convexity in $\varepsilon(u)$ and continuity in $m\in Z$, it is standard to show that $\mathcal{E}(t,\cdot) : \mathcal{Q} = \mathcal{F}\times\mathcal{Z} \to \mathbb{R}$ is lower semi-continuous with compact sublevels. Moreover $\mathcal{D} : \mathcal{Z}\times\mathcal{Z} \to [0,\infty)$ is continuous in the strong L^1-topology (or in the weak H^1-topology). Thus, existence of energetic solutions for $(\mathcal{E},\mathcal{D})$ can be obtained from Theorem 1.

Since the magnetic field H is the dual variable to B, the equation

$$\mathrm{D}_B\mathcal{E}(t,u(t),B(t),m(t))[\widehat{B}] = 0 \text{ for all } \widehat{B}\in \mathrm{L}^2_{\mathrm{div}}(\mathbb{R}^d,\mathbb{R}^d)$$

is equivalent to (5.8) in the form

$$\mathrm{div} B = 0 \;\text{ and }\; \mathrm{curl}(\tfrac{1}{\mu_0}B - H_{\mathrm{ext}} - \chi_\Omega m) = 0 \;\text{ in } \mathbb{R}^d \,.$$

It is more common to formulate the problem of magnetostriction in terms of the potential U of the magnetic field H, i.e., $H = \nabla U$. In the above formulation we may then replace B via

$$B = \mu_0(\nabla U + H_{\mathrm{ext}}(t) + \chi_\Omega m) \tag{5.9}$$

in the energy $\mathcal{E}$ to arrive at

$$\begin{aligned}\widetilde{\mathcal{E}}(t,u,U,m) = \int_\Omega W(x,\varepsilon(u),m) - \tfrac{\mu_0}{2}|m|^2 - m\cdot H_{\mathrm{ext}}(t)\,\mathrm{d}x \\ + \int_{\mathbb{R}^d} \tfrac{\mu_0}{2}|\nabla U|^2 - \tfrac{\mu_0}{2}|H_{\mathrm{ext}}(t)|^2 \,\mathrm{d}x - \langle \ell_{\mathrm{mech}}(t), u\rangle \,.\end{aligned}$$

Note that the Euler-Lagrange equation for U does not supply the desired Maxwell equation

$$\mathrm{div}(\nabla U + H_{\mathrm{ext}}(t,\cdot) + \chi_\Omega m) = 0 \;\text{ in } \mathbb{R}^d \,. \tag{5.10}$$

Thus, to derive an energetic formulation in this situation, the variable U has to be taken as a function of $m\in L^1(\Omega,Z)$ and $t\in[0,T]$ via $H_{\mathrm{ext}}(t,\cdot)$, namely $U = \mathcal{U}(t,m)$ being the solution of (5.10).

Instead of simply replacing B by the corresponding variable, we might as well perform a partial Legendre transform such that $\widetilde{W}(f, x, \varepsilon(u), B, m)$ is replaced by

$$\widehat{W}(t, x, \varepsilon, H, m) = \widetilde{W}(t, x, \varepsilon, B, m) - B \cdot \partial_B \widetilde{W}(t, x, \varepsilon, B, m)$$

where B is again eliminated by the use of (5.9). The corresponding energy $\widehat{\mathcal{E}}$ then contains the negative definite term $-\int_{\mathbb{R}^d} \frac{\mu_0}{2} |\nabla U|^2 \, \mathrm{d}x$. Thus, we may use $\mathrm{D}_U \widehat{\mathcal{E}}(t, u, U, m)[\widehat{U}] = 0$ to obtain (5.10), but the saddle point structure of $\widehat{\mathcal{E}}$ does not allow us to introduce a stability condition in terms of (u, U, m). Thus, it is not possible to derive an energetic formulation either.

6 Conclusions

The energetic formulation of rate-independent processes was developed much further via the abstract approaches described in [22, 43, 48]. The major improvement occurred through finding abstract versions of the ideas in [15] for treating a rate-independent model for crack growth. Now it is possible to deal with problems where the energy $\mathcal{E}(t, \cdot, z) : \mathcal{F} \to \mathbb{R}_\infty$ is non-convex. In general, the abstract theory is available in topological spaces without any linear structure. Thus, it is possible to treat finite-strain elasticity (cf. [22, 34]) as well as internal variables which lie in general nonconvex sets such as in magnetism (cf. Sect. 5.6) or in finite-strain plasticity, see [26]. Moreover, it is possible to include Young measures into the state space as well [38, 53].

Further developments include the abstract theory of Γ-convergence and relaxations of the energetic formulation. This allows us, for instance, to treat numerical approximations, see [4, 38, 54]. However, the numerical analysis and efficient simulations still need a lot of further developments.

The major drawback of the energetic formulation is that there are only very few results on the uniqueness of solutions, see [10, 51, 58]. Another deficiency concerns the fact that the stability condition (S) involves a global stability condition. For a better physical modeling and for numerical implementation it would be desirable to replace this condition by a suitable local stability condition. First attempts are given in [18, 52], but a reasonable general theory is not yet developed. This is closely related to the general problem how these rate-independent models can be embedded into more general dynamical problems, for instance, models including rate-dependent heat conduction, viscous effects or even kinetic terms.

On the side of material modeling there is now quite a variety of models for shape-memory materials. There it is possible to describe models on many different length scales. However, the question of upscaling and deriving effective models on larger scales needs further investigations. The relaxations and Γ-convergence results in Sect. 4 might be a good basis for doing this, see also [56]. A first step in two-scale homogenization will be developed in [62].

Moreover, evolutionary models for microstructures and textures will certainly be important future areas where the energetic formulation can be helpful.

The strength of the energetic formulation is that it can model the statics extremely well by adjusting the energy-storage functional $\mathcal{E}$ according to experiments, see, e.g., [38]. However, the modeling of the dissipation distances, which contains the only information on the dynamics, is not supported very well by experiments. In this sense, the energetic formulation provides a first mathematical step to well-posed evolutionary models for complex material behavior.

Acknowledgments. The research was supported by DFG within the Collaborative Research Center SFB 404 *Multifield Problems in Fluid and Solid Mechanics*, subproject C7. Further support by the European Union under HPRN-CT-2002-00284 *Smart Systems: New Materials, Adaptive Systems and their Nonlinearities* and by the DFG Research Center MATHEON, subproject C18. The author is grateful to G. Francfort, A. Mainik, C. Miehe, M. Ortiz, R. Rossi, T. Roubíček, and U. Stefanelli for stimulating discussions.

References

1. J. Alberty and C. Carstensen. Numerical analysis of time-depending primal elastoplasticity with hardening. *SIAM J. Numer. Anal.*, 37:1271–1294 (electronic), 2000.
2. L.-E. Andersson and A. Klarbring. A review of the theory of static and quasi-static frictional contact problems in elasticity. *R. Soc. Lond. Philos. Trans. Ser. A*, 359:2519–2539, 2001.
3. S. Aubry, M. Fago, and M. Ortiz. A constrained sequential-lamination algorithm for the simulation of sub-grid microstructure in martensitic materials. *Comput. Methods Appl. Mech. Engrg.*, 192:2823–2843, 2003.
4. F. Auricchio, A. Mielke, and U. Stefanelli. A rate-independent model for the isothermal quasi-static evolution of shape-memory materials. *In preparation*, 2006.
5. F. Auricchio and L. Petrini. Improvements and algorithmical considerations on a recent three-dimensional model describing stress-induced solid phase transformations. *Int. J. Numer. Meth. Engng*, 55:1255–1284, 2002.
6. J. Ball. Constitutive inequalities and existence theorems in nonlinear elastostatics. In *Nonlinear Analysis and Mechanics: Heriot-Watt Symposium (Edinburgh, 1976), Vol. I*, pages 187–241. Res. Notes in Math., No. 17. Pitman, London, 1977.
7. J. M. Ball and R. D. James. Fine phase mixtures as minimizers of energy. *Arch. Rational Mech. Anal.*, 100(1):13–52, 1987.
8. S. Bartels, C. Carstensen, K. Hackl, and U. Hoppe. Effective relaxation for microstructure simulations: algorithms and applications. *Comput. Methods Appl. Mech. Engrg.*, 193:5143–5175, 2004.
9. K. Bhattacharya. *Microstructure of Martensite. Why it Forms and How it Gives Rise to the Shape-Memory Effect.* Oxford University Press, New York, 2003.

10. M. Brokate, P. Krejčí, and H. Schnabel. On uniqueness in evolution quasivariational inequalities. *J. Convex Analysis*, 11:111–130, 2004.
11. M. Brokate and J. Sprekels. *Hysteresis and Phase Transitions*. Springer-Verlag, New York, 1996.
12. A. Chambolle. A density result in two-dimensional linearized elasticity and applications. *Arch. Rat. Mech. Analysis*, 167:211–233, 2003.
13. P. G. Ciarlet. *Mathematical Elasticity. Vol. I: Three-Dimensional Elasticity*. North-Holland Publishing Co., Amsterdam, 1988.
14. S. Conti and F. Theil. Single-slip elastoplastic microstructures. *Arch. Rational Mech. Analysis*, 178:125–148, 2005.
15. G. Dal Maso, G. Francfort, and R. Toader. Quasistatic crack growth in nonlinear elasticity. *Arch. Rat. Mech. Anal.*, 176:165–225, 2005.
16. A. Dorfmann and R. W. Ogden. Some problems in nonlinear magnetoelasticity. *Z. Angew. Math. Phys.*, 56:718–745, 2005.
17. M. Efendiev and A. Mielke. An energetical model for evolution in magnetostriction. *In preparation*, 2006.
18. M. Efendiev and A. Mielke. On the rate-independent limit of systems with dry friction and small viscosity. *J. Convex Analysis*, 13(1):151–167, 2006.
19. M. Efendiev, A. Mielke, R. Rossi, and G. Savaré. On a metric approach to a class of doubly nonlinear evolution equations and applications to rate-independent problems. *In preparation*, 2006.
20. G. Francfort and A. Garroni. A variational view of brittle damage evolution. *Arch. Rational Mech. Anal.*, 2006. To appear.
21. G. Francfort and J.-J. Marigo. Revisiting brittle fracture as an energy minimization problem. *J. Mech. Phys. Solids*, 46:1319–1342, 1998.
22. G. Francfort and A. Mielke. Existence results for a class of rate-independent material models with nonconvex elastic energies. *J. reine angew. Math.*, 2006. In print.
23. M. Frémond. *Non-Smooth Thermomechanics*. Springer-Verlag, Berlin, 2002.
24. M. Frémond and S. Miyazaki. *Shape Memory Alloys*. Springer-Verlag, Wien, 1996.
25. S. Govindjee, A. Mielke, and G. Hall. The free–energy of mixing for n–variant martensitic phase transformations using quasi–convex analysis. *J. Mech. Physics Solids*, 50:1897–1922, 2002. Erratum and Correct Reprinting: 51(4) 2003, pp. 763 & I-XXVI.
26. E. Gürses, A. Mainik, C. Miehe, and A. Mielke. Analytical and numerical methods for finite-strain elastoplasticity. In *SFB404-Abschlussband*. Springer-Verlag, 2006.
27. K. Hackl. Generalized standard media and variational principles in classical and finite strain elastoplasticity. *J. Mech. Phys. Solids*, 45(5):667–688, 1997.
28. K. Hackl and U. Hoppe. On the calculation of microstructures for inelastic materials using relaxed energies. In C. Miehe, editor, *IUTAM Symposium on Computational Mechanics of Solids at Large Strains*, pages 77–86. Kluwer, 2003.
29. B. Halphen and Q. S. Nguyen. Sur les matériaux standards généralisés. *J. Mécanique*, 14:39–63, 1975.
30. W. Han and B. D. Reddy. *Plasticity (Mathematical Theory and Numerical Analysis)*, volume 9 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, 1999.

31. R. D. James. Configurational forces in magnetism with application to the dynamics of a small-scale ferromagnetic shape memory cantilever. *Contin. Mech. Thermodyn.*, 14:55–86, 2002.
32. C. Johnson. Existence theorems for plasticity problems. *J. Math. Pures Appl. (9)*, 55(4):431–444, 1976.
33. M. Kamlah. Ferroelectric and ferroelastis piezoceramics-modelling of electromechanical hysteresis phenomena. *Continuum Mech.Thermodyn.*, 13(4):219–268, 2001.
34. D. Knees and A. Mielke. Energy release rate for cracks in finite-strain elasticity. *Math. Methods Applied Sciences*, 2006. Submitted (WIAS Preprint 1100).
35. M. Kočvara, A. Mielke, and T. Roubíček. A rate–independent approach to the delamination problem. *Math. Mech. Solids*, 2005. In print, online June 2005.
36. M. A. Krasnosel'skiĭ and A. V. Pokrovskiĭ. *Systems With Hysteresis*. Springer-Verlag, Berlin, 1989.
37. P. Krejčí. Evolution variational inequalities and multidimensional hysteresis operators. In *Nonlinear differential equations (Chvalatice, 1998)*, volume 404, pages 47–110. Chapman & Hall/CRC, Boca Raton, FL, 1999.
38. M. Kružík, A. Mielke, and T. Roubíček. Modelling of microstructure and its evolution in shape-memory-alloy single-crystals, in particular in cualni. *Meccanica*, 40:389–418, 2005.
39. M. Kružík and M. Luskin. The computation of martensitic microstructure with piecewise laminates. *J. Sci. Comp.*, 19:293–308, 2003.
40. M. Kružík and T. Roubíček. Mesoscopic model of microstructure evolution in shape memory alloys with applications to NiMaGa. Technical Report no. 2003, IMA Minneapolis, 2004.
41. H. Le Dret and A. Raoult. Variational convergence for nonlinear shell models with directors and related semicontinuity and relaxation results. *Arch. Ration. Mech. Anal.*, 154(2):101–134, 2000.
42. A. Mainik. A rate-independent model for phase transformations in shape-memory alloys. PhD Thesis, IADM, Universität Stuttgart, 2005.
43. A. Mainik and A. Mielke. Existence results for energetic models for rate-independent systems. *Calc. Var. PDEs*, 22:73–99, 2005.
44. J. A. C. Martins, A. Pinto da Costa, and F. M. F. Simões. Some notes on friction and instabilities. In *Friction and instabilities (Udine, 2000)*, volume 457 of *CISM Courses and Lectures*, pages 65–136. Springer, 2002.
45. G. A. Maugin. *The Thermomechanics of Plasticity and Fracture*. Cambridge University Press, Cambridge, 1992.
46. A. Mielke. Estimates on the mixture function for multiphase problems in elasticity. In A.-M. Sändig, W. Schiehlen, and W. Wendland, editors, *Multifield Problems*, pages 96–103, Berlin, 2000. Springer–Verlag.
47. A. Mielke. Deriving new evolution equations for microstructures via relaxation of variational incremental problems. *Comput. Methods Appl. Mech. Engrg.*, 193:5095–5127, 2004.
48. A. Mielke. Evolution in rate-independent systems (ch. 6). In C. Dafermos and E. Feireisl, editors, *Handbook of Differential Equations, Evolutionary Equations, vol. 2*, pages 461–559. Elsevier B.V., 2005.
49. A. Mielke. Temperature-induced phase transformations in shape-memory alloys. In preparation, 2006.
50. A. Mielke and M. Ortiz. A class of minimum principles for characterizing the trajectories of dissipative systems. *In preparation*, 2006.

51. A. Mielke and R. Rossi. Existence and uniqueness results for a class of rate-independent hysteresis problems. *M^3AS Math. Models Methods Appl. Sci.*, 2006. To appear (WIAS Preprint 1078).
52. A. Mielke, R. Rossi, and G. Savaré. On a metric approach to a class of rate-independent problems. *In preparation*, 2006.
53. A. Mielke and T. Roubíček. A rate–independent model for inelastic behavior of shape–memory alloys. *Multiscale Model. Simul.*, 1:571–597, 2003.
54. A. Mielke and T. Roubíček. Numerical approaches to rate-independent processes and applications in inelasticity. *In preparation*, 2005.
55. A. Mielke and T. Roubíček. Rate-independent damage processes in nonlinear elasticity. *M^3AS Math. Models Methods Appl. Sci.*, 16:177–209, 2006.
56. A. Mielke, T. Roubíček, and U. Stefanelli. Γ-limits and relaxations for rate-independent evolution equations. *In preparation*, 2006.
57. A. Mielke and F. Theil. A mathematical model for rate-independent phase transformations with hysteresis. In H.-D. Alber, R. Balean, and R. Farwig, editors, *Proceedings of the Workshop on "Models of Continuum Mechanics in Analysis and Engineering"*, pages 117–129. Shaker-Verlag, 1999.
58. A. Mielke and F. Theil. On rate-independent hysteresis models. *Nonl. Diff. Eqns. Appl. (NoDEA)*, 11:151–189, 2004. (Accepted July 2001).
59. A. Mielke, F. Theil, and V. I. Levitas. A variational formulation of rate-independent phase transformations using an extremum principle. *Arch. Rational Mech. Anal.*, 162:137–177, 2002.
60. A. Mielke and A. Timofte. An energetic material model for time-dependent ferroelectric behavior: existence and uniqueness. *Math. Meth. Appl. Sciences*, 2005. In print (WIAS preprint no. 1014).
61. A. Mielke and A. M. Timofte. Modeling and analytical study for ferroelectric materials. *Mech. Advanced Materials Structures*, 2006. In print.
62. A. Mielke and A. M. Timofte. Two-scale homogenization for evolutionary variational inequalities via the energetic formulation. *In preparation*, 2006.
63. J.-J. Moreau. On unilateral constraints, friction and plasticity. In *New Variational Techniques in Mathematical Physics (Centro Internaz. Mat. Estivo (C.I.M.E.), II Ciclo, Bressanone, 1973)*, pages 171–322. Edizioni Cremonese, Rome, 1974.
64. J.-J. Moreau. Application of convex analysis to the treatment of elastoplastic systems. In P. Germain and B. Nayroles, editors, *Applications of Methods of Functional Analysis to Problems in Mechanics*, pages 56–89. Springer-Verlag, 1976. Lecture Notes in Mathematics, 503.
65. S. Müller. Variational models for microstructure and phase transitions. In *Calculus of Variations and Geometric Evolution Problems (Cetraro, 1996)*, pages 85–210. Springer, Berlin, 1999.
66. M. Ortiz and E. Repetto. Nonconvex energy minimization and dislocation structures in ductile single crystals. *J. Mech. Phys. Solids*, 47(2):397–462, 1999.
67. L. Prigozhin. On the bean critical-state model in superconductivity. *European J. of Appl. Math.*, 7:237–248, 1996.
68. J. Rice. Inelastic constitutive relations for solids: an internal-variable theory and its applications to metal plasticity. *J. Mech. Phys. Solids*, 19:203–240, 1971.
69. H. Romanowski and J. Schröder. Coordinate invariant modelling of the ferroelectric hysteresis within a thermodynamically consistent framework. A mesoscopic approach. In Y. Wang and K. Hutter, editors, *Trends in Applications of Mathematics to Mechanics*, pages 419–428. Shaker Verlag, 2005.

70. T. Roubíček. Models of microstructure evolution in shape memory alloys. In P. Ponte Castaneda, J. Telega, and B. Gambin, editors, *Nonlinear Homogenization and its Applications to Composites, Polycrystals and Smart Materials*, pages 269–304. Kluwer, 2004. NATO Sci. Series II/170.
71. T. Roubíček and M. Kružík. Mircrostructure evolution model in micromagnetics. *Zeits. angew. Math. Physik*, 55:159–182, 2004.
72. T. Roubíček and M. Kružík. Mesoscopic model for ferromagnets with isotropic hardening. *Zeits. angew. Math. Physik*, 56:107–135, 2005.
73. P. Rybka and M. Luskin. Existence of energy minimizers for magnetostrictive materials. *SIAM J. Math. Anal.*, 36:2004–2019, 2005.
74. F. Schmid. A local existence result for a finite-dimensional model in dry friction. *In preparation*, 2006.
75. F. Schmid and A. Mielke. Vortex pinning in super-conductivity as a rate-independent process. *Europ. J. Appl. Math.*, 2005. In print.
76. A. Souza, E. Mamiya, and N. Zouain. Three-dimensional model for solids undergoing stress-induced phase transformations. *Europ. J. Mech., A/Solids*, 17:789–806, 1998.
77. P.-M. Suquet. Sur les équations de la plasticité: existence et régularité des solutions. *J. Mécanique*, 20(1):3–39, 1981.
78. F. Theil. Relaxation of rate-independent evolution problems. *Proc. Roy. Soc. Edinburgh Sect. A*, 132:463–481, 2002.
79. A. Visintin. *Differential Models of Hysteresis*. Springer-Verlag, Berlin, 1994.
80. H. Ziegler and C. Wehrli. The derivation of constitutive relations from the free energy and the dissipation function. In *Advances in Applied Mechanics, Vol. 25*, pages 183–237. Academic Press, Orlando, FL, 1987.

Heterogeneous Domain Decomposition for Numerical Aeroacoustics*

Jens Utzmann, Thomas Schwartzkopff, Michael Dumbser, and Claus-Dieter Munz

Institute of Aero and Gas Dynamics, University of Stuttgart, Pfaffenwaldring 21, 70569 Stuttgart, Germany
utzmann@iag.uni-stuttgart.de
schwartzkopff@iag.uni-stuttgart.de
dumbser@iag.uni-stuttgart.de
munz@iag.uni-stuttgart.de

Summary. A strategy is presented to accelerate the direct simulation of aeroacoustic problems in terms of computer time. The key idea is to introduce a heterogeneous domain decomposition. The whole computational domain is sub-divided into smaller domains. In each of these sub-domains the equations, the discretization, the mesh and the time step may be different and are adapted to the local behavior of the solution. To reduce the total number of elements we propose the use of high order methods. Here the class of ADER-Finite Volume schemes on structured meshes and ADER-Discontinuous Galerkin methods on unstructured meshes seem a good choice to us. The coupling procedure is validated and numerical results for the *Single Airfoil Gust Response Problem* from the 4th CAA Workshop on Benchmark Problems are presented, together with the acoustic scattering problem at a cylinder and at multiple objects. The coupling approach proves to be especially efficient for the propagation of sound in large domains.

Keywords: Domain decomposition, direct simulation, aeroacoustics, high-order methods, grid coupling

1 Introduction

The field of Computational Aeroacoustics (CAA) is very demanding for two main reasons: The requirements regarding the accuracy of the methods are very high and in general, quite large domains have to be computed. It was Sir Lighthill who developed the first acoustic analogy [8] and started the field of noise predictions. In the following years, the acoustic analogies have been

* Research Project C8 "Heterogeneous Domain Decomposition for Numerical Aeroacoustics"

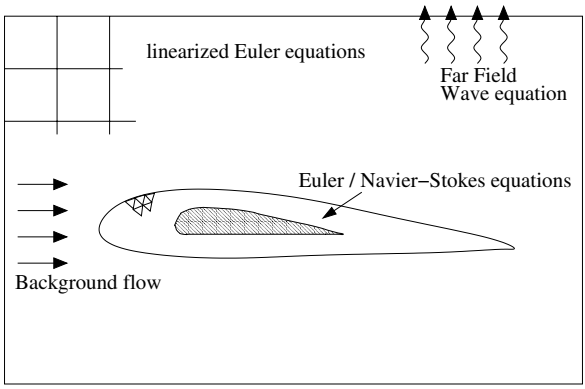


Fig. 1. Sketch of non-overlapping domain decomposition

continuously improved. However, these methods suffer from basic problems: First of all, there is no feedback from acoustics to the fluid flow. Furthermore, all of these analogies model the acoustic sources in some way, implying an uncertainty in the modeling.

The method that is proposed here does not suffer from these problems, as it is basically a direct simulation. However, an accurate and feasible direct simulation that considers both the generation of sound and its propagation into the far-field is hard to realize with one numerical method in a single computational domain. The idea is to use a non-overlapping domain decomposition method where the equations, methods, grids and time steps are adapted to meet the local requirements. Such a decomposition has already been proposed for jet noise by Freund et al. [3]. Also composite overset structured grids are related to the presented domain decomposition. Chimera methods (introduced by Steger et al. [21]) couple structured body fitted grid components with Cartesian background grids and are used in a wide range of applications. An overview is given by Meakin [9]. However, these approaches did not simplify the domains to the degree which is described here.

The previous work in subproject C8 "Heterogeneous Domain Decomposition for Numerical Aeroacoustics" concentrated on the examination of basic aspects: Which decomposition approaches are possible and which numerical methods are suitable in the different domains? How can reflections at the domain interface be avoided [11]? This was the starting point for the development of the coupling method that is decribed in the following: First we start with a short description of the basic problems of Computational Aeroacoustics (CAA). Then we briefly introduce the high order methods for structured and unstructured grids. In Sect. 4, the domain decomposition method is described and analyzed, e.g. regarding the order of convergence. In the last section several numerical examples are given.

2 Basic Problems

The basic problems in Computational Aeroacoustics are briefly described. We start at the pure physical problem with no numerics involved. In principle, the complete problem can be described by the full nonlinear compressible Navier-Stokes equations . However, it is not necessary to include all physical effects in the whole domain. There are three different areas in the domain of interest (see Fig. 1). The area around bodies or interfaces include all nonlinear as well as viscous effects. A short distance away, the viscous parts may already be neglected and the simpler system of nonlinear Euler equations can be used. Further away, nonlinear effects become less important and the linearized Euler equations become a good choice. This is a region where only small perturbations, e.g. acoustics, propagate.

Next we want to look at the multiscale problem in aeroacoustics. The interesting acoustic wavelength is often much larger than the length scale of the fluid flow phenomena which generate the wave. These phenomena are eddies which have a much smaller structure. On the other hand, eddies are advected with the fluid flow, whereas acoustic waves run with the speed of sound.

Last but not least we want to look at the numerics. In the vicinity of complex obstacles one has two major choices in discretizing the domain. First, one can use body-fitted block-structured meshes and second one can use completely unstructured meshes. We will use fully unstructured meshes because complex geometries can be handled much more easily by these grids. On the other hand, there is no need to use these meshes away from the bodies where everything is regular. Here it is much more convenient to use structured, Cartesian uniformly spaced meshes for the following reasons: They need much less memory, can be implemented much more effectively and the step to high order is much easier.

The idea of the decomposition method is now to simplify everything as much as possible. This means that the equations which are used are adapted to the relevant local phenomena. The grids and, due to the CFL condition, the time steps are chosen as large as possible and also the type of grid is flexible. The complete setup of this method is described in Sect. 4.

3 Numerical Methods

A key ingredient in our method is to use high order schemes everywhere. The recent idea of ADER schemes by Toro et al. [27, 25, 26] has been adopted and optimized for linear and nonlinear equations on structured grids in the framework of Finite Volume (FV) schemes by Schwartzkopff et al. [18, 17, 14, 12]. The same idea is used for the Discontinuous Galerkin (DG) methods by Dumbser et al. [1] which is used on unstructured grids. A direct comparison of these method is given in Table 1. The main advantage of the ADER-FV schemes on structured grids is that the first step in a FV scheme, the reconstruction,

Table 1. Comparison of Finite Volume and Discontinuous Galerkin methods

FV-methods	DG-methods
+ easy to implement	- difficult to implement
+ high order on regular, structured grids	- slow
+ fast algorithms for structured grids	+ efficient parallelization
– high order on unstructured grids is difficult	+ high order on unstructured grids is easy
– unstructured grid must be regular	+ high convergence rates on irregular grids

becomes rather simple, as all neighborhoods are known directly. Hence, the implementation of higher order becomes easier. The advantage of DG methods is that they are very local as they only need direct neighbors, thus high order on unstructured meshes can be achieved easily.

4 Heterogeneous Domain Decomposition

The aim of the present work was to develop a method which makes it possible to calculate a direct simulation of an aeroacoustic problem within reasonable time. The domain of the calculation shall cover both the area of sound generation and the area of sound propagation into the far-field . Hence, the main goal was to reduce the effort by a domain decomposition method.

First of all, unstructured grids are used only in the direct vicinity of complex bodies. As fast as possible, these grids are aligned to a straight line and are switched to structured, Cartesian meshes.

Second, we allow the cell size to jump at a domain interface. Thus, the mesh can be adapted to the local flow phenomena which means that the eddies in the flow around a body need not to be resolved in the near far field, where one is only interested in the acoustics and therefore the grid size may be much larger. The numerical methods used are always explicit time integration methods with a time step that is chosen as large as possible in each domain. As a consequence, not only the spatial grid size has a jump at an interface, but also the time step. By this, the far-field domains are updated only once in a while because large regions are covered with only a small number of cells.

Finally, we allow the underlying equations to change. The nonlinear equations are used in the core regions near sound generating obstacles or phenomena and for the surrounding area we solve only the linearized equations. For the linearized equations, the underlying method can be optimized very well, see Schwartzkopff [14].

By using these techniques, the decrease of memory and CPU time becomes immense and a direct simulation in this sense becomes more and more feasi-

ble. In the following paragraphs we explain some basic parts and properties of these methods.

4.1 Coupling Strategies

Coupling of Different Equations

Physical Considerations

The coupling of different equations is indeed a very demanding issue. The coupling of nonlinear and linear Euler equations has been discussed by Schwartzkopff et al. [15, 11]. If a conservative coupling regarding the flux at the interface is used, the reflection errors can be estimated easily. The propagation speed of the wave changes from $u + \sqrt{\frac{\gamma p}{\rho}}$ to the constant speed $u_0 + \sqrt{\frac{\gamma p_0}{\rho_0}}$. Therefore, the interface Γ can be interpreted as an artificially introduced material interface of two materials with different wave speeds. For this case, the physics of waves [24] directly gives estimates for reflections r with the refraction indices n and the refraction for a wavefront with the angle of incidence α^E:

$$n^{LE} = \frac{1}{u_0 + c_0}\,, \qquad n^E = \frac{1}{u + c}\,, \tag{1}$$

$$r = \left(\frac{n^{LE} - n^E}{n^{LE} + n^E}\right)^2 \approx \left(\frac{u' + \mathcal{O}(c')}{2\mathcal{O}(u_0) + 2\mathcal{O}(c_0)}\right)^2\,, \tag{2}$$

$$\alpha^{LE} = \arcsin\left(\frac{n^E}{n^{LE}}\sin(\alpha^E)\right)\,. \tag{3}$$

Even total reflection is possible, if the product $n^E/n^{LE} \times \sin(\alpha^E)$ gets larger than 1, which corresponds to

$$\sin(\alpha^E) = \frac{n^{LE}}{n^E} \approx \frac{u_0 + \mathcal{O}(c_0)}{u_0 + c_0} + \frac{u' + \mathcal{O}(c')}{u_0 + c_0}\,. \tag{4}$$

For example, if the perturbations $\mathbf{U}'$ of the state vector are 10% of $\mathbf{U}_0$, then the angle for total reflection will be $\alpha^E \approx 64°$!
One can conclude, that if a numerical coupling scheme is applied which simulates the artificially introduced interface like a material interface using the continuity of the flux, then there will certainly be reflections at the interface which are of the magnitude $\mathcal{O}(\mathbf{U}'^2)$. The aim is now to find a transmission condition which avoids this behavior at the artificial interface.

Coupling of Different Grids

Two domains Ω_1 and Ω_2 are coupled at their common boundary $\partial\Omega = \Gamma$ only over data in the so-called ghostcells. These ghostcells are then used by

the numerical method to update the inner cells in each domain. The number of ghostcells that are needed for the numerical method depends strongly on the method itself. The DG methods are very local and so they need only one row of ghostcells where the high order ADER-FV schemes use reconstruction operators and thus need a larger number of ghostcells.

Fully Conservative Coupling

Two cases may occur in the coupling procedure:

1. **Restrictions:** The projection of the ghostcell on the partner domain covers more than 1 cell (see left picture in Fig. 2). In this case an integral average of all cells which lie partially in the projection is used. The weight of this averaging is the ratio of the respective area of cell V_{S_i} which is in the projection and the total area of the ghostcell V_G:

$$\overline{\mathbf{U}}_G^{\Omega_g} = \frac{1}{V_G} \sum_i V_{S_i} \overline{\mathbf{U}}_i^{\Omega_f} \ . \tag{5}$$

   As it is obvious from Fig. 3, very complex and small cut-out polygons can occur, which can become intricate.

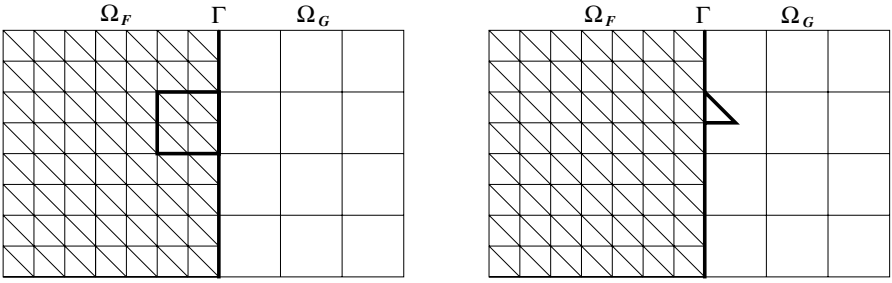


Fig. 2. Coupling different grids: Averaging and interpolation

2. **Interpolations:** The projection of the ghostcell lies completely within one cell on the partner domain (see right picture in Fig. 2). In this case a conservative interpolation from the coarse grid to the ghostcell is used. A 2-dimensional Lagrange interpolation (Fig. 4) $\mathcal{L}^{\Omega_g}(\mathbf{X})$ delivers a interpolation of the state vector $\mathbf{U}$ and its derivatives from given integral mean-values $\overline{\mathbf{U}}^{\Omega_g}$ (the values that are stored in the cells). The integral mean-values are obtained afterwards by a summation over the Gauß points with Gauß weights ω_i:

$$\overline{\mathbf{U}}_G^{\Omega_f} = \sum_{i=1}^{N_{GP}} \omega_i \mathbf{U}_{GP_i} \ , \tag{6}$$

$$\mathbf{U}_{GP_i} = \mathcal{L}^{\Omega_g}(\mathbf{X}_i) \ . \tag{7}$$

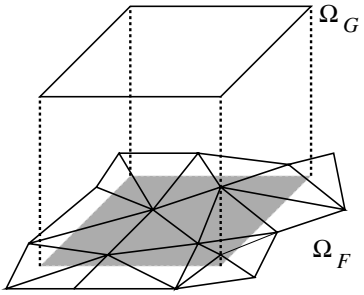


Fig. 3. Averaging by weighted volumes

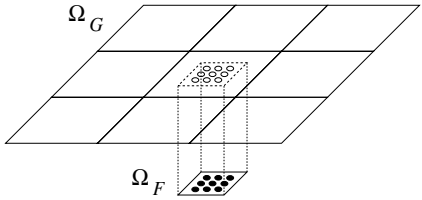


Fig. 4. Interpolation at the Gauß integration points

Gauß Point Coupling

The previously described coupling procedure was validated (see Sect. 4.2 and Schwartzkopff [13]) and proved to be working very well with both structured and unstructured grids in 2 dimensions. However, determining the cut-out polygons for the averaging case leads to complex algorithms. Furthermore, the treatment of "multidomain-ghostcells" (Fig. 7), ghostcells which overlap different neighboring domains, is very cumbersome. Those problems in 2D can become painfully complex when the coupling procedure will be extended to 3D. In the following, we propose a basic and unified coupling mechanism. The main idea is to treat both cases by interpolating the values from the neighbor-grid onto the Gauß integration points of the ghostcell and then to integrate in order to obtain the integral mean-values (Fig. 5). This allows a very generalized coupling procedure, as there is no difference between restrictions and interpolations anymore. The integration to get the integral mean-value is independent of the way the interpolation onto the Gauß point was performed. The order of interpolation is set by the spatial order of the neighbor-domain. Thus, the coupling domains can be of arbitrary size and discretization. For example, Fig. 5 shows a Cartesian ghostcell that belongs to a domain Ω_G with spatial accuracy $\mathcal{O}4$. The neighbor-domain is of spatial accuracy $\mathcal{O}5$ and uses a corresponding 5×5 stencil for the 2D Lagrange interpolation on the marked Gauß point. On the other hand, there even may be no stencil necessary: Figure 6 shows a ghostcell coupling with an unstructured Discontinuous Galerkin domain. As there is no reconstruction step for the very compact DG methods, all information can be obtained from the single neighbor-element in which the Gauß point lies. The unified coupling procedure also simplifies the treatment of multidomain cells (Figs. 7 and 8): The large cell in Fig. 8 from domain Ω_G depicts a cell that is similar to the grey colored ghostcell in Fig. 7 and overlaps two different domains Ω_{F1} and Ω_{F2}. The dashed line indicates the domain interface between the two neighbor domains Ω_{F1} and Ω_{F2}. Each Gauß point can be assigned a unique element of a neighboring grid that is the basis for the necessary interpolation stencil.

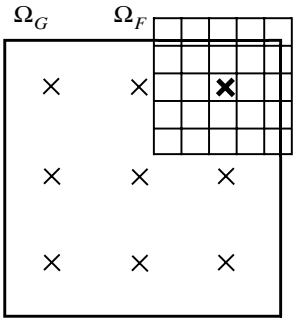


Fig. 5. Interpolation from a fine, structured (ADER-)FV grid Ω_F onto a Gauß point of the target ghostcell from domain Ω_G

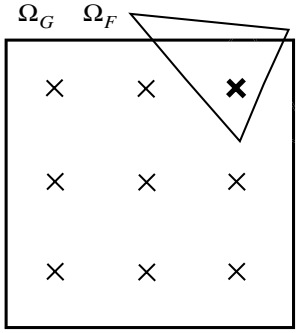


Fig. 6. Interpolation from a fine, unstructured (ADER-)DG grid Ω_F onto a Gauß point of the target ghostcell from domain Ω_G

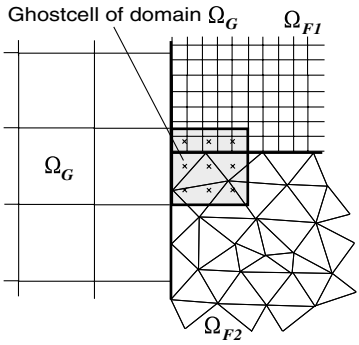


Fig. 7. A multidomain cell: A ghostcell from domain Ω_G couples with both domains Ω_{F1} and Ω_{F2}

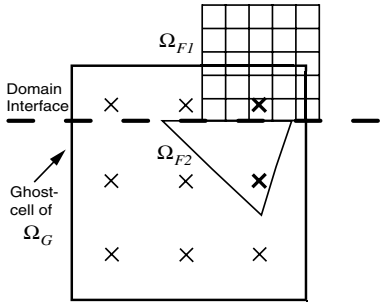


Fig. 8. Close-up of the multidomain cell: The Gauß points obtain their values from different domains

A distinction between restriction and prolongation has only to be made for the so-called Cauchy-Kovalevskaja (CK) procedure (see Sect. 4.1): If the ghostcell-domain has a smaller time step than the neighbor-domain, not only the state vector, but also its spatial derivatives have to be determined at the Gauß point location.

The coupling procedure may however not be fully conservative anymore regarding the integral mean-values in the averaging case. Not every fine cell that lies within the coarse ghostcell's projection is considered anymore for the integral mean value. An investigation of the new coupling mode's properties regarding conservation and reflections is subject to current work. The following results refer to the fully conservative coupling method.

High-Frequency Perturbations

The conservative coupling procedure is examined for a problem that arises in the case when two coupling grids have different mesh sizes:

A wave with wavelength λ is represented by $PPW = \lambda/\Delta x$ points. For domain Ω_1, Δx_1 is chosen in a way that $PPW_1 > 3$ and a high frequency wave can be resolved on the fine grid. For Ω_2, Δx_2 is chosen in a way that $PPW_2 < 3$. Hence, the wave cannot be resolved anymore on the coarse grid. Now it will be observed, what happens when the wave crosses the interface Γ from Ω_1 to Ω_2.

As initial condition, a perturbation in the pressure is prescribed according to the *Aliasing* problem by Tam [22] from the 4th CAA Workshop on Benchmark Problems:

$$\begin{bmatrix} \rho' \\ u' \\ v' \\ p' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ (2+\cos(\alpha x))\, e^{-(ln2)\left(\frac{x}{10}\right)^2} \end{bmatrix}, \tag{8}$$

with

$$\begin{bmatrix} \rho_0 \\ u_0 \\ v_0 \\ p_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{9}$$

$$\alpha = 1.7\,. \tag{10}$$

The constant values of the mean background flow are $\mathbf{U}_0$. The pressure perturbation splits into two parts running to the left and to the right. The area $\Omega = [-300 : 200]$ was divided into two domains $\Omega_1 = [-300 : 0]$ and

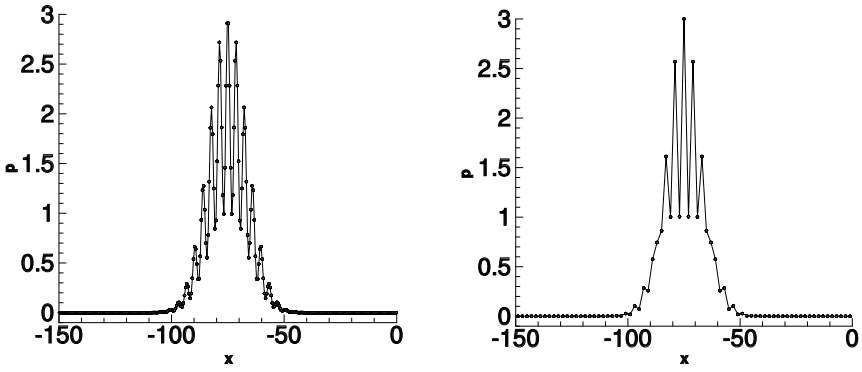


Fig. 9. Initial condition of the high-frequency perturbations; fine grid on the left side, coarse grid on the right side

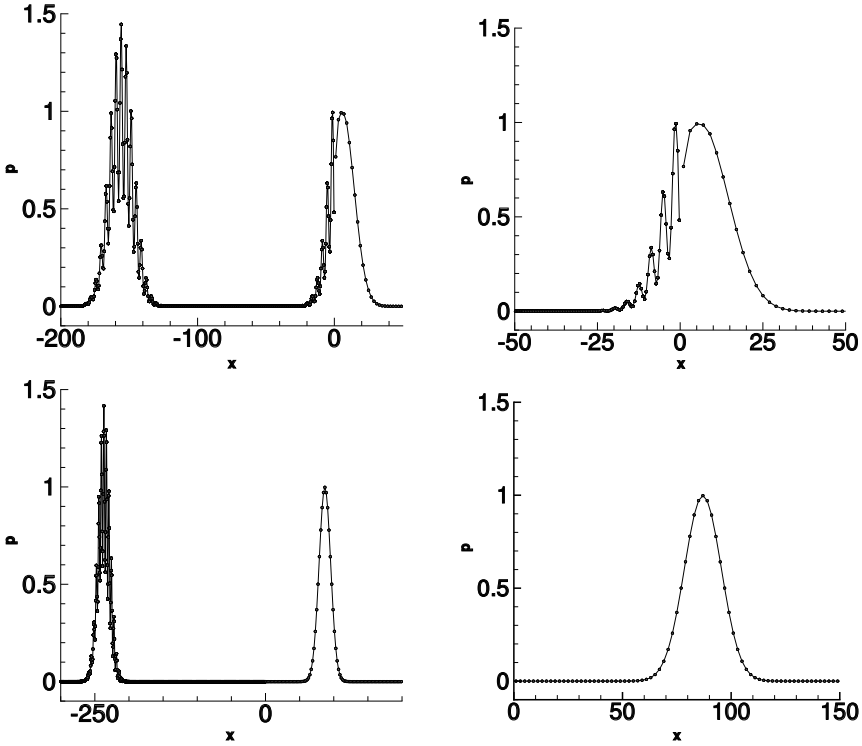


Fig. 10. Behavior regarding high-frequency perturbations; top: state after $n = 150$ (crossing the interface); bottom: after $n = 300$ (past the interface); right sides: zoom into the regions

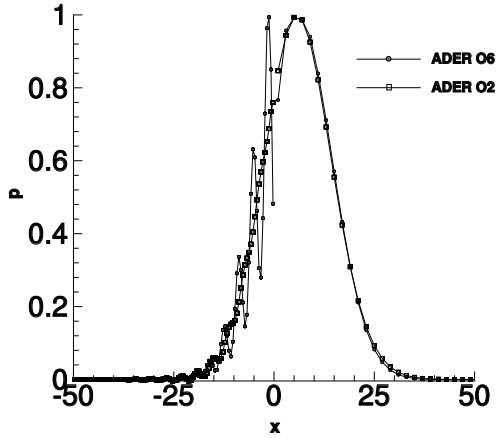


Fig. 11. Influence of the order of accuracy

$\Omega_2 = [0 : 200]$. The initial condition's pressure peak is located in Ω_1 at $x = -75$. The grid sizes are $\Delta x_1 = 0.5$ and $\Delta x_2 = 2.0$ and the ratio between the time steps is also $\Delta t_2/\Delta t_1 = 4$. The 6th order ADER-FV method was used on Cartesian grids. Figure 9 shows the initial condition both on the fine and the coarse grid. It is obvious, that the method on the coarse mesh cannot capture the high frequencies. Figure 10 depicts the right pressure wave exactly at the point of time when it crosses the interface $\Gamma = \Omega_1 \cap \Omega_2$ (top) and at a time, when the pulse is past the interface (bottom). The wave cannot be captured anymore on Ω_2, but it leaves domain Ω_1 without interacting with the interface or triggering reflections.

In Fig. 11, the initial condition for different orders of accuracy shows, that only high-order methods are capable of resolving the high frequencies from the beginning. The 2nd order method dissipates almost all parts of the high frequencies, while the long wave Gaussian pulse is reproduced very well. A considerably finer mesh would have to be used here in order to achieve a similar quality of the initial solution which was obtained with the 6th order method. An analysis of both amplitude and phase error for different PPW for the ADER-FV scheme is given in Schwartzkopff [14].

Coupling of Different Schemes

The coupling of different discretization methods causes no problems. Each method (DG and FV) uses the values in the ghostcells. We emphasize that there is a choice for the flux at the boundary Γ in domain Ω_1 to be different from the flux at the same boundary in Ω_2 as we do not define the flux but the data.

Coupling of Different Time Steps

Subcycles: Time Stepping and Data Exchange

The idea of multi-size meshes with different time steps has been examined by Tam et al. [23] for DRP schemes on particularly designed grids. This basic idea has been extended in our approach for FV and DG schemes on arbitrary meshes.

The data between the domains is exchanged at the common time level. The local CFL number in each domain is adjusted such that neighboring domains have the same time level, if the domain with the larger time step is updated once. Meanwhile, the other domain has been updated n times with $n \geq 1$. The idea of subcycling is illustrated in Fig. 12: Four domains Ω_0, Ω_1, Ω_2 and Ω_3 are coupled in a calculation. All four domains have different time steps ($\Delta t_{\Omega_0} < \Delta t_{\Omega_1} < \Delta t_{\Omega_2} < \Delta t_{\Omega_3}$) which have been adjusted in a way that each time step is a multiple of the previous one. For example, during only one time step of Ω_2, domain Ω_1 performs three steps. Depending on how many of the domains have reached the same absolute time level, subcycle 1, 2 or 3 is

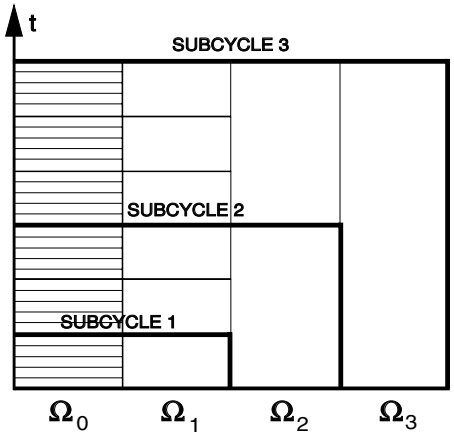


Fig. 12. Subcycles of a calculation with four differently sized domains

complete. For example, after a subcycle of type 1, domain Ω_0 has made five time steps and shares now a common time level with Ω_1 and the domains can exchange data at their boundary $\Gamma_{\Omega_0,\Omega_1}$. When subcycle 3 is complete, all of the domains – Ω_0, Ω_1, Ω_2 and Ω_3 – have arrived at the same time level.

The updating procedure itself is sketched in Fig. 13 for the exchange between Ω_1 and Ω_2. The first step is the data exchange step, the second step and third step are the data update steps in domain Ω_1, the fourth step is the data update in both domains Ω_1 and Ω_2. Then, the sequence would start again with step 1. The adaption of the CFL number for each domain is done automatically at each subcycle.

The question is what to do with the ghostcells that are only updated every $n-th$ time step. In general, these are cells in a domain with a finer grid than the domain from which the values are taken. Thus these fine-grid ghostcells are interpolated from the coarse grid by a conservative interpolation. The key idea is here to use the Cauchy-Kovalevskaja (CK) procedure to update the ghostcells. The starting point is a Taylor series in time

$$\mathbf{U}(\mathbf{X}_0, t^n + \tau) = \mathbf{U}(\mathbf{X}_0, t^n) + \tau \frac{\partial \mathbf{U}(\mathbf{X}_0, t^n)}{\partial t} + \ldots + \frac{\tau^k}{k!} \frac{\partial \mathbf{U}^k(\mathbf{X}_0, t^n)}{\partial^k t} \,. \quad (11)$$

The truncation order matches the chosen order of accuracy, e.g. third order for a $\mathcal{O}3$ method in space and time. Time derivatives are now replaced by spatial derivatives with the Cauchy-Kovalevskaja procedure [7]. This procedure is used in Lax-Wendroff type time integration methods [7, 6] and is also a main ingredient of the ADER scheme. The major advantage of this method is that once the derivatives at the point $\mathbf{X}_0$ are known, the state at an arbitrary time $t^n + \tau$ can be calculated easily. Moreover, one does not have any dependency on a spatial scale from the fine grid domain but only on the spatial scale from the interpolation on the coarse grid. As a consequence, these ghostcells

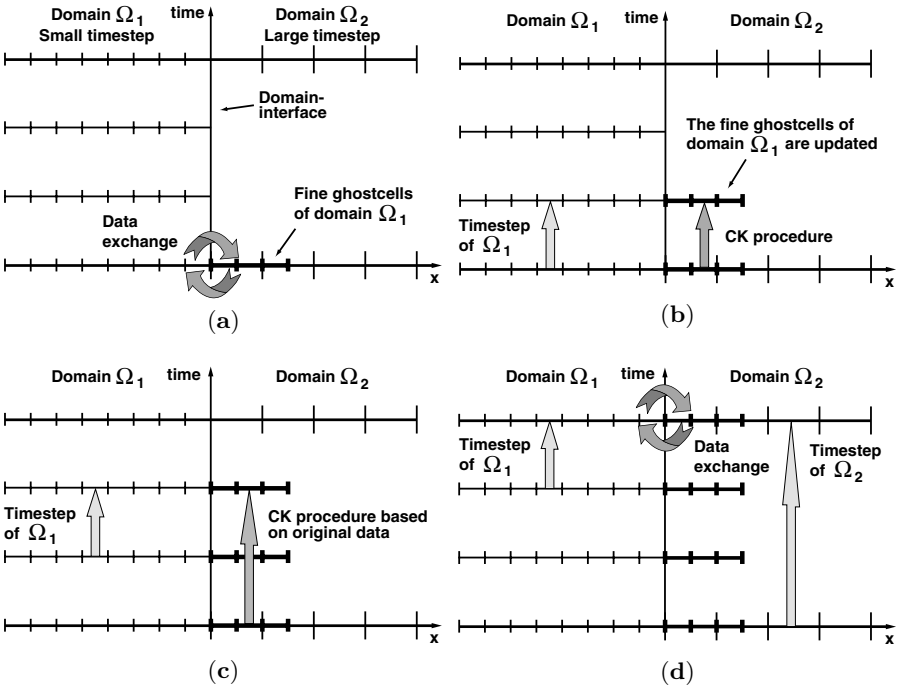


Fig. 13. Coupling different time steps: (**a**) data exchange at the common time level: the states are interpolated onto the ghostcells of Ω_1 and Ω_2; for Ω_1, also the spatial derivatives are interpolated; (**b**) time stepping of domain Ω_1; at the same time, its ghostcells are lifted to the new time level by the CK procedure; (**c**) Ω_1 makes another time step; again, its ghostcells are pushed to the new time level, based on the original data; (**d**) Ω_2 performs one single time step and meets Ω_1 at a common time level again; then, a new data exchange takes place

inherit the stability condition from the coarse grid and every evaluation of the Taylor series between the two coarse grid time steps is in the stable range, although these ghostcells may have small sizes and the CFL number built with this small size is much larger than 1. This is step 1b in Fig. 13. If Ω_1 needs more than 1 intermediate time step than the update of the ghostcells is always based on the last common time level.

The Cauchy-Kovalevskaja Procedure for Linear Systems

In general, the CK procedure has to be applied to every Gauss point of the ghostcell that is supposed to be updated. However, for linear systems of equations, there exists a way to conveniently save memory and CPU time: In case every Gauss point of the ghostcell has the same interpolation stencil on the neighbor-grid (which is often the case for prolongation elements), the cell can be integrated at an early stage at the time before the actual calculation when

the coupling information is created. This is possible because the CK procedure for linear systems with constant coefficients does not depend on the values of the single Gauss points. Hence, there remains one Gauss point per ghostcell that has to be updated. For the procedure itself, an efficient implementation can be found.

The CK procedure takes advantage of the original PDE in order to create time derivatives out of space derivatives: Considering a set of 2-dimensional linear equations

$$\mathbf{U}'_t + \mathsf{A}\,\mathbf{U}'_x + \mathsf{B}\,\mathbf{U}'_y = 0\ ,$$

it is possible to obtain the time derivatives just by reordering and deriving:

$$\begin{aligned}
\mathbf{U}'_t &= -\mathsf{A}\,\mathbf{U}'_x - \mathsf{B}\,\mathbf{U}'_y\ , && (12)\\
\mathbf{U}'_{tt} &= -\mathsf{A}\,\mathbf{U}'_{xt} - \mathsf{B}\,\mathbf{U}'_{yt}\ ,\\
\mathbf{U}'_{tx} &= -\mathsf{A}\,\mathbf{U}'_{xx} - \mathsf{B}\,\mathbf{U}'_{yx}\ ,\\
\mathbf{U}'_{ty} &= -\mathsf{A}\,\mathbf{U}'_{xy} - \mathsf{B}\,\mathbf{U}'_{yy}\ ,\\
\mathbf{U}'_{tt} &= \mathsf{A}\,\mathsf{A}\,\mathbf{U}'_{xx} + \mathsf{A}\,\mathsf{B}\,\mathbf{U}'_{yx} + \mathsf{B}\,\mathsf{A}\,\mathbf{U}'_{xy} + \mathsf{B}\,\mathsf{B}\,\mathbf{U}'_{yy}\ . && (13)
\end{aligned}$$

In short, one can write:

$$\frac{\partial^k \mathbf{U}(\mathbf{X},t)}{\partial t^k} = (-1)^k \left(\mathsf{A}\frac{1}{\partial x} + \mathsf{B}\frac{1}{\partial y}\right)^k \partial^k \mathbf{U}(\mathbf{X},t)\ .$$

The following relation can be used :

$$\begin{aligned}
\frac{\partial^{n+1,i,j}\mathbf{U}}{\partial^{n+1}t\,\partial^i x\,\partial^j y} &= -\mathsf{A}\cdot\frac{\partial^{n,i+1,j}\mathbf{U}}{\partial^n t\,\partial^{i+1}x\,\partial^j y} - \mathsf{B}\cdot\frac{\partial^{n,i,j+1}\mathbf{U}}{\partial^n t\,\partial^i x\,\partial^{j+1} y}\ , && (14)\\
&n \in [0,\ldots,\mathcal{O}-2]\ ,\\
&i \in [0,\ldots,\mathcal{O}-2-n]\ ,\\
&j \in [0,\ldots,\mathcal{O}-2-n-i]\ .
\end{aligned}$$

In order to calculate the time derivative $\partial^{n+1}/\partial t^{n+1}$ for an arbitrary order, the time derivatives of level n, $\partial^n/\partial t^n$ with one higher space derivative are needed. The derivatives for $n = 0$ are known from the interpolation step itself. For example, the calculation of a 4th order time derivative would look like this:

```
n i=0   i=1  i=2  i=3  i=4
0 0     x    xx   xxx  xxxx
   ↓↙  ↓↙  ↓↙  ↓↙  ↙
1 0t    xt   xxt  xxxt
   ↓↙  ↓↙  ↓↙  ↙
2 0tt   xtt  xxtt
   ↓↙  ↓↙  ↙
3 0ttt  xttt
   ↓↙  ↙
4 0tttt
```

The Cauchy-Kovalevskaja Procedure for Non-Linear Systems

The nonlinear hyperbolic system of the two-dimensional Euler equations reads as

$$\begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{pmatrix}_t + \begin{pmatrix} (\rho u) \\ (\rho u^2) + p \\ (\rho uv) \\ u(\rho E + p) \end{pmatrix}_x + \begin{pmatrix} (\rho v) \\ (\rho uv) \\ (\rho v^2) + p \\ v(\rho E + p) \end{pmatrix}_y = 0 \,, \tag{15}$$

with the equation of state

$$p = (\gamma - 1) \left(\rho E - \frac{1}{2} \left(\rho u^2 + \rho v^2 \right) \right) \,. \tag{16}$$

In addition to the four conservative variables density (ρ), x-momentum (ρu), y-momentum (ρv) and total energy per mass unit (ρE), we use the auxiliary variables u, v, ρu^2, ρv^2, ρuv and the pressure p during the computation of the CK procedure.

As it is even more important in the nonlinear case than in the case of linear hyperbolic systems, an efficient algorithm to carry out the CK procedure has to be found. Although this seems to become cumbersome, there exists an efficient algorithm for the two-dimensional Euler equations, proposed by Dyson [2]. It may be generalized to three space dimensions in a straight forward manner. The method is based on the generalized Leibniz rule, which gives the space-time derivatives of arbitrary order for a product of multivariate scalar functions f and g:

$$\mathcal{L}^{(a,b,c)}\left(f,g\right) := \frac{\partial^{a+b+c}\left(f\left(x,y,t\right)g\left(x,y,t\right)\right)}{\partial x^a \partial y^b \partial t^c} =$$

$$\sum_{i=0}^{a}\sum_{j=0}^{b}\sum_{k=0}^{c}\left[\binom{a}{i}\binom{b}{j}\binom{c}{k}\frac{\partial^{(a-i)+(b-j)+(c-k)}f}{\partial x^{(a-i)}\partial y^{(b-j)}\partial t^{(c-k)}} \times \frac{\partial^{i+j+k}g}{\partial x^i \partial y^j \partial t^k}\right] \,. \tag{17}$$

This rule alone is not yet sufficient to treat the nonlinear Euler equations, since there appears for example a term like $(\rho u)\, u$ in the momentum flux, but u and its derivatives are not known a priori, but only those of the conservative variables ρ and ρu. The trick now consists in reformulating the Leibniz rule (17) such as to calculate all the space-time derivatives of the auxiliary variable u from the known space-time derivatives of ρ and ρu and all previously computed space-time derivatives of lower order of the auxiliary variable u. This is achieved by a modified Leibniz rule of the form

$$\mathcal{L}_{**}^{(a,b,c)}\left(fg,f,g\right) := \frac{\partial^{a+b+c}g\left(x,y,t\right)}{\partial x^a \partial y^b \partial t^c} =$$

$$\frac{1}{f} \times \left[\frac{\partial^{a+b+c}\left(f\left(x,y,t\right)g\left(x,y,t\right)\right)}{\partial x^a \partial y^b \partial t^c} - \mathcal{L}_{*}^{(a,b,c)}\left(f,g\right)\right] \,, \tag{18}$$

with $f \neq 0$. The operator $\mathcal{L}_*^{(a,b,c)}(f,g)$ only contains lower order derivatives of g and is defined as the original Leibniz rule (17) except of the last term in the sum:

$$\mathcal{L}_*^{(a,b,c)}(f,g) = \underbrace{\sum_{i=0}^{a}\sum_{j=0}^{b}\sum_{k=0}^{c}}_{i+j+k\neq a+b+c}\left[\binom{a}{i}\binom{b}{j}\binom{c}{k}\frac{\partial^{(a-i)+(b-j)+(c-k)}f}{\partial x^{(a-i)}\partial y^{(b-j)}\partial t^{(c-k)}}\times\frac{\partial^{i+j+k}g}{\partial x^i\partial y^j\partial t^k}\right],$$

$$\text{with}\quad \mathcal{L}_*^{(0,0,0)}(f,g) := 0\,. \tag{19}$$

4.2 Validation of the Coupling Procedure

The conservative coupling approach was validated in an extensive series of numerical experiments. The behavior towards high-frequency perturbations and changing equations has already been treated in Sect. 4.1. Two additional selected examples in which the linearized Euler Equations (LEE) are solved shall be presented here. In both examples, a 2-dimensional Gaussian pulse in density is transported through the calculation area with constant advection speed. In order to eliminate the influence of the numerical errors of the methods themselves, the 9th order ADER-FV scheme is used. Convergence studies were performed in order to show the high-order accuracy of the whole system of coupled domains. The boundary conditions were also treated in 9th order. Therefore, the order of convergence should be of 9th order for the entire coupled calculation area in the ideal case.

Partitioning

First of all, the basic coupling mechanisms are examined. For this purpose, the calculation area is divided in 24 subdomains (Fig. 14), without changing anything regarding the grids, equations or methods. Some of the depicted domains are too small to be plotted correctly by the visualization program (parallel lines in the upper picture of Fig. 14). The partitioning even includes domains which consist only of one single cell! The solutions of the calculations match exactly. Besides the start-up time for creating the connectivity information between the single subdomains, the overall CPU time for the partitioned calculation (performed on a single processor computer) is higher than the CPU time for the reference calculation with one domain. The time that is necessary for the data exchange itself is about 5% of the overall time. The overhead is caused by the subroutine for the setting of the coupling ghostcells which has to be called more often as the number of additional ghostcells is strongly dependent on the number of partitions. In this extremely partitioned example, the overhead compared to the single domain solution is about 50%.

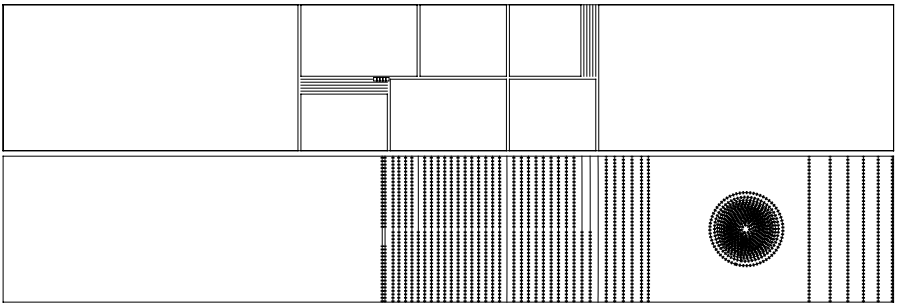


Fig. 14. Comparison of the solution on one grid with the solution on a calculation area that has been divided in 24 subdomains (the density is plotted); top: illustration of the single partitions; bottom: contour lines of the solution (dotted: solution from the partitioned solution; lines: solution from the unpartitioned reference domain)

Changing Grids with Different Time Steps

The calculation area Ω is now divided in two domains Ω_1 and Ω_2. Again, the advection of the Gaussian pulse is examined: The pulse starts in Ω_1 and gets transported to Ω_2, where the calculation ends. The error norms are determined then. Domain Ω_2 has a larger cell size than Ω_1 ($\Delta x_{\Omega_1} < \Delta x_{\Omega_2}$) in the first two cases and vice versa in a third one. Both grids are refined in an equal ratio for determination of the order of convergence. For the time steps, the same ratio as for the grid sizes is chosen ($\Delta x_2/\Delta x_1 = \Delta t_2/\Delta t_1 = 2$ in the first, $\Delta x_2/\Delta x_1 = \Delta t_2/\Delta t_1 = 4$ in the second and $\Delta x_2/\Delta x_1 = \Delta t_2/\Delta t_1 = 1/4$ in the third case). First of all, the ghostcells of the finer domain with the smaller time step is not updated and integrated in time via the CK procedure. Therefore, they

Table 2. Error norms and orders of convergence in Ω_2 for the case of changing grids with different time steps $\Delta t_2/\Delta t_1 = \Delta x_2/\Delta x_1$; the ghostcells are not integrated in time with the CK procedure

#	$\frac{\Delta x_2}{\Delta x_1}$	Ω_2	L_∞ Error	L_1 Error	L_∞ Order	L_1 Order
1	2	150× 50	9.71074E-04	2.67910E-05	–	–
2	2	225× 75	6.47382E-04	1.78773E-05	1.0	1.0
3	2	300×100	4.85993E-04	1.34097E-05	1.0	1.0
4	2	375×125	3.89048E-04	1.07284E-05	1.0	1.0
1	4	80× 26	1.10042E-03	1.09415E-04	–	–
2	4	120× 39	7.37393E-04	7.33741E-05	1.0	1.0
3	4	160× 52	5.52716E-04	5.51462E-05	1.0	1.0
4	4	200× 65	4.42873E-04	4.42760E-05	1.0	1.0
1	1/4	280×104	2.17131E-02	2.44640E-03	–	–
2	1/4	420×156	1.44238E-02	1.62613E-03	1.0	1.0
3	1/4	560×208	1.08050E-02	1.21809E-03	1.0	1.0
4	1/4	700×260	8.63941E-03	9.73842E-04	1.0	1.0

Table 3. Error norms and orders of convergence in Ω_2 for the case of changing grids with different time steps $\Delta t_2/\Delta t_1 = \Delta x_2/\Delta x_1$; the ghostcells are integrated and updated in time with the CK procedure

#	$\frac{\Delta x_2}{\Delta x_1}$	Ω_2	L_∞ Error	L_1 Error	L_∞ Order	L_1 Order
1	2	150× 50	4.56111E-06	9.04538E-08	–	–
2	2	225× 75	1.32447E-07	2.58487E-09	8.7	8.8
3	2	300×100	1.03001E-08	1.99641E-10	8.9	8.9
4	2	375×125	1.40747E-09	2.70982E-11	8.9	8.9
1	4	80× 26	1.23498E-03	2.97428E-05	–	–
2	4	120× 39	6.06183E-05	1.08607E-06	7.4	8.2
3	4	160× 52	5.00610E-06	8.95367E-08	8.7	8.7
4	4	200× 65	7.36905E-07	1.27843E-08	8.6	8.7
1	1/4	280×104	1.54786E-03	3.56406E-05	–	–
2	1/4	420×156	6.24332E-05	1.27408E-06	7.9	8.2
3	1/4	560×208	5.30053E-06	1.05903E-07	8.6	8.6
4	1/4	700×260	7.47070E-07	1.51503E-08	8.8	8.7

keep the value of the last data exchange. Table 2 shows the error norms and convergence rates: It is obvious, that only a 1st order convergence could be achieved, although both domains used 9th order methods. The experiment is repeated with the CK procedure switched on for the ghostcells. Table 3 now shows full 9th order of convergence in the error norms for the coupled calculation. Thus, it has been proved that a precise simulation in time in very high-order is possible, even for changing grid sizes and time steps. This is one of the most important features of the coupling mechanism that leads to a tremendous decrease of CPU time without sacrificing accuracy.

5 Numerical Results

5.1 Coupling ADER-DG and ADER-FV: Numerical Convergence

In the following numerical examples, we show that the same behavior towards convergence as in the validation part of Sect. 4 is obtained when ADER-DG domains are coupled to other ADER-DG domains or ADER-FV domains with different time steps.

The general computational setup is as follows: We solve the linearized Euler equations with the background flow $\rho_0 = 1$, $u_0 = 1$, $v_0 = 0$ and $p_0 = 1/\gamma$ in the two domains $\Omega_1 = [0; 50] \times [0; 50]$ and $\Omega_2 = [50; 100] \times [0; 50]$ using 4th order schemes in both domains. The initial condition at $t = 0$ with $\sigma = 5$ is

$$\rho'(0) = e^{-\frac{1}{2}\frac{(x-25)^2+(y-25)^2}{\sigma^2}}, \qquad u'(0) = v'(0) = p'(0) = 0\,. \tag{20}$$

The exact reference solution of this problem at time $t = 50$ is given by

$$\rho'(50) = e^{-\frac{1}{2}\frac{(x-75)^2+(y-25)^2}{\sigma^2}}, \qquad u'(50) = v'(50) = p'(50) = 0\,. \tag{21}$$

Table 4. Convergence rates for coupling two ADER-DG $\mathcal{O}4$ domains *with* the CK procedure on the domain interface

N_G	L_∞	L_1	L_2	$\mathcal{O}_{L_\infty}$	$\mathcal{O}_{L_1}$	$\mathcal{O}_{L_2}$
10/19	1.41554E-03	2.16305E-02	2.22552E-03			
20/38	9.12429E-05	1.47782E-03	1.60463E-04	3.96	3.87	3.79
30/57	3.45024E-05	3.00860E-04	3.45277E-05	2.40	3.93	3.79
40/76	1.07109E-05	9.62131E-05	1.10751E-05	4.07	3.96	3.95

Table 5. Convergence rates for coupling of *fast*ADER $\mathcal{O}4$ and ADER-DG $\mathcal{O}4$ domains *with* the CK procedure

N_G	L_∞	L_1	L_2	$\mathcal{O}_{L_\infty}$	$\mathcal{O}_{L_1}$	$\mathcal{O}_{L_2}$
20/10	2.03021E-01	1.05524E+01	9.14841E-01			
40/20	3.81902E-02	1.78482E+00	1.58902E-01	2.41	2.56	2.53
60/30	9.51515E-03	4.04211E-01	3.49496E-02	3.43	3.66	3.73
80/40	3.65000E-03	1.45833E-01	1.15907E-02	3.33	3.54	3.84

In our first computation, we use two ADER-DG domains with regular unstructured grids that do not conform at the boundary. In both grids we use the maximum possible time step, leading to a ratio of the time steps of 1.9:1. It can be shown that, if the CK procedure is not applied, the global order of accuracy is decreasing to 1.5 on the finest grid. In the next computation we use the CK procedure to update the ghostcells and obtain the results shown in Table 4: The method proves to be globally 4th order accurate. We perform a third computation in which we couple a Cartesian ADER-FV domain Ω_1 with an ADER-DG domain Ω_2. Due to the more restrictive CFL condition in the 4th order ADER-DG domain, the time step ratio now is about 15:1. The calculation (again with the CK procedure) globally restores almost 4th order of accuracy (Table 5).

5.2 Single Airfoil Gust Response Problem

The single airfoil gust response by Scott [20] from the 4th Computational Aeroacoustics (CAA) Workshop on Benchmark Problems (Problem 1 from Category 3) has been calculated. The results presented here are for testcase one and two, a Joukowski profile with 12% thickness ratio, a free stream Mach number of $Ma_\infty = 0.5$. In case 1, the profile has a camber ratio of zero and an angle of attack of $\alpha = 0°$, in case 2, the profile has a camber ratio of 0.02 and an angle of attack of $\alpha = 2°$. In the inflow condition a vortical gust is prescribed with:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} U_\infty \\ 0 \end{bmatrix} + \epsilon \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \cos\left(k(x+y) - k\,t\right), \tag{22}$$

$$\epsilon = 0.02\,, \tag{23}$$

$$k = \{0.1, 2.0\}\,. \tag{24}$$

Case 1 is calculated at first with a low gust frequency ($k = 0.1$) and a high gust frequency ($k = 2.0$), then case 2 with $k = 2.0$ is calculated. Among the other cases of the benchmark problem, the first and the last case can be considered to be the easiest (due to the low frequency) and the most difficult (due to the high frequency) example to test the numerical methods and the domain decomposition. In our approach, there is no difference in complexity between calculating case 2 with or without camber ratio, but in the first case, a numerical reference solution exists. The computational domain is $[-20.5, \ldots, 62] \times [-42, \ldots, 42]$. This domain is discretized with 1 unstructured domain and 11 structured domains, see Fig. 15 and Table 7. On the inner 5 domains the nonlinear Euler equations (EE) are solved and on the outer 7 domains we solve the linearized Euler (LEE) equations with 11th order of accuracy in space and time. As boundary conditions we use a further set of 4 structured domains where we apply a sponge layer (see (27) in "5.4. Acoustic Scattering of a Circular Cylinder" in this section). Figure 16 shows the numerical solution (hydrodynamic + acoustic pressure, sponge layers not visible).

In case 1 with the low-frequency gust ($k = 0.1$), the solution is compared with the solution from a linear theory (Scott [20]). The solution for case 1 matches the reference solution very well and is shown in Fig. 17. For case 1 with $k = 2.0$ (Figure 18), there are unfortunately no reference data. Due to the varying gust frequencies, case 2 with $k = 2.0$ resembles case 1 with $k = 2.0$ more than case 1 with $k = 0.1$. For case 2 with $k = 2.0$, the linearized theory cannot be applied anymore for an accurate far-field solution because of the high frequency [19]. Here, the solution is compared to a numerical reference solution by Golubev [4], which is shown in Fig. 19. The solutions correlate also nicely, although the predicted amplitudes are a little bit too small.

Golubev et al. [4] also gives CPU times for his calculation: 47 hours on a cluster of 16 × 2.4 GHz Linux machines. The calculation with the heterogeneous domain decomposition was performed on a Linux machine with a single 3.2 GHz processor. For comparison purposes, the CPU time per time and area

$$t_{factor} = \frac{CPU_{time} \times No._{Processors}}{Area \times t_{end} \times Frequency} \tag{25}$$

Table 6. Comparison of the calculations

Parameter	Golubev	Domain Decomposition
CPU_{time}	47h	115h
$No._{Processors}$	16	1
$Frequency$	2.4	3.2
$Area$	<600	14535
t_{end}	400	200
t_{factor}	$1.3E-03$	$1.2E-05$

unit is used. Also the number of processors and their clock frequency is considered (Table 6). Golubev's calculation area consists of two ellipses, which fit into a rectangle of the extents [–10:10]×[–15:15]. All in all, Golubev calculates a smaller area in less wall clock time, but with 16 processors. However, in terms of CPU time per calculated area, the heterogeneous coupling approach is faster (Table 6): If the larger domain decomposition area was calculated by Golubev's method with only one processor of the same specifications, the calculation would be 2 magnitudes of wall-clock time slower. Of course, this comparison is only a coarse approximation and does not necessarily allow the reverse if the calculation area is reduced to Golubev's domain for the domain decomposition approach. Large linear, and thus inexpensive areas would be left out, so the ratio of CPU-time per area would increase. On the other side, the comparison demonstrates the domain decomposition's potential of accelerating calculations in large domains. In this regard, Table 7 shows the percentage of total CPU time used by the single subdomains in comparison with their area contribution for the coupled calculation. It is clear, that the smallest areas with the high-order non-linear methods consume most of the CPU time, whereas the wave transport into the far-field is very inexpensive.

Table 7. Percentage of the CPU-time and of the total area of the calculation

Domain	CPU Time [%]	Area Fraction [%]	Discretization
Unstruct. 01	14.4	0.01	ADER-DG, $\mathcal{O}3$, non-linear EE
Struct. 01	11.0	0.16	ADER-FV $\mathcal{O}4$, non-linear EE
Struct. 02	12.7	0.19	ADER-FV $\mathcal{O}4$, non-linear EE
Struct. 03	8.6	0.12	ADER-FV $\mathcal{O}4$, non-linear EE
Struct. 04	14.8	0.22	ADER-FV $\mathcal{O}4$, non-linear EE
Struct. 05	1.9	3.65	fastADER-FV $\mathcal{O}11$, LEE
Struct. 06	1.9	3.65	fastADER-FV $\mathcal{O}11$, LEE
Struct. 07	1.9	3.65	fastADER-FV $\mathcal{O}11$, LEE
Struct. 08	1.9	3.65	fastADER-FV $\mathcal{O}11$, LEE
Struct. 09	13.8	30.01	fastADER-FV $\mathcal{O}11$, LEE
Struct. 10	1.9	7.64	fastADER-FV $\mathcal{O}11$, LEE
Struct. 11	1.9	7.64	fastADER-FV $\mathcal{O}11$, LEE
Struct. 12	0.2	18.57	fastADER-FV $\mathcal{O}5$, LEE
Struct. 13	0.1	6.44	fastADER-FV $\mathcal{O}5$, LEE
Struct. 14	1.9	7.96	fastADER-FV $\mathcal{O}11$, LEE
Struct. 15	0.1	6.44	fastADER-FV $\mathcal{O}5$, LEE
Coupling	11.0		Interpolation order $\mathcal{O}5$
Total	100.0	100.0	

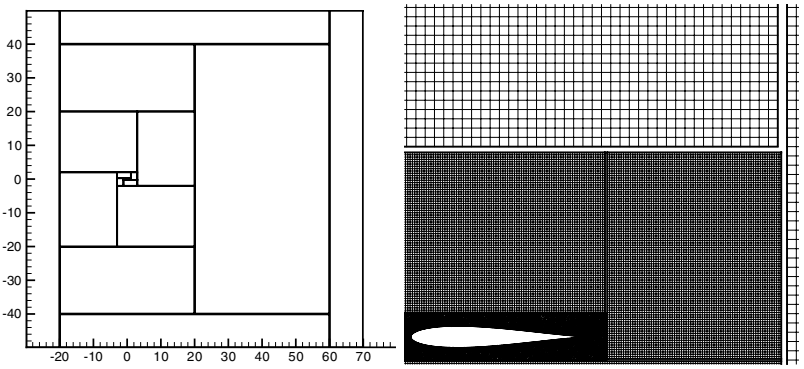


Fig. 15. Domains used for calculation; left: overview over all domains, including sponge layer domains; right: zoom of the core regions

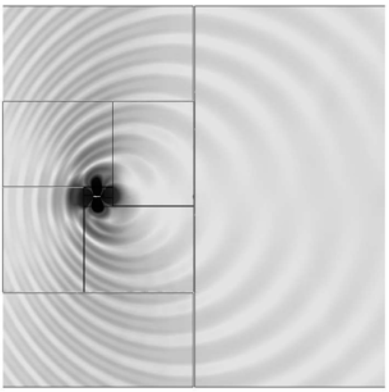


Fig. 16. Full acoustic + hydrodynamic solution for $k = 2.0$

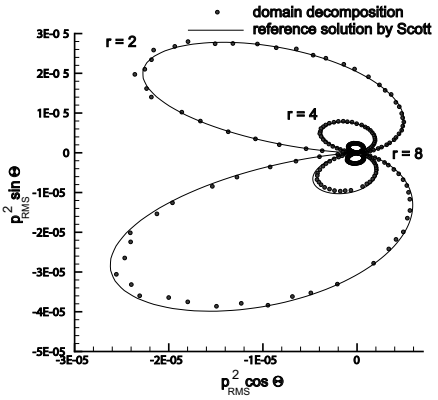


Fig. 17. Directivity pattern of the acoustic pressure for case 1 with $k = 0.1$

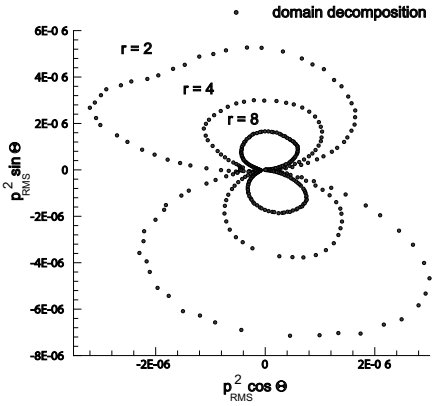


Fig. 18. Directivity pattern of the acoustic pressure for case 1 with $k = 2.0$

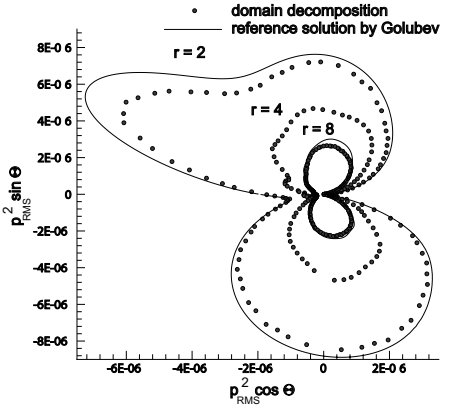


Fig. 19. Directivity pattern of the acoustic pressure for case 2 with $k = 2.0$

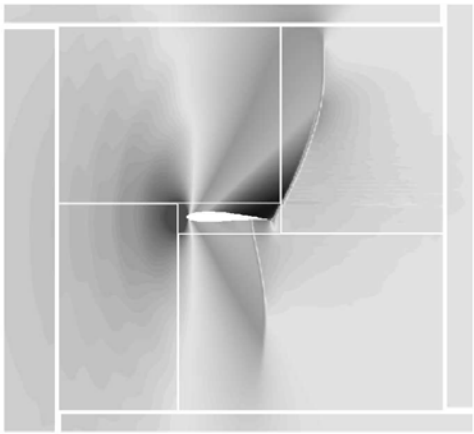


Fig. 20. Pressure contours for a transonic flow at $Ma_\infty = 0.9$; zoom into region

5.3 Transonic Flow

In addition to the previous single airfoil gust response in a subsonic regime at $Ma_\infty = 0.5$, the free stream Mach number has been raised now to $Ma_\infty = 0.9$ for case 2 with $k = 2.0$. At this flow speed, the airfoil becomes transonic and one can expect shocks. Indeed, two shocks are clearly visible on the suction and pressure side which cross the domain coupling interface without any effort (Fig. 20). Yet, the numerical schemes remain stable: An ENO-Limiter by Harten, Engquist, Osher and Chakravarthy [5] was applied to the reconstruction process for the non-linear Cartesian ADER-FV domains.

5.4 Acoustic Scattering of a Circular Cylinder

In order to validate the coupling between ADER-DG schemes on unstructured grids and ADER-FV schemes on Cartesian grids, we choose a non-trivial example for which, however, an analytical solution exists. We compute the scattering of a time harmonic plane sinusoidal acoustic pressure fluctuation incident on a single circular cylinder of radius R with perfectly reflecting walls. The plane wave travels in x-direction and has the form

$$p'_{in}(\mathbf{x}, t) = e^{-I\Omega t} e^{Ikx} = e^{-I\Omega t} e^{Ikr\cos\theta} , \tag{26}$$

with the wavenumber k, the angular frequency Ω and $I = \sqrt{-1}$.

For a derivation of the analytical solution see, e.g., Morse and Feshbach [10]. For the numerical simulation, we solve the problem in a computational

domain with extent $[-20;20] \times [-20;20]$ using the linearized Euler equations in all domains. The background flow, about which the Euler equations are linearized, is chosen as $\rho_0 = 1$, $u_0 = v_0 = 0$ and $p_0 = 1/\gamma$ leading especially to a speed of sound $c_0 = 1$. We use one unstructured domain of extent $[-2;2] \times [-2;2]$ containing the cylinder of radius $R = 1$ and four Cartesian blocks for discretizing the rest of the computational domain. On the unstructured mesh we use a 4th order ADER-DG scheme and on the Cartesian grids we apply a 6th order *fast*ADER scheme [12, 14]. The initial condition is imposed such that it contains only fluctuations in the right-moving characteristic wave with an amplitude of the characteristic fluctuations of 1e-03. The cylinder wall is perfectly reflecting and the radiation boundary condition is imposed by a sponge layer. Since outgoing waves should leave without reflections and at the same time we must allow the incoming plane wave to enter the domain, we use a special sponge layer of the following form

$$\tilde{u}_{ij}^{n+1} = (1-\sigma)\, u_{ij}^{n+1} + \sigma u_{in}\left(x_i, y_j, t^{n+1}\right) \ , \tag{27}$$

where u_{ij}^{n+1} is the time update of the numerical solution as computed by the numerical scheme, $\tilde{u}_{ij}^{n+1}$ is the modified numerical solution after applying the sponge layer, u_{in} is the entering plane wave and $0 \leq \sigma \leq 1$ is the sponge parameter defined as

$$\sigma = \begin{cases} \left(\frac{L-\delta}{L}\right)^s & \text{if } \delta \leq L \\ 0 & \text{if } \delta > L \end{cases} \ . \tag{28}$$

In (28), L denotes the thickness of the sponge layer, δ is the distance of the barycenter of cell Q_{ij} to the nearest boundary of the computational domain and s is the power of the sponge layer. In the following computations, we choose $L = 2$ and $s = 4$. Please note that if the sponge layer is applied to the cell averages according to (27), the scheme is only first order accurate at the boundaries.

We do several computations of this example, first with $k = 2$ on a fine grid with 2048 triangles in the unstructured mesh and 132×108 cells in each Cartesian block, corresponding to a resolution of about 20 points per wavelength. Figure 21 shows a zoom into the inner region ($[-5;5] \times [-5;5]$) of the fine grid domain decomposition. For the unstructured domain the primal mesh is shown, for the Cartesian grids the dual mesh is plotted, explaining the gaps between the grids. A global view of the instantaneous field of the acoustic pressure fluctuations in the entire domain is shown in Fig. 23 (Fig. 24 depicts the pressure fluctuation for $k = 4$). A very important piece of information for acousticians is the so-called directivity pattern, which shows the sound pressure level in a polar plot as a function of angle θ and distance r between observer and the center of the scatterer. Such a directivity pattern, obtained by our simulation, is shown in Fig. 25. The analytical solution is also shown in the same figure. Already in this geometrically simple case the directivity pattern has a quite complicated structure due to constructive and destructive interference between incident and scattered wave, depending on angle and

distance. We note an excellent agreement of our simulation with the exact analytical reference solution.

Next, we solve the same problem ($k = 2$) on a two-times coarser grid in each space dimension, reducing the number of triangles in a slightly enlarged unstructured domain (new extent: $[-3.5; 3.5] \times [-3.5; 3.5]$) to 1152 and the number of cells in each Cartesian block to 85×60. The resolution is now about 11 PPW. Figure 22 shows again a zoom into the inner region ($[-7; 7] \times [-7; 7]$) of the computational domain. We note that the sound pressure level is still very well predicted compared to the analytical solution, see Fig. 26.

Finally, we decrease the resolution to 5.5 PPW when solving the scattering problem for a wavenumber of $k = 4$ on the same grid as in the previous computation. The directivity pattern is compared to the analytical solution in Fig. 27 and we note that the numerical simulation globally predicts a too low sound pressure level because for the 6th order ADER-FV scheme which we used, the amplitude errors become significant in this PPW range [12, 14]. If we increase the order of accuracy of the ADER-FV scheme to 12, we get the directivity patterns depicted in Fig. 28, showing a considerable improvement concerning the magnitude of the SPL. Unfortunately, however, the SPL now is slightly overpredicted in some regions in front of the cylinder. This is most probably due to the sponge layer that reduces the boundary conditions to first order of accuracy. We emphasize that the 12th order ADER-FV scheme uses the same CPU time as a 4th order DRP scheme with 4th order Runge-Kutta timestepping, see Schwartzkopff [16].

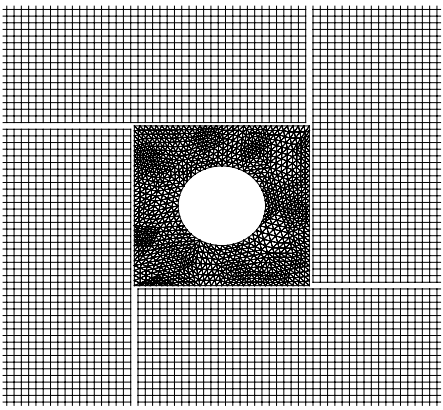


Fig. 21. Domain decomposition for the cylinder scattering problem: zoom into the fine grid

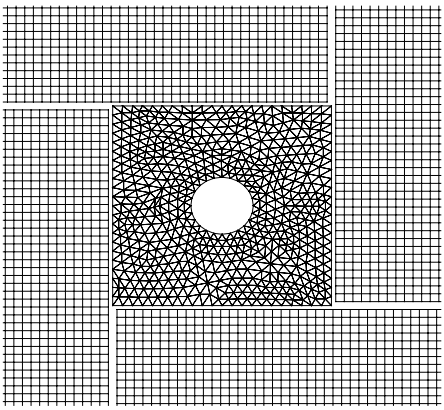


Fig. 22. Domain decomposition for the cylinder scattering problem: zoom into the coarse grid

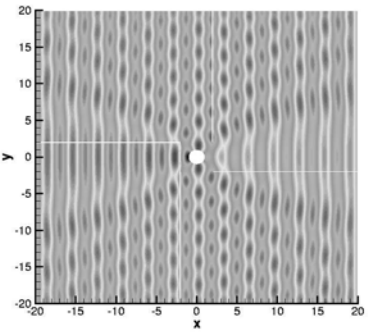


Fig. 23. Instantaneous field of the acoustic pressure fluctuation for $k = 2$ at $t = 100.0$

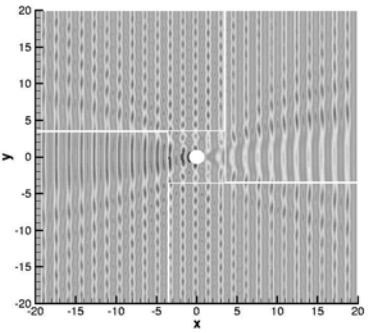


Fig. 24. Instantaneous field of the acoustic pressure fluctuation for $k = 4$ at $t = 100.0$

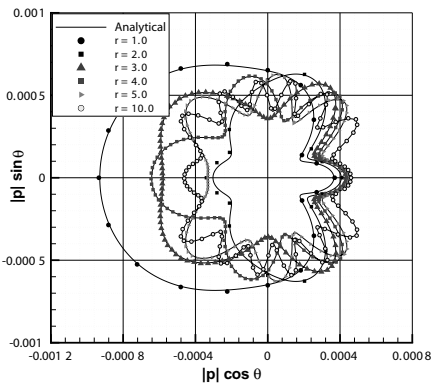


Fig. 25. Directivity patterns for $k = 2$; solution obtained on the fine grid

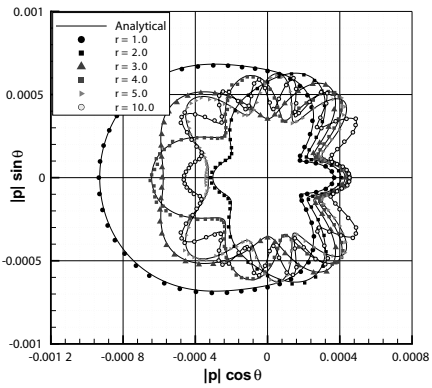


Fig. 26. Directivity patterns for $k = 2$; solution obtained on the coarse grid

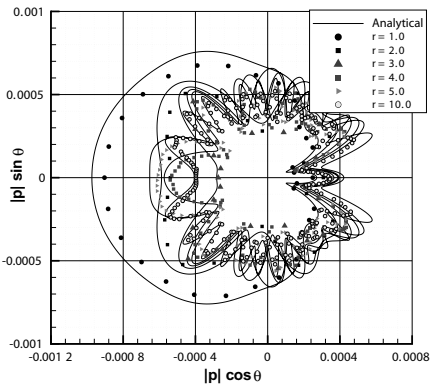


Fig. 27. Directivity patterns for $k = 4$, using a 6th order *fast*ADER scheme

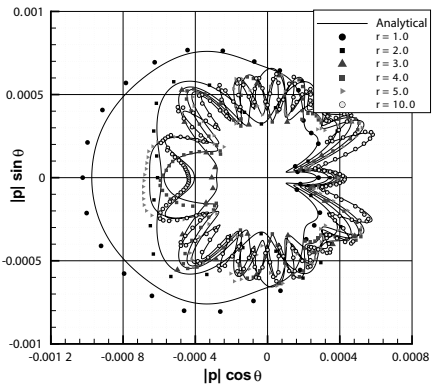


Fig. 28. Directivity patterns for $k = 4$, using a 12th order *fast*ADER scheme

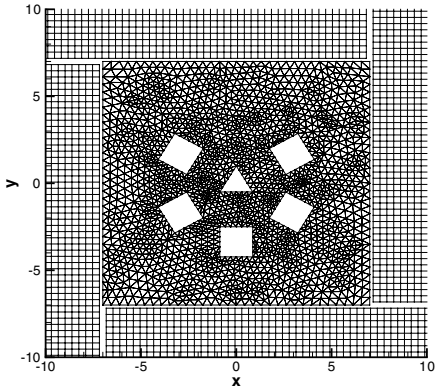


Fig. 29. Domain decomposition for the multiple scattering problem: zoom into the grid

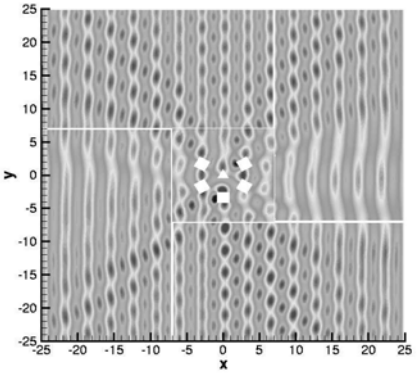


Fig. 30. Instantaneous field of the acoustic pressure fluctuation at $t = 100.0$ for the multi-body scattering problem ($k = 2$)

5.5 Multiple Acoustic Scattering

In the previous section, we validated the domain decomposition approach on a geometrically relatively simple scattering problem for which an analytical reference solution exists. In this section, we apply the approach to a more complicated multi-body scattering problem. Here, the application of a triangular grid for discretizing the inner zone containing the scatterers is very convenient because the unstructured mesh can be generated in a completely automatic manner. For the discretization of the far field, four Cartesian blocks are used.

The configuration is as follows: We solve again the linearized Euler equations with the same mean flow as in the cylinder scattering problem in a computational domain with extent $[-25; 25] \times [-25; 25]$. The unstructured domain covers the zone $[-7; 7] \times [-7; 7]$. The scatterers are supposed to have perfectly reflecting walls and they are subject to an incident plane wave of wavenumber $k = 2$. The scatterers consist of one approximately equilateral triangle and five squares, see Fig. 29. The three nodes of the triangle have the coordinates $(0; 0.5)$, $(-0.866; -0.5)$ and $(0.866; 0.5)$. The lower left corner of the square on the bottom has the coordinate $(-0.866; -4.232)$ and its side length is 1.732. The other four squares are derived from this square by rotation about the origin of $\pm 60°$ and $\pm 120°$. The configuration of the scatterers is not symmetric which should also be visible later on in the directivity patterns. The unstructured grid contains 3556 elements and the Cartesian blocks have 90×50 cells each, leading to an overall resolution of about 9 PPW. The radiation boundary conditions for the far field are again applied by a sponge

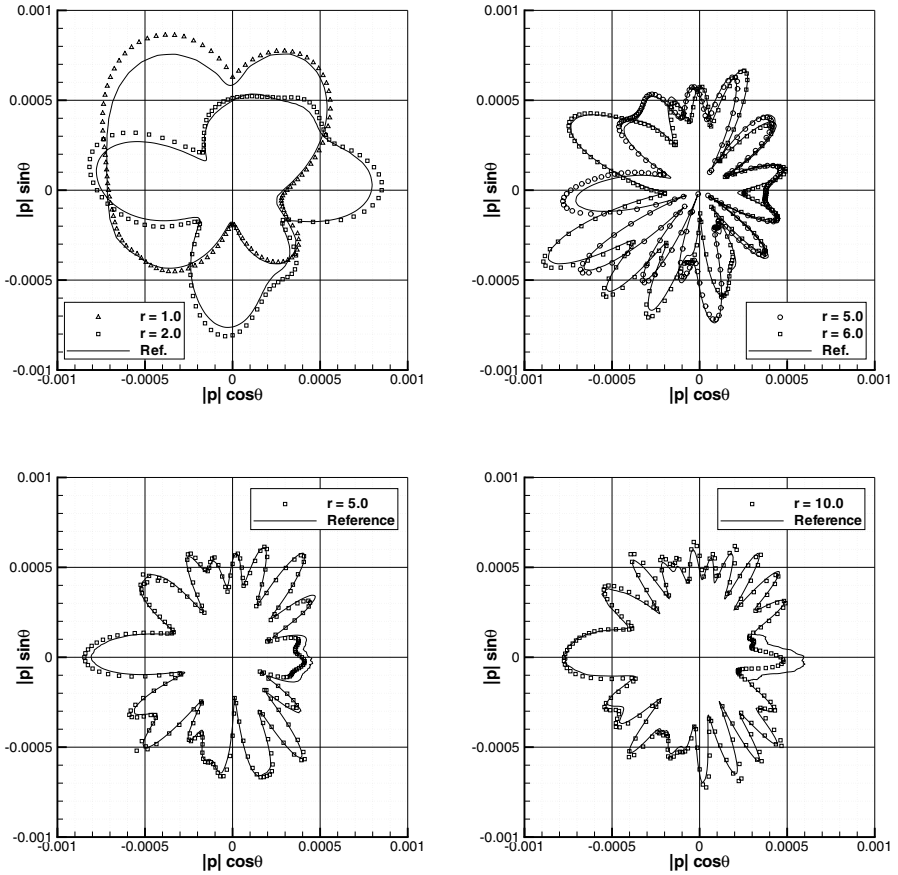


Fig. 31. Directivity patterns of the sound pressure level for the multi-body scattering problem ($k = 2$)

layer. We use a 4th order ADER-DG scheme in the unstructured domain and 6th order ADER-FV schemes on the Cartesian blocks.

The reference solution is this time not computed analytically. Instead, we solve the Helmholtz equation for the scattered wave p'_{sc} on a very fine grid with about 32000 triangles, using the PDE tool of MATLAB. The computational domain is a circle with center (0/0) and radius 25, in the far field, the radiation condition

$$\frac{\partial p'_{sc}}{\partial t} + c_0 \ \mathbf{n} \cdot \nabla p'_{sc} = 0 \tag{29}$$

is imposed, where $c_0 = 1$ is the speed of sound and $\mathbf{n}$ is the outward-pointing normal vector. On the walls we use the boundary condition

$$\frac{\partial p'_{sc}}{\partial n} = -\frac{\partial p'_{in}}{\partial n} \tag{30}$$

for the scattered wave where the right hand side is known from the incident wave (26).

The directivity pattern of our numerical simulation is compared to the one obtained with MATLAB in Fig. 31. We can see a very good agreement between both solutions, showing the capability of the domain decomposition approach to tackle this kind of wave propagation problem also in the time domain. Via the coupling of unstructured and Cartesian grids, it is at the same time possible to take into account the presence of obstacles, which might become geometrically even more complicated than the ones shown in this multiple scattering example, as well as computing large domains by switching as fast as possible from unstructured domains to Cartesian domains for the far field.

6 Conclusion

A heterogeneous domain decomposition approach for the direct simulation of aeroacoustic problems has been presented. The scheme allows the coupling of different numerical methods, grids, equations and time steps. It is capable of retaining globally the order of accuracy that is used by the high-order methods in the domains. The Cauchy-Kovalevskaja procedure proved to be essential for the coupling of regions with different time steps. Numerical examples showed that the domain decomposition decreases the computational effort of a direct numerical simulation. Especially the wave propagation becomes cheap, as fast methods can be used on Cartesian grids. The extension to 3D is straight forward with Gauss point coupling. The coupling scheme is currently prepared for parallel computing. All subdomains can be calculated independently with data exchange at the interfaces of adjacent subdomains in each time step. The subdomains can then be divided into MPI partitions. While the numerical results shown here refer to benchmark examples, it will be interesting to apply the domain decomposition method to real life problems.

References

1. M. Dumbser and C.-D. Munz. Arbitrary high order Discontinuous Galerkin schemes. In S. Cordier, T. Goudon, M. Gutnic, and E. Sonnendrücker, editors, *Numerical Methods for Hyperbolic and Kinetic Problems*, IRMA Series in Mathematics and Theoretical Physics, pages 295–333. EMS Publishing House, 2005.
2. R. W. Dyson. Technique for very high order nonlinear simulation and validation. Technical Report TM-2001-210985, NASA, 2001.
3. J. B. Freund, S. K. Lele, and P. Moin. Matching of near/far-field equation sets for direct computation of aerodynamic sound. In *15th AIAA Aeroacoustics Conference*, October 1993. AIAA-1993-4326.

4. V. V. Golubev, R. R. Makbadi, and J. R. Scott. Numerical inviscid analysis of nonlinear airfoil response to impinging high-intensity high-frequency gust. In *Proceedings of 10th AIAA/CEAS Aeroacoustics Conference*, 2004. AIAA 2004-3002.
5. A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes III. *J. Comput. Phys.*, 71:231–303, 1987.
6. C. Hirsch. *Numerical Computation of Internal and External Flows Vol I: Fundamentals of Numerical Discretisation.* Wiley, 1988.
7. P. D. Lax and B. Wendroff. Systems of conservation laws. *Communications in Pure and Applied Mathematics*, 13:217–237, 1960.
8. M. J. Lighthill. On sound generated aerodynamically - 1. General theory. *Proc. Roy. Soc. London*, 211(A1107):564–587, 1952.
9. R. L. Meakin. Composite overset structured grids. In J. Thompson, B. Soni, and N. Weatherill, editors, *Handbook of Grid Generation.* CRC Press, 1998.
10. P. M. Morse and H. Feshbach. *Methods of Theoretical Physics.* McGraw-Hill, 1953.
11. C.-D. Munz and T. Schwartzkopff. Direct simulation of sound generation and sound propagation using heterogenous domain decomposition. In *Proccedings of WCCM*, July 2002.
12. T. Schwartzkopff, M. Dumbser, and C.-D. Munz. Fast high order ADER schemes for linear hyperbolic equations and their numerical dissipation and dispersion. Technical Report 2003/35, Preprint series of SFB404, Stuttgart University, 2003.
13. T. Schwartzkopff, M. Dumbser, and C.-D. Munz. CAA using domain decomposition and high order methods on structured and unstructured meshes. In *10th AIAA/CEAS Aeroacoustic Conference*, Manchester, GB, May 2004. AIAA-2004-2964.
14. T. Schwartzkopff, M. Dumbser, and C.-D. Munz. Fast high order ADER schemes for linear hyperbolic equations. *J. Comput. Phys.*, 197:532–539, 2004.
15. T. Schwartzkopff and C.-D. Munz. Direct simulation of aeroacoustics. In W. L. Wendland and M. Efendiev, editors, *Analysis and Simulation of Multifield Problems*, volume 12 of *Lecture Notes in applied and computational Mechanics*, pages 337–342. Springer, 2003.
16. T. Schwartzkopff and C.-D. Munz. fastADER: An arbitrary high order scheme for linear acoustics. In *Proceedings of the 4th Computational Aeroacoustics (CAA) Workshop on Benchmark Problems*, September 2004. NASA/CP-2004-212954.
17. T. Schwartzkopff, C.-D. Munz, and E. F. Toro. ADER: A high order approach for linear hyperbolic systems in 2d. *Journal of Scientific Computing*, 17(1-4):231–240, 2002.
18. T. Schwartzkopff, C.-D. Munz, E. F. Toro, and R. C. Millington. The ADER approach in 2d. In T. Sonar, editor, *Discrete Modelling and Discrete Algorithms on Continuum Mechanics*, pages 207–216, Berlin, 2001. Logos Verlag.
19. J. R. Scott. Benchmark solutions for computational aeroacoustics (CAA) code validation. *IMECE2004-59865*, November 2004. Anaheim, CA.
20. J. R. Scott. Single airfoil gust response. In *Proceedings of 4th Computational Aeroacoustics (CAA) Workshop on Benchmark Problems*, September 2004. NASA/CP-2004-212954.

21. J. Steger, F. C. Dougherty, and J. Benek. A chimera grid scheme. In K. Ghia and U. Ghia, editors, *Advances in Grid Generation*, volume 5, pages 59–69. ASME FED, 1983.
22. C. K. W. Tam. Aliasing. In *Proceedings of 4th Computational Aeroacoustics (CAA) Workshop on Benchmark Problems*, September 2004. NASA/CP-2004-212954.
23. C. K. W. Tam and K. A. Kurbatskii. Multi-size-mesh multi-time-step dispersion-relation-preserving scheme for multiple-scales aeroacoustic problems. *International Journal of Computational Fluid Dynamics*, 17:119–132, 2003.
24. P. Tipler. *Physik*. Spektrum Akademischer Verlag, 1994.
25. V. A. Titarev and E. F. Toro. ADER: Arbitrary high order Godunov approach. *Journal of Scientific Computing*, 17(1-4):609–618, December 2002.
26. V. A. Titarev and E. F. Toro. ADER schemes for three-dimensional nonlinear hyperbolic systems. *Isaac Newton Institute for Mathematical Sciences Preprint Series*, 2003.
27. E. F. Toro, R. C. Millington, and L. Nejad. Towards very high order Godunov schemes. In E. F. Toro, editor, *Godunov Methods. Theory and Applications*, pages 907–940. Kluwer/Plenum Academic Publishers, 2001.

Boundary Element Tearing and Interconnecting Domain Decomposition Methods*

Günther Of[2], Olaf Steinbach[2], and Wolfgang L. Wendland[1]

[1] Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
wendland@mathematik.uni-stuttgart.de

[2] Institute for Computational Mathematics, Graz University of Technology, Steyrergasse 30, 8010 Graz, Austria
of@tugraz.at
o.steinbach@tugraz.at

Summary. The combination of non-overlapping domain decomposition methods with fast boundary element methods provides an efficient simulation tool to handle coupled boundary value problems with piecewise constant coefficients. Based on a standard boundary element domain decomposition formulation using the symmetric boundary integral representation of the Dirichlet to Neumann map, i.e., the Steklov–Poincaré operator, we have to solve a system of linear equations in parallel, where the assembled stiffness matrix is symmetric and positive definite. Instead, by using a boundary element tearing and interconnecting approach we have to solve a saddle point problem or the dual problem for finding the Lagrange multipliers. For this method, we describe different preconditioned solution strategies using both, local and global preconditioning techniques. All local boundary integral operators are realized via a fast multipole method leading to an almost optimal algorithm. Besides a rigorous mathematical analysis we give numerical examples and applications for the potential equation as well as for the system of linear elastostatics with jumping coefficients.

Keywords: Domain decomposition methods, boundary integral equations, fast multipole method

1 Introduction

Domain decomposition methods based on boundary integral equations are an efficient approach for handling partial differential equations with piecewise constant coefficients, to parallelize the implementation, and to construct appropriate preconditioners. The global boundary value problem is here replaced

* Research Project C10 "Domain Decomposition Methods"

by local boundary value problems and additional transmission or interface conditions. The Dirichlet to Neumann map of the local boundary value problems can be expressed in terms of boundary integral equations to represent the Steklov–Poincaré operator. Although the boundary integral representation of the Dirichlet to Neumann map is not unique, the so-called symmetric formulation turns out to be favorable from both, the theoretical and the numerical point of view. By employing the local Dirichlet to Neumann maps, the global boundary value problem is reformulated as a global variational problem to determine the complete Cauchy data on the skeleton of the domain decomposition.

The first boundary element domain decomposition methods were based on the weakly singular boundary integral equations only [11, 12, 42]. The boundary element discretization by collocation and Galerkin schemes were analyzed in, e.g., [87]. Based on Neumann series, alternative realizations of the Dirichlet to Neumann map were considered in [37, 38]. The symmetric formulation of boundary element domain decomposition methods was introduced in [15, 40]. Efficient preconditioned parallel iterative solution strategies then were proposed and analyzed in [13, 49, 79]. In particular, global preconditioners are needed which are robust with respect to the coefficients' jumps [6, 36, 43, 44]. In addition, efficient local preconditioners are needed for the fast iterative solution of the local subproblems. There are many different methods available. Some of them are additive or multiplicative methods [35, 56, 85], BPX-like preconditioners [7, 25, 80], multigrid methods [52, 53, 86], boundary integral operators of opposite orders [83], and the application of an approximation of the inverse by means of the $\mathcal{H}$-matrix arithmetics [2, 29].

The Galerkin approximation of the global variational formulation leads to an assembled linear system, where the local contributions are dense matrices due to the non-local boundary integral operators involved. While there exist several formulations of standard boundary element domain decomposition methods, their applicability is restricted – in particular for three-dimensional problems, where the number of boundary elements is rather large for describing structures with complicated geometries. Hence there is a serious need to combine the boundary integral equations with fast solution methods. In the last decades there was a significant development of the fast boundary element methods. In particular, now fast boundary element methods such as the fast multipole algorithm [28, 74], panel clustering [31], wavelet compression techniques [16, 33, 75] and algebraic approximation methods, i.e., Adaptive Cross Approximation [1, 3] and Hierarchical Matrices [4, 26, 29, 30], are available for the Laplace equation. Although those methods can be transfered directly to other partial differential equations, e.g., to the system of linear elastostatics, to the Helmholtz equation, and to the Maxwell system [9, 10], it is not obvious how those methods should be adopted to obtain an almost optimal realization. In fact, this depends on the particular representation of the boundary integral operators under consideration.

The Finite Element Tearing and Interconnecting (FETI) methods [21, 22] are currently widely used in engineering applications, see, e.g., [18, 19, 20, 23, 72, 78], and the analysis of these methods is well understood [8, 45, 47, 57, 58]. The main ideas of these methods consist of the tearing of the global trial functions into local ones, and the subsequent interconnecting of the local trial functions across the interfaces by means of interface constraints and Lagrange multipliers. New versions, the so-called FETI-DP [19] have recently been developed to deal with so-called floating subdomains.

Based on different boundary integral representations of the Dirichlet to Neumann map, conforming boundary element domain decomposition methods were formulated in [39, 82]. The aim of this project C10 was to replace the standard Galerkin discretization of the boundary integral operators by an approximation with fast boundary element methods, in particular by the fast multipole method [66], and to develop and to analyze a fast boundary element domain decomposition algorithm. Instead of a standard Dirichlet domain decomposition approach, the Boundary Element Tearing and Interconnecting (BETI) method in [54] was chosen as a starting point.

This paper is organized as follows. In Sect. 2, we formulate symmetric boundary element domain decomposition methods for two model problems, the potential equation and the system of linear elastostatics with jumping coefficients. The boundary element discretization is based on a standard Galerkin boundary element method. Then the fast multipole method for all discrete boundary integral operators is presented in Sect. 3. Using integration by parts, the applications of all the boundary integral operators of linear elastostatics are reduced to those of the Laplace equation. Preconditioning techniques suitable for fast boundary element methods are given in Sect. 4. In particular, efficient solution strategies for the Neumann boundary value problems are considered. In Sect. 5, we describe various solution algorithms for the boundary element tearing and interconnecting methods. Finally, numerical results are given and they are in agreement with the theoretical estimates.

2 Boundary Element Domain Decomposition Methods

We consider the mixed boundary value problem

$$\begin{aligned} (Lu)(x) &= 0 && \text{for } x \in \Omega\,, \\ \gamma_0 u(x) &= g_D(x) && \text{for } x \in \Gamma_D\,, \\ \gamma_1 u(x) &= g_N(x) && \text{for } x \in \Gamma_N \end{aligned} \tag{1}$$

of a second order scalar elliptic partial differential operator L or second order formally positive elliptic systems (see [62]) in a bounded, simply connected domain $\Omega \subset \mathbb{R}^3$ with Lipschitz boundary $\Gamma = \partial\Omega$ which is decomposed into two adjacent parts Γ_D and Γ_N, $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, where boundary conditions of

Dirichlet and Neumann type are given, respectively. As model problems, we consider the scalar differential operator

$$(Lu)(x) = -\mathrm{div}[\alpha(x)\nabla u(x)] \tag{2}$$

with some positive coefficient function $\alpha(x)$ and the system of linear elastostatics

$$(Lu)(x) = -\mathrm{div}\,\sigma(u,x)\ . \tag{3}$$

In the latter case, the stress tensor $\sigma(u)$ is related to the strain tensor $e(u)$ by Hooke's law

$$\sigma(u) = \frac{E\nu}{(1+\nu)(1-2\nu)}\mathrm{tr}\ e(u)I + \frac{E}{(1+\nu)}e(u)$$

with the Young modulus $E > 0$, the Poison ratio $\nu \in (0, 1/2)$ and the linearized strain tensor

$$e(u) = \frac{1}{2}(\nabla u^\top + \nabla u)\ .$$

In the case of linear elastostatics, the boundary conditions can be assigned component-wise. In our notation, all functions are either scalar or vector valued depending on the considered problem. For a function u given in the domain Ω, its trace on the boundary Γ is defined as

$$\gamma_0 u(x) := \lim_{\Omega \ni \widetilde{x} \to x \in \Gamma} u(\widetilde{x}) \qquad \text{for almost all } x \in \Gamma\ .$$

The associated conormal derivative is given by

$$\gamma_1 u(x) := \lim_{\Omega \ni \widetilde{x} \to x \in \Gamma} n(x) \cdot [\alpha(\widetilde{x})\nabla u(\widetilde{x})] \qquad \text{for almost all } x \in \Gamma$$

in the case of the scalar equation and by

$$\gamma_1 u(x) := \lim_{\Omega \ni \widetilde{x} \to x \in \Gamma} [\sigma(u,\widetilde{x})n(x)] \qquad \text{for almost all } x \in \Gamma$$

in the case of linear elastostatics. In both cases, $n(x)$ denotes the outer normal vector at almost all $x \in \Gamma$ with respect to the surface measure on Γ.

Here and in what follows, we consider a non-overlapping domain decomposition of the domain Ω into p mutual non-overlapping subdomains Ω_i, i.e.

$$\overline{\Omega} = \bigcup_{i=1}^{p} \overline{\Omega}_i \ \text{ where } \Omega_i \cap \Omega_j = \emptyset \quad \text{for } i \neq j\ .$$

$\Gamma_i := \partial\Omega_i$ denotes the boundary of the subdomain Ω_i. The local coupling boundaries of adjacent subdomains are given by $\Gamma_{ij} := \Gamma_i \cap \Gamma_j$ for all $i < j$. The skeleton of the domain decomposition is defined by

$$\Gamma_S := \bigcup_{i=1}^{p} \Gamma_i = \Gamma \cup \bigcup_{i<j} \overline{\Gamma}_{ij}\ .$$

We assume that the material parameters are piecewise constant, i.e., α_i, E_i and ν_i are constant in each subdomain Ω_i. Then we consider the local boundary value problems

$$\begin{aligned} (L_i u_i)(x) &= 0 && \text{for } x \in \Omega_i \ , \\ \gamma_0^i u_i(x) &= g_D(x) && \text{for } x \in \Gamma_i \cap \Gamma_D \ , \\ \gamma_1^i u_i(x) &= g_N(x) && \text{for } x \in \Gamma_i \cap \Gamma_N \end{aligned} \tag{4}$$

[3]

with the transmission conditions

$$\gamma_0^i u_i(x) = \gamma_0^j u_j(x) \quad \text{and} \quad \gamma_1^i u_i(x) + \gamma_1^j u_j(x) = 0 \quad \text{for } x \in \Gamma_{ij} \ . \tag{5}$$

γ_0^i and γ_1^i denote the local trace operators with respect to the subdomain Ω_i.

The local Steklov–Poincaré operator S_i defines the local Dirichlet to Neumann map for each of the individual subdomains Ω_i and is given as

$$\gamma_1^i u_i(x) = \left[D_i + (\frac{1}{2} I + K_i')V_i^{-1}(\frac{1}{2} I + K_i) \right] \gamma_0^i u_i(x) =: (S_i \gamma_0^i u_i)(x)$$

for $x \in \Gamma_i$, when using boundary integral operators. These are the single layer potential $V_i : H^{-1/2}(\Gamma_i) \to H^{1/2}(\Gamma_i)$

$$(V_i t)(x) = \int_{\Gamma_i \setminus \{x\}} U^*(x,y) t(y) \mathrm{d}s_y \quad \text{for } x \in \Gamma_i \ ,$$

the double layer potential $K_i : H^{1/2}(\Gamma_i) \to H^{1/2}(\Gamma_i)$

$$(K_i u)(x) = \int_{\Gamma_i \setminus \{x\}} T^*(x,y) u(y) \mathrm{d}s_y \quad \text{for } x \in \Gamma_i \ ,$$

the adjoint double layer potential $K_i' : H^{-1/2}(\Gamma_i) \to H^{-1/2}(\Gamma_i)$

$$(K_i' t)(x) = \int_{\Gamma_i \setminus \{x\}} \gamma_{1,x} U^*(x,y) t(y) \mathrm{d}s_y \quad \text{for } x \in \Gamma_i \ ,$$

and the hypersingular operator $D_i : H^{1/2}(\Gamma_i) \to H^{-1/2}(\Gamma_i)$

$$(D_i u)(x) = -\gamma_{1,x} \int_{\Gamma_i \setminus \{x\}} T^*(x,y) u(y) \mathrm{d}s_y \quad \text{for } x \in \Gamma_i \ ,$$

where $T^*(x,y) = (\gamma_{1,y} U^*(x,y))^\top$. The fundamental solution $U^*(x,y)$ is given by

$$U^*(x,y) = \frac{\alpha_i}{4\pi} \frac{1}{|x-y|}$$

[3] Here and in what follows, $x \in \Gamma_i$ is understood almost everywhere with respect to the surface measure of Γ_i.

for (2) and

$$U^*_{k\ell}(x,y) = \frac{1}{8\pi}\frac{1}{E_i}\frac{1+\nu_i}{1-\nu_i}\left[(3-4\nu_i)\frac{\delta_{k\ell}}{|x-y|} + \frac{(x_k-y_k)(x_\ell-y_\ell)}{|x-y|^3}\right]$$

with $k,\ell = 1,\dots,3$ for (3), respectively. We can now formulate the coupled boundary value problems (4) with the transmission conditions (5) as local transmission problems

$$\begin{aligned}
\gamma_1^i u_i(x) &= (S_i u_i)(x) && \text{for } x \in \Gamma_i\ , \\
u_i(x) &= g_D(x) && \text{for } x \in \Gamma_i \cap \Gamma_D\ , \\
\gamma_1^i u_i(x) &= g_N(x) && \text{for } x \in \Gamma_i \cap \Gamma_N\ , \\
u_i(x) &= u_j(x) && \text{for } x \in \Gamma_{ij}\ , \\
\gamma_1^i u_i(x) + \gamma_1^j u_j(x) &= 0 && \text{for } x \in \Gamma_{ij}
\end{aligned}$$

for $i = 1,\dots,p$. Starting from these local transmission problems, several domain decomposition methods [82] can be derived. One of these methods is the Dirichlet domain decomposition method in [13, 15, 40, 41]. This method uses a global function $u \in H^{1/2}(\Gamma_s)$ as unknown to guarantee the continuity of the solution across the coupling interfaces. The local functions $u_i := u_{|\Gamma_i}$ are defined as the restrictions of the global function u. The boundary and coupling conditions of the Neumann data $\gamma_1^i u_i(x)$ are fulfilled only in a variational sense. Using the Dirichlet to Neumann maps which determine the solution of the local subproblems and homogenization, the sum of these local variational equations gives the final variational formulation:
Find $\widehat{u} \in H_0^{1/2}(\Gamma_S,\Gamma_D)$ *such that*

$$\begin{aligned}
\sum_{i=1}^p \int_{\Gamma_i} (S_i \widehat{u}_{|\Gamma_i})(x) v_{|\Gamma_i}(x)\mathrm{d}s_x = &-\sum_{i=1}^p \int_{\Gamma_i} (S_i \widehat{g}_{D|\Gamma_i})(x) v_{|\Gamma_i}(x)\mathrm{d}s_x \\
&+ \int_{\Gamma_N} g_N(x) v_{|\Gamma_N}(x)\mathrm{d}s_x
\end{aligned} \tag{6}$$

holds for all $v \in H_0^{1/2}(\Gamma_S,\Gamma_D) := \left\{v \in H^{1/2}(\Gamma_S) : v(x) = 0 \text{ for all } x \in \Gamma_D\right\}$. Here $\widehat{g}_D \in H^{1/2}(\Gamma)$ is a suitable extension of the given data $g_D \in H^{1/2}(\Gamma_D)$. The solution is then given by $u = \widehat{u} + \widehat{g}_D$. The bilinear form defined by the left-hand side

$$a(u,v) = \sum_{i=1}^p \int_{\Gamma_i} (S_i u_{|\Gamma_i})(x) v_{|\Gamma_i}(x)\mathrm{d}s_x$$

is bounded and $H_0^{1/2}(\Gamma_S,\Gamma_D)$-elliptic [38, 39, 82]. Due to the lemma of Lax–Milgram, the variational problem (6) is uniquely solvable.

For the discretization, we use a regular globally quasi-uniform boundary element mesh $\Gamma_i = \bigcup_{\ell=1}^{N_i} \tau_\ell^i$ with the average mesh size h on Γ_S and Γ_i. The function u is approximated by piecewise linear and continuous basis functions φ_j. The trial space is denoted as $S_h^1(\Gamma_S) = \operatorname{span}\{\varphi_j\}_{j=1}^{M_S}$ and the local re-

strictions of this space are $S_h^1(\Gamma_i) = \operatorname{span}\{\varphi_j^i\}_{j=1}^{M_i}$. This leads to the discrete variational formulation or the equivalent system of linear equations

$$S_h \mathbf{u} = \sum_{i=1}^{p} A_i^\top S_{i,h} A_i \mathbf{u} = \sum_{i=1}^{p} A_i^\top \mathbf{f}_i + \sum_{i=1}^{p} A_i^\top \mathbf{f}_{N,i} \,, \tag{7}$$

with the connectivity matrices $A_i \in \mathbb{R}^{M_i \times M_S}$. These matrices map the global nodes to the local ones and the global vector $\mathbf{v}$ to the local vectors $\mathbf{v}_i = A_i \mathbf{v}$. The components of the vectors $\mathbf{f}_{N,i}$ are given by the Neumann data as

$$f_{N,i,k} = \begin{cases} \displaystyle\int_{\Gamma_N} g_N(x) \varphi^i_{k|\Gamma_N}(x) \mathrm{d}s_x & \text{for supp } \varphi_k^i \cap \Gamma_N \neq \emptyset \\ 0 & \text{else} \,, \end{cases}$$

and those of the vectors $\mathbf{f}_i$ are given by

$$f_{i,k} = -\int_{\Gamma_i} (S_i \widehat{g}_{D|\Gamma_i})(x) \varphi^i_{k|\Gamma_i}(x) \mathrm{d}s_x \,.$$

The system of linear equations (7) is uniquely solvable due to the properties of the bilinear form $a(\cdot,\cdot)$ and Cea's lemma. As the inverse operator V_i^{-1} of the single layer potential can not be represented explicitly, an approximation $\widetilde{S}_i$ of the Steklov–Poincaré operator S_i is used which is based on the inversion of the discrete matrix $V_{i,h}$ of the single layer potential operator. Now, the global system of linear equations (7) can be solved by the use of the discrete local Steklov–Poincaré operators given by

$$\widetilde{S}_{i,h} = D_{i,h} + (\frac{1}{2} M_{i,h}^\top + K_{i,h}^\top) V_{i,h}^{-1} (\frac{1}{2} M_{i,h} + K_{i,h}) \,.$$

The local Neumann data are approximated by piecewise constant basis functions ψ_ℓ^i on each triangle. The Galerkin matrices are given by

$$\begin{aligned} V_{i,h}[\ell,k] &= \langle V_i \psi_k^i, \psi_\ell^i \rangle_{\Gamma_i} \,, & K_{i,h}[\ell,m] &= \langle K_i \varphi_m^i, \psi_\ell^i \rangle_{\Gamma_i} \,, \\ K'_{i,h}[j,k] &= \langle K'_i \psi_k^i, \varphi_j^i \rangle_{\Gamma_i} \,, & D_{i,h}[j,m] &= \langle D_i \varphi_m^i, \varphi_j^i \rangle_{\Gamma_i} \,, \\ M_{i,h}[\ell,m] &= \langle \varphi_m^i, \psi_\ell^i \rangle_{\Gamma_i} \end{aligned}$$

for $k,\ell = 1,\ldots,N_i$ and $m,j = 1,\ldots,M_i$. For the solution $\widetilde{u}_h$ of the approximate version of the system of linear equations (7), the error estimate

$$\begin{aligned} \|u - \widetilde{u}_h\|^2_{H^{1/2}(\Gamma_S)} \leq c_1 \inf_{v_h \in S_h^1(\Gamma_S)} \|(u - \widehat{g}_D) - v_h\|^2_{H^{1/2}(\Gamma_S)} \\ + c_2 \sum_{i=1}^{p} \inf_{\tau_{h,i} \in S_h^0(\Gamma_i)} \left\| S_i u_{|\Gamma_i} - \tau_{h,i} \right\|^2_{H^{-1/2}(\Gamma_i)} \end{aligned} \tag{8}$$

holds [38, 82]. Under the assumption that $u \in H^2(\Gamma_S)$ and $S_i u_{|\Gamma_i} \in \widetilde{H}^1_{\mathrm{pw}}(\Gamma_i)$, this estimate and the approximation properties of the used trial spaces give an optimal convergence rate of $3/2$ [4].

[4] To the knowledge of the authors, for Lipschitz boundaries and mixed boundary conditions such regularity has not yet been proved rigorously.

3 Fast Multipole Boundary Element Method

The effort for an application of the boundary integral operators is quadratic in the number of unknowns. Goal of this section is to provide a fast realization of the boundary integral operators for each subdomain. For ease of reading, we drop the indices indicating the subdomains. The fast multipole method [28, 74] can be used as a fast boundary element method to overcome the disadvantage of the quadratic effort for the matrix times vector multiplication of a standard boundary element method. There are several other fast boundary element methods, such as panel clustering methods [31], wavelet matrix compression techniques [16, 33, 75] and the adaptive cross approximation method [1, 3]. The $\mathcal{H}$-matrices [26, 29] and the $\mathcal{H}^2$-matrices [4, 30] provide a whole arithmetic for such kind of approximation matrices. There are many results on the fast multipole method. Reference [65] tries to give an overview on the development of this topic. Here, we will only describe the fast multipole method for a matrix times vector multiplication $\mathbf{w} = V_{L,h}\mathbf{t}$ of the single layer potential of the Laplacian and piecewise constant elements,

$$w_\ell = \sum_{k=1}^{N} V_{L,h}[\ell,k]t_k = \sum_{k=1}^{N} \frac{t_k}{4\pi} \int_{\tau_\ell} \int_{\tau_k} \frac{1}{|x-y|} \mathrm{d}s_y \mathrm{d}s_x$$

for $\ell = 1, \ldots, N$ based on [27, 66, 69]. Several details concerning a fast realization will be dropped for simplicity. The effort to set up this matrix and for a matrix times vector multiplication is of order $\mathcal{O}(N^2)$. If the kernel $|x-y|^{-1}$ can be split into terms with separated variables x and y, this effort will be reduced to an almost linear one. One possibility would be to use the Taylor series expansion, but it turned out to be more efficient to use spherical harmonics. As such an expansion is only valid for $x \neq y$, it can only be used in the so-called farfield of the element τ_ℓ. The use of the truncated expansions leads to the approximation

$$\widetilde{w}_\ell = \sum_{k \in \mathrm{NF}(\ell)} V_{L,h}[\ell,k]\, t_k + \sum_{k \in \mathrm{FF}(\ell)} \frac{t_k}{4\pi} \int_{\tau_\ell} \int_{\tau_k} \sum_{n=0}^{\varrho} \sum_{m=-n}^{n} \overline{S_n^m}(y) R_n^m(x) \mathrm{d}s_y \mathrm{d}s_x \ ,$$

with the reformulated spherical harmonics ($m \geq 0$) [71, 88, 89]

$$R_n^{\pm m}(x) = \frac{1}{(n+m)!} \frac{\mathrm{d}^m}{\mathrm{d}u^m} P_n(u)\big|_{u=\widehat{x}_3} (\widehat{x}_1 \pm \mathrm{i}\widehat{x}_2)^m |x|^n \ ,$$

$$S_n^{\pm m}(y) = (n-m)! \frac{\mathrm{d}^m}{\mathrm{d}u^m} P_n(u)\big|_{u=\widehat{y}_3} (\widehat{y}_1 \pm \mathrm{i}\widehat{y}_2)^m \frac{1}{|y|^{n+1}} \ ,$$

with $\widehat{x} = x/|x|$ and $P_n(u)$ the Legendre polynomials. The coefficients

$$M_n^m(O,\ell) = \int_{\tau_\ell} R_n^m(x) \mathrm{d}s_x \qquad \text{and} \qquad L_n^m(O,k) = \int_{\tau_k} \overline{S_n^m}(y) \mathrm{d}s_y$$

can be computed either by numerical integration or for flat triangles exactly [61, 64] with respect to a local center O. Then the matrix times vector multiplication can be executed in the form

$$\widetilde{w}_\ell = \sum_{k\in \mathrm{NF}(\ell)} V_{L,h}[\ell,k]t_k + \frac{1}{4\pi}\sum_{n=0}^{\varrho}\sum_{m=-n}^{n} M_n^m(O,\ell)\widetilde{L}_n^m(\mathrm{FF}(\ell)) \,. \tag{9}$$

If the computation of the coefficients

$$\widetilde{L}_n^m(\mathrm{FF}(\ell)) = \sum_{k\in \mathrm{FF}(\ell)} t_k L_n^m(O,k) \qquad \text{for } n=0,\ldots,\varrho, m=-n,\ldots,n$$

is efficient, the matrix times vector multiplication will become very fast via (9). The main problem for the efficient computation of the coefficients $\widetilde{L}_n^m(\mathrm{FF}(\ell))$ is that they depend on the corresponding farfield $\mathrm{FF}(\ell)$ and therefore differ from each other. For their efficient computation, as many coefficients $L_n^m(O,k)$ as possible are summed up and used for as many coefficients $\widetilde{L}_n^m(\mathrm{FF}(\ell))$ as possible. A hierarchical structure built upon the boundary elements τ_k is decisive for the efficient calculation of the coefficients $\widetilde{L}_n^m(\mathrm{FF}(\ell))$. For creating the hierarchy, all boundary elements $\{\tau_k\}_{k=1}^N$ are located in a cube, which contains the domain Ω. They form the cluster ω_1^0 of the coarsest level 0. The clusters $\omega_j^{\lambda+1}$ of level $\lambda+1$ are defined as the sons of the clusters ω_i^λ of level λ. To this end, the corresponding cube of each cluster ω_i^λ is subdivided uniformly into eight sub-cubes. The boundary elements of the clusters ω_i^λ are assigned to these eight smaller cubes due to the position of their centers and they form up to eight new clusters $\omega_j^{\lambda+1}$. In the case of a quasi uniform boundary discretization, this construction is executed until a maximal level L such that the clusters of the finest level still contain some single boundary elements as their sons. Adaptive versions of the fast multipole method are described in [14, 61].

A cluster ω_i^λ is in the nearfield of a cluster ω_j^λ of the same level λ, if the condition

$$\mathrm{dist}\,\{C_i^\lambda, C_j^\lambda\} \le (d+1)\max\{r_i^\lambda, r_j^\lambda\} \tag{10}$$

is satisfied. C_i^λ is the center of the cube associated with cluster ω_i^λ and r_i^λ is the corresponding cluster radius, i.e., $r_i^\lambda = \sup_{x\in\omega_i^\lambda}|x - C_i^\lambda|$. This definition is used to define the nearfield and the farfield of the boundary elements τ_ℓ:

$$\begin{aligned}\mathrm{NF}(\ell) &:= \big\{k, 1\le k\le N \text{ and (10) holds for the cluster } \omega_i^L \text{ of } \tau_k \\ &\qquad \text{and } \omega_j^L \text{ is the cluster containing } \tau_\ell.\big\}\,, \\ \mathrm{FF}(\ell) &:= \{1,\ldots,N\}\setminus \mathrm{NF}(\ell)\,.\end{aligned}$$

The symmetric definition of the nearfield is important for the symmetry of the approximation of the matrix $V_{L,h}$ of the single layer potential [66, 69].

Now the coefficients $\widetilde{L}_n^m(\mathrm{FF}(\ell))$ are computed utilizing the cluster hierarchy. First, the multipole coefficients

$$\widetilde{M}_n^m(C_j^L, \mathrm{P}(\omega_j^L)) = \sum_{\tau_k \in \omega_j^L} t_k \int_{\tau_k} R_n^m(x) \mathrm{d}s_x \tag{11}$$

are computed for all clusters ω_j^L of the finest level L. $\mathrm{P}(\omega_j^\lambda) := \{k, \tau_k \in \omega_j^\lambda\}$ is the set of all boundary elements τ_k of the cluster ω_j^λ. The coefficients $\widetilde{M}_n^m$ then are used to compute the multipole coefficients of the next coarser level by the translation

$$\widetilde{M}_n^m(C_j^\lambda, \mathrm{P}(\omega_j^\lambda)) = \sum_{\omega_i^{\lambda+1} \in \mathrm{sons}(\omega_j^\lambda)} \sum_{s=0}^{n} \sum_{t=-s}^{s} R_s^t(\overrightarrow{C_j^\lambda C_i^{\lambda+1}}) \widetilde{M}_{n-s}^{m-t}(C_i^{\lambda+1}, \mathrm{P}(\omega_i^{\lambda+1}))$$

which is done recursively. The wanted local coefficients of a second cluster ω_i^λ in the farfield of ω_j^λ are computed from these multipole coefficients by the conversion

$$\widetilde{L}_n^m(C_i^\lambda, \mathrm{P}(\omega_j^\lambda)) = \sum_{s=0}^{\infty} \sum_{t=-s}^{s} (-1)^n \overline{S_{n+s}^{m+t}}(\overrightarrow{C_j^\lambda C_i^\lambda}) \widetilde{M}_s^t(C_j^\lambda, \mathrm{P}(\omega_j^\lambda)) \ .$$

These conversions are executed on the coarsest level on which the admissibility condition (10) is still fulfilled, i.e., between two clusters which are in their mutual farfields but their fathers are in their own mutual nearfields. These local coefficients are summed up for each cluster and now in the opposite direction are translated from each cluster ω_i^λ to its sons $\omega_j^{\lambda+1}$ by

$$\widetilde{L}_n^m(C_j^{\lambda+1}, \mathrm{FF}(\omega_i^\lambda)) = \sum_{s=n}^{\varrho} \sum_{t=-s}^{s} R_{s-n}^{t-m}(\overrightarrow{C_i^\lambda C_j^{\lambda+1}}) \widetilde{L}_s^t(C_i^\lambda, \mathrm{FF}(\omega_i^\lambda)) \ .$$

The sums of all these coefficients $\widetilde{L}_n^m(C_j^L, \cdot)$ are the coefficients $\widetilde{L}_n^m(\mathrm{FF}(\ell))$ needed for the fast matrix times vector multiplication (9), where ω_j^L is the cluster containing the boundary element τ_ℓ.

In the case of the double layer potential K_L, the normal derivative has to be applied to (11) for setting up the multipole coefficients. In the farfield expansion, the normal derivative has to be applied to the coefficients of the expansion in (9) in the case of the adjoint double layer potential. If the same expansion degree is used for the double layer potential and its adjoint, the approximations will keep the transposedness. The bilinear form of the hypersingular operator D_L can be written as the sum of bilinear forms of the single layer potential V_L [59, 63] as

$$\langle D_L u, v \rangle_\Gamma = \sum_{k=1}^{3} \langle V_L \mathbf{curl}_\Gamma u\big|_k, \mathbf{curl}_\Gamma v\big|_k \rangle_\Gamma \ ,$$

where

$$\mathbf{curl}_\Gamma u(x) := n_x \times \nabla u^*_\Gamma(x) \quad \text{for } x \in \Gamma$$

is the surface curl of a function $u \in C^1(\Gamma)$ and, where $u^*_\Gamma(x)$ is a constant extension of the function $u(x)$ along the normal direction n_x into a small neighborhood of Γ. In the case of piecewise linear, continuous basis functions $\varphi(x)$ and plane triangles as boundary elements, the vector $\mathbf{curl}_\Gamma \varphi(x)$ is piecewise constant and constant on each boundary element. Therefore, the matrix $V_{L,h}$ of the single layer potential can be reused.

Next, we will briefly describe a realization of the boundary integral operators of linear elastostatics, for details see [67, 70]. There are some other versions [24, 34, 73, 89, 90] of a fast multipole boundary element method available in linear elastostatics.

The fundamental solution of linear elastostatics can be rewritten as

$$\begin{aligned} U^*_{k\ell}(x-y) = \frac{1+\nu}{2E(1-\nu)}\frac{1}{4\pi}\Bigg[& (3-4\nu)\frac{\delta_{k\ell}}{|x-y|} \\ & - \frac{1}{2}x_\ell \frac{\partial}{\partial x_k}\frac{1}{|x-y|} - \frac{1}{2}y_\ell \frac{\partial}{\partial y_k}\frac{1}{|x-y|} \\ & - \frac{1}{2}x_k \frac{\partial}{\partial x_\ell}\frac{1}{|x-y|} - \frac{1}{2}y_k \frac{\partial}{\partial y_\ell}\frac{1}{|x-y|}\Bigg] . \end{aligned}$$

This representation guarantees the symmetry of the fast multipole approximation within one matrix block and between the matrix blocks. Based on this representation, the single layer potential V_E is realized by the same techniques as used for the single and double layer potential of the Laplacian. All other boundary integral operators of linear elastostatics are reduced to single and double layer potentials of the Laplacian and to the single layer potential of linear elastostatics. The double layer potential K_E of linear elastostatics can be written as [48]

$$\begin{aligned} \left(K_E u\right)(x) = & \frac{1}{4\pi}\int_\Gamma u(y)\frac{\partial}{\partial n_y}\frac{1}{|x-y|}\mathrm{d}s_y - \frac{1}{4\pi}\int_\Gamma \frac{1}{|x-y|}\left(\mathcal{M}u\right)(y)\mathrm{d}s_y \\ & + 2\mu\left(V_E\left(\mathcal{M}u\right)\right)(x) \end{aligned}$$

where

$$\mathcal{M} = \begin{pmatrix} 0 & -\mathbf{curl}_{\Gamma,3} & \mathbf{curl}_{\Gamma,2} \\ \mathbf{curl}_{\Gamma,3} & 0 & -\mathbf{curl}_{\Gamma,1} \\ -\mathbf{curl}_{\Gamma,2} & \mathbf{curl}_{\Gamma,1} & 0 \end{pmatrix} \quad \text{and} \quad \mu = \frac{E}{2(1+\nu)} .$$

This representation can be reused for the bilinear form of the adjoint double layer potential K'_E such that

$$\langle K'_E t, v\rangle_\Gamma = \langle K'_L t, v\rangle_\Gamma - \langle V_L t, \mathcal{M}v\rangle_\Gamma + 2\mu\langle V_E t, \mathcal{M}v\rangle_\Gamma$$

holds. As in the case of the Laplacian, the bilinear form of the hypersingular operator can be rewritten by integration by parts as a sum of bilinear forms of single layer potentials [32],

$$\langle D_E u, v\rangle_\Gamma = \int_\Gamma \int_\Gamma \frac{\mu}{4\pi} \frac{1}{|x-y|} \left(\sum_{k=1}^{3} (\mathcal{M}_{k+2,k+1} v)(x) \cdot (\mathcal{M}_{k+2,k+1} u)(y) \right) \mathrm{d}s_y \mathrm{d}s_x$$
$$+ \int_\Gamma \int_\Gamma (\mathcal{M} v)^\top (x) \left(\frac{\mu}{2\pi} \frac{I}{|x-y|} - 4\mu^2 U^*(x,y) \right) (\mathcal{M} u)(y) \mathrm{d}s_y \mathrm{d}s_x$$
$$+ \int_\Gamma \int_\Gamma \sum_{i,j,k=1}^{3} (\mathcal{M}_{k,j} v_i)(x) \frac{\mu}{4\pi} \frac{1}{|x-y|} (\mathcal{M}_{k,i} u_j)(y) \mathrm{d}s_y \mathrm{d}s_x \, .$$

Consequently, it is sufficient to have a fast realization of the single and double layer potential. If the expansion degree ϱ of the fast multipole approximation is chosen proportional to $\log N$ [67, 69], the approximation of the boundary integral operators will still guarantee a quasi optimal error estimate like (8) for the perturbed solution. This choice of the expansion degree ϱ leads to an effort of $\mathcal{O}(N \log^2 N)$ for each application of a boundary integral operator by the fast multipole method [67, 69]. The memory requirements are of order $\mathcal{O}(N \log^2 N)$, too.

If the Cauchy data $\gamma_0 u$ and $\gamma_1 u$ are known on the whole boundary Γ, the solution will be given by the representation formula

$$u(x) = \int_\Gamma \gamma_{0,y} U^*(x,y) \gamma_1 u(y) \mathrm{d}s_y - \int_\Gamma \left(\gamma_{1,y} U^*(x,y) \right)^\top \gamma_0 u(y) \mathrm{d}s_y \qquad (12)$$

for $x \in \Omega$. Again, the fast multipole algorithm provides a very fast evaluation of the representation formula even for large sets of evaluation points.

Remark 1. These ideas can also be used for collision detection by utilizing the property

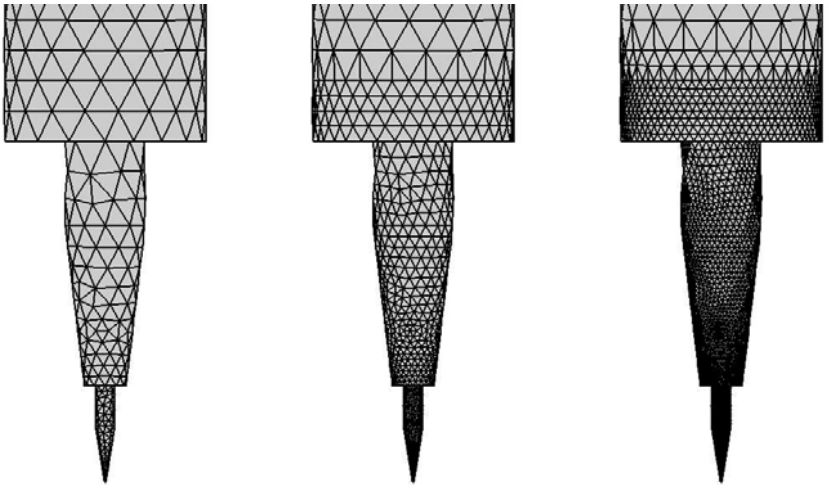
$$-\frac{1}{4\pi} \int_\Gamma n_y \cdot \nabla_y \frac{1}{|x-y|} \mathrm{d}s_y = \begin{cases} 1 & \text{for } x \in \Omega \, , \\ 0 & \text{for } x \notin \Omega \cup \Gamma \, , \\ \frac{1}{2} & \text{for } x \in \Gamma \end{cases}$$

of the double layer potential of the Laplacian and the fast multipole method. This technique is compared with the ray crossing method for collision detection [17].

The fast multipole boundary element method as described in this paper is now also used as a simulation tool in engineering and industrial applications, see for example [67, 69]. One of the tested examples is the experimental setup of a lance in a Faraday cage, in the context of the simulation of spray painting. Dirichlet boundary conditions with zero potential are required on the Faraday cage. Neumann boundary conditions are imposed on the lance where the flux is zero. Only the brass shell and the metal tip of the lance have a given potential of $-60000V$. Even though the structure of the experimental setup is quite simple, the initial boundary mesh is highly adaptive with a mesh ratio $h_{\max}/h_{\min} \approx 1220$ due to the small size of the brass shell and the metal tip. The second reason for the highly adaptive mesh is the fast descent of the

potential and the related rapid drop of the electric field, which both have to be resolved by the boundary mesh.

As this adapted initial mesh still does not resolve this problem around the brass shell and the metal tip satisfactorily, we have run the simulation for this mesh and in addition for two adaptively refined meshes as shown in Fig. 1.



32006 triangles 34424 triangles 44168 triangles

Fig. 1. Adaptive refinement at the tip

The computational results are compared in Table 1. The numbers of entries of the part of the system matrix related to the nearfield grow slightly faster than theoretically expected since we have omitted an extra recompression of the nearfield in favor of a simpler implementation of the adaptive version of the fast multipole method. This also effects the numbers for setting up and solving the system of linear equations.

Table 1. Results for the experimental setup

N	# entries	Setup(s)	Solving(s)	it
32006	4435610	443	127	79
34424	5576398	571	155	75
44168	7889093	860	247	83

The adaptive refinement is based on the defect $e_h(x) = u_h(x) - \widetilde{u}(x)$ of the Cauchy data $u_h(x)$ and the evaluation $\widetilde{u}(x)$ of the representation formula at the boundary Γ:

$$\widetilde{u}(x) = Vt_h(x) + (\frac{1}{2}I - K)u_h(x) \qquad \text{for } x \in \Gamma \ .$$

This defect is reduced significantly for the adaptively refined meshes. The results of the once adaptively refined mesh is of the same quality as the results of a uniform refined mesh but for significantly lower computational costs. The second adaptive refinement has further reduced the defect. The defect $e_h(x)$ is related to the a posteriori error estimator developed in [76, 77]. In the case of a Neumann boundary value problem, for example, the exact error $u - u_h$ of the approximate solution u_h and the exact solution u solves the boundary integral equation of the second kind

$$(\frac{1}{2}I + K)(u - u_h)(x) = e_h(x) \qquad \text{for almost all } x \in \Gamma \ .$$

A first simple error estimator is given by [77]

$$\frac{1}{1 + c_K} \|e_h\|_D \leq \|u - u_h\|_D \leq \frac{1}{1 - c_K} \|e_h\|_D \ ,$$

with the energy norm $\|\cdot\|_D$ induced by the hypersingular operator D. The constant

$$c_K = \frac{1}{2} + \sqrt{\frac{1}{2} - c_1^V c_1^D} < 1$$

is the contraction constant of the double layer potential [84]

$$\|(\frac{1}{2}I + K)v\|_{V^{-1}} \leq c_K \|v\|_{V^{-1}} \ .$$

c_1^V and c_1^D are the ellipticity constants of the single layer potential V and the hypersingular operator D.

4 Preconditioning Techniques

For an efficient domain decomposition method, we need not only fast realizations of the local boundary integral operators but also efficient preconditioners for the local problems. Here, we consider the local Neumann boundary value problem and use the formulation $Su = f$ with the local Steklov–Poincaré operator S. The right-hand side f has to be in the image$(S) = (\ker(S))^0 := \{v \in H^{-1/2}(\Gamma) : \langle v, w\rangle_\Gamma = 0 \text{ for all } w \in \ker(S)\}$. Next, we consider the equivalent variational formulation:
Find $u \in \mathcal{H}$ such that

$$\langle Su, v\rangle_\Gamma = \langle f, v\rangle_\Gamma \qquad \text{for all } v \in \mathcal{H} \tag{13}$$

holds.

This problem is uniquely solvable [68, 81] in the subspace $\mathcal{H}$ of all functions in $H^{1/2}(\Gamma_i)$, which are orthogonal to the rigid body motions with respect to the scalar product induced by the inverse single layer potential, up to the rigid body motions given by $\mathcal{R} = \text{span}\,\{1\}$ for the Laplacian and by

$$\mathcal{R} = \text{span}\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -x_3 \\ x_2 \end{pmatrix}, \begin{pmatrix} x_3 \\ 0 \\ -x_1 \end{pmatrix} \right\}$$

for linear elastostatics. These rigid body motions are denoted by v_k. For the resulting system of linear equations, we can use a conjugate gradient method. Therefore, only the application of the matrix times vector multiplication is needed and the fast multipole method can be applied. The concept of using boundary integral operators of opposite order [83] is used to construct an efficient preconditioner.

Here, we will use the single layer potential of the Laplacian for the preconditioning of the Steklov–Poincaré operator of the potential equation as well as of linear elastostatics provided $\nu < 1/2$. The blockwise application of the single layer potential of the Laplacian is much cheaper than using the single layer potential of linear elastostatics. It can be shown that the variational problem (13) is uniquely solvable in $\mathcal{H} = H_L^{1/2}(\Gamma)$ [67] with

$$H_L^{1/2}(\Gamma) = \left\{ v \in H^{1/2}(\Gamma) : \langle V_L^{-1} v, v_k \rangle_\Gamma = 0 \text{ for all } v_k \in \mathcal{R} \right\} .$$

For a simple realization, we use a modified hypersingular operator and the correspondingly stabilized variational formulation:
Find $u \in H^{1/2}(\Gamma)$ *such that*

$$\langle \widehat{S} u, v \rangle_\Gamma := \langle Su, v \rangle_\Gamma + \sum_{k=1}^{\dim \mathcal{R}} \beta_k \langle u, \widetilde{w}_k \rangle_\Gamma \langle v, \widetilde{w}_k \rangle_\Gamma = \langle f, v \rangle_\Gamma \tag{14}$$

holds for all $v \in H^{1/2}(\Gamma)$.
The parameters β_k will be chosen later appropriately. The functions $\widetilde{w}_k$ form an orthogonal basis of the set $\mathcal{R}$ of the rigid body motions with respect to the scalar product induced by the inverse single layer potential V_L^{-1} and are constructed by the Gram–Schmidt orthogonalization method. The use of the single layer potential V_L of the Laplacian makes this construction much faster, since only four inversions of the single layer potential of the Laplacian are needed instead of six inversions for the one of linear elastostatics. This stabilization technique guarantees that the solution u of the modified variational formulation is in $H_L^{1/2}(\Gamma)$. The solution of the original problem (13) is then given by

$$u + \sum_{\ell=1}^{\dim \mathcal{R}} \gamma_\ell v_\ell \, .$$

Further, we show in [67] that the single layer potential V_L of the Laplacian can be used as a preconditioner of the modified Steklov–Poincaré operator $\widehat{S}$ due to spectral equivalence inequalities

$$c_1^{V_L}\widetilde{c}_1^{D}\langle V_L^{-1}v,v\rangle_\Gamma \le \langle \widehat{S}v,v\rangle_\Gamma \le (\frac{1}{4}+c_K)c_E\langle V_L^{-1}v,v\rangle_\Gamma \quad \text{for all } v\in H^{1/2}(\Gamma)\ .$$

For these estimates the parameters β_k are chosen as

$$\beta_k = \frac{1}{4}c_E\frac{1}{\langle \widetilde{v}_k,\widetilde{w}_k\rangle_\Gamma}\ .$$

$c_1^{V_L}$ is the ellipticity constant of the single layer potential V_L of the Laplacian. $\widetilde{c}_1^{D}$ denotes the ellipticity constant of the hypersingular operator D in $H_L^{1/2}(\Gamma)$. In the case of the Laplacian we have $c_E = 1$, while in linear elastostatics c_E is given by

$$c_E = \frac{E}{1-2\nu}\frac{1-\nu}{1+\nu}\ .$$

In each application of the matrix $\widetilde{S}_h$ of the approximate Steklov–Poincaré operator $\widetilde{S}$, the inverse of the matrix V_h is needed. This can be done by a conjugate gradient method each time. Therefore, an efficient preconditioning is suitable. One technique was suggested in [80] as an artificial multilevel boundary element preconditioner. Another preconditioning technique is an algebraic multigrid method for boundary elements [52, 53] using the adaptive cross approximation method [3]. There the coarsening is based on suitable criteria on functions of the distance between the midpoints of the boundary elements. The corresponding matrices of the coarser grids are constructed by applying the corresponding restriction operators to the fine grid matrices. This coarsening strategy is not suitable for the fast multipole method. The number of conversions would stay very large on the coarser grids. As the conversions account for a big part of the total effort, this would lead to a non-optimal method. If the clusters and the coarse grid elements coincide, then an optimal method can be constructed. In the simplest case, the clusters constructed in Sect. 3 form the coarse grid elements. Due to our realization, we now consider the single layer potential of the Laplacian discretized by piecewise constant elements. The constructed algebraic multigrid preconditioner can also be applied component-wise to the single layer potential of linear elastostatics. The matrix entries of the coarser grids can be computed directly from the finer grid by summing up the entries of the sons

$$V_\ell[i,j] = \sum_{m\in \mathrm{sons}(i)}\ \sum_{n\in \mathrm{sons}(j)} V_{\ell+1}[m,n]\ . \tag{15}$$

In the case of the fast multipole method, only the nearfield entries are known. As the nearfield of a coarser grid is larger than the one of the finer grid, it is not possible to compute all nearfield entries of the coarser grid by (15). But

these unknown nearfield entries of the coarser grid can be computed from the conversions of the fast multipole method on the corresponding cluster level. First, the multipole coefficients $\widehat{M}_n^m$ of each cluster ω_j^ℓ and the corresponding coarse grid element τ_j^ℓ are computed recursively as the sums of the translations of the multipole coefficients of their sons by

$$\widehat{M}_n^m(C_j^\ell, \mathrm{P}(\omega_j^\ell)) = \sum_{\omega_i^{\ell+1} \in \mathrm{sons}(\omega_j^\ell)} \sum_{s=0}^{n} \sum_{t=-s}^{s} R_s^t(\overrightarrow{C_j^\ell C_i^{\ell+1}}) \widehat{M}_{n-s}^{m-t}(C_i^{\ell+1}, \mathrm{P}(\omega_i^{\ell+1})) \; .$$

The coefficients of the clusters ω_i^L of the finest level L are given by

$$\widehat{M}_n^m(C_i^L, \mathrm{P}(\omega_i^L)) = \sum_{\tau_k \in \omega_i^L} M_n^m(C_i^L, k) \; .$$

These coefficients are used to evaluate the converted expansions, too. Then the unknown matrix entry of the two coarse grid elements τ_j^ℓ and τ_k^ℓ is computed by

$$V_\ell[j,k] = \sum_{n=0}^{\varrho} \sum_{m=-n}^{n} \widehat{M}_n^m(C_j^\ell, \mathrm{P}(\omega_j^\ell)) \sum_{s=0}^{\varrho} \sum_{t=-s}^{s} (-1)^n \overline{S_{n+s}^{m+t}}(\overrightarrow{C_k^\ell C_j^\ell}) \widehat{M}_s^t(C_k^\ell, \mathrm{P}(\omega_k^\ell)) \; .$$

The farfield part of the matrix times vector multiplication of each level is realized by the fast multipole method in the usual way. Therefore, only the operations down to level $\ell - 1$ of the cluster tree have to be executed. The coefficients $\widehat{M}_n^m$ are again used for setting up the coefficients and for the evaluation of the local expansions.

A detailed analysis in [67] shows that the effort to construct the nearfield parts of the coarse grid matrices and the total memory requirements are of order $\mathcal{O}(N \log^2 N)$ with the number N of boundary elements. The effort of an application of the constructed algebraic multigrid preconditioner using a V-cycle is now also of order $\mathcal{O}(N \log^2 N)$. The expansion degree used ϱ in (9) for the matrix approximation on the finest grid is still sufficient to guarantee the accuracy of the coarse grid approximations.

5 Boundary Element Tearing and Interconnecting Methods

Instead of the global system of linear equations (7)

$$\widetilde{S}_h \mathbf{u} = \sum_{i=1}^{p} A_i^\top \widetilde{S}_{i,h} A_i \mathbf{u} = \sum_{i=1}^{p} A_i^\top \mathbf{f}_i + \sum_{i=1}^{p} A_i^\top \mathbf{f}_{N,i} \; ,$$

we consider the equivalent minimization problem for the functional

$$F(\mathbf{u}) = \min_{\mathbf{v}\in\mathbb{R}^M} \sum_{i=1}^{p} \left[\frac{1}{2}(\widetilde{S}_{i,h} A_i\mathbf{v}, A_i\mathbf{v}) - (\mathbf{f}_i, A_i\mathbf{v}) - (\mathbf{f}_{N,i}, A_i\mathbf{v}) \right] . \tag{16}$$

Starting from this minimization problem, one obtains the Boundary Element Tearing and Interconnecting (BETI) method in [54]. The BETI method is the counterpart of the boundary element method to the Finite Element Tearing and Interconnecting (FETI) methods [21, 22], which are widely used in engineering applications [18, 19, 20, 23, 72, 78] and are based on a well established analysis [8, 45, 47, 57, 58]. The combination of the new BETI with FETI methods is presented in [55]. The first step to derive the FETI and BETI methods is the tearing of the global trial function u into local trial functions u_i or local vectors $\mathbf{u}_i = A_i\mathbf{u}$. Instead of the global minimization problem (16), now the corresponding local minimization problems are considered. In the second step, the local vectors $\mathbf{u}_i$ are interconnected along the coupling interfaces by the constraints

$$\sum_{i=1}^{p} B_i\mathbf{u}_i = \mathbf{0} \tag{17}$$

to reinforce the continuity across the interfaces. Each constraint and each row of the matrix $B = (B_1, \ldots, B_p)$, respectively, is connected with a pair of matching nodes on the coupling interfaces. Here, for strong continuity, the entries of such a row are 1 and -1 for the indices corresponding to the matching nodes on the coupling interface $\Gamma_C = \Gamma_S \setminus \Gamma$ and 0 otherwise. In the case of linear elastostatics, these constraints have to be used for each component. Here, we use the so-called non-redundant version, which assumes that the number of constraints at some global node is equal to the number of matching subdomains minus one. For the redundant version see, e.g., [45].

Now, we have to solve p independent minimization problems subject to the constraints (17). Using Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{M_L}$ we can set up a system of linear equations

$$\begin{pmatrix} \widetilde{S}_{1,h} & & & -B_1^\top \\ & \ddots & & \vdots \\ & & \widetilde{S}_{p,h} & -B_p^\top \\ B_1 & \ldots & B_p & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_p \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_1 + \mathbf{f}_{N,1} \\ \vdots \\ \mathbf{f}_p + \mathbf{f}_{N,p} \\ \mathbf{0} \end{pmatrix} . \tag{18}$$

Subdomains with sufficient Dirichlet boundary conditions are called non-floating, otherwise floating. Non-floating subdomains are those subdomains for which the local Steklov–Poincaré operators $\widetilde{S}_{i,h}$ are invertible

$$\mathbf{u}_i = \widetilde{S}_{i,h}^{-1}(\mathbf{f}_i + \mathbf{f}_{N,i} + B_i^\top\boldsymbol{\lambda}) . \tag{19}$$

The local Steklov–Poincaré operator matrices $\widetilde{S}_{i,h}$ of the floating subdomains Ω_i, $i = 1, \ldots, q$, which lack sufficient Dirichlet conditions, are singular. According to Sect. 4 a pseudo-inverse of such a local Steklov–Poincaré operator

$\widetilde{S}_{i,h}$ is defined via the discrete version of (14). The solution of the local problem is now given by

$$\mathbf{u}_i = \widehat{S}_{i,h}^{-1}(\mathbf{f}_i + \mathbf{f}_{N,i} + B_i^\top \boldsymbol{\lambda}) + \sum_{k=1}^{\dim \mathcal{R}} \gamma_{k,i} \mathbf{v}_{k,i} \qquad \text{for } i = 1, \ldots, q \tag{20}$$

with unknown constants $\gamma_{k,i} \in \mathbb{R}$ which have to be determined from the global problem. In addition, the compatibility conditions

$$(\mathbf{f}_i + \mathbf{f}_{N,i} + B_i^\top \boldsymbol{\lambda}, \mathbf{v}_{k,i}) = 0 \qquad \text{for all } k = 1, \ldots, \dim \mathcal{R} \tag{21}$$

have to be fulfilled. Substituting the local solutions in the last line of the system of linear equations (18) leads to the modified BETI system

$$\begin{aligned} F\boldsymbol{\lambda} + G\boldsymbol{\gamma} &= \sum_{i=1}^{q} B_i \widehat{S}_{i,h}^{-1} B_i^\top \boldsymbol{\lambda} + \sum_{i=q}^{p} B_i \widetilde{S}_{i,h}^{-1} B_i^\top \boldsymbol{\lambda} + G\boldsymbol{\gamma} \\ &= -\sum_{i=1}^{p} B_i \widehat{S}_{i,h}^{-1}(\mathbf{f}_i + \mathbf{f}_{N,i}) - \sum_{i=1}^{p} B_i \widetilde{S}_{i,h}^{-1}(\mathbf{f}_i + \mathbf{f}_{N,i}) = \mathbf{d} \,, \end{aligned} \tag{22}$$

where $G = (B_1 \mathbf{v}_{1,1}, \ldots, B_1 \mathbf{v}_{\dim \mathcal{R},1}, \ldots, B_q \mathbf{v}_{\dim \mathcal{R},q})$. The compatibility conditions (21) can be rewritten as

$$G^\top \boldsymbol{\lambda} = -\left((\mathbf{f}_i + \mathbf{f}_{N,i}, \mathbf{v}_{k,i})\right)_{k=1:\dim \mathcal{R}, i=1:q} =: \mathbf{e} \,. \tag{23}$$

The projection

$$P = I - QG(G^\top QG)^{-1} G^\top \tag{24}$$

separates the computation of the Lagrange multipliers $\boldsymbol{\lambda}$ and of the constants $\boldsymbol{\gamma}$, since $P^\top G\boldsymbol{\gamma} = 0$ holds; Q is a suitable diagonal scaling matrix [8, 45]. First, the solution of the system of linear equations

$$P^\top F\boldsymbol{\lambda} = P^\top \mathbf{d}$$

is determined by a conjugate gradient method. If the initial guess in the conjugate gradient method is chosen as

$$\boldsymbol{\lambda}_0 = QG(G^\top QG)^{-1}\mathbf{e}$$

then the compatibility condition (23) will be guaranteed. Next, the constants $\boldsymbol{\gamma}$ are computed as

$$\boldsymbol{\gamma} = (G^\top QG)^{-1} G^\top Q(\mathbf{d} - F\boldsymbol{\lambda}) \,.$$

The local solutions $\mathbf{u}_i$ are given by (19) and (20), respectively. In the conjugate gradient method, a scaled hypersingular BETI preconditioner

$$C_{\mathrm{BETI}}^{-1} = (BC_\alpha^{-1}B^\top)^{-1} BC_\alpha^{-1} D_h C_\alpha^{-1} B^\top (BC_\alpha^{-1}B^\top)^{-1} \tag{25}$$

is used with suitable diagonal matrices C_α [8, 45]. Following [54], the condition number of this system can be estimated by

$$\kappa(C_{\mathrm{BETI}}^{-1}F) \le c\left(1+\log\frac{H}{h}\right)^2$$

with a constant c independent of h, H, p and the jumps in the coefficients α_i for the potential equation. In linear elastostatics this constant is independent of the values $E_i/(1+\nu_i)$. Here, H denotes the maximum of the local diameters $H_i := \operatorname{diam}\Omega_i$ of the subdomains.

A drawback of the BETI formulation (22) is that we have to distinguish between floating and non-floating subdomains. This might get rather complicated in linear elastostatics since the number of local rigid body motions may vary from subdomain to subdomain. One possibility to overcome this difficulty is the dual primal formulation of the FETI method called FETI-DP [18, 19, 46]. In this formulation, some global nodes and the corresponding primal variables are reintroduced to guarantee the invertibility of the local Steklov–Poincaré operators. The choice of these global variables is important for the robustness and the efficiency of the method. Again, this approach seems to be rather involved for linear elastostatics [46].

Alternatively, we present an all-floating formulation [67] of the BETI method. The difference to the standard formulation is that the function u of the Dirichlet data is not split into the known part g_D and the unknown part $\widehat{u}$, but is considered as unknown on the complete boundary. Then we use some additional constraints to guarantee that the approximations $u_{i,h}$ satisfy the Dirichlet boundary conditions. The constraints $u_{i,k} = g(x_k)$ are added to the BETI constraints of the coupling interfaces. This defines new constraint matrices $\widetilde{B}_i$ as extensions of the already defined matrices B_i. The resulting system of linear equations now reads

$$\begin{pmatrix} \widetilde{S}_{1,h} & & & -\widetilde{B}_1^\top \\ & \ddots & & \vdots \\ & & \widetilde{S}_{p,h} & -\widetilde{B}_p^\top \\ \widetilde{B}_1 & \dots & \widetilde{B}_p & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_p \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_{N,1} \\ \vdots \\ \mathbf{f}_{N,p} \\ \mathbf{g} \end{pmatrix} \tag{26}$$

for the all-floating formulation. Now, all subdomains are floating and the local Steklov–Poincaré operators are inverted by (20). The system (22) is reduced to

$$\sum_{i=1}^{p} B_i \widehat{S}_{i,h}^{-1} B_i^\top \boldsymbol{\lambda} + G\boldsymbol{\gamma} = \mathbf{g} - \sum_{i=1}^{p} B_i \widehat{S}_{i,h}^{-1} \mathbf{f}_{N,i}\,. \tag{27}$$

This system can be solved as described for the system (22). The number of unknowns is increased, but the treatment of the subdomains is unified and therefore simplified. Another advantage of the all-floating formulation is that the preconditioning of the local Steklov–Poincaré operators by the

single layer potential is optimal for floating subdomains, whereas in the case of non-floating subdomains, the condition number of the inversion of the local Steklov–Poincaré operators is growing logarithmically [60]. Therefore the all-floating formulation shows a better asymptotic complexity.

Both formulations of the BETI method can be solved as Schur complement systems and as saddle point problems. To avoid the inner inversion of the matrix of the single layer potential, we introduce the local Neumann data

$$\mathbf{t}_i := V_{i,h}^{-1}(\frac{1}{2}M_{i,h} + K_{i,h})\mathbf{u}_i$$

and consider the twofold saddle point problems

$$\begin{pmatrix} V_{1,h} & & & -\widehat{K}_{1,h} & & & \\ & \ddots & & & \ddots & & \\ & & V_{p,h} & & & -\widehat{K}_{p,h} & \\ \widehat{K}_{1,h}^\top & & & D_{1,h} & & & -B_1^\top \\ & \ddots & & & \ddots & & \vdots \\ & & \widehat{K}_{p,h}^\top & & & D_{p,h} & -B_p^\top \\ & & & B_1 & \dots & B_p & 0 \end{pmatrix} \begin{pmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_p \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_p \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{f}_1 + \mathbf{f}_{N,1} \\ \vdots \\ \mathbf{f}_p + \mathbf{f}_{N,p} \\ \mathbf{0} \end{pmatrix} \tag{28}$$

and

$$\begin{pmatrix} V_{1,h} & & & -\widehat{K}_{1,h} & & & \\ & \ddots & & & \ddots & & \\ & & V_{p,h} & & & -\widehat{K}_{p,h} & \\ \widehat{K}_{1,h}^\top & & & \widehat{D}_{1,h} & & & -\widetilde{B}_1^\top \\ & \ddots & & & \ddots & & \vdots \\ & & \widehat{K}_{p,h}^\top & & & \widehat{D}_{p,h} & -\widetilde{B}_p^\top \\ & & & \widetilde{B}_1 & \dots & \widetilde{B}_p & 0 \end{pmatrix} \begin{pmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_p \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_p \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{f}_{N,1} \\ \vdots \\ \mathbf{f}_{N,p} \\ \mathbf{g} \end{pmatrix}, \tag{29}$$

where $\widehat{K}_{i,h} := \frac{1}{2}M_{i,h} + K_{i,h}$. All systems are solved by a parallel, preconditioned conjugate gradient method using the projection P (24). The saddle point problems are transformed into symmetric and positive definite systems by suitable transformation [5, 50, 51, 91]. All matrix times vector products are executed by the fast multipole method. As preconditioners, we use the algebraic multigrid method for the local single layer potentials, the technique of boundary integral operators of opposite orders for the local Steklov–Poincaré operators and the BETI preconditioner (25). If we assume that the algebraic multigrid preconditioner is optimal, we can estimate the total efforts for solving these twofold saddle point problems. The inversion of the standard BETI formulation (29) needs less than $\mathcal{O}((1+\log(H/h))^2)$ iteration steps and $\mathcal{O}((H/h)^2(1+\log(H/h))^4)$ arithmetic operations. The all-floating formulation (28) needs only $\mathcal{O}(1+\log(H/h))$ iteration steps and $\mathcal{O}((H/h)^2(1+\log(H/h))^3)$ arithmetic operations [67].

6 Computational Examples

For testing our domain decomposition algorithms we use a three-dimensional unit cube that we subdivide into eight similar cubes or subdomains. This domain decomposition is still simple to be created by hand but already contains a lot of cross points. We choose a mixed boundary value problem in linear elastostatics as test example. The lower and upper side of the unit cube are the parts of the boundary with Dirichlet boundary conditions. Neumann boundary conditions are imposed on the rest of the boundary. Table 2 shows the numbers of unknowns and levels of subdivisions.

Table 2. Number of unknowns of several formulations

	Schur-CG	B-Schur	B-SPP	B-SPP2	A-Schur	A-SPP	A-SPP2
L	(7)	(22)	(18)	(28)	(27)	(26)	(29)
0	111	105	321	897	225	561	1137
1	537	351	1239	3543	663	1863	4167
2	2397	1275	4947	14163	2259	6915	16131
3	10149	4851	19851	56715	8331	26811	63675
4	41781	18915	79611	227067	31995	105771	253227
5	169557	74691	318939	908763	125403	420363	1010187

First, we compare the three systems of the standard BETI method, see Table 3. The numbers of iterations in brackets are the numbers of the inner inversions in each of the outer iteration steps. The numbers of iterations of the inversion of the BETI system and of the inner inversion of the local Steklov–Poincaré operators grow logarithmically as expected from theory. The numbers of iterations are increased when changing to the saddle point problem (18) and once again for the twofold saddle point problem (28), since the numbers of unknowns are increasing. The computational times are reduced when skipping to the saddle point problems since now inner inversions are avoided. The fastest formulation is the twofold saddle point problem (28).

Table 3. Results of the BETI method for the mixed problem

	BETI-Schur (22)		BETI-SPP (18)		BETI-SPP2 (28)	
L	t_2	It.	t_2	It.	t_2	It.
0	20	15(21(12))	11	63(11)	4	77
1	176	22(36(15))	38	76(15)	17	97
2	1916	29(44(16))	240	86(16)	97	106
3	12643	34(51(17))	1228	90(16)	496	116
4	101357	38(56(16))	8056	87(16)	3699	131
5			72276	99(17)	27531	150

In Table 4, the results of the all-floating formulation are given. The numbers of iterations are less than those of the standard BETI method in Table 3. The asymptotic bound of the numbers of iterations for the inversion of the local Steklov–Poincaré operators is almost attained, but there is still a small growth. In the case of the Laplacian this bound is already reached on these refinement levels. It takes some more refinement steps to reach the asymptotic case in linear elastostatics, as we only use the single layer potential of the Laplacian for preconditioning. Again the twofold saddle point problem (29) is the fastest version with less than one third of the computational time for solving the system compared with the saddle point problem (26).

Table 4. Results of the all-floating formulation for the mixed problem

	Allfl.-Schur (27)		Allfl.-SPP (26)		Allfl.-SPP2 (29)	
L	t_2	It.	t_2	It.	t_2	It.
0	22	18(17(11))	10	53(11)	4	63
1	145	19(30(15))	34	70(15)	13	81
2	1291	23(35(16))	191	69(16)	76	80
3	8383	28(39(17))	984	72(16)	358	83
4	70985	33(42(16))	7544	81(16)	2673	93
5			64093	87(17)	18935	102

We compare the computational times of the two twofold saddle point problems with those of the primal domain decomposition method (7) in Table 5. The numbers of iterations of the all-floating formulation are lower than those of the standard BETI formulation. Therefore the times t_2 for solving the system of linear equations are less. Even the times t_1 for setting up the systems of linear equations are less for the finer grids. Both BETI versions are faster than the primal domain decomposition method (7). The all-floating formulation is about twice as fast.

Table 5. Comparison of the methods for the mixed problem

	Schur-CG (7)			BETI-SPP2 (28)			Allfl.-SPP2 (29)		
L	t_1	t_2	It.	t_1	t_2	It.	t_1	t_2	It.
0	2	3	25(12)	2	4	77	3	4	63
1	4	14	60(15)	5	17	97	6	13	81
2	11	99	74(16)	11	97	106	12	76	80
3	30	575	85(16)	29	496	116	26	358	83
4	193	4128	92(16)	193	3699	131	157	2673	93
5	1353	37146	101(17)	1360	27531	150	1045	18935	102

As second example, we consider the domain decomposition of Fig. 2 with jumping coefficients. Two subdomains of steel are surrounded by 16 subdomains of concrete. The bottom side is fixed while a displacement is imposed to the upper side. The rest of the boundary has vanishing Neumann data.

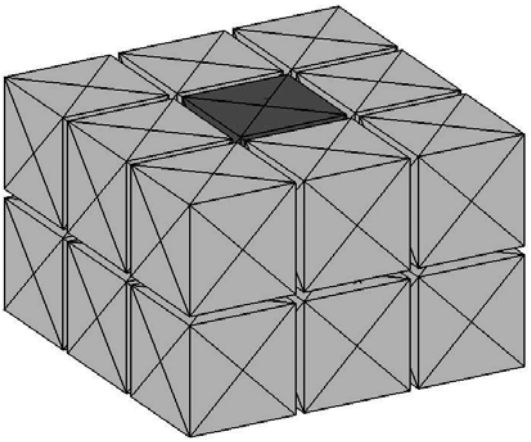


Fig. 2. Domain decomposition of a piece of steel in concrete

Each cube is refined up to 24576 boundary elements in the global refinement. A uniform extension of this mesh into the domain would create a finite element mesh consisting of more than 14 millions tetrahedrons. In Table 6, we compare the twofold saddle point problems (28) and (29) of the two BETI versions with the primal domain decomposition method (7). The standard BETI method is about twice as fast as the primal method. The all-floating formulation is even three times faster. The numbers show the superior asymptotic behavior of the BETI-method and especially of the all-floating formulation.

Table 6. Comparison of the methods for jumping coefficients

	Schur-CG (7)			BETI-SPP2 (28)			Allfl.-SPP2 (29)		
L	t_1	t_2	It.	t_1	t_2	It.	t_1	t_2	It.
0	5	7	53(10)	5	7	78	4	8	65
1	7	25	110(14)	7	19	100	7	19	82
2	13	181	130(14)	15	112	114	15	115	85
3	30	986	148(14)	30	562	129	27	476	95
4	191	6902	154(14)	189	4352	153	159	3119	105
5	1332	59264	166(16)	1334	31645	172	1060	23008	120

7 Conclusions

In this contribution we presented efficient parallel solvers for the Laplacian and for the Lamé system of linear elastostatics with jumping coefficients based on fast multipole Galerkin boundary element methods. This approach is used for solving and simulating practical problems from engineering and industry involving complicated structures and strongly adaptive boundary element meshes. In the same way, one may handle exterior boundary value problems for the Helmholtz equation, see Research Project B5, where one also has to take care of the spurious modes. Moreover, the efficient evaluation of the double layer potential can be used in collision detection, see Research Project B7.

The coupling of the presented boundary element domain decomposition methods with finite elements seems to be straightforward. Hence, the given solution strategies can be transfered to coupled finite and boundary element tearing and interconnecting domain decomposition methods for solving more general boundary value problems with jumping coefficients where in the finite element domain also nonlinearities may be treated. In general, those methods can also be used to handle complex coupled problems with different physical fields. Then, also non-matching grids need to be considered.

Acknowledgment

This research was supported by the German Research Foundation (DFG) within the Collaborative Research Center (SFB) 404 "Multifield Problems in Solid and Fluid Mechanics", subproject C10.

References

1. M. Bebendorf. *Effiziente numerische Lösung von Randintegralgleichungen unter Verwendung von Niedrigrang-Matrizen.* Doctoral thesis, Universität des Saarlandes, 2001.
2. M. Bebendorf. Hierarchical LU decomposition-based preconditioners for BEM. *Computing*, 74(3):225–247, 2005.
3. M. Bebendorf and S. Rjasanow. Adaptive low-rank approximation of collocation matrices. *Computing*, 70(1):1–24, 2003.
4. S. Börm. Approximation of integral operators by $\mathcal{H}^2$-matrices with adaptive bases. *Computing*, 74(3):249–271, 2005.
5. J. H. Bramble and J. E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comput.*, 51(183):387–388, 1988.
6. J. H. Bramble, J. E. Pasciak, and A. H. Schatz. The construction of preconditioners for elliptic problems by substructuring. I. *Math. Comput.*, 47:103–134, 1986.

7. J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comput.*, 55(191):1–22, 1990.
8. S. C. Brenner. An additive Schwarz preconditioner for the FETI method. *Numer. Math.*, 94(1):1–31, 2003.
9. J. Breuer. *Schnelle Randelementmethoden zur Simulation von elektrischn Wirbelstromfeldern sowie ihrer Wärmeproduktion und Kühlung.* Doctoral thesis, Universität Stuttgart, 2005.
10. J. Breuer, W. L. Wendland, O. Steinbach, and Z. Andjelic. A fast BEM formulation for high energy devices – simulation of eddy current fields, heat production and cooling. *Comp. Vis. Sci.*, submitted.
11. F. Brezzi and C. Johnson. On the coupling of boundary integral and finite element methods. *Calcolo*, 16:189–201, 1979.
12. F. Brezzi, C. Johnson, and J. C. Nedelec. On the coupling of boundary integral and finite element methods. In *Basic Problems of Numerical Mathematics, 4th Symp., Plezn/CSSR 1978*, pages 103–114. Charles Univ., 1978.
13. C. Carstensen, M. Kuhn, and U. Langer. Fast parallel solvers for symmetric boundary element domain decomposition equations. *Numer. Math.*, 79(3):321–347, 1998.
14. H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *J. Comput. Phys.*, 155(2):468–498, 1999.
15. M. Costabel. Symmetric methods for the coupling of finite elements and boundary elements. In *Boundary Elements IX, Vol. 1*, pages 411–420. Comput. Mech., Southampton, 1987.
16. W. Dahmen, S. Prößdorf, and R. Schneider. Wavelet approximation methods for pseudodifferential equations. II: Matrix compression and fast solution. *Adv. Comput. Math.*, 1(3–4):259–335, 1993.
17. P. Eberhard, B. Muth, G. Of, and O. Steinbach. Collision detection for complicated polyhedra using the fast multipole method or ray crossing. In preparation.
18. C. Farhat, M. Lesoinne, P. LeTallec, K. Pierson, and D. Rixen. FETI-DP: A dual-primal unified FETI method. I: A faster alternative to the two-level FETI method. *Int. J. Numer. Methods Eng.*, 50(7):1523–1544, 2001.
19. C. Farhat, M. Lesoinne, and K. Pierson. A scalable dual-primal domain decomposition method. *Numer. Linear Algebra Appl.*, 7(7–8):687–714, 2000.
20. C. Farhat, A. Macedo, and M. Lesoinne. A two-level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems. *Numer. Math.*, 85(2):283–308, 2000.
21. C. Farhat and F.-X. Roux. A method of finite element tearing and interconnecting and its parallel solution algorithm. *Int. J. Numer. Methods Eng.*, 32(6):1205–1227, 1991.
22. C. Farhat and F.-X. Roux. Implicit parallel processing in structural mechanics. *Comput. Mech. Adv.*, 2(1):1–124, 1994.
23. Y. Fragakis and M. Papadrakakis. A unified framework for formulating domain decomposition methods in structural mechanics. Technical report, National Technical University, Athens, 2002.
24. Y. Fu, K. J. Klimkowski, G. J. Rodin, E. Berger, J. C. Browne, J. K. Singer, R. A. van de Geijn, and K. S. Vemaganti. A fast solution method for three-dimensional many-particle problems of linear elasticity. *Int. J. Numer. Methods Eng.*, 42(7):1215–1229, 1998.
25. S. A. Funken and E. P. Stephan. The BPX preconditioner for the single layer potential operator. *Appl. Anal.*, 67(3–4):327–340, 1997.

26. L. Grasedyck. *Theorie und Anwendungen Hierarchischer Matrizen*. Doctoral thesis, Universität Kiel, 2001.
27. L. Greengard. *The Rapid Evaluation of Potential Fields in Particle Systems*. The MIT Press, Cambridge, MA, 1987.
28. L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73:325–348, 1987.
29. W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices. I: Introduction to $\mathcal{H}$-matrices. *Computing*, 62(2):89–108, 1999.
30. W. Hackbusch, B. Khoromskij, and S. A. Sauter. On $\mathcal{H}^2$-matrices. In H.-J. Bungartz et al., editor, *Lectures on Applied Mathematics. Proceedings of the Symposium organized by the Sonderforschungsbereich 438 on the Occasion of Karl-Heinz Hoffmann's 60th Birthday, Munich, Germany, June 30-July 1, 1999*, pages 9–29. Berlin: Springer, 2000.
31. W. Hackbusch and Z. P. Nowak. On the fast matrix multiplication in the boundary element method by panel clustering. *Numer. Math.*, 54(4):463–491, 1989.
32. H. Han. The boundary integro-differential equations of three-dimensional Neumann problem in linear elasticity. *Numer. Math.*, 68(2):269–281, 1994.
33. H. Harbrecht. *Wavelet Galerkin Schemes for the Boundary Element Method in Three Dimensions*. Doctoral thesis, Technische Universität Chemnitz, 2001.
34. K. Hayami and S. A. Sauter. A panel clustering method for 3-D elastostatics using spherical harmonics. In B. Bertram et al., editor, *Integral Methods in Science and Engineering. Proceedings of the 5th International Conference, IMSE'98*, volume 418 of *Chapman Hall/CRC Res. Notes Math.*, pages 179–184. Boca Raton, FL: Chapman & Hall/CRC, 2000.
35. N. Heuer, E. P. Stephan, and T. Tran. Multilevel additive Schwarz method for the *h-p* version of the Galerkin boundary element method. *Math. Comput.*, 67(222):501–518, 1998.
36. G. Hsiao, B. Khoromskij, and W. Wendland. Preconditioning for boundary element methods in domain decomposition. *Eng. Anal. Bound. Elem.*, 25(4–5):323–338, 2001.
37. G. C. Hsiao, E. Schnack, and W. L. Wendland. A hybrid coupled finite-boundary element method in elasticity. *Comput. Methods Appl. Mech. Eng.*, 173(3–4):287–316, 1999.
38. G. C. Hsiao, E. Schnack, and W. L. Wendland. Hybrid coupled finite-boundary element methods for elliptic systems of second order. *Comput. Methods Appl. Mech. Eng.*, 190(5–7):431–485, 2000.
39. G. C. Hsiao, O. Steinbach, and W. L. Wendland. Domain decomposition methods via boundary integral equations. *J. Comput. Appl. Math.*, 125(1–2):521–537, 2000.
40. G. C. Hsiao and W. L. Wendland. Domain decomposition in boundary element methods. In *Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, Proc. Symp., Moscow/Russ. 1990* , pages 41–49. SIAM, Philadelphia, PA, 1991.
41. G. C. Hsiao and W. L. Wendland. Domain decomposition via boundary element methods. In H. Alder et al., editor, *Numerical Methods in Engineering and Applied Sciences, Part I*, pages 198–207. CIMNE, Barcelona, 1992.
42. C. Johnson and J. C. Nedelec. On the coupling of boundary integral and finite element methods. *Math. Comput.*, 35:1063–1079, 1980.

43. B. N. Khoromskij and W. L. Wendland. Spectrally equivalent preconditioners for boundary equations in substructuring techniques. *East-West J. Numer. Math.*, 1(1):1–26, 1993.
44. B. N. Khoromskij and G. Wittum. *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface.* Lecture Notes in Computational Science and Engineering 36. Berlin: Springer, 2004.
45. A. Klawonn and O. B. Widlund. FETI and Neumann-Neumann iterative substructuring methods: Connections and new results. *Commun. Pure Appl. Math.*, 54(1):57–90, 2001.
46. A. Klawonn and O. B. Widlund. Dual-Primal FETI methods for linear elasticity. Technical Report TR2004-855, Courant Institute of Mathematical Sciences, Department of Computer Science, September 2004.
47. A. Klawonn, O. B. Widlund, and M. Dryja. Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients. *SIAM J. Numer. Anal.*, 40(1):159–179, 2002.
48. V. D. Kupradze, T. G. Gegelia, M. O. Baseleisvili, and T. V. Burculadze. *Three-dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity.* North-Holland Series in applied Mathematics and Mechanics. Vol. 25. Amsterdam, New York, Oxford: North-Holland Publishing Company, 1979.
49. U. Langer. Parallel iterative solution of symmetric coupled FE/BE-equations via domain decomposition. In A. Quarteroni et al., editor, *Domain Decomposition Methods in Science and Engineering. The sixth International Conference on Domain Decomposition, Como, Italy, June 15–19, 1992*, pages 335–344. Providence, RI: American Mathematical Society. Contemp. Math., 1994.
50. U. Langer, G. Of, O. Steinbach, and W. Zulehner. Inexact data-sparse boundary element tearing and interconnecting methods. Technical Report 2005/06, SFB 404 Mehrfeldprobleme in der Kontinuumsmechanik, Universität Stuttgart, 2005.
51. U. Langer, G. Of, O. Steinbach, and W. Zulehner. Inexact Fast Multipole Boundary Element Tearing and Interconnecting Methods. In *Proceedings of the 16th International Conference on Domain Decomposition, New York, USA, January 12–15, 2005. Lecture Notes in Computational Science and Engineering.* Berlin: Springer, 2005. published online.
52. U. Langer and D. Pusch. Data-sparse algebraic multigrid methods for large scale boundary element equations. *Appl. Numer. Math.*, 54(3–4):406–424, 2005.
53. U. Langer, D. Pusch, and S. Reitzinger. Efficient preconditioners for boundary element matrices based on grey-box algebraic multigrid methods. *Int. J. Numer. Methods Eng.*, 58(13):1937–1953, 2003.
54. U. Langer and O. Steinbach. Boundary element tearing and interconnecting methods. *Computing*, 71(3):205–228, 2003.
55. U. Langer and O. Steinbach. Coupled boundary and finite element tearing and interconnecting methods. In R. Kornhuber et al., editor, *Domain Decomposition Methods in Science and Engineering. Selected Papers of the 15th International Conference on Domain Decomposition, Berlin, Germany, July 21–25, 2003. Lecture Notes in Computational Science and Engineering 40* , pages 83–97. Berlin: Springer, 2005.
56. M. Maischak, E. P. Stephan, and T. Tran. Multiplicative Schwarz algorithms for the Galerkin boundary element method. *SIAM J. Numer. Anal.*, 38(4):1243–1268, 2000.
57. J. Mandel and R. Tezaur. Convergence of a substructuring method with Lagrange multipliers. *Numer. Math.*, 73(4):473–487, 1996.

58. J. Mandel and R. Tezaur. On the convergence of a dual-primal substructuring method. *Numer. Math.*, 88(3):543–558, 2001.
59. A. Maue. Zur Formulierung eines allgemeinen Beugungsproblems durch eine Integralgleichung. *Z. Phys.*, 126:601–618, 1949.
60. W. McLean and O. Steinbach. Boundary element preconditioners for a hypersingular integral equation on an interval. *Adv. Comput. Math.*, 11(4):271–286, 1999.
61. K. Nabors, F. T. Korsmeyer, F. T. Leighton, and J. White. Preconditioned, adaptive, multipole-accelerated iterative methods for three-dimensional first-kind integral equations of potential theory. *SIAM J. Sci. Comput.*, 15(3):713–735, 1994.
62. J. Nečas. *Les Méthodes Directes en Théorie des Équations Elliptiques*. Masson et Cie, Éditeurs, Paris, 1967.
63. J. C. Nedelec. Integral equations with non integrable kernels. *Integral Equations Oper. Theory*, 5:562–572, 1982.
64. J. N. Newman. Distributions of sources and normal dipoles over a quadrilateral panel. *Journal of Engineering Mathematics*, 20:113–126, 1986.
65. N. Nishimura. Fast multipole accelerated boundary integral equation methods. *Applied Mechanics Reviews*, 55(4):299–324, 2002.
66. G. Of. Die Multipolmethode für Randintegralgleichungen. Diploma thesis, Universität Stuttgart, 2001.
67. G. Of. *BETI-Gebietszerlegungsmethoden mit schnellen Randelementverfahren und Anwendungen*. Doctoral thesis, Universität Stuttgart, 2006.
68. G. Of and O. Steinbach. A fast multipole boundary element method for a modified hypersingular boundary integral equation. In W. Wendland et al., editor, *Analysis and Simulation of Multifield Problems. Selected papers of the International Conference on Multifield Problems, Stuttgart, Germany, April 8–10, 2002. Lect. Notes Appl. Comput. Mech. 12* , pages 163–169. Berlin: Springer, 2003.
69. G. Of, O. Steinbach, and W. Wendland. The fast multipole method for the symmetric boundary integral formulation. *IMA J. Numer. Anal.*, 26(2):272–296, 2006.
70. G. Of, O. Steinbach, and W. L. Wendland. Applications of a fast multipole Galerkin boundary element method in linear elastostatics. *Comput. Visual. Sci.*, 8:201–209, 2005.
71. J. M. Perez-Jorda and W. Yang. A concise redefinition of the solid spherical harmonics and its use in the fast multipole methods. *J. Chem. Phys.*, 104(20):8003–8006, 1996.
72. K. H. Pierson, P. Raghaven, and G. M. Reese. Experiences with FETI-DP in a production level finite element application. In I. Herrera, D. E. Keyes, O. Widlund, and R. Yates, editors, *Proceedings of the 14th International Conference on Domain Decomposition Methods*, pages 233–240. DDM.org, 2003.
73. V. Popov and H. Power. An $O(N)$ Taylor series multipole boundary element method for three-dimensional elasticity problems. *Eng. Anal. Bound. Elem.*, 25(1):7–18, 2001.
74. V. Rokhlin. Rapid solution of integral equations of classical potential theory. *J. Comput. Phys.*, 60:187–207, 1985.
75. R. Schneider. *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur effizienten Lösung großer vollbesetzter Gleichungssysteme*. Advances in Numerical Mathematics. B. G. Teubner, Stuttgart, 1998.

76. H. Schulz and O. Steinbach. A new a posteriori error estimator in adaptive direct boundary element methods: The Dirichlet problem. *Calcolo*, 37(2):79–96, 2000.
77. H. Schulz and O. Steinbach. A new a posteriori error estimator in adaptive direct boundary element methods. The Neumann problem. In A.-M. Sändig et al., editor, *Multified Problems. State of the Art*, pages 201–208. Berlin: Springer, 2000.
78. D. Stefanica. A numerical study of FETI algorithms for mortar finite element methods. *SIAM J. Sci. Comput.*, 23(4):1135–1160, 2001.
79. O. Steinbach. *Gebietszerlegungsmethoden mit Randintegralgleichungen und effiziente numerische Lösungsverfahren für gemischte Randwertprobleme*. Doctoral thesis, Universität Stuttgart, 1996.
80. O. Steinbach. Artificial multilevel boundary element preconditioners. *Proc. Appl. Math. Mech.*, 3:539–542, 2003.
81. O. Steinbach. *Numerische Näherungsverfahren für elliptische Randwertprobleme. Finite Elemente und Randelemente*. B.G. Teubner, Stuttgart, Leipzig, Wiesbaden, 2003.
82. O. Steinbach. *Stability Estimates for Hybrid Coupled Domain Decomposition Methods*. Lecture Notes in Mathematics 1809. Berlin: Springer, 2003.
83. O. Steinbach and W. L. Wendland. The construction of some efficient preconditioners in the boundary element method. *Adv. Comput. Math.*, 9(1–2):191–216, 1998.
84. O. Steinbach and W. L. Wendland. On C. Neumann's method for second-order elliptic systems in domains with non-smooth boundaries. *J. Math. Anal. Appl.*, 262(2):733–748, 2001.
85. E. P. Stephan. Multilevel methods for the h-, p-, and hp-versions of the boundary element method. *J. Comput. Appl. Math.*, 125(1–2):503–519, 2000.
86. T. von Petersdorff and E. Stephan. Multigrid solvers and preconditioners for first kind integral equations. *Numer. Methods Partial Differ. Equations*, 8(5):443–450, 1992.
87. W. L. Wendland. On asymptotic error estimates for combined BEM and FEM. In *Finite Element and Boundary Element Techniques from Mathematical and Engineering Point of View*, volume 301 of *CISM Courses and Lectures*, pages 273–333. Springer-Verlag, Vienna, 1988.
88. C. A. White and M. Head-Gordon. Derivation and efficient implementation of the fast multipole method. *J. Chem. Phys.*, 101(8):6593–6605, 1994.
89. K. Yoshida. *Applications of Fast Multipole Method to Boundary Integral Equation Method*. Doctoral thesis, Kyoto University, 2001.
90. K. Yoshida, N. Nishimura, and S. Kobayashi. Application of fast multipole Galerkin boundary integral equation method to elastostatic crack problems in 3D. *Int. J. Numer. Methods Eng.*, 50(3):525–547, 2001.
91. W. Zulehner. Analysis of iterative methods for saddle point problems: A unified approach. *Math. Comput.*, 71(238):479–505, 2002.

Analytical and Numerical Methods for Finite-Strain Elastoplasticity*

Ercan Gürses[1], Andreas Mainik[2], Christian Miehe[1], and Alexander Mielke[3,4]

[1] Institute of Applied Mechanics, Faculty of Civil and Environmental Engineering, University of Stuttgart, Pfaffenwaldring 7, 70569 Stuttgart, Germany
cm@mechbau.uni-stuttgart.de
guerses@mechbau.uni-stuttgart.de

[2] Institute of Analysis, Dynamics and Modeling, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart
mainik@mathematik.uni-stuttgart.de

[3] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany
mielke@wias-berlin.de

[4] Department of Mathematics, Humboldt-Universität zu Berlin, Rudower Chaussee 25, 12489 Berlin-Adlershof, Germany

Summary. An important class of finite-strain elastoplasticity is based on the multiplicative decomposition of the strain tensor $\mathbf{F} = \mathbf{F}_{\rm el}\mathbf{F}_{\rm pl}$ and hence leads to complex geometric nonlinearities. This survey describes recent advances in the analytical treatment of time-incremental minimization problems with or without regularizing terms involving strain gradients. For a regularization controlling all of $\nabla\mathbf{F}_{\rm pl}$ we provide an existence theory for the time-continuous rate-independent evolution problem, which is based on a recently developed energetic formulation for rate-independent systems in abstract topological spaces.

In systems without gradient regularization one encounters the formation of microstructures, which can be described by sequential laminates or more general gradient Young measures. We provide a mathematical framework for the evolution of such microstructures and discuss algorithms for solving the associated space-time discretizations. In a finite-step-sized incremental setting of standard dissipative materials (also called generalized standard materials) we outline also details of relaxation-induced microstructure developments for strain softening von Mises plasticity and single-slip crystal plasticity. The numerical implementations are based on simplified assumptions concerning the complexity of the microstructures.

Keywords: Multiplicative plasticity, energetic formulation, time-incremental minimization, microstructure, energy relaxation

* Project C11 "Mathematical Models of Plasticity with Finite Deformations"

1 Introduction

We study the theory of elastoplasticity in the case of finite strains in applications such as visualized in Fig. 1, where the deformation gradient $\mathbf{F} = \nabla\boldsymbol{\varphi}$ is considered as a matrix with positive determinant. Moreover, we work under the basic assumption that the multiplicative decomposition

$$\mathbf{F} = \nabla\boldsymbol{\varphi} = \mathbf{F}_{\mathrm{el}}\,\mathbf{P} \quad \text{with } \mathbf{P} = \mathbf{F}_{\mathrm{pl}}$$

can be used to describe the elastic properties via the elastic part $\mathbf{F}_{\mathrm{el}}$ of the deformation tensor and the plastic evolution via the plastic tensor $\mathbf{P}$. In contrast to this, the additive decomposition $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_{\mathrm{el}} + \boldsymbol{\varepsilon}_{\mathrm{pl}}$ is well-established in small-strain elastoplasticity and has nice mathematical features since it can be easily combined with convexity tools. The assumption of finite strains and the multiplicative split destroy classical convexity properties and the more general notions of poly- and quasi-convexity need to be employed for the energy-storage potential

$$\psi(\mathbf{F}, \mathbf{P}, p) = \widetilde{\psi}(\mathbf{F}\,\mathbf{P}^{-1}, p)\,,$$

where p are additional hardening variables.

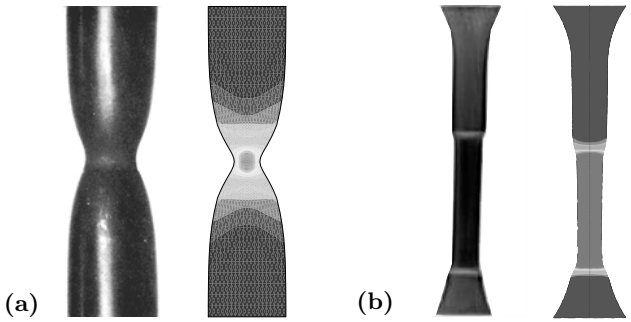


Fig. 1. Experiments and numerical simulations of finite plastic deformations. **(a)** Necking of a polycrystalline material. **(b)** Neck propagation in a tensile test of an amorphous glassy polymer

1.1 Mathematical Analysis and Existence Results

The subsequent mathematical analysis as well as the numerical implementations are based on the time-incremental minimization problems introduced in Sect. 2 which are phrased in terms of the full stored energy

$$\mathcal{E}(t, \boldsymbol{\varphi}, \mathbf{P}, p) = \int_{\mathcal{B}} \psi(\nabla\boldsymbol{\varphi}, \mathbf{P}, p) + U(\mathbf{P}, p, \nabla\mathbf{P}, \nabla p)\,\mathrm{d}x - \langle \Pi_{\mathrm{ext}}(t), \boldsymbol{\varphi}\rangle$$

and a dissipation distance

$$\mathcal{D}((\mathbf{P}_0,p_0),(\mathbf{P}_1,p_1)) = \int_{\mathcal{B}} D((\mathbf{P}_0,p_0),(\mathbf{P}_1,p_1))\,\mathrm{d}x\,.$$

For a given partition $0 = t_0 < t_1 < \cdots < t_N = T$ of the time interval $[0,T]$ the *time-incremental minimization problem* has the form

$$\textbf{(IP)} \qquad (\boldsymbol{\varphi}_j,\mathbf{P}_j,p_j) \in \operatorname*{Arg\,Min}_{\widetilde{\boldsymbol{\varphi}},\widetilde{\mathbf{P}},\widetilde{p}} \Big(\mathcal{E}(t_j,\widetilde{\boldsymbol{\varphi}},\widetilde{\mathbf{P}},\widetilde{p}) + \mathcal{D}((\mathbf{P}_{j-1},p_{j-1}),(\widetilde{\mathbf{P}},\widetilde{p}))\Big)\,.$$

In Sect. 2.2 (cf. [51]) it is shown that this incremental problem occurs naturally as the fully implicit (backward Euler) scheme for the *energetic formulation* (S) & (E), which is a weak formulation of the time-continuous problem consisting of the elastic equilibrium together with the plastic flow law, see (2.3). In Sect. 2.3 we discuss the arising nonlinearities, which are best understood when considering the matrix groups $\mathrm{GL}_+(d) = \{\,\mathbf{F}\in\mathbb{R}^{d\times d} \mid \det\mathbf{F} > 0\,\}$ and $\mathrm{SL}(d) = \{\,\mathbf{P}\in\mathbb{R}^{d\times d} \mid \det\mathbf{P} = 1\,\}$ as Lie groups.

In Sect. 3 several existence results are surveyed. In the situation without any length scale (i.e., the term U involving $\nabla\mathbf{P}$ in $\mathcal{E}$ is not present) the variables $\mathbf{P}$ and p can be minimized pointwise for each $x\in\mathcal{B}$ in the incremental problem (IP). This leads to the condensed potential

$$W^{\mathrm{cond}}((\mathbf{P}_0,p_0);\mathbf{F}) = \min\{\,\psi(\mathbf{F},\mathbf{P}_1,p_1) + D((\mathbf{P}_0,p_0),(\mathbf{P}_1,p_1)) \mid (\mathbf{P}_1,p_1)\,\}\,,$$

which plays a fundamental rôle in the existence theory in Sect. 3.1. Under the assumptions that $W^{\mathrm{cond}}((\mathbf{I},p_0);\cdot)$ is polyconvex and that it satisfies the usual coercivity assumptions, an existence theory for (IP) was derived in [54]. If polyconvexity of W^{cond} fails, then existence of solutions is not to be expected because of the formation of microstructure. In this situation the relaxation techniques of Sect. 4 have to be used to derive effective properties.

In Sect. 3.2 a regularization of (IP) is considered which involves the geometric dislocation tensor $\mathbf{G}_\mathbf{P} = (\operatorname{curl}\mathbf{P})\mathbf{P}^\top$ via the potential U in $\mathcal{E}$, namely $U(\mathbf{P},\nabla\mathbf{P}) = V(\mathbf{G}_\mathbf{P})$. In [59] it was observed that the multiplicative decomposition $\mathbf{F}_{\mathrm{el}} = \nabla\boldsymbol{\varphi}\,\mathbf{P}^{-1}$ is perfectly suited to be controlled in the sense of polyconvexity, if $\operatorname{curl}\mathbf{P}$ can be bounded by the energy. Hence, the solvability of (IP) can be proved under suitable assumptions on the dissipation distance.

Finally, in Sect. 3.3 we discuss work in progress (cf. [36]) which uses a full regularization of the internal variables $(\mathbf{P},p)$ in the energy-storage functional $\mathcal{E}$, i.e., $U(\mathbf{P},p,\nabla\mathbf{P},\nabla p) \geq c|(\nabla\mathbf{P},\nabla p)|^r$. Using the abstract theory for rate-independent systems developed in [35, 55, 18], it is possible to show first existence of solutions for (IP) and then to pass to the limit for time step going to 0. The limit function obtained along a carefully chosen subsequence can finally be identified as a solution of the original energetic formulation (S) & (E).

1.2 Modeling and Computation of Microstructures

In Sect. 4 we define material instabilities in rate-independent standard dissipative solids based on *finite-step-sized incremental energy minimization principles* as outlined in Table 1 and apply the results in Sects. 5 and 6 to the prediction of deformation microstructures in strain-softening von Mises and single-slip crystal plasticity. The formulation offers two important perspectives. First, the *definition of material stability* of standard dissipative materials is based on weak convexity conditions of incremental stress potentials in analogy to finite hyper-elasticity. Second, *microstructure developments* in unstable inelastic solids such as visualized in Fig. 2 are associated with non-convex incremental stress potentials similar to elastic phase transformation problems. These deformation microstructures can be resolved by a *relaxation of incremental energy functionals* based on a convexification of the non-convex stress potential.

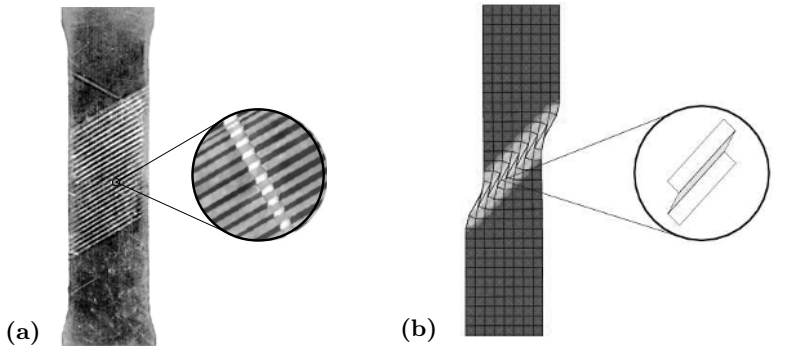


Fig. 2. Experiment and numerical simulation of microstructures in finite plastic deformations. **(a)** Experimentally observed microstructure. **(b)** Numerical simulation based on rank-one laminate microstructure

An incremental variational formulation for standard dissipative materials is outlined in the works [39, 49, 51], which generalized treatments on the deformation theory of plasticity [38] and its application to a finite-step-sized incremental setting [66, 68]. It describes the response of an inelastic material by only two scalar functionals: the energy storage function and the dissipation functional. The general set up of this generic type of material models can be traced back to the works [7, 76, 19]. It covers a broad spectrum of models in viscoelasticity, plasticity and damage mechanics. For this class of materials we define a variational formulation, where a quasi-hyperelastic stress potential at discrete time is obtained from a local *minimization problem of the constitutive response*, in Table 1 denoted by problem (C). Algorithms for a discrete setting of this constitutive minimization problem are outlined in

Table 1. Incremental Minimization Principles for Standard Dissipative Solids

(M) Constitutive Model. $\mathbf{F} \in GL_+(3)$ at $x \in \mathcal{B}$ is the local deformation and $\boldsymbol{\mathcal{I}} \in Z$ a generalized vector of internal variables. Set of local material equations has the structure

stresses $\quad \mathbb{P} = \partial_{\mathbf{F}} \psi(\mathbf{F}, \boldsymbol{\mathcal{I}})$

evolution equation $\quad \mathbf{0} \in \partial_{\boldsymbol{\mathcal{I}}} \psi(\mathbf{F}, \boldsymbol{\mathcal{I}}) + \partial_{\dot{\boldsymbol{\mathcal{I}}}} \phi(\dot{\boldsymbol{\mathcal{I}}}, \boldsymbol{\mathcal{I}})$, $\boldsymbol{\mathcal{I}}(0) = \boldsymbol{\mathcal{I}}_0$

defined in terms of an energy storage and a dissipation function ψ, ϕ.

(C) Incremental Variational Formulation of Constitutive Model. In a finite time increment $[t_n, t_{n+1}]$, the *minimization problem of the constitutive response*

stresses $\quad \mathbb{P}_{n+1} = \partial_{\mathbf{F}} W(\mathbf{F}_{n+1})$

stress potential $\quad W(\mathbf{F}_{n+1}) = \inf_{\boldsymbol{\mathcal{I}}} \int_{t_n}^{t_{n+1}} [\, \dot{\psi} + \phi \,] \, \mathrm{d}t$, $\boldsymbol{\mathcal{I}}(t_n) = \boldsymbol{\mathcal{I}}_n$

determines the current internal state $\boldsymbol{\mathcal{I}}_{n+1} \in Z$ and provides a potential for the stresses at time t_{n+1}.

(S) Stability of Incremental Constitutive Response. In $[t_n, t_{n+1}]$ the *material is locally stable if the incremental stress potential W is quasi-convex*

stable response $\quad W(\mathbf{F}_{n+1}) \leq \inf_{\mathbf{w}} \frac{1}{|D|} \int_D W(\mathbf{F}_{n+1} + \nabla \mathbf{w}(\mathbf{y})) \, \mathrm{d}x$

for all possible fluctuations $\mathbf{w}(\mathbf{y})$ on the domain D.

(R) Microstructure Development in Non-Stable Materials. For an unstable non-convex response, the *incremental minimization problem of convexification*

macro-stresses $\quad \mathbb{P}_{Qn+1} = \partial_{\mathbf{F}} W_Q(\mathbf{F}_{n+1})$

relaxation $\quad W_Q(\mathbf{F}_{n+1}) = \inf_{\mathbf{w}} \frac{1}{|D|} \int_D W(\mathbf{F}_{n+1} + \nabla \mathbf{w}(\mathbf{y})) \, \mathrm{d}x$

provides a relaxed quasi-convex hull W_Q of W and determines the current microstructure fluctuation field $\mathbf{w}(\mathbf{y})$.

the works [68, 40, 41, 47, 48] for different approaches to finite plasticity. A computational formulation of brittle fracture is considered in [43].

A key advantage of the outlined variational formulation is the opportunity to define the stability of the incremental inelastic response in terms of terminologies used in elasticity theory, see e.g. [13, 31, 10, 37, 70]. Here, a necessary condition for the existence of minimizers forces the energy functional to be sequentially weakly lower semicontinuous (s.w.l.s.). An important implication of this desired property is the quasiconvexity of the stored elastic energy function, a terminology introduced in [64]. The above outlined constitutive variational formulation enables us to extend these results to the finite-step-sized incremental response of inelasticity. To this end, we introduce an *incremental energy minimization principle for standard dissipative solids* that contains the incremental stress potential. The inelastic solid is then considered to be stable if this potential is quasiconvex, see condition (S) in Table 1. However, quasiconvexity is a global integral condition which is hard to verify in practice. More manageable is the slightly weaker rank-one convexity that is considered to be a close approximation of quasiconvexity. As presented in [45, 44, 46], classical conditions of material stability of elastic-plastic solids outlined in [73, 27, 69]

are consistent with the infinitesimal form of the rank-one convexity, i.e. the strong ellipticity or Legendre-Hadamard condition.

As pointed out in the recent papers [33, 45, 44, 46, 21] the incremental variational formulation for the inelastic response opens up the opportunity to resolve a developing microstructure in unstable standard dissipative solids by a relaxation of the associated non-convex incremental variational problem, in Table 1 denoted by problem (R). If the above outlined material stability analysis detects a non-convex incremental stress potential, an energy-minimizing deformation microstructure is assumed to develop such as indicated in Fig. 2. A relaxation is associated with a convexification of the non-convex stress potential by constructing its quasi-or rank-one convex envelope. We refer to [30, 13, 70, 65, 15] for a sound mathematical basis. The concept of relaxation has been applied to elastic phase decomposition problems in [29, 34, 9, 14, 20, 2, 32], single crystal plasticity in [66, 67, 46], strain-softening von Mises plasticity in [33, 45, 44] and damage mechanics in [21]. We comment on these results in Sects. 5 and 6.

2 Modeling of Rate-Independent Elastoplasticity

2.1 Mathematical Modeling of Standard Dissipative Materials

We consider a body with reference configuration $\mathcal{B} \subset \mathbb{R}^d$. The deformation is denoted by $\boldsymbol{\varphi} : \mathcal{B} \to \mathbb{R}^d$, and the deformation gradient $\mathbf{F} = \nabla\boldsymbol{\varphi}$ is called strain tensor. Additionally, in the sense of standard dissipative materials (also called *generalized standard materials*, cf. [77, 23, 57]) we consider a set of internal variables $\boldsymbol{\mathcal{I}} : \mathcal{B} \to Z$, where Z is a suitable finite-dimensional manifold. The theory is based on the elastic potential ψ and the dissipation potential ϕ as the underlying constitutive functions

$$\psi = \psi(\mathbf{F}, \boldsymbol{\mathcal{I}}) \text{ and } \phi = \phi(\boldsymbol{\mathcal{I}}, \dot{\boldsymbol{\mathcal{I}}}) \geq 0 \, .$$

Finite-strain elastoplasticity is based on the multiplicative decomposition $\mathbf{F} = \nabla\boldsymbol{\varphi} = \mathbf{F}_{\text{el}}\,\mathbf{F}_{\text{pl}}$ of the deformation gradient, where $\mathbf{F}_{\text{el}}$ is the elastic part of the strain tensor and $\mathbf{P} := \mathbf{F}_{\text{pl}}$ the plastic part, shortly the *plastic tensor*. The internal variable takes the form $\boldsymbol{\mathcal{I}} = (\mathbf{P}, p) \in Z$, where $p \in \mathbb{R}^m$ denotes some hardening variable. For simplicity we neglect any dependence on the material point $x \in \mathcal{B}$.

The deformation gradient $\mathbf{F}$ is best considered as an element of the Lie group $\mathrm{GL}_+(d) = \{\, \mathbf{F} \in \mathbb{R}^{d\times d} \mid \det \mathbf{F} > 0 \,\}$ and the plastic tensor $\mathbf{P}$ is usually assumed to have determinant 1, i.e. $\mathbf{P}$ is an element of the special linear group $\mathrm{SL}(d) = \{\, \mathbf{P} \in \mathbb{R}^{d\times d} \mid \det \mathbf{P} = 1 \,\}$. We will investigate the arising geometric nonlinearities in Sect. 2.3. Consequently, ϕ is defined on the tangent bundle $\mathrm{T}Z$ of the manifold Z of internal variables. The multiplicative decomposition or equivalently the axiom of plastic indifference (cf. (Sy2) on p. 359 in [51]) means

$$\psi = \psi(\mathbf{F}, \boldsymbol{\mathcal{I}}) = \psi(\mathbf{F}, \mathbf{P}, p) = \widetilde{\psi}(\mathbf{F}_{\rm el}, p) = \widetilde{\psi}(\mathbf{F}\,\mathbf{P}^{-1}, p)\ , \tag{2.1}$$

$$\phi = \phi(\boldsymbol{\mathcal{I}}, \dot{\boldsymbol{\mathcal{I}}}) = \phi(\mathbf{P}, p, \dot{\mathbf{P}}, p) = \widetilde{\phi}(p, \dot{\mathbf{P}}\,\mathbf{P}^{-1}, \dot{p})\ . \tag{2.2}$$

Rate-independency is expressed in the fact, that ϕ is homogeneous of degree 1 in the rate $\dot{\boldsymbol{\mathcal{I}}} = (\dot{\mathbf{P}}, \dot{p})$, i.e., $\phi(\boldsymbol{\mathcal{I}}, \delta\dot{\boldsymbol{\mathcal{I}}}) = \delta\phi(\boldsymbol{\mathcal{I}}, \dot{\boldsymbol{\mathcal{I}}})$ for all $\delta \geq 0$.

The local balance laws involve the conjugated forces

$$\mathbb{P} = \partial_{\mathbf{F}}\psi(\mathbf{F}, \boldsymbol{\mathcal{I}}) = \partial_{\mathbf{F}_{\rm el}}\widetilde{\psi}(\mathbf{F}_{\rm el}, \boldsymbol{\mathcal{I}})\mathbf{P}^{-\top} \text{ and } \mathcal{F} = -\partial_{\boldsymbol{\mathcal{I}}}\psi(\mathbf{F}, \boldsymbol{\mathcal{I}}) \in \mathrm{T}^*_{\boldsymbol{\mathcal{I}}}Z$$

and take the following form

$$-\mathrm{div}\mathbb{P} = f_{\rm ext} \quad \text{and} \quad 0 \in \partial_{\dot{\boldsymbol{\mathcal{I}}}}\phi(\boldsymbol{\mathcal{I}}, \dot{\boldsymbol{\mathcal{I}}}) - \mathcal{F} \quad \text{in } \mathcal{B}\ . \tag{2.3}$$

The first equation, together with suitable boundary conditions, is the elastic equilibrium equation and the second is the plastic flow law which is defined on $\mathrm{T}^*_{\boldsymbol{\mathcal{I}}}Z$.

2.2 Energetic Formulation Using Dissipation Distances

We now use the abstract theory developed in [63, 62] in the Banach space setting and in [35, 55, 18] in the fully nonlinear setting to formulate the time-continuous problem, which contains the full initial-boundary value problem of elastoplasticity. Because of the rate-independence and the strong nonconvexities we cannot expect that the rates $\dot{\boldsymbol{\mathcal{I}}}$ exist and hence we need a derivative-free formulation.

The function ϕ can be understood as an infinitesimal metric on $\boldsymbol{\mathcal{I}}$ which defines a (global) distance D, called *dissipation distance* in the sequel:

$$D(\boldsymbol{\mathcal{I}}_0,\boldsymbol{\mathcal{I}}_1) = \inf\{\, \textstyle\int_0^1 \phi(\boldsymbol{\mathcal{I}}(s),\dot{\boldsymbol{\mathcal{I}}}(s))\mathrm{d}s \mid \boldsymbol{\mathcal{I}}\in \mathrm{C}^1([0,1],Z),\, \boldsymbol{\mathcal{I}}(0)=\boldsymbol{\mathcal{I}}_0,\, \boldsymbol{\mathcal{I}}(1)=\boldsymbol{\mathcal{I}}_1 \,\}\ .$$

The definition provides immediately the triangle inequality

$$D(\boldsymbol{\mathcal{I}}_1, \boldsymbol{\mathcal{I}}_3) \leq D(\boldsymbol{\mathcal{I}}_1, \boldsymbol{\mathcal{I}}_2) + D(\boldsymbol{\mathcal{I}}_2, \boldsymbol{\mathcal{I}}_3) \text{ for all } \boldsymbol{\mathcal{I}}_1, \boldsymbol{\mathcal{I}}_2, \boldsymbol{\mathcal{I}}_3 \in Z\ . \tag{2.4}$$

The plastic indifference (2.2) provides the invariance

$$D((\mathbf{P}_0\mathbf{P}_*, p_0), (\mathbf{P}_1\mathbf{P}_*, p_1)) = D((\mathbf{P}_0, p_0), (\mathbf{P}_1, p_1)) \tag{2.5}$$

for all $\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_*$ and p_0, p_1.

For deformations $\boldsymbol{\varphi} : \mathcal{B} \mapsto \mathbb{R}^d$ and internal states $\boldsymbol{\mathcal{I}} : \mathcal{B} \mapsto Z$ we define global *energy functionals* by integration over the whole body $\mathcal{B}$ as follows

$$\begin{aligned} &\mathcal{E}(t, \boldsymbol{\varphi}, \boldsymbol{\mathcal{I}}) = \textstyle\int_{\mathcal{B}} \psi(\nabla\boldsymbol{\varphi}(x), \boldsymbol{\mathcal{I}}(x))\,\mathrm{d}x - \langle \mathit{\Pi}_{\rm ext}(t), \boldsymbol{\varphi}\rangle\ , \\ &\mathcal{D}(\boldsymbol{\mathcal{I}}_0, \boldsymbol{\mathcal{I}}_1) = \textstyle\int_{\mathcal{B}} D(\boldsymbol{\mathcal{I}}_0(x), \boldsymbol{\mathcal{I}}_1(x))\,\mathrm{d}x\ , \end{aligned} \tag{2.6}$$

where $\mathit{\Pi}_{\rm ext}(t)$ denotes the external loading depending on the process time $t \in [0, T]$. Here, $\mathcal{E}(t, \boldsymbol{\varphi}, \mathbf{P}, p)$ is the Gibbs energy at time t associated with the state

$(\boldsymbol{\varphi}, \boldsymbol{\mathcal{I}}) : \mathcal{B} \to \mathbb{R}^d \times Z$ and and is also called the *energy storage functional*. The dissipation distance $\mathcal{D}(\boldsymbol{\mathcal{I}}_0, \boldsymbol{\mathcal{I}}_1)$ is the minimal amount of dissipation occurring when the internal state $\boldsymbol{\mathcal{I}}_0$ is changed continuously into $\boldsymbol{\mathcal{I}}_1$.

A pair $(\boldsymbol{\varphi}, \boldsymbol{\mathcal{I}}) : [0,T] \times \mathcal{B} \mapsto \mathbb{R}^d \times Z$ is called an *energetic solution* for the functionals $(\mathcal{E}, \mathcal{D})$, if it satisfies for all $t \in [0,T]$ the following *stability condition* (S) and the *energy balance* (E):

(S) Stability: For all comparison states $(\widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\mathcal{I}}})$ we have

$$\mathcal{E}(t, \boldsymbol{\varphi}(t), \boldsymbol{\mathcal{I}}(t)) \le \mathcal{E}(t, \widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\mathcal{I}}}) + \mathcal{D}(\boldsymbol{\mathcal{I}}(t), \widetilde{\boldsymbol{\mathcal{I}}}) \,.$$

(E) Energy balance:

$$\mathcal{E}(t, \boldsymbol{\varphi}(t), \boldsymbol{\mathcal{I}}(t)) + \mathrm{Diss}_{\mathcal{D}}(\boldsymbol{\mathcal{I}}, [0,t]) \le \mathcal{E}(0, \boldsymbol{\varphi}(0), \boldsymbol{\mathcal{I}}(t)) - \int_0^t \langle \dot{Il}_{\mathrm{ext}}(s), \boldsymbol{\varphi}(s) \rangle \, \mathrm{d}s \,.$$

The dissipated energy $\mathrm{Diss}_{\mathcal{D}}(\boldsymbol{\mathcal{I}}, [r,s])$ along a process $\boldsymbol{\mathcal{I}} : [0,T] \times \mathcal{B} \mapsto Z$ is

$$\mathrm{Diss}_{\mathcal{D}}(\boldsymbol{\mathcal{I}}, [r,s]) = \sup\{ \textstyle\sum_1^N \mathcal{D}(\boldsymbol{\mathcal{I}}(t_{j-1}), \boldsymbol{\mathcal{I}}(t_j)) \mid r \le t_0 < t_1 < \cdots < t_N \le s \}$$

and coincides with $\int_r^s \int_{\mathcal{B}} \phi(\boldsymbol{\mathcal{I}}(x,t), \dot{\boldsymbol{\mathcal{I}}}(x,t)) \, \mathrm{d}x \, \mathrm{d}s$ for smooth processes.

The *energetic formulation* (S) & (E) characterizes the process completely and it does neither involve derivatives of $\mathbf{F} = \nabla \boldsymbol{\varphi}$ and $\boldsymbol{\mathcal{I}}$ with respect to t or x nor derivatives of the constitutive functions ψ and ϕ. It is shown in [51, 55] that the energetic formulation is consistent with the classical local balance laws (2.3), i.e., they are satisfied for any sufficiently smooth energetic solution. Moreover, in smooth and convex cases we have uniqueness of energetic solutions if a suitable initial state $(\boldsymbol{\varphi}_0, \boldsymbol{\mathcal{I}}_0)$ is specified.

The energetic formulation is intrinsically linked to the time-incremental problem, which has the major advantage that it is a minimization problem. For a given partition $0 = t_0 < t_1 \cdots < t_N = T$ of the time interval $[0,T]$ and a given initial value $(\boldsymbol{\varphi}_0, \boldsymbol{\mathcal{I}}_0)$ we define the *incremental problem*

(IP) Incremental minimization problem:
Find iteratively $(\boldsymbol{\varphi}_j, \boldsymbol{\mathcal{I}}_j)$ for $j = 1, ..., N$ such that

$$(\boldsymbol{\varphi}_j, \boldsymbol{\mathcal{I}}_j) \in \mathrm{Arg\,min}\{ \mathcal{E}(t_j, \widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\mathcal{I}}}) + \mathcal{D}(\boldsymbol{\mathcal{I}}_{j-1}, \widetilde{\boldsymbol{\mathcal{I}}}) \mid \text{ all } (\widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\mathcal{I}}}) \} \,.$$

This is a fully backward, hence fully implicit scheme which is difficult to solve numerically. Moreover, the dissipation distance $D : Z \times Z \to [0, \infty]$, which defines $\mathcal{D}$, is usually not known explicitly. Hence, the algorithms discussed in [39, 41, 49] are suitable variants of (IP).

The big advantage of (IP) is its mathematical consistency which arises from $\mathcal{D}$ satisfying the triangle inequality (2.4). Just using this and the minimization property, we obtain that every solution $(\boldsymbol{\varphi}_j, \boldsymbol{\mathcal{I}}_j)_{j=1,...,N}$ of (IP) satisfies for $j = 1, ..., N$ the following discretized versions of (S) and (E):

$$\mathcal{E}(t_j, \boldsymbol{\varphi}_j, \boldsymbol{\mathcal{I}}_j) \le \mathcal{E}(t_j, \widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\mathcal{I}}}) + \mathcal{D}(\boldsymbol{\mathcal{I}}_j, \widetilde{\boldsymbol{\mathcal{I}}}) \text{ for all } (\widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\mathcal{I}}}) \,, \tag{2.7}$$

$$\begin{aligned}&\mathcal{E}(t_j,\boldsymbol{\varphi}_j,\boldsymbol{\mathcal{I}}_j)+\mathcal{D}(\boldsymbol{\mathcal{I}}_{j-1},\boldsymbol{\mathcal{I}}_j)\\&\le \mathcal{E}(t_{j-1},\boldsymbol{\varphi}_{j-1},\boldsymbol{\mathcal{I}}_{j-1})+\textstyle\int_{t_{j-1}}^{t_j}\partial_s\mathcal{E}(s,(\boldsymbol{\varphi}_{j-1},\boldsymbol{\mathcal{I}}_{j-1}))\,\mathrm{d}s\,.\end{aligned}\tag{2.8}$$

These estimates will be crucial for the subsequent analysis.

2.3 Lie Groups and Geometric Nonlinearities

Before dealing with an existence theory for the energetic formulation we work out a little more the geometry which arises from the fact that we are dealing with finite strains and that we are using the multiplicative decomposition. In finite-strain elasticity the stored-energy density ψ should be considered as a mapping from the Lie group

$$\mathfrak{G} := \mathrm{GL}_+(d) = \{\,\mathbf{F}\in\mathbb{R}^{d\times d}\mid \det\mathbf{F}>0\,\}\,.$$

The plastic tensor $\mathbf{P}$ is assumed to lie in the Lie subgroup

$$\mathfrak{P} := \mathrm{SL}(d) = \{\,\mathbf{P}\in\mathbb{R}^{d\times d}\mid \det\mathbf{P}=1\,\}\,,$$

or even a smaller subgroup. Note that $\mathfrak{P}$ can be seen as the matrix group that maps the crystal lattice onto itself. We write $Z=\mathfrak{P}\times H$ for the manifold of internal states, where $\boldsymbol{\mathcal{I}}=(\mathbf{P},p)$ with $p\in H$.

The conjugated forces have the following properties

$$\begin{aligned}\mathbb{P}&=\partial_{\mathbf{F}}\psi(\mathbf{F},\mathbf{P},p)\in \mathrm{T}^*_{\mathbf{F}}\mathfrak{G}\,,\\ \mathbf{Q}&=-\partial_{\mathbf{P}}\psi(\mathbf{F},\mathbf{P},p)\in \mathrm{T}^*_{\mathbf{P}}\mathfrak{P}\,,\\ q&=-\partial_p\psi(\mathbf{F},\mathbf{P},p)\in \mathrm{T}^*_pH\,.\end{aligned}$$

However, using the multiplicative structure of the Lie groups it is more advantageous to use the multiplicative derivatives defined via

$$\begin{aligned}&\mathbf{K}{:}\mathbf{A}=\tfrac{\mathrm{d}}{\mathrm{d}s}\psi(\mathrm{e}^{s\mathbf{A}}\mathbf{F},\mathbf{P},p)|_{s=0}=\partial_{\mathbf{F}_{\mathrm{el}}}\widetilde{\psi}(\mathbf{F}_{\mathrm{el}},p)\mathbf{F}_{\mathrm{el}}^{\top}{:}\mathbf{A}\ \text{for}\ \mathbf{A}\in\mathrm{T}_{\mathbf{I}}\mathfrak{G}\,,\\&\mathbf{M}{:}\mathbf{B}=-\tfrac{\mathrm{d}}{\mathrm{d}s}\psi(\mathbf{F},\mathrm{e}^{s\mathbf{B}}\mathbf{P},p)|_{s=0}=\mathbf{F}_{\mathrm{el}}^{\top}\partial_{\mathbf{F}_{\mathrm{el}}}\widetilde{\psi}(\mathbf{F}_{\mathrm{el}},p)\mathbf{P}^{-\top}{:}\mathbf{B}\ \text{for}\ \mathbf{B}\in\mathrm{T}_{\mathbf{I}}\mathfrak{P}\,.\end{aligned}$$

Hence, we find stress tensors in the dual Lie algebras $\mathfrak{g}$ and $\mathfrak{p}$:

$$\mathbf{K}=\partial_{\mathbf{F}_{\mathrm{el}}}\widetilde{\psi}(\mathbf{F}_{\mathrm{el}},p)\mathbf{F}_{\mathrm{el}}^{\top}\in\mathfrak{g}:=\mathrm{T}^*_{\mathbf{I}}\mathfrak{G}\ \text{ and }\ \mathbf{M}=\mathbf{F}_{\mathrm{el}}^{\top}\partial_{\mathbf{F}_{\mathrm{el}}}\widetilde{\psi}(\mathbf{F}_{\mathrm{el}},p)\in\mathfrak{p}:=\mathrm{T}^*_{\mathbf{I}}\mathfrak{P}\,.$$

The tensors are known as the Kirchhoff stress tensor $\mathbf{K}=\mathbb{P}\mathbf{F}^T$ and the Mandel stress tensor $\mathbf{M}=\mathbf{Q}\mathbf{P}^{\top}$.

The first fact about these tensors is that we obtain another insight into the flow law $0\in\partial_{\dot{\boldsymbol{\mathcal{I}}}}\phi(\boldsymbol{\mathcal{I}};\dot{\boldsymbol{\mathcal{I}}})-\mathcal{F}$ which is a differential inclusion on $\mathrm{T}^*_{\boldsymbol{\mathcal{I}}}Z$. Using the plastic invariance of ϕ we define the elastic domain at $\mathbf{P}=\mathbf{I}$ via

$$\begin{aligned}&\mathbb{Q}(p)=\partial_{(\mathbf{P},p)}\phi((\mathbf{I},p),(0,0))\\&=\{\,(\mathbf{M},q)\mid\forall\dot{\boldsymbol{\mathcal{I}}}\in Z\colon\phi((\mathbf{I},p),\widetilde{\boldsymbol{\mathcal{I}}})\ge(\mathbf{M},q){:}\widetilde{\boldsymbol{\mathcal{I}}}\,\}\subset\mathfrak{p}\times\mathrm{T}^*_pH\end{aligned}$$

and obtain, with $\mathbf{M}=\mathbf{Q}\mathbf{P}^{\top}$, the flow law in invariant form

$$(\dot{\mathbf{P}}\mathbf{P}^{-1},\dot{p})\in\mathrm{N}_{(\mathbf{M},q)}\mathbb{Q}(p)=\partial\mathcal{X}_{\mathbb{Q}(p)}(\mathbf{M},q)\subset\mathfrak{p}\times\mathrm{T}_pH\,.$$

The second fact about these tensors is that they satisfy much better estimates in terms of the energy potential ψ. In fact, following [4] it is reasonable to work with the following multiplicative stress control estimates:

$$\exists\, C_{\mathbf{K}} > 0\ \forall\, \mathbf{F}_{\mathrm{el}} \in \mathfrak{G}:\ |\mathbf{K}(\mathbf{F}_{\mathrm{el}}, p)| \leq C_{\mathbf{K}}\big(\psi(\mathbf{F}_{\mathrm{el}}, p)+1\big)\ , \tag{2.9}$$

$$\exists\, C_{\mathbf{M}} > 0\ \forall\, \mathbf{F}_{\mathrm{el}} \in \mathfrak{G}:\ |\mathbf{M}(\mathbf{F}_{\mathrm{el}}, p)| \leq C_{\mathbf{M}}\big(\psi(\mathbf{F}_{\mathrm{el}}, p)+1\big)\ , \tag{2.10}$$

In fact, (2.10) implies (2.9) but not vice versa. These conditions are satisfied by polyconvex potentials ψ that go to ∞ for $\det \mathbf{F}_{\mathrm{el}} \to 0$. In fact, most Ogden materials satisfy both conditions. For instance consider

$$\psi(\mathbf{F}_{\mathrm{el}}, p) = c_1|\mathbf{F}_{\mathrm{el}}|^{r_1} + c_2(\det \mathbf{F}_{\mathrm{el}})^{-r_2} + c_3(\det \mathbf{F}_{\mathrm{el}})^{r_3} + \gamma(p)$$

with $c_j, r_j > 0$ and $\gamma(p) \geq 0$. Using $\partial_{\mathbf{F}} \det \mathbf{F} = \operatorname{cof} \mathbf{F}$ and $(\operatorname{cof} \mathbf{F})\mathbf{F}^{\top} = \mathbf{F}^{\top} \operatorname{cof} \mathbf{F} = (\det \mathbf{F})\mathbf{I}$ it is easy to see that (2.9) and (2.10) hold. A similar estimate does not hold for $\mathbb{P}(\mathbf{F}_{\mathrm{el}}, p)$, since $(\det \mathbf{F}_{\mathrm{el}})^{-1-r_2} \operatorname{cof} \mathbf{F}$ cannot be estimated by $(\det \mathbf{F}_{\mathrm{el}})^{-r_2}$. It was observed in [18, 28] that these estimates can be used effectively in rate-independent systems to control the power of the external forces.

On the Lie groups $\mathfrak{G}$ it is possible to introduce right-invariant distance

$$d_{\mathfrak{G}}(\mathbf{F}_0, \mathbf{F}_1) = \inf\{\, \textstyle\int_0^1 |\dot{\mathbf{F}}(s)\mathbf{F}(s)^{-1}|\mathrm{d}s \mid \mathbf{F}\in \mathrm{C}^1([0,1],\mathfrak{G}), \mathbf{F}(0)=\mathbf{F}_0, \mathbf{F}(1)=\mathbf{F}_1 \,\}\ ,$$

which satisfies $d_{\mathfrak{G}}(\mathbf{F}_0, \mathbf{F}_1) = d_{\mathfrak{G}}(\mathbf{F}_0\mathbf{F}_1^{-1}, \mathbf{I})$. Only in very few cases $d_{\mathfrak{G}}$ can be calculated explicitly, see [50, 25]. Conditions (2.9) or (2.10) now imply that $\log(\psi+1)$ is globally Lipschitz continuous

$$\Big|\log\big(\psi(\mathbf{F}, p)+1\big) - \log\big(\psi(\widetilde{\mathbf{F}}, p)+1\big)\Big| \leq C_{\mathrm{Lip}} d_{\mathfrak{G}}(\mathbf{F}, \widetilde{\mathbf{F}}) \text{ for all } \mathbf{F}, \widetilde{\mathbf{F}} \in \mathfrak{G}\ .$$

Since $d_{\mathfrak{G}}(\mathbf{F}, \mathbf{I}) \approx \big|\log(\mathbf{F}^{\top}\mathbf{F})\big|$, the energy potential ψ satisfies the upper estimate

$$\psi(\mathbf{F}, p) \leq C_{\mathrm{upp}}(p)\big(|\mathbf{F}|+|\mathbf{F}^{-1}|\big)^{\gamma}.$$

This upper bound is consistent with the lower estimates also called coercivity:

$$\psi(\mathbf{F}_{\mathrm{el}}, p) \geq c_1|\mathbf{F}_{\mathrm{el}}|^{r_{\mathbf{F}}}+c_2|p|^{r_p}-C_2 \text{ for all } (\mathbf{F}_{\mathrm{el}}, p) \in \mathfrak{G} \times H\ . \tag{2.11}$$

We will need that the dissipation distance $D : Z \times Z \to [0, \infty]$, which is associated with the dissipation potential $\phi : \mathrm{T}Z \to [0, \infty]$, is coercive as well, namely

$$D((\mathbf{P}, p), (\mathbf{I}, p_*)) \geq c_3\big(|\mathbf{P}|+|\mathbf{P}^{-1}|\big)^{r_{\mathbf{P}}}-C_3 \text{ for all } (\mathbf{P}, p) \in Z = \mathfrak{P}\times H\ . \tag{2.12}$$

To see that this coercivity estimate needs a significant amount of hardening we treat the simplest example with a scalar hardening parameter $p \geq 0$ and a hardening function $h : [0, \infty) \to (0, \infty)$:

$$\widetilde{\phi}(p,\mathbf{V},\dot{p}) = \begin{cases} h'(p)\dot{p} & \text{if } \dot{p} \geq |\mathbf{V}|_{\mathfrak{p}} \ , \\ \infty & \text{else} \ . \end{cases}$$

According to [51] we find

$$D((\mathbf{P}_0,p_0),(\mathbf{P}_1,p_1)) = \begin{cases} h(p_1)-h(p_0) & \text{if } p_1 \geq p_0+d_{\mathfrak{P}}(\mathbf{P}_0,\mathbf{P}_1) \ , \\ \infty & \text{else} \ . \end{cases}$$

Thus, assuming $p_* = 0$ and $h(0) = 0$ we obtain the lower estimate

$$D((\mathbf{P},p),(\mathbf{I},p_*)) \geq h(d_{\mathfrak{G}}(\mathbf{I},\mathbf{P})) \geq c_4\big(|\mathbf{P}|+|\mathbf{P}^{-1}|\big)^{r_{\mathbf{P}}} - C_4$$

only if $h(p) \geq c_5 \mathrm{e}^{\gamma p} - C_5$ for some $c_5, \gamma > 0$, since $d_{\mathfrak{G}}(\mathbf{I},\mathbf{P})$ grows only logarithmically.

3 Existence Results

The existence results discussed in this section concern solutions without microstructure. These solutions relate to classical meso and macroscopic models for finite-strain elastoplasticity which are used for describing deep drawing or other plastic processes involving large strains. In the highly nonconvex situation we have to find assumptions on the constitutive laws which are compatible with the above geometric nonlinearities and still are good enough to prevent the formation of microstructure, which turns out to be a rather common feature in finite-strain elastoplasticity, see [66, 8, 44, 46, 24, 53, 5, 12].

We choose function spaces and functionals. The admissible deformations are supposed to lie in the set

$$\mathcal{W} = \{\, \boldsymbol{\varphi} \in \mathrm{W}^{1,r_{\boldsymbol{\varphi}}}(\mathcal{B},\mathbb{R}^d) \mid \boldsymbol{\varphi}|_{\Gamma_{\mathrm{Dir}}} = \mathrm{id} \,\} \ .$$

For the internal variables we choose the set

$$\mathcal{Z} = \{\, (\mathbf{P},p) \in \mathrm{L}^{r_{\mathbf{P}}}(\mathcal{B},\mathbb{R}^{d\times d}) \times \mathrm{L}^{r_p}(\mathcal{B},\mathbb{R}^m) \mid (\mathbf{P}(x),p(x)) \in \mathfrak{P} \times H \text{ a.e. in } \mathcal{B} \,\} \ .$$

The choice of the Lebesgue exponents $r_{\boldsymbol{\varphi}}, r_{\mathbf{F}}, r_{\mathbf{P}}$ and r_p will be a crucial step in the further analysis.

All our existence results will be based on the notion of *polyconvexity*, which means that there exists a convex and lower semi-continuous function $g : \mathbb{R}^{m_d} \to [0,\infty]$ such that $\psi(\mathbf{F}) = g(\mathcal{M}(\mathbf{F}))$ holds, where $\mathcal{M}(\mathbf{F})$ is the vector of all minors (subdeterminants) of $\mathbf{F} \in \mathbb{R}^{d\times d}$. The more general condition of *quasiconvexity* is not developed enough to handle integrands ψ which take the value $+\infty$. In fact, in the quasiconvex case the lower semi-continuity results are usually based on the upper bound $\psi(\mathbf{F},p) \leq C(1+|\mathbf{F}|)^r$ for all $\mathbf{F} \in \mathbb{R}^{d\times d}$. This clearly contradicts finite-strain elasticity where $\psi(\mathbf{F},p) = +\infty$ for $\det \mathbf{F} \leq 0$ is imposed to prevent local interpenetration. In contrast, the multiplicative stress-control estimates (2.9) and (2.10) only lead to upper estimates on $\mathfrak{G}$.

3.1 Existence Results for the Incremental Problem

We survey the results in [54], where the incremental problem for systems without any regularization is investigated. The energy functional $\mathcal{E}_0$ and the dissipation distance $\mathcal{D}$ are as defined via (2.6) with the specification of ψ and D as above. We added the subscript "$_0$" to $\mathcal{E}$ to indicate that no regularization is added.

A central rôle in this formulation plays by the *condensed energy density*

$$W^{\mathrm{cond}}((\mathbf{P}_0, p_0); \mathbf{F}) = \min\{\, \psi(\mathbf{F}, \mathbf{P}_1, p_1) + D((\mathbf{P}_0, p_0), (\mathbf{P}_1, p_1)) \mid (\mathbf{P}_1, p_1) \,\}\,,$$

which contains the condensed information on the interplay of energy storage via ψ and energy dissipation via ϕ. Its importance derives from the fact that the minimization of $\int_{\mathcal{B}} \psi(\nabla\boldsymbol{\varphi}\mathbf{P}^{-1}, p) + D((\mathbf{P}_j, p_j), (\mathbf{P}, p))\,\mathrm{d}x$ can be done pointwise in $\boldsymbol{\mathcal{I}} = (\mathbf{P}, p)$ under the integral giving the definition of W^{cond}. Starting from Sect. 4 (see Table 1) the condensed stored energy W^{cond} is replaced by the *incremental stress potential* W, which differs from W^{cond} by a constant only.

First, for any solution process the deformation $\boldsymbol{\varphi}(t) : \mathcal{B} \mapsto \mathbb{R}^d$ must be a minimizer of the condensed functional

$$\mathcal{E}^{\mathrm{cond}}(\boldsymbol{\mathcal{I}}(t); t, \boldsymbol{\varphi}) := \textstyle\int_{\mathcal{B}} W^{\mathrm{cond}}(\boldsymbol{\mathcal{I}}(t, x); \nabla\boldsymbol{\varphi}(x))\,\mathrm{d}x - \langle \Pi_{\mathrm{ext}}(t), \boldsymbol{\varphi}\rangle\,,$$

which follows from the stability condition (S). Hence, W^{cond} contains significant information on the possibility of formation of microstructure (via loss of quasiconvexity [66, 8, 24, 51, 46]) or failure via fracture or localization, see [45, 44, 33]. In Sect. 4 and thereafter $\mathcal{E}^{\mathrm{cond}}$ is replaced by $\mathcal{E}$, which is obtained as $\mathcal{E}^{\mathrm{cond}}$ when W^{cond} is replaced by W. Hence, the two definitions just differ by a constant, and thus have the same minimizers.

Second, the incremental problem (IP) can be reduced to the following condensed incremental problem:

$$\text{(CIP)}\qquad \begin{aligned} &\text{Find iteratively } (\boldsymbol{\varphi}_j, \boldsymbol{\mathcal{I}}_j) \in \mathcal{W} \times \mathcal{Z} \text{ as follows:}\\ &\boldsymbol{\varphi}_j \in \operatorname{Arg\,min}\{\, \mathcal{E}^{\mathrm{cond}}(\boldsymbol{\mathcal{I}}_{j-1}; t_j, \boldsymbol{\varphi}) \mid \boldsymbol{\varphi} \in \mathcal{W} \,\}\\ &\boldsymbol{\mathcal{I}}_j(x) \in \operatorname{Arg\,min}\{\, \psi(\nabla\boldsymbol{\varphi}(x), \boldsymbol{\mathcal{I}}) + D(\boldsymbol{\mathcal{I}}_{j-1}(x), \boldsymbol{\mathcal{I}}) \mid \boldsymbol{\mathcal{I}} \in Z \,\}\,. \end{aligned}$$

Thus, to guarantee existence of minimizers for (CIP) we impose the very restrictive condition, namely

$$W^{\mathrm{cond}}(\boldsymbol{\mathcal{I}}; t, \cdot) : \mathbb{R}^{d\times d} \to [0, \infty] \text{ is polyconvex}\,. \tag{3.1}$$

Theorem 3.3 in [54] provides the following existence result for (IP).

Theorem 1. *Let $\mathcal{W}$, $\mathcal{Z}$, $\mathcal{E}_0$ and $\mathcal{D}$ be defined as above. Assume that W^{cond} satisfies* (3.1)*. Further let the coercivity assumptions* (2.11) *and* (2.12) *be satisfied with $r_{\boldsymbol{\varphi}}, r_{\mathbf{P}}$ and r_p such that*

$$\frac{1}{r_{\mathbf{F}}} + \frac{1}{r_{\mathbf{P}}} =: \frac{1}{r_{\boldsymbol{\varphi}}} < \frac{1}{d} \,. \tag{3.2}$$

If additionally $\Pi_{\mathrm{ext}} \in \mathrm{C}^1([0,T], \mathrm{W}^{1,r_{\boldsymbol{\varphi}}}(\mathcal{B},\mathbb{R}^d)^*)$*, then (IP) associated with* $(\mathcal{E}_0, \mathcal{D})$ *has, for each initial datum* $\boldsymbol{\mathcal{I}}_0 \in \mathcal{Z}$ *with* $\mathcal{D}((\mathbf{I}, p_*), \boldsymbol{\mathcal{I}}_0) < \infty$ *and each partition* $0 = t_0 < t_1 < \cdots < t_N = T$*, at least one solution* $(\boldsymbol{\varphi}_j, \boldsymbol{\mathcal{I}}_j)_{j=1,\dots,N}$ *in* $\mathcal{W} \times \mathcal{Z}$*. Moreover, there exists a constant* C *(depending on the data only) such that all solutions satisfy, for* $j = 1, ..., N$*,*

$$\|\boldsymbol{\varphi}_j\|_{1,r_{\boldsymbol{\varphi}}} + \|\mathbf{P}_j\|_{r_{\mathbf{P}}} + \|\mathbf{P}_j^{-1}\|_{r_{\mathbf{P}}} + \|p_j\|_{r_p} + \mathcal{E}_0(t_j, \boldsymbol{\varphi}_j, \boldsymbol{\mathcal{I}}_j) + \sum_{k=1}^{j} \mathcal{D}(\boldsymbol{\mathcal{I}}_{k-1}, \boldsymbol{\mathcal{I}}_k) \le C \,.$$

The proof relies on solving (CIP) with a careful bookkeeping based on the a priori estimates (2.8). The necessary coercivity of W^{cond} follows from those of ψ and D, after employing the invariance from $W^{\mathrm{cond}}((\mathbf{P},p);\mathbf{F}) = W^{\mathrm{cond}}((\mathbf{I},p_0);\mathbf{F}\mathbf{P}^{-1})$ and the Hölder inequality

$$|\mathbf{F}\mathbf{P}^{-1}|^{r_{\mathbf{F}}} \ge \big(|\mathbf{F}|/|\mathbf{P}|\big)^{r_{\mathbf{F}}} \ge c_r|\mathbf{F}|^{r_{\boldsymbol{\varphi}}} - d_r|\mathbf{P}|^{r_{\mathbf{P}}} \,.$$

The major drawback of the present theory is that the polyconvexity condition (3.1) is extremely difficult to check. The function W^{cond} is defined implicitly via ψ and D, but D itself is defined implicitly from ϕ. Hence, there are only very few cases where W^{cond} can be calculated explicitly. One case is in dimension $d = 1$ and another case is treated in [54]. It is an isotropic situation in dimension $d = 2$ using an abstract characterization of [56] for isotropic, polyconvex energy densities.

3.2 Partially Regularized Incremental Problems

The second result concerns a model which uses a partial regularization which is based on the so-called geometric dislocation tensor

$$\mathbf{G}_{\mathbf{P}} = \frac{1}{\det \mathbf{P}} \big(\operatorname{curl} \mathbf{P}\big)\mathbf{P}^{\top} \ \in \mathbb{R}^{3\times 3}$$

where the "curl" of a matrix is applied row-wise. Because of our standing assumption $\det \mathbf{P} = 1$ we can use a simpler form. The energy now reads

$$\mathcal{E}_{\mathrm{curl}}(t, \boldsymbol{\varphi}, \mathbf{P}, p) = \mathcal{E}_0(t, \boldsymbol{\varphi}, \mathbf{P}, p) + \textstyle\int_{\mathcal{B}} V\big((\operatorname{curl}\mathbf{P})\mathbf{P}^{\top}\big)\,\mathrm{d}x \,,$$

where the potential $V : \mathbb{R}^{3\times 3} \to [0,\infty]$ satisfies

$$V \text{ is convex and } V(\mathbf{G}) \ge c_6|\mathbf{G}|^{r_{\mathbf{G}}} - C_6 \text{ on } \mathbb{R}^{3\times 3} \,. \tag{3.3}$$

In [59] a general lower semi-continuity result is derived for general functionals of the form

$$\mathcal{I}(\boldsymbol{\varphi}, \mathbf{P}) = \int_{\mathcal{B}} U(\nabla\boldsymbol{\varphi}\mathbf{P}^{-1}, \mathbf{P}, (\operatorname{curl}\mathbf{P})\mathbf{P}^{\top})\,\mathrm{d}x \,.$$

Under the assumption that $U : \mathbb{R}^{3\times3} \times \mathbb{R}^{3\times3} \times \mathbb{R}^{3\times3} \to [0,\infty]$ is polyconvex in the first two arguments and convex in the third argument and that U is suitably coercive it is shown that $\mathcal{I}$ is weakly lower semicontinuous on the associated Sobolev spaces.

Consider a weakly converging sequence $(\boldsymbol{\varphi}_j, \mathbf{P}_j) \rightharpoonup (\boldsymbol{\varphi}, \mathbf{P})$. Along sequences with bounded energies $\mathcal{I}(\boldsymbol{\varphi}_j, \mathbf{P}_j) \le C$ the terms $\nabla\boldsymbol{\varphi}_j\mathbf{P}_j^{-1}$, $\mathbf{P}_j$, $\mathbf{G}_{\mathbf{P}_j}$ are controlled in suitable Lebesgue spaces. This implies a bound on $\operatorname{curl}\mathbf{P}_j$ and thus, a suitable version of the div-curl lemma can be used to show that $\mathcal{M}(\mathbf{P}_j) \rightharpoonup \mathcal{M}(\mathbf{P})$ and $\mathbf{G}_{\mathbf{P}_j} \rightharpoonup \mathbf{G}_{\mathbf{P}}$. The special form of the multiplicative decomposition $\nabla\boldsymbol{\varphi}\mathbf{P}^{-1}$ together with $\det\mathbf{P} = 1$ provide the minor relations

$$\mathbf{F}\mathbf{P}^{-1} = \mathbf{F}(\operatorname{cof}\mathbf{P})^{\top},\ \operatorname{cof}(\mathbf{F}\mathbf{P}^{-1}) = (\operatorname{cof}\mathbf{F})\,\mathbf{P}^{\top},\ \det(\mathbf{F}\mathbf{P}^{-1}) = \det\mathbf{F}\,.$$

Hence, again applying the div-curl lemma we obtain also the convergence $\mathcal{M}(\mathbf{F}_j\mathbf{P}_j^{-1}) \rightharpoonup \mathcal{M}(\mathbf{F}\mathbf{P}^{-1})$ and the weak lower semi-continuity follows using (poly-) convexity.

This result is then applied to the incremental problem (IP) associated with $\mathcal{E}_{\mathrm{curl}}$ and $\mathcal{D}$. Again, a condensation, like in Sect. 3.1, is done for the variable $p \in H$, which does not have a derivative. We assume $\psi(\mathbf{F}_{\mathrm{el}}, p) = \psi_{\mathrm{el}}(\mathbf{F}_{\mathrm{el}}) + \psi_{\mathrm{hard}}(p)$ and let

$$D^{\mathrm{cond}}((\mathbf{P}_0,p_0);\mathbf{P}) := \min\{\,\psi_{\mathrm{hard}}(p) + D((\mathbf{P}_0,p_0),(\mathbf{P},p)) \mid p \in H\,\}\,.$$

Then, the incremental problem involves the integrand $U(\mathbf{F}_{\mathrm{el}},\mathbf{P},\mathbf{G}) = \psi_{\mathrm{el}}(\mathbf{F}_{\mathrm{el}}) + D^{\mathrm{cond}}(\boldsymbol{\mathcal{I}}_{j-1}(x);\mathbf{P}) + V(\mathbf{G})$. Thus, the crucial assumption we have to make is that

$$D^{\mathrm{cond}}(\boldsymbol{\mathcal{I}};\cdot) : \mathbb{R}^{3\times3} \to [0,\infty] \text{ is polyconvex}\,. \tag{3.4}$$

The following result is established in [59].

Theorem 2. *Let $\mathcal{W}$, $\mathcal{Z}$, $\mathcal{E}_{\mathrm{curl}}$ and $\mathcal{D}$ be defined as above. Assume that V and D^{cond} satisfy* (3.3) *and* (3.4)*, respectively . Further let the coercivity assumptions* (2.11) *and* (2.12) *be satisfied with $r_{\boldsymbol{\varphi}}, r_{\mathbf{P}}, r_p$ and $r_{\mathbf{G}}$ such that*

$$\frac{1}{r_{\mathbf{F}}} + \frac{1}{r_{\mathbf{P}}} =: \frac{1}{r_{\boldsymbol{\varphi}}} < \frac{1}{d}\,,\quad \frac{1}{r_{\mathbf{G}}} + \frac{2}{r_{\mathbf{P}}} < 1\,,\ \textit{and}\ r_{\mathbf{G}} > d\,. \tag{3.5}$$

If additionally $\mathit{\Pi}_{\mathrm{ext}} \in \mathrm{C}^1([0,T], \mathrm{W}^{1,r_{\boldsymbol{\varphi}}}(\mathcal{B},\mathbb{R}^d)^*)$, then (IP) associated with $(\mathcal{E}_{\mathrm{curl}},\mathcal{D})$ has, for each initial datum $\boldsymbol{\mathcal{I}}_0 \in \mathcal{Z}$ with $\mathcal{D}((\mathbf{I},p_*),\boldsymbol{\mathcal{I}}_0) < \infty$ and each partition $0 = t_0 < t_1 < \cdots < t_N = T$, at least one solution $(\boldsymbol{\varphi}_j,\boldsymbol{\mathcal{I}}_j)_{j=1,\dots,N}$ in $\mathcal{W}\times\mathcal{Z}$. Moreover, there exists a constant C (depending on the data only) such that all solutions satisfy, for $j = 1,...,N$,

$$\|\boldsymbol{\varphi}_j\|_{1,r_{\boldsymbol{\varphi}}} + \|\mathbf{P}_j\|_{r_{\mathbf{P}}} + \|\mathbf{P}_j^{-1}\|_{r_{\mathbf{P}}} + \|p_j\|_{r_p} + \mathcal{E}_0(t_j,\boldsymbol{\varphi}_j,\boldsymbol{\mathcal{I}}_j) + \sum_{k=1}^{j}\mathcal{D}(\boldsymbol{\mathcal{I}}_{k-1},\boldsymbol{\mathcal{I}}_k) \le C\,.$$

Again the polyconvexity condition (3.4) for the condensed dissipation distance is hard to satisfy. However, we have considerably more freedom than in the case of the condensed energy potential W^{cond}. Here the condition is based on the dissipation distance only, and we are able to take any polyconvex function ψ_{el} for the elastic storage. Examples are given in Sect. 4 of [59].

However, the theory is still restrictive as we do not have good examples of dissipation distances and we do not know what type of hardening leads to polyconvexity. In the light of the example at the end of Sect. 2.3 it is a natural question to ask whether the functions

$$\mathfrak{P} \ni \mathbf{P} \mapsto \exp\big(\gamma d_{\mathfrak{P}}(\mathbf{I}, \mathbf{P})\big) \ ,$$

if extended by $+\infty$ outside of $\mathfrak{P}$, are polyconvex for sufficiently large $\gamma > 0$. It is clear that this can only hold if $d_{\mathfrak{P}}$ is locally Lipschitz continuous with respect to the classical metric in $\mathbb{R}^{d\times d}$. Thus, sub-Riemannian or sub-Finslerian metrics are not allowed.

3.3 Strain-Gradient Plasticity

In [36] a theory is developed for the case that the full gradient $(\nabla\mathbf{P}, \nabla p)$ is used for regularization. For micromechanically motivated nonlocal crystal plasticity models, see [6, 17, 22, 42, 71, 72]. This case relates to the regularized theory that was developed for other rate-independent material models like shape-memory materials, damage, brittle fracture, magnetostriction or piezoelectricity. We refer to the survey [57] in this volume.

In the present theory the incremental problem will be used as a tool to construct piecewise constant solutions for partitions with smaller and smaller step sizes. We are then able to extract a subsequence which converges to a solution of the time-continuous problem (S) and (E) as derived in Sect. 2.2. The analysis follows closely the abstract approach for general rate-independent systems on topological spaces as developed in [35, 55, 18].

We only treat the simplest case and consider the energy functional

$$\mathcal{E}_{\mathrm{reg}}(t, \boldsymbol{\varphi}, \mathbf{P}, p) = \mathcal{E}_0(t, \boldsymbol{\varphi}, \mathbf{P}, p) + \textstyle\int_{\mathcal{B}} c_1|\nabla\mathbf{P}|^{r_1} + c_2|\nabla p|^{r_2}\,\mathrm{d}x \ ,$$

where $c_1, c_2 > 0$ and $r_1, r_2 > 1$. The dissipation distance $\mathcal{D}$ is kept as above.

For the admissible deformations $\boldsymbol{\varphi}$ we keep the function space $\mathcal{W} \subset \mathrm{W}^{1,r_{\boldsymbol{\varphi}}}(\mathcal{B}, \mathbb{R}^d)$ equipped with the weak topology. For the internal variables we now set $\mathcal{Z}_{\mathrm{reg}} = \mathcal{Z}_{\mathbf{P}} \times \mathcal{Z}_p$ with

$$\begin{aligned} \mathcal{Z}_{\mathbf{P}} &:= \{\, \mathbf{P} \in \mathrm{W}^{1,r_1}(\mathcal{B}, \mathbb{R}^{d\times d}) \mid \mathbf{P}(x) \in \mathfrak{P} \text{ a.e. on } \mathcal{B} \,\} \text{ and} \\ \mathcal{Z}_p &:= \{\, p \in \mathrm{W}^{1,r_2}(\mathcal{B}, \mathbb{R}^m) \mid p(x) \in H \text{ a.e. on } \mathcal{B} \,\} \ , \end{aligned}$$

where $\mathcal{Z}_{\mathrm{reg}}$ carries the weak topology of $\mathrm{W}^{1,r_1}(\mathcal{B}, \mathbb{R}^{d\times d}) \times \mathrm{W}^{1,r_1}(\mathcal{B}, \mathbb{R}^m)$.

Using polyconvexity of $\mathbf{F} \mapsto \psi(\cdot, p)$ and the above coercivity assumptions it is possible to show that the incremental problem (IP) associated with $(\mathcal{E}_{\mathrm{reg}}, \mathcal{D})$

has at least one solution $((\boldsymbol{\varphi}_j^k, \boldsymbol{\mathcal{I}}_j^k))_{j=1,\dots,N_k}$, where we already assumed that we have a sequence of partitions indexed by $k \in \mathbb{N}$ such that the fineness $\phi_k = \max\{\, t_j^k - t_{j-1}^k \mid j = 1, ..., N_k \,\}$ tends to 0. We define the piecewise constant interpolants $(\overline{q}_k, \overline{\boldsymbol{\mathcal{I}}}_k) : [0,T] \to \mathcal{W} \times \mathcal{Z}_{\rm reg}$ with

$$\begin{aligned}&(\overline{q}_k(t), \overline{\boldsymbol{\mathcal{I}}}_k(t)) = (\boldsymbol{\varphi}_{j-1}^k, \boldsymbol{\mathcal{I}}_{j-1}^k) \text{ for } t \in [t_{j-1}^k, t_j^k)\\ &\text{and } (\overline{q}_k(T), \overline{\boldsymbol{\mathcal{I}}}_k(T)) = (\boldsymbol{\varphi}_{N_k}^k, \boldsymbol{\mathcal{I}}_{N_k}^k)\,.\end{aligned}$$

According to (2.7) these piecewise constant solutions satisfy the stability conditions (S) on each point of the partition, i.e., $(\overline{q}_k(t_j^k), \overline{\boldsymbol{\mathcal{I}}}_k(t_j^k)) \in \mathcal{S}(t_j^k)$ with

$$\mathcal{S}(t) := \{\, (\boldsymbol{\varphi}, \boldsymbol{\mathcal{I}}) \mid \forall\, (\widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\mathcal{I}}})\colon \mathcal{E}_{\rm reg}(t, \boldsymbol{\varphi}, \boldsymbol{\mathcal{I}}) \le \mathcal{E}_{\rm reg}(t, \widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\mathcal{I}}}) + \mathcal{D}(\boldsymbol{\mathcal{I}}, \widetilde{\boldsymbol{\mathcal{I}}}) \,\}\,.$$

Moreover, the energy estimate (2.8) provides the energy bounds

$$\begin{aligned}&\mathcal{E}_{\rm reg}(t_j^k, \overline{\boldsymbol{\varphi}}_k(t_j^k), \overline{\boldsymbol{\mathcal{I}}}_k(t_j^k)) + {\rm Diss}_{\mathcal{D}}(\overline{\boldsymbol{\mathcal{I}}}_k, [0, t_j^k])\\ &\le \mathcal{E}_{\rm reg}(0, \boldsymbol{\varphi}_0, \boldsymbol{\mathcal{I}}_0) + \textstyle\int_0^{t_j^k} \partial_s \mathcal{E}_{\rm reg}(s, \overline{\boldsymbol{\varphi}}_k(s), \overline{\boldsymbol{\mathcal{I}}}_k(s))\,{\rm d}s\,.\end{aligned} \tag{3.6}$$

They give rise to the bounds

$$\begin{aligned}&\|(\overline{\boldsymbol{\varphi}}_k, \overline{\boldsymbol{\mathcal{I}}}_k)\|_{{\rm L}^\infty([0,T], {\rm W}^{1,r_{\boldsymbol{\varphi}}} \times {\rm W}^{1,r_1} \times {\rm W}^{1,r_2})} \le C\,,\\ &\sup_{t\in[0,T]} \mathcal{E}_{\rm reg}(t_j^k, \overline{\boldsymbol{\varphi}}_k(t_j^k), \overline{\boldsymbol{\mathcal{I}}}_k(t_j^k)) \le C, \quad {\rm Diss}_{\mathcal{D}}(\overline{\boldsymbol{\mathcal{I}}}_k, [0,T]) \le C\,.\end{aligned}$$

Thus, by using a suitable version of Helly's selection principle (cf., [35]) it is possible to extract a subsequence and to find a limit process $(\boldsymbol{\varphi}, \boldsymbol{\mathcal{I}}) : [0,T] \to \mathcal{W} \times \mathcal{Z}_{\rm reg}$, which is a candidate for an energetic solution.

Using weak lower semi-continuity the energy bound (3.6) easily supplies the upper energy estimate

$$\mathcal{E}_{\rm reg}(t,\boldsymbol{\varphi}(t),\boldsymbol{\mathcal{I}}(t)) + {\rm Diss}_{\mathcal{D}}(\boldsymbol{\mathcal{I}}, [0,t]) \le \mathcal{E}_{\rm reg}(0,\boldsymbol{\varphi}_0,\boldsymbol{\mathcal{I}}_0) + \int_0^t \partial_s \mathcal{E}_{\rm reg}(s,\boldsymbol{\varphi}(s),\boldsymbol{\mathcal{I}}(s))\,{\rm d}s\,.$$

The crucial step in the convergence proof is to show that the sets $\mathcal{S}(t)$ of stable states are sequentially closed in the weak Banach space topology. This step is easy if $\mathcal{D}$ is weakly continuous but it also works in more realistic cases with hardening, which is irreversible, see [36]. If this step is done we know that the limit process satisfies (S) and, moreover, a general abstract proposition yields the lower energy estimate and hence (E) holds as well.

We summarize the result as follows.

Theorem 3. *Let $\mathcal{W}$, $\mathcal{Z}_{\rm reg}$, $\mathcal{E}_{\rm reg}$ and $\mathcal{D}$ be given as above with ψ and D satisfying the coercivity estimates* (2.11) *and* (2.12) *with $\frac{1}{r_{\mathbf{F}}} + \frac{1}{r_{\mathbf{P}}} = \frac{1}{r_{\boldsymbol{\varphi}}} < \frac{1}{d}$. Moreover, assume $\Pi_{\rm ext} \in {\rm C}^1([0,T], {\rm W}^{1,r_{\boldsymbol{\varphi}}}(\mathcal{B}, \mathbb{R}^d)^*)$. Then, for each stable initial state $(\boldsymbol{\varphi}_0, \boldsymbol{\mathcal{I}}_0) \in \mathcal{S}(0)$ the energetic formulation (S) and (E) has at least one solution $(\boldsymbol{\varphi}, \boldsymbol{\mathcal{I}}) : [0,T] \to \mathcal{W} \times \mathcal{Z}_{\rm reg}$. All solutions satisfy*

$$(\boldsymbol{\varphi}, \mathbf{P}, p) \in {\rm L}^\infty([0,T], {\rm W}^{1,r_{\boldsymbol{\varphi}}}(\mathcal{B}, \mathbb{R}^d) \times {\rm W}^{1,r_1}(\mathcal{B}, \mathbb{R}^{d\times d}) \times {\rm W}^{1,r_2}(\mathcal{B}, \mathbb{R}^m))$$

and ${\rm Diss}_{\mathcal{D}}((\mathbf{P}, p), [0,T]) < \infty$.

3.4 Time-Dependent Boundary Conditions

The existence results of Sects. 3.1 to 3.3 rely on the fact that the space $\mathcal{W}$ of admissible deformations is independent of time. For many applications one needs to generalize this assumption. For the incremental problem (IP) it is not too difficult to work with $\mathcal{W}(t)$, however for the energetic formulation it is not clear how to define the power $\partial_t \mathcal{E}(t,q)$ of the external loadings that are due to changes of $\mathcal{W}(t)$.

The usual way to implement time-dependent Dirichlet data is to substract a sufficiently smooth function that has the correct boundary value and then try to find the homogeneous part. In the case of small strain, when working with $\mathbf{u} : x \mapsto \boldsymbol{\varphi}(x) - x$ this means $\mathbf{u}(t,x) = \mathbf{u}_{\mathrm{Dir}}(t,x) + \mathbf{v}(t,x)$ with $\mathbf{v}(t,\cdot)|_{\Gamma_{\mathrm{Dir}}} = 0$. We let $\mathcal{W} = \{\, \mathbf{v} \in \mathrm{W}^{1,p}(\mathcal{B},\mathbb{R}^d) \mid \mathbf{v}|_{\Gamma_{\mathrm{Dir}}=0} \,\}$ and define the shifted energy $\widetilde{\mathcal{E}}(t,\mathbf{v},\boldsymbol{\mathcal{I}}) = \mathcal{E}(t,\mathbf{u}_{\mathrm{Dir}}(t)+\mathbf{v},\boldsymbol{\mathcal{I}})$. The power of the external loading now takes the form

$$\begin{aligned}\partial_t \widetilde{\mathcal{E}}(t,\mathbf{v},\boldsymbol{\mathcal{I}}) = &\int_{\mathcal{B}} \partial_{\mathbf{F}}\psi(\nabla(\mathbf{u}_{\mathrm{Dir}}(t)+\mathbf{v}),\boldsymbol{\mathcal{I}}):\nabla\dot{\mathbf{u}}_{\mathrm{Dir}}(t)\,\mathrm{d}x\\ &-\langle \dot{\mathit{\Pi}}_{\mathrm{ext}}(t), \mathbf{u}_{\mathrm{Dir}}(t)+\mathbf{v}\rangle - \langle \mathit{\Pi}_{\mathrm{ext}}(t), \dot{\mathbf{u}}_{\mathrm{Dir}}(t)\rangle\ .\end{aligned}$$

However, in the case of finite-strain elasticity we cannot guarantee that the integrand $\partial_{\mathbf{F}}\psi$ lies in $\mathrm{L}^1(\mathcal{B})$, since we cannot control the Piola-Kirchoff stress $\mathbb{P} = \partial_{\mathbf{F}}\psi$ by ψ itself.

In the case of finite-strain elasticity the stored energy density ψ takes the value $+\infty$ and $\partial_{\mathbf{F}}\psi(\mathbf{F},\boldsymbol{\mathcal{I}})$ exists only on $\mathfrak{G}$. In order to use the multiplicative stress control (2.9) for the Kirchhoff stress $\mathbf{K}$ we assume that time-dependent Dirichlet data $\boldsymbol{\varphi}_{\mathrm{Dir}}$ are given. We then decompose the desired solution $\boldsymbol{\varphi}$ via composition of functions

$$\boldsymbol{\varphi}(t,x) = \boldsymbol{\varphi}_{\mathrm{Dir}}(t,\boldsymbol{\xi}(t,x)) = (\boldsymbol{\varphi}_{\mathrm{Dir}}(t,\cdot)\circ\boldsymbol{\xi}(t,\cdot))(x)\ .$$

Here, we assume that $\boldsymbol{\varphi}_{\mathrm{Dir}}$ can be extended such that $\boldsymbol{\varphi}_{\mathrm{Dir}} \in \mathrm{C}^2([0,T]\times\mathbb{R}^d,\mathbb{R}^d)$ and that $\nabla_x\boldsymbol{\varphi}_{\mathrm{Dir}}$ and $(\nabla_x\boldsymbol{\varphi}_{\mathrm{Dir}})^{-1}$ are bounded on $[0,T]\times\mathbb{R}^d$. The set of admissible deformations is now $\mathcal{W} = \{\, \boldsymbol{\xi} \in \mathrm{W}^{1,p}(\mathcal{B},\mathbb{R}^d) \mid \boldsymbol{\xi}|_{\Gamma_{\mathrm{Dir}}} = \mathrm{id} \,\}$ with $p > d$ and the shifted energy is $\widehat{\mathcal{E}}(t,\boldsymbol{\xi},\boldsymbol{\mathcal{I}}) = \mathcal{E}(t,\boldsymbol{\varphi}_{\mathrm{Dir}}(t)\circ\boldsymbol{\xi},\boldsymbol{\mathcal{I}})$. Using the classical chain rule formula

$$\nabla_x(\boldsymbol{\varphi}_{\mathrm{Dir}}(t)\circ\boldsymbol{\xi})(x) = \nabla_y\boldsymbol{\varphi}_{\mathrm{Dir}}(t,\boldsymbol{\xi}(x))\nabla_x\boldsymbol{\xi}(x)$$

and the definition of $\mathbf{K}$ in Sect. 2.3 we find the expression for the power

$$\begin{aligned}\partial_t\widehat{\mathcal{E}}(t,\boldsymbol{\xi},\boldsymbol{\mathcal{I}}) = &\int_{\mathcal{B}} \mathbf{K}(\nabla\boldsymbol{\varphi}_{\mathrm{Dir}}\nabla\boldsymbol{\xi},\boldsymbol{\mathcal{I}}):(\nabla\boldsymbol{\varphi}_{\mathrm{Dir}})^{-1}\nabla\dot{\boldsymbol{\varphi}}_{\mathrm{Dir}}\,\mathrm{d}x\\ &-\langle \dot{\mathit{\Pi}}_{\mathrm{ext}}(t), \boldsymbol{\varphi}_{\mathrm{Dir}}\circ\boldsymbol{\xi}\rangle - \langle \mathit{\Pi}_{\mathrm{ext}}(t), \dot{\boldsymbol{\varphi}}_{\mathrm{Dir}}\circ\boldsymbol{\xi}\rangle\ .\end{aligned}$$

Here, for $\widetilde{\mathcal{E}}(t,q) < \infty$ we may conclude via (2.9) that $\mathbf{K} \in \mathrm{L}^1(\mathcal{B},\mathbb{R}^{d\times d})$ while $(\nabla\boldsymbol{\varphi}_{\mathrm{Dir}})^{-1}\nabla\dot{\boldsymbol{\varphi}}_{\mathrm{Dir}}$ lies in $\mathrm{C}^0(\overline{\mathcal{B}},\mathbb{R}^{d\times d})$. Hence, the right-hand side is indeed well defined and the power control

$$|\partial_t \widetilde{\mathcal{E}}(t, \boldsymbol{\xi}, \boldsymbol{\mathcal{I}})| \leq c_1^E \big(\widetilde{\mathcal{E}}(t, \boldsymbol{\xi}, \boldsymbol{\mathcal{I}}) + c_0^E\big)$$

can be established easily. We refer to Sect. 5 in [18] for more details concerning the full existence result for energetic solutions in the case of time-dependent Dirichlet data.

In [28] very similar ideas are used to derive formulas for the energy-release rate in crack propagation for the case of finite-strain elasticity. Also a very restricted case of temperature dependence can be treated by this method of energy control, see [58] and Sect. 5.4 in [57].

4 Modeling of Microstructure via Relaxation

In principle, the time incremental problem (IP) and the energetic formulation (S) & (E) introduced in Sect. 2.2 is a very flexible tool to treat the relaxation as well. We refer to [52, 53, 60] for some recent developments. However, the analytical methods are not yet adapted to the specific nonlinearities involved in finite-strain elastoplasticity. In particular, there is no theory which combines the theory of gradient Young measures with finite-strain plasticity. Thus, the evolutionary theory for gradient Young measures used in models for shape-memory alloys in [32] cannot be generalized to the present situation. Despite of the lacking mathematical tools in this area, the following sections show that the algorithmic approach of relaxation for the multiscale problems has advanced considerably over the last decade.

4.1 Incremental Stability of Standard Dissipative Solids

As pointed out in [45, 44, 46] a key advantage of the variational formulation outlined briefly in Table 1 is the opportunity to analyze the incremental stability of inelastic solids in terms of terminologies used in finite elasticity. In the following we define the material stability of standard dissipative solids based on global weak convexity properties of the incremental stress potential.

The existence of the constitutive minimization problem allows the introduction of an incremental minimization formulation of the boundary-value problem of finite inelasticity for standard dissipative solids. Now consider a functional $\mathcal{E}$ of the current deformation field $\boldsymbol{\varphi}_{n+1}$at the right boundary of the increment $[t_n, t_{n+1}]$:

$$\mathcal{E}(\boldsymbol{\varphi}_{n+1}) = \int_{\mathcal{B}} W(\mathbf{F}_{n+1})\, \mathrm{d}x - [\, \Pi_{\text{ext}}(\boldsymbol{\varphi}_{n+1}) - \Pi_{\text{ext}}(\boldsymbol{\varphi}_n)\,]\,, \tag{4.1}$$

with the global load potential function $\Pi_{\text{ext}}(\boldsymbol{\varphi}) = \int_{\mathcal{B}} \boldsymbol{\varphi} \cdot \boldsymbol{\gamma}\, \mathrm{d}x + \int_{\partial \mathcal{B}_t} \boldsymbol{\varphi} \cdot \mathbf{t}\, \mathrm{d}x$ of dead body forces $\boldsymbol{\gamma}(x,t)$ in $\mathcal{B}$ and surface tractions $\mathbf{t}(x,t)$ on $\partial \mathcal{B}_t$. As outlined in Sect. 3.1, see also [45, 44, 46], the current deformation map of inelastic standard dissipative materials can then be determined by a *principle of minimum incremental energy for standard dissipative solids*

$$\mathcal{E}(\boldsymbol{\varphi}^*_{n+1}) = \inf_{\boldsymbol{\varphi}_{n+1}\in\mathcal{W}} \mathcal{E}(\boldsymbol{\varphi}_{n+1}) \,, \tag{4.2}$$

subject to the essential boundary conditions of a prescribed deformation $\bar{\boldsymbol{\varphi}}$ on $\partial\mathcal{B}_{\boldsymbol{\varphi}}$, written in the form $\boldsymbol{\varphi}_{n+1} \in \mathcal{W} := \{\ \boldsymbol{\varphi} \in \mathrm{W}^{1,p}(\mathcal{B}) \mid \boldsymbol{\varphi}(x) = \bar{\boldsymbol{\varphi}}(x) \text{ on } \partial\mathcal{B}_{\boldsymbol{\varphi}}\ \}$. As usual, we consider a decomposition of the surface into a part where the deformation is prescribed and a part where the tractions are given, i.e. $\partial\mathcal{B} = \partial\mathcal{B}_{\boldsymbol{\varphi}} \cup \partial\mathcal{B}_t$ and $\partial\mathcal{B}_{\boldsymbol{\varphi}} \cap \partial\mathcal{B}_t = \emptyset$. The minimization problem (4.2) governs the response of the inelastic solid in the finite increment $[t_n, t_{n+1}]$ in a structure identical to the principle of minimum potential energy in finite elasticity.

Quasiconvexity of the Incremental Stress Potential

Extending results of the existence theory in finite elasticity as summarized in [3, 10, 13, 37, 70] to the incremental response of standard dissipative solids in the finite step $[t_n, t_{n+1}]$, we consider the *sequentially weakly lower semicontinuity* (s.w.l.s.) of the functional (4.1) as the key property for the *existence of sufficiently regular minimizers* of the variational problem (4.2). The internal part of the functional (4.1) is sequentially weakly lower semicontinuous, if the incremental stress potential defined by the constitutive minimization problem is quasiconvex and also it satisfies some technical growth condition, see e.g. [13, 1, 70]. We regard the *quasiconvexity* introduced in [64] of the incremental stress potential W as the fundamental criterion for the incremental material stability of the inelastic solid. W is said to be quasiconvex at $\mathbf{F}_{n+1}$ if condition

$$W(\mathbf{F}_{n+1}) \le \inf_{\mathbf{w}\in\mathcal{W}_0} \frac{1}{|D|} \int_D W(\mathbf{F}_{n+1} + \nabla\mathbf{w}(\mathbf{y}))\mathrm{d}x \,, \tag{4.3}$$

holds with $\mathbf{y} \in D$ being subject to the constraint $\mathbf{w} \in \mathcal{W}_0 := \{\mathbf{w} \in \mathrm{W}^{1,\infty}(D) | \mathbf{w} = \mathbf{0} \text{ on } \partial D\}$ providing a support on ∂D. Here, $D \subset \mathbb{R}^3$ is an arbitrarily chosen part of the inelastic solid. The condition states that for all fluctuations $\mathbf{w}$ on D with support on ∂D the homogeneous deformation given by $\mathbf{F}_{n+1}$ provides an absolute minimizer of the incremental potential in D. Thus the condition rules out internal buckling, the development of local fine-scale microstructures and phase decomposition of a homogeneous local deformation state. This mechanical interpretation is visualized in Fig. 3. The material is stable if the superimposed fluctuation field of Fig. 3(b) with $\mathbf{w} = \mathbf{0}$ on ∂D yields a higher energy level than the homogeneous deformation $\mathbf{F}_{n+1}$ of Fig. 3(a).

The well-motivated concept of quasiconvexity is based on a global integral condition in space which is hard to verify in practice. The central difficulty is to find the fluctuation field $\mathbf{w} \in \mathcal{W}_0$ on D that minimizes the integral in (4.3). However, recall that weak convexity conditions are related via

$$\text{convexity} \Rightarrow \text{polyconvexity} \Rightarrow \text{quasiconvexity} \Rightarrow \text{rank-one convexity} \,, \tag{4.4}$$

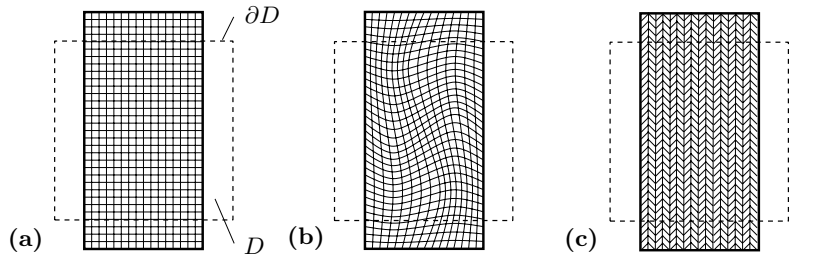


Fig. 3. Interpretation of incremental energetic stability conditions of an inelastic material. A given homogeneous deformation state $\mathbf{F}_{n+1}$ of the material such as the pure shear mode of Figure **(a)** is stable if superimposed fine-scale fluctuation patterns **(b)** (*quasiconvexity*) with support on the boundary ∂D or first-order laminates **(c)** (*rank-one convexity*) increase the averaged incremental stress potential on D

and that the slightly weaker rank-one convexity condition is considered as a close approximation of the quasiconvexity condition, see e.g. [13]. In what follows, we focus on the rank-one convexity as a criterion for material stability.

Rank-One Convexity of the Incremental Stress Potential

The definition of *rank-one convexity* can be traced back to the work of Corall and Graves, see e.g. [70]. The incremental stress potential W is said to be *rank-one convex* at $\mathbf{F}_{n+1}$ if the condition

$$W(\mathbf{F}_{n+1}) \leq \inf_{\xi,\mathbf{F}^+,\mathbf{F}^-} \{\, \xi W(\mathbf{F}^+) + (1-\xi) W(\mathbf{F}^-) \,\} \,, \tag{4.5}$$

holds for the *laminate deformations* $\mathbf{F}^+$ and $\mathbf{F}^-$ which satisfy the conditions

$$\mathbf{F}_{n+1} = \xi \mathbf{F}^+ + (1-\xi)\mathbf{F}^- \quad \text{and} \quad \operatorname{rank}[\mathbf{F}^+ - \mathbf{F}^-] \leq 1 \,, \tag{4.6}$$

in terms of the volume fraction $\xi \in [0,1]$. Condition $(4.6)_1$ states that the volume average of the micro-deformations $\mathbf{F}^{\pm}$ yields the macroscopic homogeneous deformation $\mathbf{F}_{n+1}$. The compatibility of the micro-phases $(\pm)$ along their interface is ensured by $(4.6)_2$. The rank-one convexity condition (4.5) rules out the development of local fine-scale microstructures in the form of first-order laminates defined by a rank-one deformation tensor. The material is stable if the superimposed first-order laminate-type fluctuation field of Fig. 3(c) yields a higher energy level than the homogeneous deformation $\mathbf{F}_{n+1}$ of Fig. 3(a). A qualitative picture of a non-convex, unstable incremental response is given in Fig. 4. Observe carefully that (4.5) is a *global stability criterion* that needs knowledge of the global range of instability between $\mathbf{F}^-$ and $\mathbf{F}^+$. The material stability cannot be directly decided in terms of a given local deformation $\mathbf{F}_{n+1}$, but needs the rank-one convex hull construction governed by $\mathbf{F}^-$ and $\mathbf{F}^+$. The *local form of the rank-one convexity condition* is the classical *Legendre-Hadamard* or *ellipticity condition*

$$(\mathbf{M} \otimes \mathbf{N}) : \partial^2_{\mathbf{FF}} W(\mathbf{F}_{n+1}) : (\mathbf{M} \otimes \mathbf{N}) \geq 0 \,, \tag{4.7}$$

in terms of the consistent tangent modulus for arbitrary unit vectors $\mathbf{M}$ and $\mathbf{N}$, see [26, 74]. As shown in [45, 44], classical conditions of material stability of elastic-plastic solids outlined in [73, 27, 69] are consistent with this local convexity condition, which is often motivated by considering wave propagation in solids. As shown in Fig. 4, the associated range of instability is different from the one predicted by the global condition (4.5). Recall that both conditions are mathematical definitions related to the existence of regular solutions of the variational problem (4.2). The question whether the global or local conditions (4.5) and (4.7) are relevant depends on the physical ability of an inelastic solid material to develop deformation microstructures in the associated unstable ranges. This can only be clarified by experimental investigations.

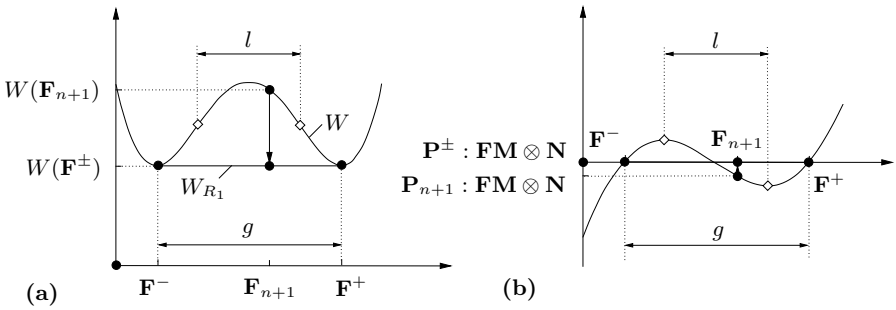


Fig. 4. Qualitative representation of a non-convex incremental stress potential and its convexification. l and g characterize the ranges where the *local* and the *global* convexity criterion are not satisfied, respectively. **(a)** At $\mathbf{F}_{n+1}$ the stress potential W is not rank-one convex (dashed). As a consequence, the macroscopic deformation state $\mathbf{F}_{n+1}$ is not stable and decomposes into micro-phases $\mathbf{F}^{\pm}$ which determine the rank-one convex envelope (solid). **(b)** The relaxed stress-strain relation characterizes a snap-through behavior between the micro-phases $\mathbf{F}^{\pm}$ due to the constant slope of the rank-one convex envelope

In what follows we rewrite the rank-one convexity condition (4.5) for two-dimensional problems. To this end, we introduce the ansatz

$$\mathbf{F}^{\pm} := \mathbf{F}_{n+1}\mathbf{L}^{\pm} \quad \text{with} \quad \begin{cases} \mathbf{L}^{+} := \boldsymbol{1} + (1-\xi) d\, \mathbf{M} \otimes \mathbf{N} \,, \\ \mathbf{L}^{-} := \boldsymbol{1} - \xi d\, \mathbf{M} \otimes \mathbf{N} \,, \end{cases} \tag{4.8}$$

for the two deformation phases that satisfy the conditions (4.6). It models a first-order laminate in terms of the two Lagrangian unit vectors $\mathbf{M}$ and $\mathbf{N}$, which correspond with those used in the Hadamard condition (4.7). For two-dimensional problems, these vectors can be parameterized by two angles φ and χ, i.e. $\mathbf{M}(\varphi) = [\cos\varphi \;\; \sin\varphi]^T$ and $\mathbf{N}(\chi) = [\cos\chi \;\; \sin\chi]^T$. The scalar d

describes the intensity of the bifurcation on the micro-scale. ξ is the volume fraction of the phase (+) and can be understood as a probability measure in the sense of [75]. Hence, for a two-dimensional description of the rank-one laminate, deformation microstructures are characterized by *four micro-variables* $\mathbf{q} = [\xi, d, \varphi, \chi]^T \in \mathcal{Q}$, which are constrained to lie in the admissible domain $\mathcal{Q} := \{\mathbf{q} \mid 0 \leq \xi \leq 1\,,\ d \geq 0\,,\ 0 \leq \varphi \leq \pi\,,\ 0 \leq \chi \leq \pi\}$. With this notation at hand, we write the global rank-one convexity condition (4.5) for two-dimensional problems as the minimization problem

$$W(\mathbf{F}_{n+1}) \leq \inf_{\mathbf{q}\in\mathcal{Q}} \{\bar{W}^h(\mathbf{F}_{n+1}, \mathbf{q})\}\ , \tag{4.9}$$

in terms of the function

$$\bar{W}^h(\mathbf{F}_{n+1}, \mathbf{q}) = \xi W(\mathbf{F}^+(\mathbf{F}_{n+1}, \mathbf{q})) + (1-\xi) W(\mathbf{F}^-(\mathbf{F}_{n+1}, \mathbf{q}))\ , \tag{4.10}$$

that represents the volume average of the potentials in the two deformation phases. Figure 4(a) provides a visual demonstration for a non-convex incremental stress potential W. The incremental stress potential $W(\mathbf{F}_{n+1})$ is greater than the interpolation of the potentials $W(\mathbf{F}^+)$ and $W(\mathbf{F}^-)$ of the phases. As a consequence, the homogeneous deformation state is not stable and decomposes into the micro-deformations $\mathbf{F}^{\pm}$ which minimize the function $\bar{W}^h$. In a typical incremental analysis of an inelastic solid, the accompanying check of incremental rank-one convexity in $[t_n, t_{n+1}]$ needs the solution of the local minimization problem (4.9)

$$\inf_{\mathbf{q}\in\mathcal{Q}} \{\bar{W}^h(\mathbf{F}_{n+1}, \mathbf{q})\} \begin{cases} = W(\mathbf{F}_{n+1}) : \text{rank-one convex at } \mathbf{F}_{n+1} \\ < W(\mathbf{F}_{n+1}) : \text{not rank-one convex at } \mathbf{F}_{n+1} \end{cases}\ , \tag{4.11}$$

for the four variables $\mathbf{q}$ defined before. The necessary condition of the minimization problem

$$\bar{W}^h_{,\mathbf{q}} = \mathbf{0}\ , \tag{4.12}$$

is a nonlinear equation for the determination of the micro-variables $\mathbf{q}$. Note that $\bar{W}^h$ is not convex and for the solution of (4.12) the Newton iteration cannot be applied directly. We refer to [45, 44, 46] for solution procedures.

4.2 Relaxation of a Non-convex Constitutive Response

As pointed out in the recent papers [33, 45, 44, 46], the incremental variational formulation for the constitutive response opens up the opportunity to resolve the developing microstructure in non-stable standard dissipative materials by a relaxation of the associated non-convex incremental variational problem. If the above outlined material stability analysis detects a non-convex incremental stress potential W, an energy-minimizing *deformation microstructure* is assumed to develop such as indicated in Fig. 3. A *relaxation* is associated with a *convexification* of the non-convex function W by constructing its *convex envelopes* $W_{\mathcal{Q}}$. The convexification is concerned with the determination of a developing microstructure. This section develops a framework for a first-order rank-one relaxation of standard dissipative solids.

Quasi-convexified Relaxed Incremental Variational Problem

If material instabilities are detected at a point $\mathbf{X} \in \mathcal{B}$ of the solid by a failure of conditions (4.5) or (4.9), we face a non-convexity of the incremental potential W in some region of the inelastic solid. If the incremental potential function W is not quasiconvex, the internal part of the functional (4.1) is assumed to be not sequentially weakly lower semicontinuous. Then the existence of solutions of (4.2) is not ensured. In other words, the minimum of the incremental boundary-value problem (4.2) is not attained. Following [13, 1] we consider the *relaxed energy functional*

$$\mathcal{E}_Q(\boldsymbol{\varphi}_{n+1}) = \int_{\mathcal{B}} W_Q(\mathbf{F}_{n+1})\, \mathrm{d}x - [\, \Pi_{\mathrm{ext}}(\boldsymbol{\varphi}_{n+1}) - \Pi_{\mathrm{ext}}(\boldsymbol{\varphi}_n)\,] \,, \tag{4.13}$$

where the internal part of the relaxed energy functional is obtained by replacing the non-convex integrand W in (4.1) by its *quasiconvex envelope* W_Q. The current deformation field of the elastic-plastic solid is then determined by the relaxed incremental variational principle

$$\mathcal{E}_Q(\boldsymbol{\varphi}^*_{n+1}) = \inf_{\boldsymbol{\varphi}_{n+1} \in \mathcal{W}} \mathcal{E}_Q(\boldsymbol{\varphi}_{n+1}) \,, \tag{4.14}$$

that *minimizes the relaxed incremental potential energy* $\mathcal{E}_Q$. The quasiconvexified incremental stress potential W_Q is defined by the minimization problem

$$W_Q(\mathbf{F}_{n+1}) = \inf_{\mathbf{w} \in \mathcal{W}_0} \frac{1}{|D|} \int_D W(\mathbf{F}_{n+1} + \nabla \mathbf{w}(\mathbf{y}))\mathrm{d}x \tag{4.15}$$

with respect to the microscopic fluctuation field $\mathbf{w}$ that constitutes the development of a deformation microstructure, subject to a boundary condition providing a support on ∂D. The first and second derivatives of the relaxed potential W_Q function define relaxed stresses and tangent moduli

$$\bar{\mathbb{P}}_{n+1} := \partial_{\mathbf{F}} W_Q(\mathbf{F}_{n+1}) \quad \text{and} \quad \bar{\mathbb{C}}_{n+1} := \partial^2_{\mathbf{F}\mathbf{F}} W_Q(\mathbf{F}_{n+1}) \,. \tag{4.16}$$

The relaxed problem (4.14) is considered to be a *well-posed problem* as close as possible to the unstable problem (4.2). The minimization problem (4.15) determines a micro-fluctuation field $\mathbf{w}$ as indicated in Fig. 3(b). However, as already mentioned the basic difficulty is the detection of relevant functions $\mathbf{w}$ which define the minimizing microstructure.

Rank-One-Convexified Relaxed Incremental Variational Problem

A failure of rank-one convexity conditions (4.5) or (4.9) indicates the instability of the homogeneous deformation state $\mathbf{F}_{n+1}$ and the development of a pattern of first- and higher-order laminates as indicated in Fig. 3(c). We consider the *relaxed energy functional*

$$\mathcal{E}_R(\boldsymbol{\varphi}_{n+1}) = \int_{\mathcal{B}} W_R(\mathbf{F}_{n+1})\, \mathrm{d}x - [\, \Pi_{\mathrm{ext}}(\boldsymbol{\varphi}_{n+1}) - \Pi_{\mathrm{ext}}(\boldsymbol{\varphi}_n)\,] \,, \tag{4.17}$$

where the internal part of the relaxed energy functional is obtained by replacing the non-convex integrand W in (4.1) by its *rank-one-convex envelope* W_R, which is considered to be close to the quasi-convex envelope W_Q. The current deformation field of the elastic-plastic solid is then determined by the relaxed incremental variational principle

$$\mathcal{E}_R(\boldsymbol{\varphi}^*_{n+1}) = \inf_{\boldsymbol{\varphi}_{n+1} \in \mathcal{W}} \mathcal{E}_R(\boldsymbol{\varphi}_{n+1}) \,, \tag{4.18}$$

that *minimizes the relaxed incremental potential energy* $\mathcal{E}_R$ for the admissible deformation field. In [30] a construction was proposed to characterize the rank-one convexification based on a recursion formula. Starting with $W_{R_0}(\mathbf{F}_{n+1}) = W(\mathbf{F}_{n+1})$, one computes the functions

$$W_{R_k}(\mathbf{F}_{n+1}) = \inf_{\xi^+, \xi^-, \mathbf{F}^+, \mathbf{F}^-} \{\xi^+ W_{R_{k-1}}(\mathbf{F}^+) + \xi^- W_{R_{k-1}}(\mathbf{F}^-)\} \quad \text{with} \quad k \geq 1 \,, \tag{4.19}$$

for the scales $k = 1, 2, 3....$ After an infinite number of steps $k \to \infty$ the exact rank-one convexified incremental stress potential

$$W_R(\mathbf{F}_{n+1}) = \lim_{k \to \infty} W_{R_k}(\mathbf{F}_{n+1}) \tag{4.20}$$

is obtained. Similar to (4.16), relaxed stresses and tangent moduli are defined as $\bar{\mathbb{P}}_{n+1} := \partial_{\mathbf{F}} W_R(\mathbf{F}_{n+1})$ and $\bar{\mathbb{C}}_{n+1} := \partial^2_{\mathbf{F}\mathbf{F}} W_R(\mathbf{F}_{n+1})$. According to recursive approach by [30] any phase of order $k-1$ decomposes into two phases $(+)$ and $(-)$ of order k and minimizes the average of the corresponding incremental stress potentials. The developing micro-phases form a *sequential laminate*.

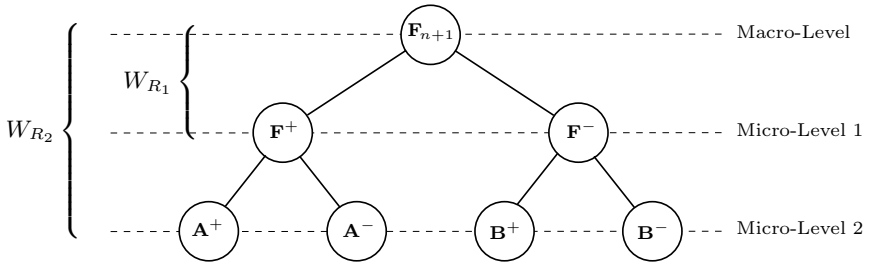


Fig. 5. Rank-one convexification and development of sequential laminates. The rank-one convexification $W_{R_k}(\mathbf{F}_{n+1})$ based on Kohn-Strang's recursion formula implies the development of a sequential laminate. Starting from the homogeneous deformation state $\mathbf{F}_{n+1}$ any phase of level $k-1$ decomposes into two phases $(+)$ and $(-)$ of level k. As a consequence, a typical binary tree structure emerges

Table 2. First-Order Rank-One Convexification of Incremental Response

1. Database $\{\mathbf{F}_{n+1}, \boldsymbol{\mathcal{I}}_n^+, \boldsymbol{\mathcal{I}}_n^-\}$ and starting value $\mathbf{q}_0 := \{\xi, d, \mathbf{N}, \mathbf{M}\}_0$ given.
2. Set micro-deformation phases
$$\mathbf{F}^{\pm} := \mathbf{F}_{n+1}\mathbf{L}^{\pm} \quad \text{with} \quad \begin{cases} \mathbf{L}^+ := \mathbf{1} + (1-\xi) d\, \mathbf{M} \otimes \mathbf{N} \,, \\ \mathbf{L}^- := \mathbf{1} - \xi d\, \mathbf{M} \otimes \mathbf{N} \,. \end{cases}$$
3. Evaluate the potential $\bar{W}^h(\mathbf{F}_{n+1}, \mathbf{q}) = \xi W(\mathbf{F}^+) + (1-\xi) W(\mathbf{F}^-)$ and its derivatives $\bar{W}^h_{,\mathbf{F}}$, $\bar{W}^h_{,\mathbf{q}}$, $\bar{W}^h_{,\mathbf{FF}}$, $\bar{W}^h_{,\mathbf{qq}}$, $\bar{W}^h_{,\mathbf{q}\bar{\mathbf{F}}}$.
4. Convergence check: If $(\| \bar{W}^h_{,\mathbf{q}} \| \leq \text{tol})$ go to 6.
5. Newton update of micro-variables $\mathbf{q} \Leftarrow \mathbf{q} - [\, \bar{W}^h_{,\mathbf{qq}} \,]^{-1} [\, \bar{W}^h_{,\mathbf{q}} \,]$.
6. Set relaxed macro-stresses and tangent macro-moduli
$$\bar{\mathbb{P}}_{n+1} = \bar{W}^h_{,\bar{\mathbf{F}}} \quad \text{and} \quad \bar{\mathbb{C}}_{n+1} = \bar{W}^h_{,\bar{\mathbf{F}}\bar{\mathbf{F}}} - [\, \bar{W}^h_{,\bar{\mathbf{F}}\mathbf{q}} \,][\, \bar{W}^h_{,\mathbf{qq}} \,]^{-1} [\, \bar{W}^h_{,\mathbf{q}\bar{\mathbf{F}}} \,] \,.$$

Figure 5 shows the typical binary tree structure of a rank-2 laminate. The unstable macroscopic deformation state $\mathbf{F}_{n+1}$ decomposes into two micro-phases $\mathbf{F}^+$ and $\mathbf{F}^-$ of micro-level 1 which again split into two pairs of micro-phases $\mathbf{A}^+$, $\mathbf{A}^-$ and $\mathbf{B}^+$, $\mathbf{B}^-$ of micro-level 2. The rank-one convexified potential W_{R_2} then consists of the volume average of the stress potentials W at the root of the tree, i.e. $W_{R_2}(\mathbf{F}_{n+1}) = \xi^{\mathbf{F}^+}[\xi^{\mathbf{A}^+} W(\mathbf{A}^+) + \xi^{\mathbf{A}^-} W(\mathbf{A}^-)] + \xi^{\mathbf{F}^-}[\xi^{\mathbf{B}^+} W(\mathbf{B}^+) + \xi^{\mathbf{B}^-} W(\mathbf{B}^-)]$. In the context of subgrain dislocation structures in single crystal plasticity, [66, 67] relax the incremental constitutive description of the material based on the explicit construction of microstructures by recursive lamination and their subsequent equilibration. However, they apply, based on physical arguments, a strong approximation by freezing the orientation of the laminates and the volume fractions during the deformation process. Such a strong assumption has also been applied by [45, 44] for the analysis of microstructure development in strain-softening von Mises plasticity. In contrast to these approaches, in [46, 2] a rank-one convexification has been proposed that determines both the developing orientation of the laminates as well as the volume fraction.

First-Order Rank-One-Convexified Incremental Problem

We approximate the exact rank-one convexification procedure outlined above by a *two-phase analysis* that takes into account only the first micro-level of Fig. 5. Hence, an unstable macro-deformation $\mathbf{F}_{n+1}$ decomposes into the two phases $\mathbf{F}^+$ and $\mathbf{F}^-$ modeled by ansatz (4.8). Then the first-order rank-one convexification of the non-convex function W is obtained for two-dimensional problems by the minimization problem

$$W_{R_1}(\mathbf{F}_{n+1}) = \inf_{\mathbf{q} \in \mathcal{Q}} \bar{W}^h(\mathbf{F}_{n+1}, \mathbf{q}) \,, \tag{4.21}$$

for the function $\bar{W}^h$ defined in (4.10) with respect to the set of micro-variables $\mathbf{q}$. A problem similar to (4.21) was solved in [33] for a one-dimensional strain-softening elastic-plastic bar. The solution of the minimization problem (4.21) yields solutions of ξ, d, φ, χ, which in the two-dimensional context determine two stable phases. The relaxed stresses and moduli are obtained by evaluation of derivatives of the function (4.10) with respect to $\mathbf{F}$. The first derivative of (4.21) with respect to the deformation $\mathbf{F}_{n+1}$ at the solution point $\mathbf{q}^*$ reads

$$\partial_{\mathbf{F}} W_{R_1} = \bar{W}^h_{,\mathbf{F}} + [\,\bar{W}^h_{,\mathbf{q}}\,][\,\mathbf{q}_{,\mathbf{F}}\,] \,. \tag{4.22}$$

Here, the last term vanishes due to the necessary condition (4.12) of the minimization problem. Thus we identify the macro-stresses

$$\bar{\mathbb{P}}_{n+1} = \bar{W}^h_{,\mathbf{F}} \,. \tag{4.23}$$

The second derivative of the potential reads

$$\partial^2_{\mathbf{FF}} W_{R_1} = \bar{W}^h_{,\mathbf{FF}} + [\,\bar{W}^h_{,\mathbf{Fq}}\,][\,\mathbf{q}_{,\mathbf{F}}\,] \,. \tag{4.24}$$

Here, the sensitivity of the fluctuation with respect to the macro-deformation is obtained by taking the linearization of (4.12), i.e. $\mathbf{q}_{,\mathbf{F}} = -[\bar{W}^h_{,\mathbf{qq}}]^{-1}[\bar{W}^h_{,\mathbf{qF}}]$. Insertion into (4.24) finally specifies the relaxed moduli to

$$\bar{\mathbb{C}}_{n+1} = \bar{W}^h_{,\mathbf{FF}} - [\,\bar{W}^h_{,\mathbf{Fq}}\,][\,\bar{W}^h_{,\mathbf{qq}}\,]^{-1}[\,\bar{W}^h_{,\mathbf{qF}}\,] \,. \tag{4.25}$$

Observe that the relaxed moduli consist of the volume average of the moduli of the phases and a softening part. The latter is the consequence of the flexibility of the rank-one laminate due to the phase decay. The algorithm of first-order rank-one convexification is summarized in Table 2.

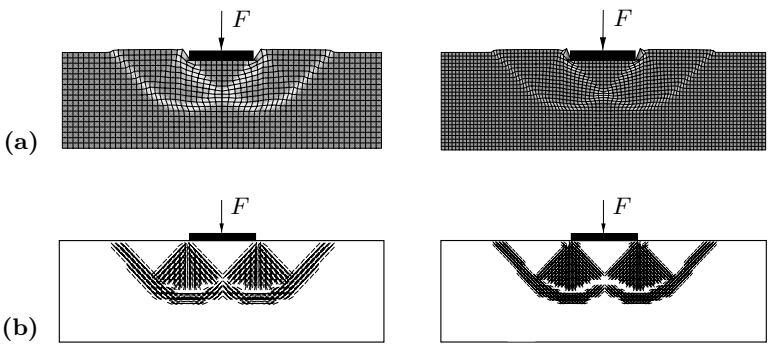


Fig. 6. Localization of Indentation Test at Plane Strain. Comparison of the 30×18 and the 45×27 element meshes. **(a)** Deformed meshes with equivalent plastic strains, **(b)** Relevant localization directions

5 Relaxation of Strain Softening Isotropic Plasticity

The relaxation technique outlined in Sect. 4 is applied to the treatment of shear band localizations in strain-softening isotropic elastoplasticity. The softening response of the model causes localization phenomena which is interpreted as microstructure developments on multiple scales associated with non-convex incremental stress potentials. Strain softening inelastic materials with non-convex incremental stress potentials have been investigated in the context of one-dimensional elastic-plastic bar in [33], and in isochoric damage mechanics in [21].

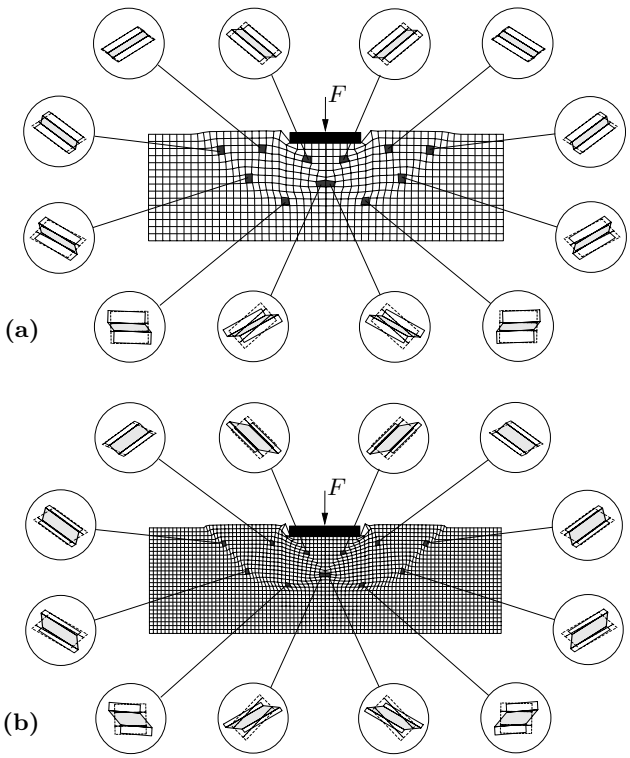


Fig. 7. Indentation Test at Plane Strain. Deformed mesh with zoomed-out microstructures of shaded elements at the center Gauß point for the **(a)** 25×15 and **(b)** 45×27 element mesh

The main goals of the numerical investigations are the analysis of the *developing microstructures* and the demonstration of the *mesh-invariance* of the relaxation technique proposed. We refer to [45, 44] for details of the relaxation algorithm. The elastic energy storage function has the following form

$$\psi(\mathbf{F}_{\rm el},\alpha)=\frac{\mu}{2}[\,\|\mathbf{F}_{\rm el}\|^2-3\,]+\frac{\mu^2}{\lambda}[\,J^{-\lambda/\mu}-1\,]+\frac{1}{2}h\alpha^2\;, \tag{5.1}$$

with $J:=\det\mathbf{F}_{\rm el}=\det\mathbf{F}$, the shear modulus $\mu>0$, the Lamé constant $\lambda>0$ and the softening modulus $h<0$. The level set function is given as

$$\mathbb{E}=\{\,(\boldsymbol{\Sigma},\beta)\,|\;\|\boldsymbol{\Sigma}\|+\sqrt{\frac{2}{3}}\beta\le c\,\}\;, \tag{5.2}$$

where $\boldsymbol{\varphi}$ is the Mandel stress, β is the conjugate force to the hardening variable α and c is a material parameter. Then the dissipation function for the isotropic von Mises plasticity with softening can be formulated as

$$\phi(\mathbf{L}_{\rm pl},\dot\alpha)=\sup_{(\boldsymbol{\Sigma},\beta)\in\mathbb{E}}\{\boldsymbol{\Sigma}:\mathbf{L}_{\rm pl}+\beta\dot\alpha\}\;, \tag{5.3}$$

in terms of the plastic velocity gradient $\mathbf{L}_{\rm pl}:=\dot{\mathbf{P}}\mathbf{P}^{-1}$ and the rate of hardening variable $\dot\alpha$.

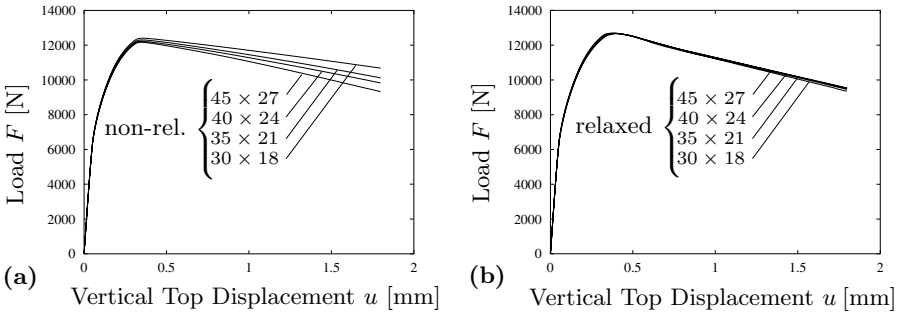


Fig. 8. Localization of Indentation Test at Plane Strain. Load displacement curves for different finite element meshes **(a)** with non-relaxed formulation and **(b)** proposed relaxation technique

Here, we approximate the minimization problem (4.21) by introduction of an a priori length scale δ representing the width of micro-shearband. Then, in the finite element context the volume fraction ξ at each integration point is described as a function of the length scale δ and a characteristic geometric parameter g of the finite element. A further simplification to the minimization problem is obtained by fixing the laminate orientation angle χ to the critical direction $\theta_{\rm cr}$ obtained from the acoustic tensor

$$\mathbf{Q}(\alpha):=\mathbf{L}(\alpha)\cdot\partial^2_{\mathbf{FF}}W(\mathbf{F}_{n+1})\cdot\mathbf{L}(\alpha)\;, \tag{5.4}$$

where $\mathbf{L}(\alpha)=[\cos(\alpha)\;\sin(\alpha)]^T$ is a unit vector to describe the localization direction. The material stability is controlled by the following minimization problem for the determinant of the acoustic tensor

$$\min_{\alpha}\{\det[\mathbf{Q}(\alpha)]\} \begin{cases} > 0: & \text{stable at } \mathbf{F}_{n+1} \\ \leq 0: & \text{unstable at } \mathbf{F}_{n+1} \end{cases}, \tag{5.5}$$

and if the determinant becomes zero or negative then the critical angle θ_{cr} and the laminate orientation χ are determined as

$$\chi = \theta_{\mathrm{cr}} = \arg\{\min_{\alpha}[\det \mathbf{Q}(\alpha)]\} . \tag{5.6}$$

Furthermore we consider $\mathbf{M} \cdot \mathbf{N} = 0$ which characterizes a shear band type failure. Then, the approximated relaxed energy is obtained by a minimization with respect to one scalar variable d,

$$W_{R_1}(\mathbf{F}_{n+1}) = \inf_{d} [\, \xi W^{+}(\mathbf{F}_{n+1}, d) + (1-\xi) W^{-}(\mathbf{F}_{n+1}, d) \,] . \tag{5.7}$$

Having computed W_{R_1} the relaxed stresses $\bar{\mathbb{P}}$ and the relaxed moduli $\bar{\mathbb{C}}$ can be computed from (4.23) and (4.25), respectively.

As a representative example, we consider next a plane strain indentation test where a localization zone in the form of curved shear bands is observed experimentally. The equivalent plastic strains and the formation of shear bands with corresponding localization directions are plotted in Fig. 6. In Fig. 7 the development of microstructures is visualized at the selected integration points for two discretizations. In order to prove the mesh objectivity of the proposed relaxation algorithm load-deflection curves are plotted in Fig. 8 for four different mesh densities. The non-relaxed formulation in Fig. 8(a) shows a clear mesh dependence whereas the proposed relaxation algorithm in Fig. 8(b) exhibits no mesh dependence in the post-critical regime.

6 Relaxation of Non-convex Single-Slip Plasticity

We now point out details for the first-order rank-one convexification analysis introduced in Sect. 4 for the model problem of single slip plasticity. Contrary to the strain softening example discussed in Sect. 5, the non-convexity appears in the single-slip plasticity as a result of geometric constraints related to the orientation of a slip-system. The model problem of single slip crystal plasticity has already been investigated in several works, see [8, 5, 53, 12, 11, 46].

The main goals of the numerical investigations are the analysis of the *developing microstructures* and the demonstration of the *mesh-invariance* of the relaxation technique proposed. We refer to [46] for details of the relaxation algorithm based on first-order rank-one convexification. As a concrete form, we apply a compressible Neo-Hookean material

$$\psi(\mathbf{F}_{\mathrm{el}}) = \frac{\mu}{2}\Big[\|\mathbf{F}_{\mathrm{el}}\|^2 - 3] + \frac{\kappa}{4}[J^2 - 2(1 + 2\frac{\mu}{\kappa}) \ln J - 1\Big] , \tag{6.1}$$

with $J := \det\mathbf{F}_{\mathrm{el}} = \det\mathbf{F}$. $\kappa > 0$ and $\mu > 0$ denote the bulk and the shear moduli, respectively. The dissipation function for the linear hardening model of single-slip plasticity is

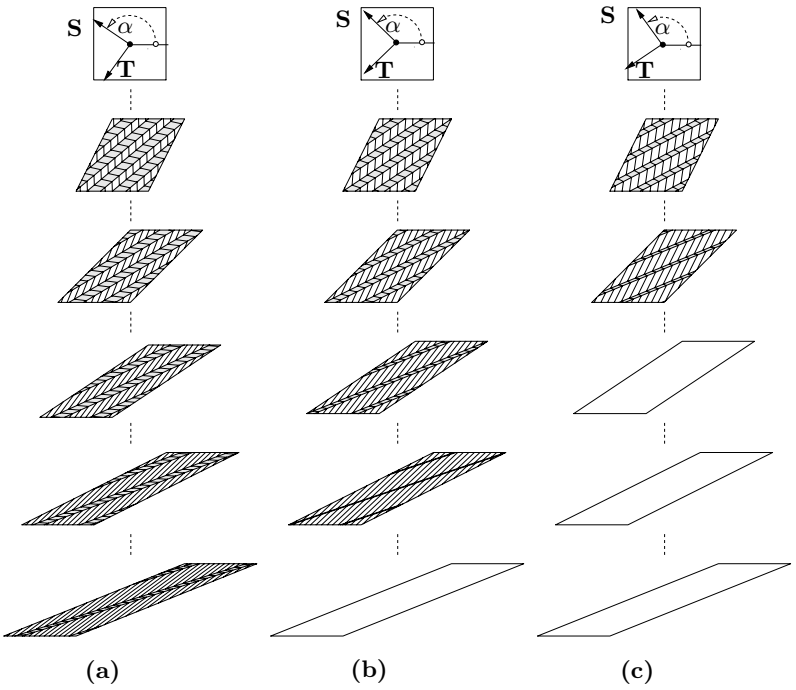


Fig. 9. Simple shear test. Comparison of evolution of microstructures for simple shear tests with three different slip system **(a)** $\alpha = 145^\circ$, **(b)** $\alpha = 135^\circ$, **(c)** $\alpha = 125^\circ$. After loss of material stability microstructures develop which are modeled as first-order rank-one laminates

$$\phi(\mathbf{L}_{\rm pl}) = [\tau_0 + h\gamma] \; |\mathbf{L}_{\rm pl} : (\mathbf{S} \otimes \mathbf{T})| \; , \tag{6.2}$$

in terms of the Schmid stress τ associated with the slip system of single-slip plasticity and the linear hardening modulus h. The slip system is described by the slip direction $\mathbf{S}$ and the slip normal $\mathbf{T}$ with $\mathbf{S} \cdot \mathbf{T} = 0$.

Here, a key contribution is the derivation of a *semi-analytical solution* that reduces for two-dimensional problems the independent micro-variables from four in $\mathbf{q}$ to just one variable. Recall the necessary conditions (4.12) of the minimization problem of relaxation

$$\left.\begin{aligned}
\bar{W}^h_{,\xi} &= W^+ - W^- - d\,[\xi\mathbb{P}^+ + (1-\xi)\mathbb{P}^-] : (\mathbf{FM} \otimes \mathbf{N}) = 0 \\
\bar{W}^h_{,d} &= \xi(1-\xi) \quad [\mathbb{P}^+ - \mathbb{P}^-] : (\mathbf{FM} \quad \otimes \mathbf{N}) \quad = 0 \\
\bar{W}^h_{,\varphi} &= \xi(1-\xi)\; d\,[\mathbb{P}^+ - \mathbb{P}^-] : (\mathbf{FM}_{,\varphi} \otimes \mathbf{N}) \quad = 0 \\
\bar{W}^h_{,\chi} &= \xi(1-\xi)\; d\,[\mathbb{P}^+ - \mathbb{P}^-] : (\mathbf{FM} \quad \otimes \mathbf{N}_{,\chi}) \quad = 0
\end{aligned}\right\} \; , \tag{6.3}$$

in terms of the four micro-variables $\mathbf{q} := [\xi, d,\ \varphi, \chi]^T$. Note that the first two conditions in (6.3) are the physical and the configurational force equilibrium

conditions on the interface between two phases. In the sequel, we will evaluate these conditions and derive a semi-analytical solution for the minimizing laminate $\mathbf{F}^{\pm}$. The plastic deformation $\mathbf{P}^{\pm}$ and the hardening variable in the

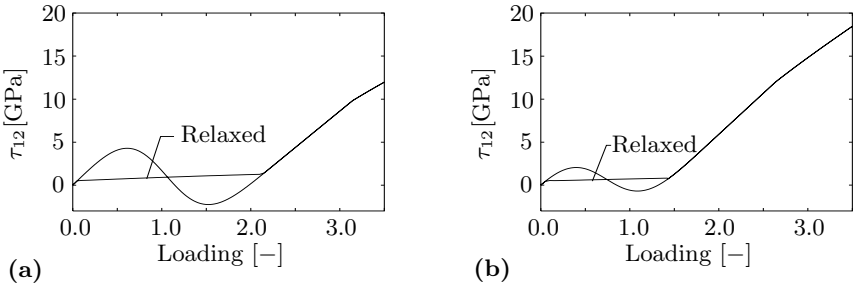


Fig. 10. Simple shear test. Comparison of relaxed and non-relaxed Kirchhoff stress components for simple shear tests with two different slip systems. **(a)** $\alpha = 135°$, **(b)** $\alpha = 125°$. The shape of the governing stress coordinate τ_{12} represents a snap-through behavior within the non-convex range. After recovery of the stable homogeneous state the relaxed and the non-relaxed stress responses coincide again

phases ($\pm$) are denoted

$$\mathbf{P}^{\pm} = \mathbf{P}^{\star}(\mathbf{1} \pm \Delta\gamma^{\pm}\mathbf{S} \otimes \mathbf{T}) \quad \text{and} \quad \gamma^{\pm} = \gamma^{\star} + \Delta\gamma^{\pm} \,, \tag{6.4}$$

where $\Delta\gamma^{\pm} = (\gamma - \gamma_n)^{\pm}$ are the incremental plastic arc lengths. $\mathbf{P}^{\star}$ and $\gamma^{\star}$ are the plastic deformation and $\gamma^{\star}$ the hardening variable of the last stable homogeneous state, respectively. Equation (6.4) points out the cause of the phase decay for the model problem of single slip plasticity that results from the *bifurcation of the plastic deformation* starting from $\mathbf{P}^{\star}$ with $\Delta\gamma^{\pm}$. The equilibrium of the Schmid stresses $\tau^{+} = \tau^{-}$ yields the identity $\Delta\gamma^{+} = \Delta\gamma^{-} = \Delta\gamma$ of the incremental slips. If one postulates the preservation of the volumetric deformation $\det[\mathbf{F}^{+}] = \det[\mathbf{F}^{-}] = \det[\mathbf{F}]$ it turns out that the Lagrangian laminate vectors are orthogonal, i.e. $\mathbf{N} \cdot \mathbf{M} = 0$. This result allows for the parameterization of these vectors in terms of the vectors of the slip system $\mathbf{N} = \cos\theta\ \mathbf{S} - \sin\theta\ \mathbf{T}$ and $\mathbf{M} = \sin\theta\ \mathbf{S} + \cos\theta\ \mathbf{T}$ where θ is an in-plane orientation angle. Exploitation of these results leads to the identification of the inclination angle and a formula for the micro-intensity

$$\tan\theta = -\mathbf{P}^{\star} : \mathbf{S} \otimes \mathbf{T} \quad \text{and} \quad d = \frac{2\Delta\gamma}{\cos^2\theta\ (1 + \Delta\gamma^2)} \,. \tag{6.5}$$

Insertion of the above obtained results in the necessary conditions $(6.3)_{1,4}$ yields an expression for the volume fraction

$$\xi = \frac{1}{2} + d^{-1} \left[\frac{c_{NM}}{c_{MM}} + \tan\theta \right] \,, \tag{6.6}$$

where we have introduced the abbreviation $c_{XY} = \mathbf{X} \cdot \mathbf{C} \cdot \mathbf{Y}$. The incremental plastic multiplier $\Delta\gamma$ can be determined by algebraic manipulations as

$$\Delta\gamma = \frac{2d + E}{\cos^2\theta \; d^2 + F} \; , \tag{6.7}$$

in terms of the coefficients $E = -4(h\gamma^\star + c)/(\mu c_{MM})$ and $F = [4h/\mu + 4 \; \cos^2\theta \; (c_{NN} - c_{NM}^2/c_{MM})]/c_{MM}$. Note, that the incremental slip is only a function of the micro-intensity.

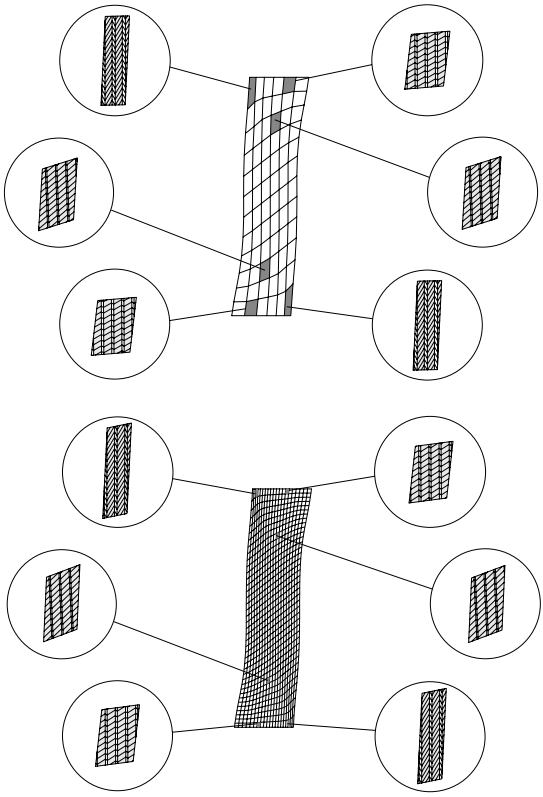


Fig. 11. Rectangular specimen in plane strain tension. Visualization of microstructure developments at selected Gauß points for discretizations with **(a)** 6×12 and **(b)** 20×40 elements

As a consequence, insertion of (6.7) into $(6.5)_2$ leads to a polynomial of degree five $p(d)$ which *only depends on the micro-intensity*. The solutions of the polynomial $p(d)$ is the relevant micro-intensity d^* that minimizes the volume average of the stress-potentials in the two micro-phases

$$d^* = \arg\{ \inf_{d \in \mathcal{D}} [\bar{W}^h] \; \} \quad \text{with} \quad \mathcal{D} \in \{d | p(d) = 0\} \; . \tag{6.8}$$

Having computed d, the volume fraction ξ, the relaxed stresses $\bar{\mathbb{P}}$ and the relaxed moduli $\bar{\mathbb{C}}$ can be computed from (6.6), (4.23) and (4.25), respectively.

First, we investigate a homogeneous simple shearing with different slip systems shown in Fig. 9. Because of a specific choice of the orientation of the slip-systems the material stability of the homogeneous deformation can be lost and microstructures may arise. The development of the first-order rank-one laminate type microstructures is plotted in Fig. 9 for various levels of deformation. During the macro deformation the plastic slip-systems start to rotate and align to the principal loading mode. The stronger the blocking of the principal deformation the longer the non-convex range. The shear component of the Kirchhoff stress for the relaxed and the non-relaxed solutions are plotted in Fig. 10 where the range of the non-convex domain is clearly dependent on the chosen slip system orientation.

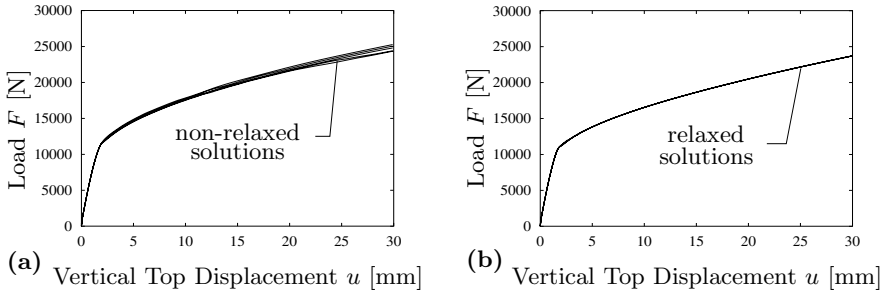


Fig. 12. Rectangular specimen in tension. Load-displacement curves for five different finite element meshes in terms of **(a)** the non-relaxed (non-objective) formulation (the finer the mesh the softer the response) **(b)** the proposed relaxation technique

Next, a plane strain tension test is considered where the slip direction vector is taken to be 10° counterclockwise from the horizontal. In Fig. 11 the development of microstructures is visualized for two different mesh densities. The specific orientation of the slip system causes the non-convex incremental potential which is relaxed by the proposed algorithm in terms of first-order laminates. In order to prove the mesh objectivity of the proposed relaxation algorithm load-deflection curves are plotted in Fig. 12 for different mesh densities. Although there is no softening in the model, the non-relaxed formulation shows mesh dependence due to non-convexity in the problem whereas the proposed relaxation algorithm exhibits no mesh dependence.

The final example is concerned with a rectangular specimen in shear where the slip direction is chosen to be 135° counterclockwise from the horizontal. In Fig. 13, the development of microstructures is visualized by considering two

different levels of deformation where the evolution of volume fractions and laminate orientations can be seen.

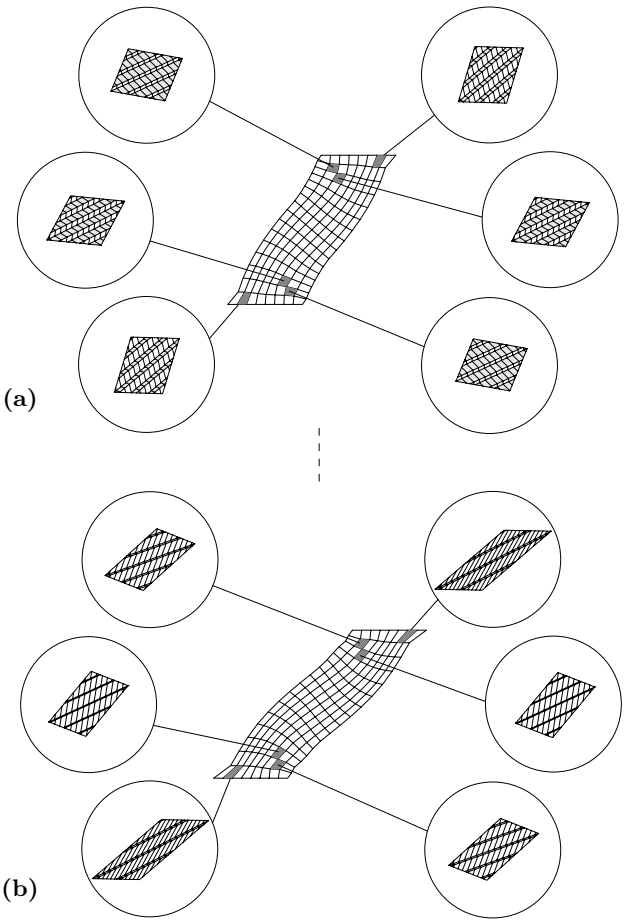


Fig. 13. Rectangular specimen in shear. Visualization of microstructure developments at selected Gauß points at **(a)** $u = 20\,\mathrm{mm}$ **(b)** $u = 30\,\mathrm{mm}$

7 Conclusions

The energetic formulation for finite-strain elastoplasticity has been proved as a very flexible mathematical tool that links the heavily used time-incremental minimization problem to a suitable weak time-continuous problem. Moreover,

the theory of the calculus of variations can be used to provide existence results for the incremental problem as well as for the time-continuous one. For the latter case we still need to assume spatial regularizations to prevent the formation of microstructure. At present the global existence theory has proved to be successful in the simplest situations, but further developments are needed to explore the capability of the method for providing classical solutions, i.e., without microstructure. Moreover, it will be essential to derive reliable and efficient numerical algorithms in the spirit of [32].

The energetic formulation has the major drawback that the stability condition is a *global* condition, whereas a *local* condition would be more physical and better for numerical purposes. First results to understand rate-independent systems as limits of systems with small viscosity are presented in [16], but this theory is restricted to finite-dimensional Hilbert spaces. Generalizations to infinite dimensions including abstract metric spaces are developed in [61], but their applicability in elastoplasticity is still out of reach.

References

1. E. Acerbi and N. Fusco. Semicontinuity problems in the calculus of variations. *Archive of Rational Mechanics and Analysis*, 83:125–145, 1984.
2. S. Aubry, M. Fago, and M. Ortiz. A constrained sequential-lamination algorithm for the simulation of sub-grid microstructure in martensitic materials. *Computer Methods in Applied Mechanics and Engineering*, 192:2823–2843, 2003.
3. J. M. Ball. Convexity conditions and existence theorems in nonlinear elasticity. *Archive of Rational Mechanics and Analysis*, 63:337–403, 1977.
4. J. M. Ball. Some open problems in elasticity. In P. Newton, P. Holmes, and A. Weinstein, editors, *Geometry, Mechanics, and Dynamics*, pages 3–59. Springer, New York, 2002.
5. S. Bartels, C. Carstensen, K. Hackl, and U. Hoppe. Effective relaxation for microstructure simulations: algorithms and applications. *Computer Methods in Applied Mechanics and Engineering*, 193:5143–5175, 2004.
6. M. Becker. *Incompatibility and Instability Based Size Effects in Crystals and Composites at Finite Elastoplastic Strains*. PhD thesis, Institut für Mechanik (Bauwesen), Report No. I-18, Universität Stuttgart, 2006.
7. M. A. Biot. *Mechanics of Incremental Deformations*. John Wiley & Sons Inc., New York, 1965.
8. C. Carstensen, K. Hackl, and A. Mielke. Non-convex potentials and microstructures in finite-strain plasticity. *Proceedings of the Royal Society London, Series A*, 458:299–317, 2002.
9. C. Carstensen and P. Plecháč. Numerical solution of the scalar double-well problem allowing microstructures. *Mathematics of Computation*, 66:997–1026, 1997.
10. P. G. Ciarlet. *Mathematical Elasticity*. Elsevier Science Publishers B.V., Amsterdam, 1988.
11. S. Conti and M. Ortiz. Dislocation microstructures and the effective behavior of single crystals. *Archive of Rational Mechanics and Analysis*, 176:103–147, 2005.

12. S. Conti and F. Theil. Single-slip elastoplastic microstructures. *Archive of Rational Mechanics and Analysis*, 178:125–148, 2005.
13. B. Dacorogna. *Direct Methods in the Calculus of Variations*. Springer-Verlag, Berlin Heidelberg, 1989.
14. A. DeSimone and G. Dolzmann. Material instabilities in nematic elastomers. *Physica D*, 136:175–191, 2000.
15. G. Dolzmann. *Variational Methods for Crystalline Microstructure – Analysis and Computation*. Springer-Verlag, Berlin Heidelberg, 2003.
16. M. Efendiev and A. Mielke. On the rate-independent limit of systems with dry friction and small viscosity. *Journal of Convex Analysis*, 13(1):151–167, 2006.
17. N. A. Fleck, G. M. Müller, M. F. Ashby, and J. Hutchinson. Strain gradient plasticity: theory and experiment. *Acta Materialia*, 42:475–487, 1994.
18. G. Francfort and A. Mielke. Existence results for a class of rate-independent material models with nonconvex elastic energies. *J. Reine Angew. Math.*, 2006. In print.
19. P. Germain. *Cours de Mécanique des Milieux Continus*. Masson et Cie, Paris, 1973.
20. S. Govindjee, A. Mielke, and G. J. Hall. The free-energy of mixing for n-variant martensitic phase transformations using quasi-convex analysis. *Journal of the Mechanics and Physics of Solids*, 50:1897–1922, 2002.
21. E. Gürses and C. Miehe. Relaxation analysis of material instabilities in damage mechanics based on incremental convexification techniques. *Submitted to International Journal of Solids and Structures*, 2006.
22. M. E. Gurtin. A gradient theory of single-crystal viscoplasticity that accounts for geometrically necessary dislocations. *Journal of the Mechanics and Physics of Solids*, 50:5–32, 2002.
23. K. Hackl. Generalized standard media and variational principles in classical and finite strain elastoplasticity. *Journal of the Mechanics and Physics of Solids*, 45(5):667–688, 1997.
24. K. Hackl and U. Hoppe. On the calculation of microstructures for inelastic materials using relaxed energies. In C. Miehe, editor, *IUTAM Symposium on Computational Mechanics of Solids at Large Strains*, pages 77–86. Kluwer, 2003.
25. K. Hackl, A. Mielke, and D. Mittenhuber. Dissipation distances in multiplicative elastoplasticity. In W. Wendland and M. Efendiev, editors, *Analysis and Simulation of Multifield Problems*, pages 87–100. Springer-Verlag, 2003.
26. J. Hadamard. *Leçons sur la propagation des ondes et les équations de l'hydrodynamique*. Hermann, Paris, 1903.
27. R. Hill. Acceleration waves in solids. *Journal of the Mechanics and Physics of Solids*, 10:1–16, 1962.
28. D. Knees and A. Mielke. Energy release rate for cracks in finite-strain elasticity. *Mathematical Methods in the Applied Sciences*, 2006. Submitted (WIAS Preprint 1100).
29. R. V. Kohn. The relaxation of a double-well problem. *Continuum Mechanics and Thermodynamics*, 3:193–236, 1991.
30. R. V. Kohn and G. Strang. Optimal design and relaxation of variational problems i, ii, iii. *Communications on Pure and Applied Mathematics*, 39:113–137, 139–182, 353–377, 1986.
31. A. Krawietz. *Materialtheorie: Mathematische Beschreibung des phänomenologischen thermomechanischen Verhaltens*. Springer-Verlag, Berlin, 1986.

32. M. Kružík, A. Mielke, and T. Roubíček. Modelling of microstructure and its evolution in shape-memory-alloy single-crystals, in particular in CuAlNi. *Meccanica*, 40:389–418, 2005.
33. M. Lambrecht, C. Miehe, and J. Dettmar. Energy relaxation of non-convex incremental stress potentials in a strain-softening elastic-plastic bar. *International Journal of Solids and Structures*, 40:1369–1391, 2003.
34. M. Luskin. On the computation of crystalline microstructure. *Acta Numerica*, 36:191–257, 1996.
35. A. Mainik and A. Mielke. Existence results for energetic models for rate-independent systems. *Calc. Var. PDEs*, 22:73–99, 2005.
36. A. Mainik and A. Mielke. An existence result for rate-independent strain-gradient plasticity at finite strains. In preparation, 2006.
37. J. E. Marsden and T. J. R. Hughes. *Mathematical Foundations of Elasticity.* Dover Publications Inc., New York, 1994.
38. J. B. Martin. *Plasticity. Fundamentals and General Results.* MIT press, Cambridge, Massachusetts, 1975.
39. C. Miehe. Strain-driven homogenization of inelastic microstructures and composites based on an incremental variational formulation. *International Journal for Numerical Methods in Engineering*, 55:1285–1322, 2002.
40. C. Miehe and N. Apel. Anisotropic elastic-plastic analysis of shells at large strains. A comparison of multiplicative and additive approaches to enhanced finite element design and constitutive modeling. *International Journal for Numerical Methods in Engineering*, 61:2067–2113, 2004.
41. C. Miehe, N. Apel, and M. Lambrecht. Anisotropic additive plasticity in the logarithmic strain space: Modular kinematic formulation and implementation based on incremental minimization principles for standard materials. *Computer Methods in Applied Mechanics and Engineering*, 191:5383–5425, 2002.
42. C. Miehe and M. Becker. Incompatibility based strain gradient crystal plasticity. *Submitted to Computer Methods in Applied Mechanics and Engineering*, 2006.
43. C. Miehe and E. Gürses. A robust algorithm for configurational-force-driven brittle crack propagation with r-adaptive mesh alignment. *Submitted to International Journal for Numerical Methods in Engineering*, 2006.
44. C. Miehe and M. Lambrecht. Analysis of microstructure development in shear-bands by energy relaxation of incremental stress potentials: Large-strain theory for standard dissipative solids. *International Journal for Numerical Methods in Engineering*, 58:1–41, 2003.
45. C. Miehe and M. Lambrecht. A two-scale finite element relaxation analysis of shear bands in non-convex inelastic solids: Small-strain theory for standard dissipative materials. *Computer Methods in Applied Mechanics and Engineering*, 192:473–508, 2003.
46. C. Miehe, M. Lambrecht, and E. Gürses. Analysis of material instabilities in inelastic solids by incremental energy minimization and relaxation methods: Evolving deformation microstructures in finite plasticity. *Journal of the Mechanics and Physics of Solids*, 52:2725–2769, 2004.
47. C. Miehe and J. Schotte. Anisotropic finite elastoplastic analysis of shells: Simulation of earing in deep-drawing of single- and polycrystalline sheets by taylor-type micro-to-macro transitions. *Computer Methods in Applied Mechanics and Engineering*, 193:25–57, 2004.

48. C. Miehe and J. Schotte. Crystal plasticity and evolution of polycrystalline microstructure. In E. Stein, R. de Borst, and J. R. Hughes, editors, *Encyclopedia of Computational Mechanics*, chapter 8, pages 267–289. John Wiley & Sons, 2004.
49. C. Miehe, J. Schotte, and M. Lambrecht. Homogenization of inelastic solid materials at finite strains based on incremental minimization principles. application to the texture analysis of polycrystals. *Journal of the Mechanics and Physics of Solids*, 50:2123–2167, 2002.
50. A. Mielke. Finite elastoplasticity, Lie groups and geodesics on SL(d). In P. Newton, A. Weinstein, and P. J. Holmes, editors, *Geometry, Dynamics, and Mechanics*, pages 61–90. Springer-Verlag, 2002.
51. A. Mielke. Energetic formulation of multiplicative elasto-plasticity using dissipation distances. *Continuum Mechanics and Thermodynamics*, 15:351–382, 2003.
52. A. Mielke. Evolution of rate-independent inelasticity with microstructure using relaxation and Young measures. In C. Miehe, editor, *IUTAM Symposium on Computational Mechanics of Solid Materials at Large Strains (Stuttgart Aug. 2001)*, pages 33–44. Kluwer, 2003.
53. A. Mielke. Deriving new evolution equations for microstructures via relaxation of variational incremental problems. *Computer Methods in Applied Mechanics and Engineering*, 193:5095–5127, 2004.
54. A. Mielke. Existence of minimizers in incremental elasto-plasticity with finite strains. *SIAM Journal on Mathematical Analysis*, 36:384–404, 2004.
55. A. Mielke. Evolution in rate-independent systems (ch. 6). In C. Dafermos and E. Feireisl, editors, *Handbook of Differential Equations, Evolutionary Equations, vol. 2*, pages 461–559. Elsevier B.V., 2005.
56. A. Mielke. Necessary and sufficient conditions for polyconvexity of isotropic functions. *Journal of Convex Analysis*, 12:291–314, 2005.
57. A. Mielke. A mathematical framework for generalized standard materials in the rate-independent case. In *SFB404-Abschlussband*. Springer-Verlag, 2006.
58. A. Mielke. Temperature-induced phase transformations in shape-memory alloys. In preparation, 2006.
59. A. Mielke and S. Müller. Lower semicontinuity and existence of minimizers for a functional in elastoplasticity. *ZAMM Zeitschrift für Angewandte Mathematik und Mechanik*, 86:233–250, 2006.
60. A. Mielke and M. Ortiz. A class of minimum principles for characterizing the trajectories of dissipative systems. *In preparation*, 2006.
61. A. Mielke, R. Rossi, and G. Savaré. On a metric approach to a class of rate-independent problems. *In preparation*, 2006.
62. A. Mielke and F. Theil. On rate-independent hysteresis models. *Nonlinear Partial Differential Equations and their Applications (NoDEA)*, 11:151–189, 2004. (Accepted July 2001).
63. A. Mielke, F. Theil, and V. I. Levitas. A variational formulation of rate-independent phase transformations using an extremum principle. *Archive of Rational Mechanics and Analysis*, 162:137–177, 2002.
64. C. B. Morrey. Quasiconvexity and the semicontinuity of multiple integrands. *Pacific Journal of Mathematics*, 2:25–53, 1952.

65. S. Müller. Variational models for microstructure and phase transisitions. In S. Hilderbrandt and M. Struwe, editors, *Calculus of Variation and Geometric Evolution Problems, Lecture Notes in Mathematics 1713*, pages 85–210. Springer Verlag, Berlin Heidelberg, 1999.
66. M. Ortiz and E. Repetto. Nonconvex energy minimization and dislocation structures in ductile single crystals. *Journal of the Mechanics and Physics of Solids*, 47(2):397–462, 1999.
67. M. Ortiz, E. A. Repetto, and L. Stainier. A theory of subgrain dislocation structures. *Journal of the Mechanics and Physics of Solids*, 48:2077–2114, 2000.
68. M. Ortiz and L. Stainier. The variational formulation of viscoplastic constitutive updates. *Computer Methods in Applied Mechanics and Engineering*, 171:419–444, 1999.
69. J. R. Rice. The localization of plastic deformation. In W. T. Koiter, editor, *Theoretical and Applied Mechanics*, pages 207–220. North-Holland, Amsterdam, 1976.
70. M. Šilhavý. *The Mechanics and Thermodynamics of Continuous Media*. Springer-Verlag, Berlin Heidelberg New York, 1997.
71. P. Steinmann. Views on multiplicative elastoplasticity and the continuum theory of dislocations. *International Journal of Engineering Science*, 34:1717–1735, 1996.
72. B. Svendsen. Continuum thermodynamic models for crystal plasticity including the effects of geometrically-necessary dislocations. *Journal of the Mechanics and Physics of Solids*, 50:1297–1329, 2002.
73. T. Y. Thomas. *Plastic Flow and Fracture in Solids*. Academic Press, London, 1961.
74. C. Truesdell and W. Noll. The nonlinear field theories of mechanics. In S. Flügge, editor, *Handbuch der Physik*, chapter Bd. III/3. Springer-Verlag, Berlin, 1965.
75. L. C. Young. *Lectures on the Calculus of Variations and Optimal Control Theory*. Saunders, London, 1969.
76. H. Ziegler. Some extremum principles in irreversible thermodynamics with application to continuum mechanics. In I. N. Sneddon and H. R., editors, *Progress in Solid Mechanics, Vol. IV*. Springer Verlag, Berlin Heidelberg, 1963.
77. H. Ziegler and C. Wehrli. The derivation of constitutive relations from the free energy and the dissipation function. In *Advances in applied mechanics, Vol. 25*, pages 183–237. Academic Press, Orlando, FL, 1987.

Nonconforming Discretization Techniques for Coupled Problems *

Bernd Flemisch and Barbara I. Wohlmuth

Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
flemisch@ians.uni-stuttgart.de
wohlmuth@ians.uni-stuttgart.de

Summary. Multifield problems yield coupled problem formulations for which nonconforming discretizations schemes and problem-adapted solvers can be used to develop efficient numerical algorithms. Of crucial importance are numerically robust transmission operators based on weak continuity conditions. This paper presents the construction of such operators by means of dual discrete Lagrange multipliers for higher order discretizations and for general quadrilateral triangulations of possibly curved interfaces. Various applications are considered, including aero-acoustics, elasto-acoustics, contact and heat transfer.

Keywords: Domain decomposition, non-matching grids, dual Lagrange multipliers, mortar finite elements, nonconforming discretizations

1 Introduction

The approximative solution of multifield problems is characterized by the necessity of being able to combine different model equations, discretizations, spatial and temporal scales, triangulations, and/or spatial dimensions. The main goal of the project C12 "Nonconforming Discretization Techniques for Coupled Problems" is to cope with this necessity by providing general construction principles for nonconforming discretization techniques based on a geometrical decomposition of the computational domain corresponding to the different interacting fields. The two most important parts for achieving this goal are a rigorous mathematical analysis of the underlying weak formulations and the development of efficient numerical algorithms for the solution of the resulting discrete problems.

Being one of the youngest projects within the Collaborative Research Center 404, C12 started in May 2002. At the beginning of the last funding period in

* Research Project C12 "Nonconforming Discretization Techniques for Coupled Problems"

January 2004, we could already resort to several mathematically well-founded discretization methods and solution strategies. However, they had been employed mostly for comparatively simple applications such as stationary and linear model equations in two space dimensions in combination with low order finite elements and very basic domain and interface geometries. The main task for the last funding period was to get rid of these shortcomings. We performed this task by developing and successfully applying nonconforming coupling schemes for the solution of transient problem settings such as elasto-dynamics, acoustic wave propagation, elasto-acoustics, heat conduction, and electro-magnetism, as well as for nonlinear model equations such as nonlinear structural mechanics including contact problems. Moreover, we could extend the framework of mortar finite elements using dual Lagrange multipliers from low order elements and straight interfaces towards higher order elements and curvilinear interfaces.

In the following report, we can provide a closer look only on some of the topics considered within this project. In particular, after presenting some model problems and the concept of dual Lagrange multipliers in Sect. 2, we address the extension of biorthogonal bases to higher order elements and general surface meshes in Sects. 3, 4 and 5. Section 6 is devoted to implementational issues, while in Sect. 7 we present several application examples dealing with aero-acoustics, elasto-acoustics, contact and heat transfer. In Sect. 8, we conclude by listing all articles which contributed to achieving the goals of project C12.

2 Variational Setting

We first introduce the model settings which we use for several of our numerical illustrations. After that, we illustrate why dual Lagrange multipliers are an important key to efficiently solve problems discretized by mortar finite elements.

2.1 Model Problems

For the ease of notation and to avoid technicalities, we restrict ourselves to the case of two non-overlapping open subdomains Ω^{m} and Ω^{s} sharing a common interface Γ, their union giving the global domain Ω, $\overline{\Omega} = \overline{\Omega^{\mathrm{m}}} \cup \overline{\Omega^{\mathrm{s}}}$. By taking into account the standard modifications at the cross-points or at the wire-basket of more than two subdomains, the following considerations apply analogously to decompositions into many subdomains, [4]. For scalar problems, we focus on Poisson's equation. In particular, we seek a scalar function u as the solution of

$$-\Delta u = f \text{ in } \Omega\,, \tag{1}$$

with appropriate boundary conditions on $\partial\Omega$. The Lagrange multiplier λ is chosen to be the normal flux through the interface Γ, i.e., $\lambda = -\partial u/\partial \mathbf{n}$,

with $\mathbf{n}$ denoting the unit outward normal vector field with respect to Ω^{s}. The spaces X^{s} and X^{m}, which are needed for the upcoming weak formulation, are subsets of $H^1(\Omega^{\mathrm{s}})$ and $H^1(\Omega^{\mathrm{m}})$, respectively, such that given Dirichlet conditions on the boundary of the global domain Ω are respected. The product space $X = X^{\mathrm{m}} \times X^{\mathrm{s}}$ is equipped with the broken H^1-norm. The Lagrange multiplier space M is associated with the dual of the trace space of X^{s} on Γ, equipped with the dual norm.

Additionally, we consider linear and nonlinear elasticity problems. For the linear setting, we consider the problem of finding a displacement vector field $\mathbf{u}$ such that

$$-\operatorname{div}\sigma(\mathbf{u}) = \mathbf{f} \text{ in } \Omega \,, \tag{2}$$

supplemented by boundary conditions, by the Saint–Venant Kirchhoff law

$$\sigma(\mathbf{u}) = \lambda_{\mathrm{L}}(\operatorname{tr}\varepsilon(\mathbf{u}))\mathrm{Id} + 2\mu_{\mathrm{L}}\varepsilon(\mathbf{u}) \,, \tag{3}$$

with the Lamé constants $\lambda_{\mathrm{L}}, \mu_{\mathrm{L}}$ and by the linearized strain tensor

$$\varepsilon(\mathbf{u}) = \frac{1}{2}(\operatorname{grad}\mathbf{u} + [\operatorname{grad}\mathbf{u}]^{\mathrm{T}}) \,. \tag{4}$$

Here, the Lagrange multiplier $\boldsymbol{\lambda}$ corresponds to the surface tractions on Γ, namely, $\boldsymbol{\lambda} = -\sigma(\mathbf{u})\mathbf{n}$. The spaces X and M consist of vector fields with component functions being in the corresponding spaces for the scalar case.

However, the validity of the linearized elasticity equations (2)–(4) is restricted to small strains and small deformations. In order to correctly capture large strains and deformations, we employ the Mooney–Rivlin law, [44],

$$\mathsf{S} = \lambda_{\mathrm{L}} s'(j) j \mathsf{C}^{-1} + \mu_{\mathrm{L}} \left((1 - c_m)(\mathsf{Id} - \mathsf{C}^{-1}) + c_m(\operatorname{tr}\mathsf{C}\,\mathsf{Id} - \mathsf{C} - \mathsf{C}^{-1}) \right) \,, \tag{5}$$

defining the second Piola–Kirchhoff stress tensor S with $\mathsf{F} = \mathsf{Id} + \operatorname{grad}\mathbf{u}$ the deformation gradient and $\mathsf{C} = \mathsf{F}^{\mathrm{T}}\mathsf{F}$ the right Cauchy–Green strain tensor, c_m a material constant, $j = \det(\mathsf{F})$ and $s(j) = (j^2 - 1 - 2\ln j)/4$. The basic equilibrium condition is given by

$$-\operatorname{div}(\mathsf{FS}) = \mathbf{f} \,, \tag{6}$$

constituting the partial differential equation to solve, which has to be complemented by appropriate boundary conditions.

The strong formulations (1) and (2)–(4) yield saddle point problems of the following structure, [3]: find a primal variable $u = (u^{\mathrm{m}}, u^{\mathrm{s}}) \in X$ and a Lagrange multiplier $\lambda \in M$ such that

$$a(u, v) + b(v, \lambda) = f(v) \,, \qquad v \in X \,, \tag{7a}$$

$$b(u, \mu) \qquad\quad = 0 \,, \qquad \mu \in M \,, \tag{7b}$$

with a bilinear form $a(\cdot,\cdot) = \sum_{k=\mathrm{m,s}} a_k(\cdot,\cdot)$ and a coupling bilinear form

$$b(v,\mu) = \langle [v], \mu \rangle_{M' \times M} \ , \tag{8}$$

where $[v] = v^{\mathrm{s}} - v^{\mathrm{m}}$ denotes the jump across the interface Γ, and $\langle \cdot, \cdot \rangle_{M' \times M}$ stands for the duality pairing on $M' \times M$. In case of (5)–(6), the form $a(\cdot,\cdot)$ in (7) is nonlinear in the first argument. Although we have reserved boldface to indicate vectorial quantities, we assume that, unless clarified explicitly, standard notation applies equally to corresponding scalar and vectorial quantities. We require that $a(\cdot,\cdot)$ is elliptic on the constrained space

$$V = \{ v \in X : b(v,\mu) = 0, \ \mu \in M \} \ ,$$

where in case of the Laplace operator, [21], and of the linear elasticity setting, [22], it is well known that the ellipticity constant does not depend on the number of subdomains. The saddle point problem (7) can be equivalently reformulated as the positive definite problem of finding $u \in V$ such that

$$a(u,v) = f(v) \ , \qquad v \in V \ . \tag{9}$$

2.2 Dual Lagrange Multipliers

For the numerical solution of (7), we employ the usual Galerkin approach. The required approximation of X and M by finite element spaces $X_h = X_h^{\mathrm{m}} \times X_h^{\mathrm{s}}$ and M_h is based on two triangulations $\mathcal{T}_{\mathrm{m}}$ of Ω^{m} and $\mathcal{T}_{\mathrm{s}}$ of Ω^{s}. Using superscripts, we indicate by $\mathcal{T}^{\mathrm{m}}$ and $\mathcal{T}^{\mathrm{s}}$ the corresponding surface grids meeting the interface Γ. The finite element nodes on $\mathcal{T}^{\mathrm{m}}$ and $\mathcal{T}^{\mathrm{s}}$ are called master and slave nodes, respectively, all remaining nodes are indicated as inner nodes. The discrete Lagrange multiplier space M_h is associated with the mesh $\mathcal{T}^{\mathrm{s}}$ on the slave side. The corresponding discretization of problem (7) can be written as

$$\begin{pmatrix} A_{\mathrm{ii}} & A_{\mathrm{im}} & A_{\mathrm{is}} & 0 \\ A_{\mathrm{mi}} & A_{\mathrm{mm}} & 0 & -M^{\mathrm{T}} \\ A_{\mathrm{si}} & 0 & A_{\mathrm{ss}} & D^{\mathrm{T}} \\ 0 & -M & D & 0 \end{pmatrix} \begin{pmatrix} u_h^{\mathrm{i}} \\ u_h^{\mathrm{m}} \\ u_h^{\mathrm{s}} \\ \lambda_h \end{pmatrix} = \begin{pmatrix} f^{\mathrm{i}} \\ f^{\mathrm{m}} \\ f^{\mathrm{s}} \\ 0 \end{pmatrix} \ , \tag{10}$$

where the subscripts i, m, and s represent the inner, master and slave nodes, respectively. We emphasize that, in this paper, we do not consider any modifications on $\partial\Gamma$ which are required if Γ meets a part of a Dirichlet boundary. The entries of the coupling matrices M and D are assembled from integrals of the form

$$(\phi_p^{\mathrm{m}}, \mu_q)_\Gamma \ , \quad \text{and} \quad (\phi_p^{\mathrm{s}}, \mu_q)_\Gamma \ , \tag{11}$$

respectively, where $(\cdot,\cdot)_\Gamma$ denotes the L^2-inner product on Γ, ϕ_p^k, $k = \mathrm{m}, \mathrm{s}$, indicates the scalar nodal basis function of the trace space $W_h^k = X_h^k|_\Gamma$ of the finite element space on $\mathcal{T}^k$ associated with the node p, and μ_q stands for the scalar basis function of the discrete Lagrange multiplier space M_h associated with the node q. There exist several possibilities for choosing the basis

functions spanning M_h, most of them with equal mathematical stability and approximation properties. In Fig. 1, four types of basis functions for a 1D interface using linear or bilinear finite elements are presented: (a) the standard ones coinciding with the trace space W_h^{s}, [4], (b) the dual ones spanned by piecewise linear discontinuous basis functions, [46], (c) the dual continuous ones where the discontinuous dual basis functions are modified by cubic polynomials [47], (d) the piecewise constant ones spanned by basis functions which are constant from one edge midpoint to the next. In particular, we speak of *dual* basis functions μ_q, if they satisfy the biorthogonality relation

$$(\phi_p^{\mathrm{s}}, \mu_q)_\Gamma = \delta_{pq}(\phi_p^{\mathrm{s}}, 1)_\Gamma \,. \tag{12}$$

From the examples mentioned above, (b) and (c) satisfy (12) while (a) and (d) don't. The importance of (12) comes into play when attempting to solve the discrete problem (10). There exist various possibilities for solving the problem efficiently by iterative solvers. The development of positive definite discrete formulations, which are equivalent to (10) and for which multigrid schemes can be applied, always involves the elimination of the discrete Lagrange multipliers from the indefinite system (10), [48]. This elimination is performed in terms of the discrete projection operator $\widehat{M} = D^{-1}M$, which enters into the positive definite system matrix. The same operator plays an essential role if Dirichlet–Neumann solvers are applied, [23]. Depending on the structure of D, this projection can be carried out locally or it has to be carried out globally, represented by a sparse or a dense matrix $\widehat{M}$, respectively. In particular, if the biorthogonality relation (12) is satisfied, the matrix D is diagonal, and therefore, $\widehat{M}$ is sparse and can be easily calculated. We emphasize that the applicability of dual Lagrange multipliers is not restricted to linear stationary problems like (7). In more general cases, one has to face a linear system of the structure (10) in each iteration step of a time integration and/or a nonlinear solution method. For example, the advantages of the dual approach have been fully exploited for the solution of contact problems, [24]. We refer to the report of the project B8 within this collection for more details.

Within the project C12, we could extend the framework of mortar finite elements using dual Lagrange multipliers from low order elements and straight interfaces towards higher order elements and curvilinear interfaces. We will

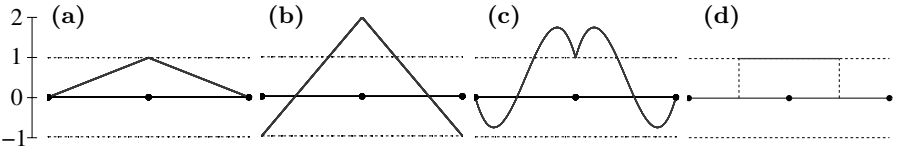


Fig. 1. Basis functions: **(a)** standard, **(b)** discontinuous dual, **(c)** continuous dual, **(d)** constant

address these extensions in the next three sections, starting with the higher order case. Following this, we address the local construction in case of general quadrilaterals and curvilinear interfaces.

3 Higher Order Dual Lagrange Multipliers

In the following, we present dual Lagrange multiplier spaces based on Gauß–Lobatto nodes. We concentrate on two-dimensional problem settings. For the 3D case, an alternative way of constructing a higher order biorthogonal basis is given in [29]. There, we generalize the concept of dual bases by relaxing the condition that the trace space of the approximation space at the slave side and the Lagrange multiplier space have the same dimension. However, the presentation of this new theoretical framework would be beyond the scope of this report.

3.1 Construction Principle

In [33], we construct basis functions which are biorthogonal to conforming one-dimensional nodal finite element basis functions of order p. In contrast to earlier approaches, [42], we require that these basis functions have the same support as the conforming nodal basis functions. If the set of nodes $\mathcal{S}_p = \{-1 = x_1 < x_2 < \cdots < x_{p+1} = 1\}$ on the reference interval $I = [-1, 1]$ with associated nodal basis $\Phi_p = \{\phi_1^{\mathrm{s}}, \ldots, \phi_{p+1}^{\mathrm{s}}\}$ is given, the dual basis $\Lambda_p = \{\mu_1, \ldots, \mu_{p+1}\}$ with $\operatorname{span} \Lambda_p = \operatorname{span} \Phi_p$ is already uniquely determined by the requirement

$$(\phi_p^{\mathrm{s}}, \mu_q)_I = \delta_{pq} (\phi_p^{\mathrm{s}}, 1)_\Gamma \, . \tag{13}$$

In order to obtain optimal a priori results of order h^p, a sufficient condition is that the global biorthogonal basis functions span a finite element space M_h^p which includes the conforming finite element space $W_h^{\mathrm{s},\,p-1}$ of order $p-1$, [26]. Unfortunately, equidistant nodes on each edge fail for $p \geq 3$. In [33], we prove that taking Gauß–Lobatto nodes on each edge guarantees optimal a priori estimates.

Theorem 1. *$V_h^{p-1} \subset M_h^p$ if and only if the finite element basis Φ_p is based on Gauß–Lobatto points $\mathcal{S}_p$.*

The basis for the global interface is given by gluing together the contributions from each edge, using standard rules to transform the integrals to the reference interval I. We note that special care has to be taken of the cross-points of more than two subdomains, where some simple modifications are necessary. In Fig. 2, the resulting basis functions are plotted for $p = 3$ and $p = 4$. As for the lower order case $p = 1, 2$ and equidistant nodes, the nodal basis functions are globally continuous, while the dual ones exhibit discontinuities at the endpoints of their support. Working with a tensor product finite element

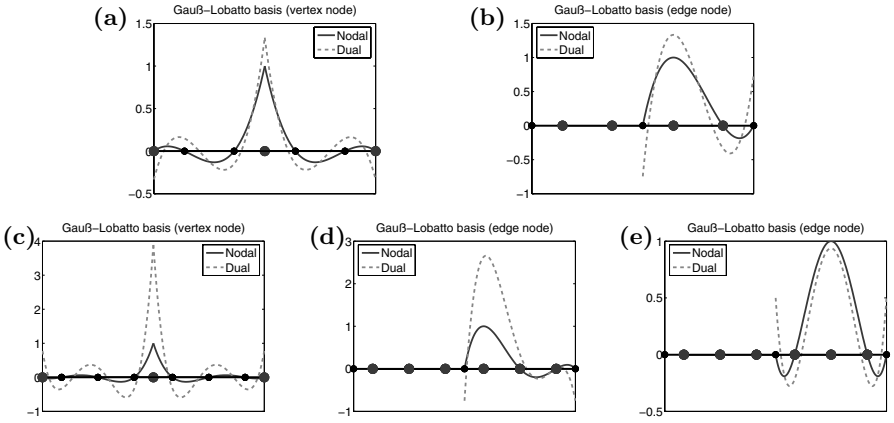


Fig. 2. Higher order nodal and dual basis functions based on Gauß–Lobatto nodes: **(a)**–**(b)** cubic, **(c)**–**(e)** quartic

space, we can apply this one-dimensional construction in a straightforward way to two-dimensional interfaces. However, this approach is restricted to interface grids consisting purely of parallelograms. For tetrahedral and more general hexahedral meshes, one has to apply different techniques, [29].

3.2 Numerical Results

We compare the results for the cubic mortar finite elements with dual Lagrange multiplier spaces based on equidistant nodes and on Gauß–Lobatto nodes. In this example, we decompose the domain $\Omega = (-1,1) \times (0,1)$ into two squares $\Omega^{\mathrm{s}} = (-1,0) \times (0,1)$ and $\Omega^{\mathrm{m}} = (0,1) \times (1,0)$ with the interface $\Gamma = \{0\} \times (0,1)$. The initial non-matching triangulation consists of 4×4 square elements on Ω^{s}, and of 2×2 squares divided into 8 simplicial elements on Ω^{m}. The problem for this example is given by (1), together with the boundary conditions $\partial u / \partial \mathbf{n}|_{\Gamma_{\mathrm{N}}} = g_{\mathrm{N}}$ and $u|_{\Gamma_{\mathrm{D}}} = g_{\mathrm{D}}$, where $\Gamma_{\mathrm{N}} = \{(x,0) \in \mathbb{R}^2 : -1 < x < 1\} \cup \{(x,1) \in \mathbb{R}^2 : -1 < x < 1\}$, and $\Gamma_{\mathrm{D}} = \partial\Omega \backslash \Gamma_{\mathrm{N}}$. The functions f, g_{N}, and g_{D} are calculated from the exact solution $u = e^x(x^2 - 1)(y^2 - y)$. In order to compare the two mortar situations, we visualize the error decay in the L^2- and in the H^1-norm in Fig. 3. For both norms, it becomes evident that the solution for the cubic case with equidistant nodes exhibits suboptimal convergence rates, whereas the cubic case with Gauß–Lobatto nodes yields an optimal behavior. The convergence rate in the L^2- and H^1-norms is of order 3 and 4, respectively. In contrast to optimal convergence rates attained by the cubic mortar finite elements with Gauß–Lobatto nodes, we only observe the convergence rate of order 2.5 and 1.5 for the use of equidistant nodes. Thus, especially for higher order, it is of crucial importance to employ appropriate discrete Lagrange multipliers.

4 General Quadrilaterals

In this section, we consider three-dimensional problem settings and introduce dual Lagrange multipliers for arbitrary planar quadrilateral interface grids $\mathcal{T}^{\mathrm{s}}$ and $\mathcal{T}^{\mathrm{m}}$. We investigate the scalar case and remark that the extension to the vectorial case is straightforward. For the moment, we focus on one planar interface Γ, and ignore any potential necessity for modifications on $\partial\Gamma$. In case of parallelograms or triangles, the following considerations reduce to the already known standard case. The discrete Lagrange multiplier space M_h is simply defined as the span of all nodal basis functions μ_p^g, where p is a vertex of the slave side grid. As usual, each basis function μ_p^g is defined element-wise as

$$\mu_p^g = \sum_{T \in \mathcal{T}^p} \mu_{p,T} \tag{14}$$

with local supports $\mathcal{T}^p = \{T \in \mathcal{T}^{\mathrm{s}} : p \text{ is a vertex of } T\}$ for μ_p^g and T for $\mu_{p,T}$. Here and in the sequel, we will abuse the notation and indicate by p either a global vertex of the triangulation or a local node number within an element T, depending on the context. Moreover, we will usually write μ_p instead of $\mu_{p,T}$ when there is no ambiguity involved. It is sufficient for the dual approach that the local multiplier functions μ_p satisfy a biorthogonality relation with the element basis functions ϕ_q^{s} of the trace space W_h^{s}, namely,

$$(\phi_p^{\mathrm{s}}, \mu_q)_T \;=\; \delta_{pq} (\phi_p^{\mathrm{s}}, 1)_T \;, \tag{15}$$

which instantly implies (12). As usual, the integration on the left side of (15) is performed via a transformation to the reference element $\widehat{T}$. For a simplex T, the corresponding reference element $\widehat{T}$ is the triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$, while for quadrilaterals, $\widehat{T}$ is set to be the unit square $(0,1)^2$. We remark that, within the considered setting, it is not sufficient to simply choose the Lagrange multiplier μ_p as $\widehat{\mu}_p \circ F_T^{-1}$ with $F_T : \widehat{T} \to T$ indicating the element mapping and $\widehat{\mu}_p$ respecting a biorthogonality relation with the shape functions $\widehat{\phi}_q^{\mathrm{s}}$ on the reference element $\widehat{T}$. This is due to the fact that,

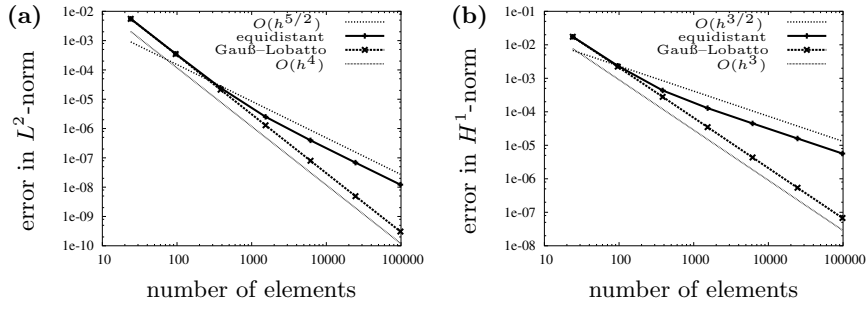


Fig. 3. Decay of the error measured in the **(a)** L^2- and in the **(b)** H^1-norm

for quadrilaterals, F_T is not necessarily an affine, i.e. $P1$-mapping, but an isoparametric $Q1$-mapping. This yields a surface element $\mathrm{d}T = |\det F_T'|\mathrm{d}\widehat{T}$ with a linear contribution $\det F_T'$, where F_T' indicates the Jacobian of F_T. We note that the expression $\det F_T'$ abuses the notation since F_T maps from $\widehat{T} \subset \mathbb{R}^2$ to $\mathbb{R}^3$, and the Jacobian F_T' is not a square matrix, see [19]. Transforming the required integral to the reference element, we obtain

$$(\phi_p^{\mathrm{s}}, \mu_q)_T \;=\; (\widehat{\phi}_p^{\mathrm{s}}, \widehat{\mu}_q |\det F_T'|)_{\widehat{T}} \;, \tag{16}$$

from which we obviously cannot expect that (15) is satisfied. In what follows, we will provide one way of defining M_h yielding (15). Our approach relies on the solution of local subproblems on each element. Alternatively, one can use a special transformation to eliminate $|\det F_T'|$ from (16), as it was carried out in [19]. Concerning the theoretical and numerical results, there is no mentionable difference between both approaches.

4.1 Local Subproblems

We indicate by D_T and $M_T \in \mathbb{R}^{n_{\mathrm{s}} \times n_{\mathrm{s}}}$ the diagonal matrix and the element mass matrix, respectively, with entries given by

$$d_{pp} = (\phi_p^{\mathrm{s}}, 1)_T \;, \qquad m_{pq} = (\phi_p^{\mathrm{s}}, \phi_q^{\mathrm{s}})_T \;. \tag{17}$$

With $A_T = D_T M_T^{-1}$, we define

$$\mu_{p,T} = \sum_q a_{pq} \phi_q^{\mathrm{s}} \;, \tag{18}$$

and obtain the biorthogonality (15) by

$$(\phi_p^{\mathrm{s}}, \mu_{q,T})_T = \sum_k a_{qk} m_{kp} = (A_T M_T)_{qp} = d_{pq} = \delta_{pq} (\phi_p^{\mathrm{s}}, \chi_T)_T \;.$$

Above and in the sequel, we always assume that the summation index runs from 1 to n_{s}, the number of element vertices, unless another index set is given. We especially focus on quadrilateral surface grids, i.e., $n_{\mathrm{s}} = 4$. For simplicial grids, the following considerations are also valid and reduce to already well known results. In [19], we prove the following lemma.

Lemma 1. *Let M_h be constructed from* (14) *and* (18)*. Then $\mathbb{P}_0 \subset M_h$.*

By using the L^2-stability of the mortar-projection, it can be easily shown that M_h satisfies an approximation property, [26].

We note that the entries of the global coupling matrix M can be easily assembled by the local contributions of $(M_{\mathrm{sm}})_{pq} = \int_{T^{\mathrm{sm}}} \phi_p^{\mathrm{s}} \phi_q^{\mathrm{m}} \, \mathrm{d}T^{\mathrm{sm}}$, where T^{sm} denotes the intersection of a slave and a master element. Formally, this gives

$$M = \sum_{T^{\mathrm{sm}} = T^{\mathrm{s}} \cap T^{\mathrm{m}}} R_{T^{\mathrm{s}}} D_{T^{\mathrm{s}}} M_{T^{\mathrm{s}}}^{-1} M_{\mathrm{sm}} R_{T^{\mathrm{m}}}^{\mathrm{T}} \;,$$

where R_{T^k} denotes the matrix which maps the local node numbers of the element T^k to the global ones with respect to $\mathcal{T}^k$, $k = \mathrm{m}, \mathrm{s}$.

4.2 Numerical Results

The theoretical results of Sect. 4.1 yield an optimal a priori estimate for the error in the broken H^1-norm. Here, we are interested in the quantitative numbers. In addition to the approach considered above, we employ two other methods for comparison: one using standard basis functions $\mu_p = \phi_p^{\mathrm{s}}$ not satisfying (15), and a "naive" one, where μ_p is chosen as $\widehat{\mu}_p \circ F_T^{-1}$ for the coupling with the master side, but it is set to be $(|T|/|\det F_T'|)\widehat{\mu}_p \circ F_T^{-1}$ for the coupling with the slave side. We note that the latter approach satisfies (15), and coincides with the original dual method for simplices and parallelepipeds. We also point out that by a suitable redefinition of the coupling bilinear form $b(\cdot,\cdot)$, the approach could be reformulated with respect to only one discrete Lagrange multiplier space M_h. The choice is motivated by the fact that nothing has to be modified for the coupling on the master side, and only a minimal modification is necessary for the coupling on the slave side. However, constants are not preserved due to the choice of $|T|$ as weights. Therefore, the approximation property is lost, and optimal convergence cannot be guaranteed anymore.

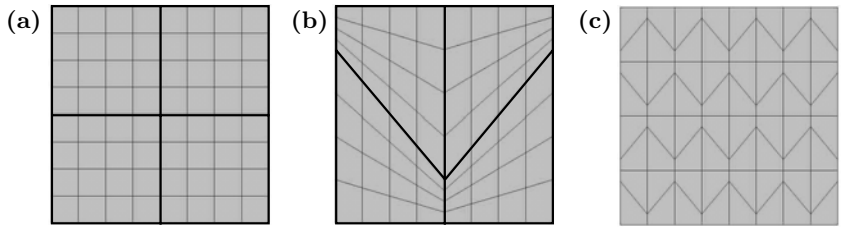


Fig. 4. Surface grids $\mathcal{T}^{\mathrm{s}}$: **(a)** square, **(b)** asymptotically parallelogram, **(c)** trapezoidal

We consider a simple test example with two cubes Ω^{m} and Ω^{s} of length and width 1 and height 0.2, sharing as interface the unit square of edge length 1 at $z = 0$. We solve (1) with right-hand side f derived from the exact solution $u(x, y, z) = yz\mathrm{e}^{-x^2}$. On the planes $z = \pm 0.2$, Dirichlet boundary data are considered, while on the remaining part of the boundary, we employ Neumann data. The Lagrange multiplier space M_h is associated with the grid on the lower cube. For the surface meshes $\mathcal{T}^{\mathrm{s}}$ on the slave side, we compare three different sequences as in [1]: square, asymptotically parallelogram, and trapezoidal, see Fig. 4 (a), (b), and (c), respectively. For the first two sequences, the initial triangulation is indicated by thick lines, and the subsequent grids are simply obtained by uniform refinements. For the trapezoidal grids, the same initial triangulation as for the asymptotically parallelogram grid is used, but instead of employing a standard uniform refinement procedure, the surface is partitioned into congruent trapezoids at each step, all similar to the trapezoid with vertices $(0, 0)$, $(0.5, 0)$, $(0.5, 0.2)$, and $(0, 0.8)$. The thin lines in the

pictures of Fig. 4 indicate the slave side grids after two refinements. On the master side, the meshes $\mathcal{T}^{\mathrm{m}}$ consist of squares twice the size of the elements on the slave side.

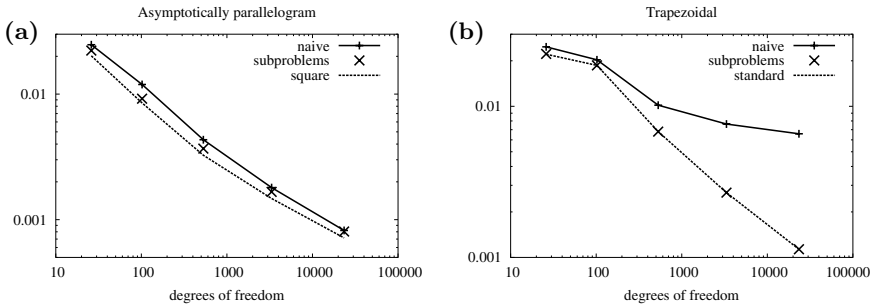


Fig. 5. Error decays, measured in the H^1-norm: **(a)** asymptotically parallelogram, **(b)** trapezoidal

In Fig. 5, the error decays measured in the H^1-norm are plotted for different grid sequences and different Lagrange multipliers. In particular, we compare two approaches: the naive dual one and the one introduced in Sect. 4.1. For the results of the asymptotically parallelogram grid sequence illustrated in Fig. 5(a), we choose the results from the uniform square grids as reference. All approaches give qualitatively the same and quantitatively almost the same results. We remark that the same quality is obtained by the use of the standard basis functions. This observation changes drastically when employing the sequence of trapezoidal grids, as illustrated in Fig. 5(b). The naive approach fails completely, the error remains almost static after a few refinement steps. Our method behaves as predicted by the theory. Moreover, the error visually coincides with the approach using standard basis functions which is taken as a reference here.

5 Curvilinear Interfaces

Our numerical results above illustrate how sensitively the quality of the mortar approximation depends on the choice of the Lagrange multiplier space. In this section, we consider variational crimes resulting from curvilinear interfaces. We focus on the 3D case and remark that the following considerations equally apply to the 2D case. In three dimensions, we consider an interface to be curvilinear when its triangulation by means of quadrilaterals yields a surface element whose four vertices are not contained in a two-dimensional linear hyper-plane. The proofs of the results in this section can be found in [15, 16, 19]. We focus on the case of vector fields. In general, the surface grids $\mathcal{T}^{\mathrm{m}}$ and $\mathcal{T}^{\mathrm{s}}$ cannot resolve a curvilinear interface Γ. Geometrically, they form piecewise

$P1$ or $Q1$ interpolations Γ_h^{m} and Γ_h^{s} of the exact interface Γ. The meshes from both sides can partially overlap or even exhibit gaps. Therefore, in order to pose a discrete problem formulation, the coupling bilinear form $b(\cdot,\cdot)$ given by (8) has to be suitably approximated by a form $b_h(\cdot,\cdot) : X_h \times M_h \to \mathbb{R}$. For the coupling of the Lagrange multipliers with the slave side, no modifications are necessary, since the corresponding functions are all associated with the same triangulation $\mathcal{T}^{\mathrm{s}}$. However, the coupling with the master side is much more involved. As in [16], we define a suitable linear and stable projection operator P_{s} onto the slave side,

$$P_{\mathrm{s}} : (L^2(\Gamma_h^{\mathrm{m}}))^3 \to (L^2(\Gamma_h^{\mathrm{s}}))^3 \ , \quad \mathbf{v}_{\mathrm{m}} \mapsto P_{\mathrm{s}}\mathbf{v}_{\mathrm{m}} \ , \tag{19}$$

a mesh dependent jump $[\cdot]_h$ by

$$[\mathbf{v}]_h = \mathbf{v}_{\mathrm{s}} - P_{\mathrm{s}}\mathbf{v}_{\mathrm{m}} \ ,$$

and the approximate coupling bilinear form $b_h(\cdot,\cdot)$ as

$$b_h(\mathbf{v},\boldsymbol{\mu}) = (\,[\mathbf{v}]_h,\, \boldsymbol{\mu}\,)_{L^2(\Gamma_h^{\mathrm{s}})} \ , \quad (\mathbf{v},\boldsymbol{\mu}) \in X_h \times M_h \ . \tag{20}$$

In [15], we show that the analogous approach for the scalar case yields optimal a priori estimates. These estimates can be transfered to the vectorial case by standard arguments. However, for the vectorial case, if dual Lagrange multipliers are chosen with respect to the coarse grid, one can observe a preasymptotic misbehavior in form of unintentional oscillations, even for a very simple linear elasticity model problem, see [16]. There, we present a remedy for the two-dimensional case which preserves the advantages of the dual approach. In [19], we present two alternatives for dimension $d \in \{2,3\}$. Both have in common that only the coupling of the Lagrange multipliers to the master side is changed, namely, $(P_{\mathrm{s}}\mathbf{v}_{\mathrm{m}},\, \boldsymbol{\mu})_{L^2(\Gamma_h^{\mathrm{s}})}$. To this end, the coupling bilinear form $b_h(\cdot,\cdot)$ is replaced by a modification $b_h^{\mathrm{mod}}(\cdot,\cdot)$. In the first alternative to be given, we replace the L^2-scalar product $(\cdot,\cdot)_{L^2(\Gamma_h^{\mathrm{s}})}$ by a discrete one, whereas in the second alternative the Lagrange multiplier $\boldsymbol{\mu} \in M_h$ as seen by the master side is replaced by $\boldsymbol{\mu} + \Delta\boldsymbol{\mu}$. Both approaches reduce to the original one in the case of a planar interface. In this report, we will only address the second alternative.

Before we introduce the modification, we focus on an important requirement. With each node p on the smooth interface Γ, an orthonormal basis $B_p \in \mathbb{R}^{3\times 3}$ is associated, given by

$$B_p = (\mathbf{b}_p^{(1)}, \mathbf{b}_p^{(2)}, \mathbf{b}_p^{(3)}) \tag{21}$$

with $\mathbf{b}_p^{(1)} = \mathbf{n}_p$ being the unit normal vector on Γ in p, and the remaining columns being corresponding unit tangent vectors. Our modification will be given in terms of the difference between two bases B_p, B_q, where p and q refer to slave nodes which are both vertices of one slave side element. For an optimal a priori estimate, it is mandatory to require that

$$\|B_p - B_q\|_\infty = O(h) \ , \tag{22}$$

provided that $\|\mathbf{p} - \mathbf{q}\|_\infty = O(h)$ where $\mathbf{p}$ and $\mathbf{q}$ are the coordinate vectors of the nodes p and q, respectively. We refer to [19] where we present an approach in terms of Householder transformations yielding (22).

5.1 Momentum Preserving Modification

We focus on a slave element T^{s} of the surface grid $\mathcal{T}^{\mathrm{s}}$ with n_{s} denoting the numbers of its vertices. Requiring that the modified Lagrange multipliers still preserve a lowest order momentum, the modification $\Delta\boldsymbol{\mu}$ will be given in terms of

$$\Delta\phi_{pq} = \phi_p^{\mathrm{s}} d_{pp}^{-1} - \phi_q^{\mathrm{s}} d_{qq}^{-1} \ , \tag{23}$$

where d_{pp} and d_{qq} are given by (17). It is obvious that $\int_{T^{\mathrm{s}}} \Delta\phi_{pq}\,\mathrm{d}T^{\mathrm{s}} = 0$, thus, a lowest order momentum will be preserved. On T^{s}, any given discrete Lagrange multiplier $\boldsymbol{\mu} \in M_h$ can be written as $\boldsymbol{\mu}|_{T^{\mathrm{s}}} = \sum_{p=1}^{n_{\mathrm{s}}} \boldsymbol{\alpha}_p \mu_p$ with coefficients $\boldsymbol{\alpha}_p \in \mathbb{R}^3$, $p = 1, \dots n_{\mathrm{s}}$. Its modification $\Delta\boldsymbol{\mu}$ is defined by

$$\Delta\boldsymbol{\mu}|_{T^{\mathrm{s}}} = \frac{1}{2} \sum_{p,q} \gamma_{pq} \Delta\phi_{pq} \left(\sum_{i=1}^{3} (\boldsymbol{\alpha}_p \cdot \mathbf{b}_p^{(i)} + \boldsymbol{\alpha}_q \cdot \mathbf{b}_q^{(i)}) \Delta\mathbf{b}_{pq}^{(i)} \right) \ , \tag{24}$$

yielding the modified multiplier $\boldsymbol{\mu}^{\mathrm{mod}} = \boldsymbol{\mu} + \Delta\boldsymbol{\mu}$. In the formula above, the coefficients γ_{pq} are the elements of a modification matrix $G \in \mathbb{R}^{n_{\mathrm{s}} \times n_{\mathrm{s}}}$ which has yet to be defined, and $\Delta\mathbf{b}_{pq}^{(i)} = \mathbf{b}_p^{(i)} - \mathbf{b}_q^{(i)}$, $i = 1, \dots, 3$.

In order to motivate our choice for the modification $\Delta\boldsymbol{\mu}$, we introduce the matrix $\mathbf{N}_{\mathrm{s}} \in \mathbb{R}^{3 \times n_{\mathrm{s}}}$ by

$$\mathbf{N}_{\mathrm{s}} = (\mathbf{n}_1^{\mathrm{s}}, \dots, \mathbf{n}_{n_{\mathrm{s}}}^{\mathrm{s}}) \ ,$$

where $\mathbf{n}_p^{\mathrm{s}}$ indicates the unit normal vector on Γ in the slave node p, $p = 1, \dots, n_{\mathrm{s}}$. Moreover, the symbolic vectors Φ_{s} and Λ_{s} of length n_{s} are given by

$$\Phi_{\mathrm{s}} = (\phi_1^{\mathrm{s}}, \dots, \phi_{n_{\mathrm{s}}}^{\mathrm{s}})^{\mathrm{T}} \ , \quad \Lambda_{\mathrm{s}} = (\mu_1, \dots, \mu_{n_{\mathrm{s}}})^{\mathrm{T}} \ ,$$

where, as before, ϕ_p^{s} and μ_p denote scalar nodal basis functions of the corresponding spaces, $p = 1, \dots, n_{\mathrm{s}}$. We require that the modification guarantees a discrete preservation of quantities which are constant in normal and tangential direction when transferring between the trace space W_h^{s} and the Lagrange multiplier space M_h. This idea is motivated by the observation that $\sum_p \mathbf{n}_p \phi_p^s$ yields a quite good approximation of the normal field on Γ whereas $\sum_p \mathbf{n}_p \mu_p^s$ gives a bad result, see [16]. Because of the duality between μ_p and ϕ_q^s this does not affect the surface traction on the slave side but on the master side. In particular, focusing on the normal direction, this requirement can be expressed element-wise by demanding that $\mathbf{N}_{\mathrm{s}} \Lambda_{\mathrm{s}} = \mathbf{N}_{\mathrm{s}} \Phi_{\mathrm{s}}$. However, when usual dual basis functions are used for M_h, this cannot be achieved. But for our modification (24) with suitably defined coefficients γ_{pq}, we can show that

$$(\mathbf{N}_{\mathrm{s}} \Lambda_{\mathrm{s}})^{\mathrm{mod}} = \mathbf{N}_{\mathrm{s}} \Phi_{\mathrm{s}} \ . \tag{25}$$

Lemma 2. *Let the scalar dual basis functions be defined as in Sect. 4.1, namely,* $\Lambda_s = D_{T^s} M_{T^s}^{-1} \Phi_s$. *Then, the choice*

$$G = \frac{1}{2} D_{T^s} M_{T^s}^{-1} D_{T^s} \tag{26}$$

yields (25).

For the analysis, we work on

$$V_h^{\mathrm{mod}} = \left\{ \mathbf{v} \in X_h : \left(\boldsymbol{\mu}_p^i, \mathbf{v}_{\mathrm{s}} - (\boldsymbol{\mu}_p^i + \Delta\boldsymbol{\mu}_p^i)\mathbf{v}_{\mathrm{m}}\right)_{L^2(\Gamma_h^{\mathrm{s}})} = 0 \,, \; p \in \mathcal{V}^{\mathrm{s}}, \; i = 1, \ldots, 3 \right\} .$$

In the definition above, $\boldsymbol{\mu}_p^i = \mu_p \mathbf{e}_i$ denotes the vectorial basis function of the Lagrange multiplier space M_h in direction x_i associated with the slave node p, $\Delta\boldsymbol{\mu}_p^i$ stands for its modification according to (24), and $\mathcal{V}^{\mathrm{s}}$ denotes the set of all element vertices of $\mathcal{T}^{\mathrm{s}}$.

Lemma 3. *Let the modification be given by* (24) *and* (26). *Then, for an arbitrary* $\mathbf{v} = (\mathbf{v}_m, \mathbf{v}_s) \in V_h$, *there exists* $\Delta\mathbf{v} \in X_h$ *such that*

$$\mathbf{v}^{\mathrm{mod}} = \mathbf{v} + \Delta\mathbf{v} \in V_h^{\mathrm{mod}} \,, \tag{27a}$$

$$\|\Delta\mathbf{v}\|_{X_h} \leq C h^{t+1/2} |\mathbf{v}_m|_{t,\Gamma_h^m} \,, \qquad t \in [0,1] \,, \tag{27b}$$

$$\|\Delta\mathbf{v}\|_{0,\Gamma_h^s} \leq C h^{3/2} |\mathbf{v}_m|_{1/2,\Gamma_h^m} \,, \tag{27c}$$

$$\|\mathbf{v}\|_{X_h} \sim \|\mathbf{v}^{\mathrm{mod}}\|_{X_h} \,, \qquad h \text{ small enough} \,. \tag{27d}$$

The analogous statement holds for arbitrary $\mathbf{v}^{\mathrm{mod}} \in V_h^{\mathrm{mod}}$.

By using Lemma 3, an optimal a priori estimate can be easily obtained by using the fact that the unmodified approach is already optimal. Moreover, it can be seen that the modification only enters in terms of $O(h^{3/2})$.

5.2 Numerical Results

In order to present the effect of our modifications, we investigate a 3D example which is analogous to the 2D example given in [16]. The global domain is a spherical shell with inner radius $r_i = 0.9$ and outer radius $r_o = 1.1$, its material data given by $E = 1.0$ and $\nu = 0.3$, as depicted in the left picture of Fig. 6(a). The outer boundary $\{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| = 1.1\}$ is fixed by homogeneous Dirichlet boundary conditions, whereas on the inner boundary $\{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| = 0.9\}$, a uniform radial pressure of magnitude -1 is applied. The symmetry of the domain and the problem data yields the exact solution $u(r) = a/r^2 + br$, depending only on $r(\mathbf{x}) = |\mathbf{x}|$, with $b = 1/(3\lambda + 2\mu + 4\mu r_o^3/r_i^3)$ and $a = -br_o^3$. In order to keep a full 3D setting, we exploit the radial symmetry only partially for the numerical simulation, namely, by considering only the octant $O_1 = \{\mathbf{x} \in \mathbb{R}^3 : x_i > 0, \; i = 1, 2, 3\}$. The interface Γ is set to be the unit sphere

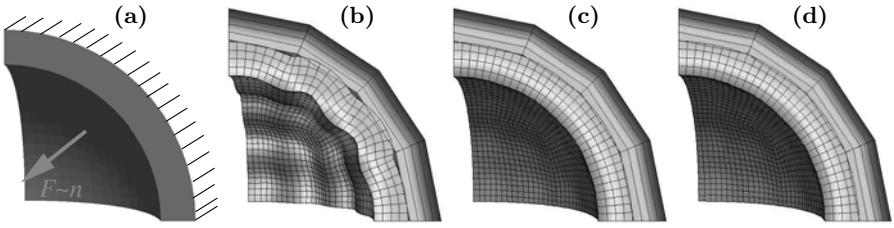


Fig. 6. (a) problem setting. Ratio $h_\mathrm{s}/h_\mathrm{m} = 8/1$, distorted domains: **(b)** unmodified dual, **(c)** modified dual, **(d)** standard Lagrange multipliers

intersected by O_1, yielding the subdomains $\Omega^\mathrm{m} = \{\mathbf{x} \in O_1 : |\mathbf{x}| \in (r_i, 1)\}$ and $\Omega^\mathrm{s} = \{\mathbf{x} \in O_1 : |\mathbf{x}| \in (1, r_o)\}$. The conditions on the symmetry boundaries $\Sigma_i = (\overline{\Omega^\mathrm{m}} \cup \overline{\Omega^\mathrm{s}}) \cap \{\mathbf{x} \in \mathbb{R}^3 : x_i = 0\}$, $i = 1, \ldots, 3$, are given by $\mathbf{u}\cdot\mathbf{n}_i = 0$ and $\sigma_\mathrm{t} = 0$, where $\mathbf{n}_i = -\mathbf{e}_i$ is the corresponding normal vector and σ_t indicates the tangential part of the surface traction $\sigma(\mathbf{u})\mathbf{n}_i$. A detailed account on the handling of the Lagrange multiplier nodes on $\Sigma_i \cap \overline{\Gamma}$ is given in [18].

Undesired oscillations occur only when the surface grid $\mathcal{T}^\mathrm{s}$ is considerably coarser than the grid $\mathcal{T}^\mathrm{m}$. To this end, we take a ratio of $h_\mathrm{s}/h_\mathrm{m} = 8/1$, and the corresponding surface grids consist of 12 and 768 elements for $\mathcal{T}^\mathrm{s}$ and $\mathcal{T}^\mathrm{m}$, respectively. In radial direction, we take four elements for each subdomain, giving a total of 3120 volume elements. In Fig. 6, the deformed domain is visualized for three different approaches: (b) the unmodified dual one, (c) the modified one as introduced above, and, as a reference, (d) the one taking standard basis functions. The solution of the unmodified dual method is subject to oscillations. The modification gives a good result, the surface tractions and the displacements, which are both constant in normal direction, are interchanged between the grids in the expected correct way. Moreover, we see a reasonable agreement with the method using standard Lagrange multipliers. In Fig. 7, the decay of the energy error under uniform refinement is visualized, this time for a ratio of $h_\mathrm{s}/h_\mathrm{m} = 4/1$. Comparing the unmodified

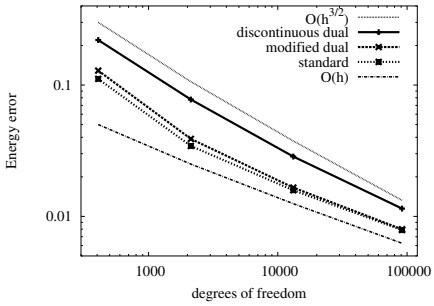


Fig. 7. Decay of the energy error

approach with the modified and with the standard one, the decays confirm the impressions obtained by the deformed domains. The error for the modified dual approach almost coincides with the error for the standard approach. As pointed out above, we see that the modification enters with $O(h^{3/2})$.

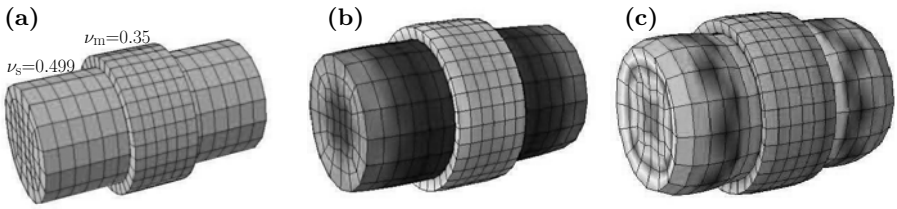


Fig. 8. Compression of a cylinder: **(a)** setting and grid, **(b)** locking, **(c)** locking-free

In our next example, we employ the mortar approach to couple a nearly incompressible with a compressible material. In particular, we consider the compression of a soft and nearly incompressible cylinder Ω^s, $E_s = 1000$, $\nu_s = 0.499$, which is partly enclosed in a hard and compressible cylinder ring Ω^m, $E_m = 6000$, $\nu_m = 0.35$, as illustrated in Fig. 8(a). On both subdomains, we solve a weak form of the nonlinear equations (5)–(6). Figure 8(b) shows the deformed domains if a standard displacement Q_1-formulation is used. Clearly, the soft material exhibits an unphysical volume locking. Alternatively, we use a modified displacement Q_1-formulation based on a Hu–Washizu formulation, [6], for the cylinder Ω^s. For Ω^m, we keep the standard formulation. The deformed domain obtained by this approach is visualized in Fig. 8(c). Now, the soft material deforms as expected. The mortar approach is suitable to couple different model equations in a robust and stable manner, [28].

6 Implementation

In what follows, we give a more detailed account of the assembly of the coupling matrices defined in (11). Omitting the subscripts p and q in (11), we have to evaluate the integral

$$(\phi^m, \mu)_\Gamma \tag{28}$$

for all basis functions μ and ϕ^m defined on the $(d-1)$-dimensional grids $\mathcal{T}^s$ and $\mathcal{T}^m$, respectively. As usual, the assembly can be performed element-wise. One possible realization is given by Algorithm 1. We remark that the naive implementation of this algorithm for 2D-problems is of order $O(n)$, but for 3D-problems, it is of order $O(n^{4/3})$ with n denoting the total number of unknowns. However, it is possible to obtain a better complexity in both cases by incorporating neighbor-ship relations of the surface elements and/or inheritance relations from an underlying geometrical multigrid hierarchy, as well as

Algorithm 1 Assembly of the coupling matrix

```
for all slave elements e_i^n, i = 1,...,n_n do
  for all master elements e_j^m, j = 1,...,n_m do
    determine intersection area T^ms = e_i^n ∩ e_j^m
    if e_ij ≠ ∅ then
      for all basis functions φ^m, μ with support ∩ e_ij ≠ ∅ do
        add (φ^m, μ)_{T^ms} to M
      end for
    end if
  end for
end for
```

by employing quadtree/octree data structures for organizing the mesh data. The crucial point in Algorithm 1 is the determination of the intersection area of two elements from the different grids. In the remainder of this section, we will address this issue for different situations.

6.1 Straight Interfaces

In order to determine whether two surface elements T^{s} and T^{m} intersect, we loop over all vertices p_k^{m}, $k = 1, \ldots, n_{\mathrm{m}}$, of T^{m}, where n_{m} denotes the corresponding number of vertices. The two elements intersect, if we find at least one vertex p_k^{m} which lies inside T^{s}. An easy way to justify this is to use the transformation of global to local coordinates, which is usually available in any finite element code. If all local coordinates of p_k^{m} with respect to the element T^{s} are within the correct ranges, the elements T^{s} and T^{m} intersect. Otherwise, it may still be possible that the element T^{s} is completely covered by T^{m}. Therefore, one has to repeat the procedure interchanging the roles of T^{s} and T^{m}. If still no vertex is found, the two elements do not intersect.

For 2D meshes and for 3D structured hexahedral grids, the determination of the intersection is quite easy, [11]. For 3D unstructured meshes, the situation is somewhat more involved. By performing a decomposition into triangles, the following considerations can be applied to quadrilateral meshes. The polygonal intersection T^{ms} of two arbitrary triangles T^{s} and T^{m} can be anything between a triangle and a 2D-hexahedral, see Fig. 9. However, the vertices of T^{ms} can be determined straightforward by including

- all vertices of T^{s} lying inside T^{m} (marked with a cross in Fig. 9),

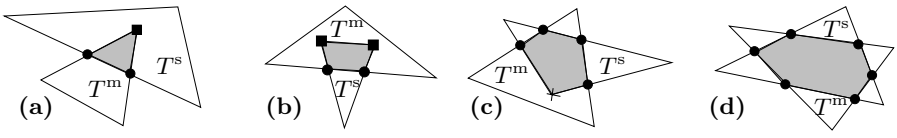


Fig. 9. 3D unstructured: Intersecting elements T^{s} and T^{m}

- all vertices of T^{m} lying inside T^{s} (filled box),
- all intersection points of edges of T^{s} with edges of T^{m} (filled circle).

For the evaluation of (28), the polygon T^{ms} can be subdivided into triangles by connecting its barycenter with the vertices and applying an appropriate quadrature formula on each subtriangle.

6.2 Curvilinear Interfaces

For curvilinear interior boundaries, the interface grids $\mathcal{T}^{\mathrm{s}}$ and $\mathcal{T}^{\mathrm{m}}$ are both not in any straight hyper-surface. In the case of nonconforming grids, this results in the fact that possibly intersecting elements T^{s} and T^{m} are not coplanar, see Fig. 10. One possible way of dealing with this situation in the case of affine-equivalent triangulations is to project the element T^{m} onto the plane of the element T^{s} [43]. Then, the same techniques as described above can be applied in order to determine the intersection T^{ms}. For the evaluation of the basis function ϕ^{m}, the quadrature points in T^{s} have to be projected back onto T^{s}. For more general element transformations as resulting from unstructured hexahedral meshes, it is possible to perform the intersection by projecting both elements onto an intermediate plane obtained by, e.g., a least squares approximation, or by simply being the plane defined by the normal vector located at the barycenter of T^{s}.

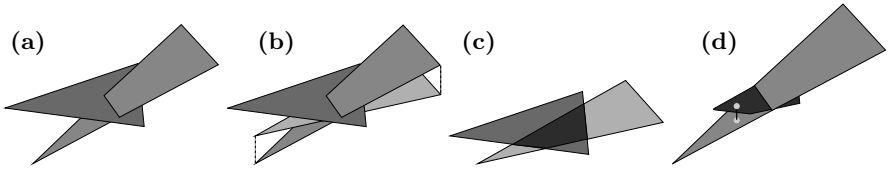


Fig. 10. 3D curvilinear: Intersecting elements and projection onto one plane

7 Applications

We present several applications for nonconforming discretization methods for coupled problems. In particular, we consider aero-acoustics in Sect. 7.1, elasto-acoustics in Sect. 7.2, contact in Sect. 7.3, and heat transfer in Sect. 7.4.

7.1 (Aero-)Acoustics

In computational aero-acoustics, a very common approach to obtain a solution for the acoustic field is Lighthill's analogy, [36]. This amounts to calculating an inhomogeneous source term for the wave equation by the fluid flow data

within the fluid region. To obtain reliable results, the discretization of the wave equation within this domain has to be very fine. However, outside the flow region, the homogeneous wave equation is solved, and one could have a relatively coarse mesh, [11].

For the description of the acoustic wave propagation, we use the wave equation for the velocity potential ψ, i.e., $\mathbf{v}^{\mathrm{a}} = -\operatorname{grad}\psi$ with $\mathbf{v}^{\mathrm{a}}$ denoting the acoustic velocity field. Thus, in each subdomain we have to find $\psi_i : \Omega^i \times (0,T) \to \mathbb{R}$ such that

$$\frac{1}{c^2}\,\ddot{\psi}_i - \Delta\,\psi_i = f_i\,, \quad \text{in } \Omega^i \times (0,T),\ i = \mathrm{m},\mathrm{s}\,, \tag{29}$$

where c indicates the speed of sound. The above equations are completed by appropriate initial conditions at time $t = 0$ and boundary conditions on the global boundary $\partial\Omega$. We emphasize that all of the following considerations apply equally to the formulation of the linear acoustic equations in terms of the acoustic pressure. For the coupling procedure, we proceed exactly like in the stationary case of Poisson's equation (1), i.e. introducing the Lagrange multiplier $\lambda = -\partial\psi/\partial\mathbf{n}$ representing the flux across the interface, and realizing the continuity condition in the primal variable in a weak sense. This gives the symmetric evolutionary saddle point problem of finding $\psi = (\psi^{\mathrm{m}}, \psi^{\mathrm{s}}) \in L^2(0,T;\,X)$ and $\lambda \in L^2(0,T;\,M)$ such that for all times $t \in (0,T)$

$$(\ddot{\psi}(t), c^{-2}w) + a(\psi(t), w) + b(w, \lambda(t)) = (f(t), w)\,, \qquad w \in X\,, \tag{30a}$$

$$b(\psi(t), \mu) = 0\,, \qquad \mu \in M\,. \tag{30b}$$

A suitable general functional framework for (30) is presented in [2], consisting of a combination of the theory of evolutionary variational equations, [8], and the theory of stationary saddle point problems, [5].

The spatial discretization of (30) yields a system of differential equations of the form

$$\mathbf{M}\underline{\ddot{y}} + \boldsymbol{S}\underline{y} = \underline{f}\,, \tag{31}$$

where $\mathbf{M}$ contains the mass matrices associated with the discretizations of Ω^{m} and Ω^{s}, and $\boldsymbol{S}$ has the usual saddle point structure as given by (10). Starting with (31), one can employ a suitable integration scheme for second order ordinary differential equations, as for example Newmark's method, [25].

As numerical example, we investigate the approximation of a single acoustic spherical pulse of frequency 1000 Hz and magnitude 1. The computational domain in the (r,z)-plane is shown in Fig. 11(a). The pulse is imposed in form of an essential boundary condition on $\Gamma_{\mathrm{D}} \subset \partial\Omega^{\mathrm{s}}$. We use $2{\times}40{\times}40 = 3200$ elements on Ω^{s}, as depicted in Fig. 11(b). In order to compare the conforming method with the nonconforming one, we take 6400 elements on Ω^{m} in both cases, choosing $n_x = 80$, $n_y = 40$ for the conforming and $n_x = 160$, $n_y = 20$ for the nonconforming case. In Fig. 12, the isolines for the velocity potential at time $t = 1.6$ ms are visualized. Whereas the conforming method exhibits numerical noise before and behind the pulse, the nonconforming approach is

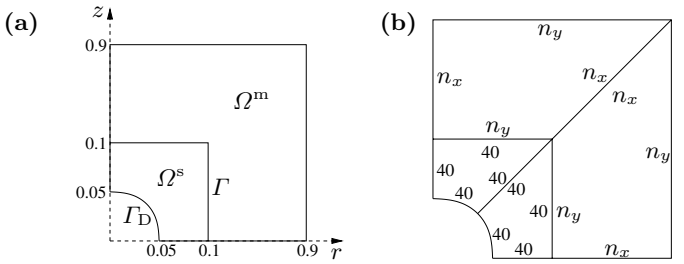


Fig. 11. Spherical pulse: **(a)** domain specifications (dimensions in meter), **(b)** grid parameters

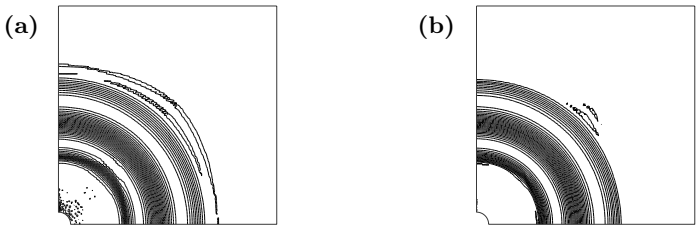


Fig. 12. Isolines of the acoustic velocity potential at time $t = 1.6\,\mathrm{ms}$: **(a)** matching, **(b)** non-matching grids

much closer to the expected solution. Inside the pulse, the radial symmetry of the isolines from the conforming method is observably disturbed. The poor quality of the conforming method can be easily explained by the fact that, in order to obtain matching interface grids, the mesh on Ω^{m} is simply too coarse in direction of large gradients of the pulse to correctly resolve the solution. In order to examine the transient behavior more closely, we visualize the evolution of the acoustic potential at the point $(0, 0.1\,\mathrm{m})^{\mathrm{T}}$ in Fig. 13. In addition to the comparison of both approaches, we employ a reference solution obtained with a uniformly fine grid of 54500 elements. As expected from the observations above, the conforming approach exhibits quite strong oscillations, while the behavior of the nonconforming method is considerably smoother and visually coinciding with the reference solution. We remark that even the reference solution is subject to small unphysical oscillations directly after the pulse, as it can be observed in Fig. 13(c).

7.2 Elasto-Acoustics

In many technical applications a sensor or/and actuator is immersed in an acoustic fluid. In most cases, the discretization within the structure has to be much finer than the one we need for the acoustic wave propagation in the fluid. Thus, this setting is ideally suited for the use of non-matching grids.

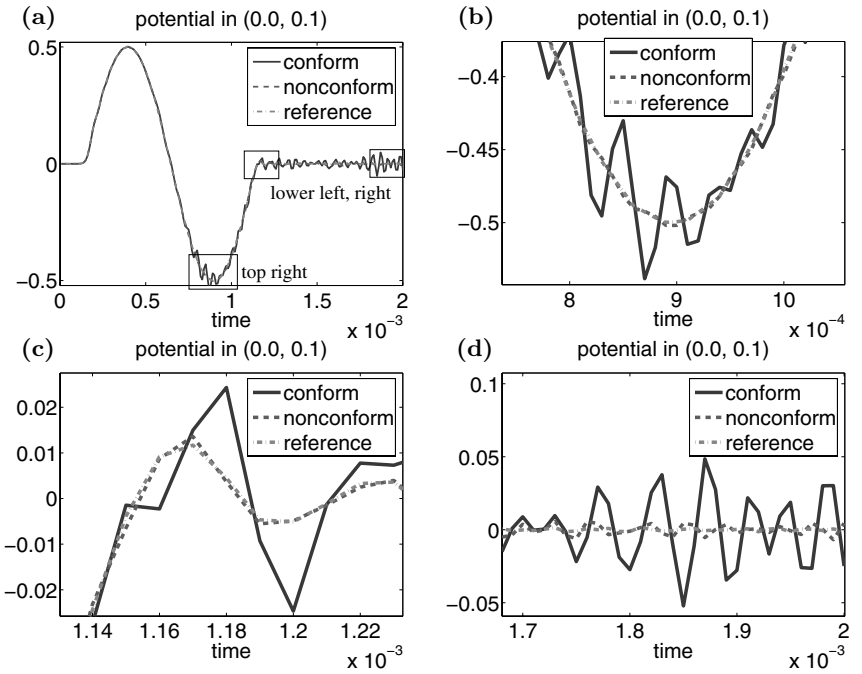


Fig. 13. Evolution of the acoustic potential at the point $(0, 0.1)^{\mathrm{T}}$

Concerning the structural part, we investigate the deformation of an elastic body Ω^{e} with density ρ_{e} under a given time dependent volume force $\mathbf{f}$. The strong formulation for elasto-dynamic problems then reads as follows: find $\mathbf{u} : \Omega^{\mathrm{e}} \times (0,T) \to \mathbb{R}^d$ such that

$$\rho_{\mathrm{e}}\ddot{\mathbf{u}} - \operatorname{div}\sigma(\mathbf{u}) = \mathbf{f} \quad \text{in } \Omega^{\mathrm{e}} \times (0,T)\,, \tag{32}$$

with (3), (4), and appropriate boundary and initial conditions. The corresponding variational formulation is given by: find $\mathbf{u} \in L^2(0,T;\, X^{\mathrm{e}})$ such that for all times $t \in (0,T)$

$$(\rho_{\mathrm{e}}\ddot{\mathbf{u}}(t),\, \mathbf{v})_{\mathrm{e}} + a_{\mathrm{e}}(\mathbf{u}(t),\, \mathbf{v}) = (\mathbf{f}(t),\, \mathbf{v})_{\mathrm{e}}\,, \quad \mathbf{v} \in X^{\mathrm{e}}\,. \tag{33}$$

At a solid/fluid interface, the continuity requires that the normal component of the mechanical surface velocity of the solid must coincide with the normal component of the acoustic velocity of the fluid. Thus, the following relation between the velocity $\mathbf{v}^{\mathrm{e}}$ of the solid expressed by the mechanical displacement $\mathbf{u}$ and the acoustic particle velocity $\mathbf{v}^{\mathrm{a}}$ expressed by the acoustic scalar potential ψ arises

$$0 \;=\; \mathbf{n}\cdot(\mathbf{v}^{\mathrm{e}} - \mathbf{v}^{\mathrm{a}}) \;=\; \mathbf{n}\cdot\dot{\mathbf{u}} \;+\; \frac{\partial\psi}{\partial\mathbf{n}}\,. \tag{34}$$

In addition, one has to consider the fact that the ambient fluid causes a surface force $-\rho_{\mathrm{a}}\mathbf{n}\dot{\psi}$, where ρ_{a} denotes the density of the fluid. This surface force acts like a pressure load on the solid, thus, a second coupling condition is given by

$$[\boldsymbol{\sigma}]\cdot\mathbf{n} \;+\; \rho_{\mathrm{a}}\mathbf{n}\dot{\psi} \;=\; 0\,. \tag{35}$$

Let us consider a setup of a coupled mechanical-acoustic problem, where the global domain consists of the structure Ω^{e} and the acoustic fluid Ω^{a}. Within Ω^{e}, the equation for the mechanical field (32), within Ω^{a}, the equation for the acoustic field (see (29)), and along the interface Γ the coupling conditions according to (34) and (35) have to be satisfied. Transforming to the weak form, we obtain: find $(\mathbf{u}, \psi) \in L^2(0, T;\, X^{\mathrm{e}} \times X^{\mathrm{a}})$ such that

$$(\rho_{\mathrm{e}}\ddot{\mathbf{u}},\, \mathbf{v})_{\mathrm{e}} + a_{\mathrm{e}}(\mathbf{u},\, \mathbf{v}) + (\rho_{\mathrm{a}}\dot{\psi}, \mathbf{v}\cdot\mathbf{n})_{\Gamma} = (\mathbf{f},\, \mathbf{v})_{\mathrm{e}}\,, \quad \mathbf{v} \in X^{\mathrm{e}}\,, \tag{36}$$

$$(\rho_{\mathrm{a}}\ddot{\psi},\, c^{-2}w)_{\mathrm{a}} + a_{\mathrm{a}}(\rho_{\mathrm{a}}\psi,\, w) - (\rho_{\mathrm{a}}w, \dot{\mathbf{u}}\cdot\mathbf{n})_{\Gamma} = (f, w)_{\mathrm{a}}\,, \quad w \in X^{\mathrm{a}}\,. \tag{37}$$

In contrast to the problem settings considered before, no additional Lagrange multiplier has to be introduced.

For the following numerical example, the structure Ω^{e} consists of 25 cylindrical silicon plates with diameter 50 µm and height 1 µm each. They are placed as a (5×5)-array, each plate having a distance of 50 µm to its nearest neighbors. An excitation force with frequency 1 MHz is applied on their lower end. For the acoustic domain Ω^{a} which is assumed to be air, a cuboid of length and width 1200 µm and height 400 µm is chosen. Due to symmetry reasons, we use as computational domain one quarter of the original one. In Fig. 14(a), a part of the finite element grid is shown. If one had to employ

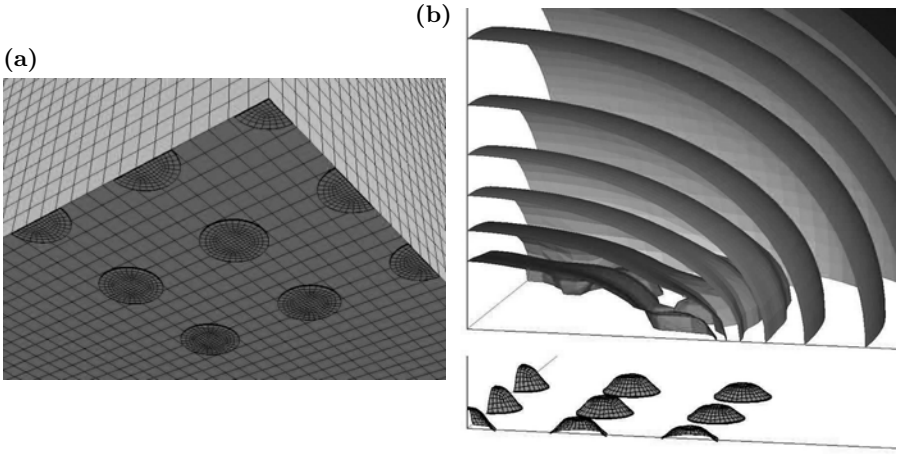


Fig. 14. **(a)** cylindrical plates attached to the fluid domain, **(b)** isosurfaces of the acoustic potential, deformed plates

matching grids, it would be quite difficult to generate them, and if the mesh-width could not be very small over the whole domain, the resulting element shapes would possibly result in a poor approximation of the solution. The nonconforming approach admits to use the grid desired for each subdomain regardless of the grids for the other subdomains. Moreover, it is very easy to add more plates or to change their position. Only the corresponding part of the coupling matrix would have to be (re-)calculated. Figure 14(b) shows the solution after 100 time steps of 3.5 ns, in the upper part isosurfaces of the velocity potential, in the lower part the deformed structure, where the deformations are magnified by a factor of 1000. Since we do not have any moving bodies involved, we note that in each time step, the system matrix which has to be inverted is the same. Thus, it is possible to factor this system matrix only once, and then to reuse the factorization in each step.

7.3 Contact

Our next application will be given in terms of an overlapping domain decomposition with nested subdomains, as depicted in Fig. 15(a). Our goal is to improve a given finite element solution u_H on the global domain Ω by calculating a better solution u_h on the patch ω. It can be achieved by employing mortar techniques to project u_H onto the patch boundary Γ yielding Dirichlet boundary conditions for u_h. For Γ, we exclude the regions $\partial\omega \cap \partial\Omega \neq \emptyset$, where u_h respects the boundary conditions of the global problem. In [17], within the context of elliptic variational equalities, we prove the following a priori estimate for standard conforming finite elements of order r and s on $\mathcal{T}_H$ and $\mathcal{T}_h$, respectively.

Theorem 2. *Let $B \supset \omega^c$ such that $d = \operatorname{dist}(\partial B \setminus \partial\Omega, \partial\omega^c \setminus \partial\Omega) > 0$. Then for H small enough and u regular enough, there exists a constant C depending on d such that*

$$\|u - u_H\|_{1,\omega^c} + \|u - u_h\|_{1,\omega} \leq Ch^s |u|_{s+1,\omega} + CH^r |u|_{r+1,B} + CH^{r+1} |u|_{r+1,\Omega} \,. \tag{38}$$

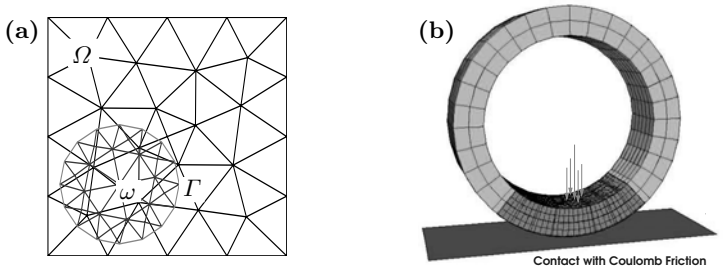


Fig. 15. (a) two nested domains, **(b)** unilateral contact problem: global domain Ω, overlapping patch ω

We note that the last term in (38) is the fundamental difference of our approach to the estimates obtained by standard adaptive finite element methods. It is due to the fact that in our one-directionally coupled approach no pollution effect is taken into account. We also would like to refer to [20] for a similar approach, where the last term vanishes at the expense of a bi-directional volume coupling. In our approach, due to the facts that the coupling matrix is based on a "volume-to-surface-"coupling, and that u_h has no influence on u_H, the assembly and the solution of the system of linear equations can be performed faster.

For the following example, we apply our approach to the solution of a unilateral frictional contact problem, as illustrated in the right picture of Fig. 15(b). A cylindrical ring Ω of outer radius 0.4, inner radius 0.3, and height 0.25 is subject to a surface traction concentrated on a small part of its inner boundary. It is pressed against a planar obstacle which constitutes a tangent plane prior to the contact. On Ω, we have to solve a variational inequality of the following form: find $(\mathbf{u}_H, \boldsymbol{\lambda}_H) \in X_H \times M_H^+$ such that

$$a_H(\mathbf{u}_H, \mathbf{v}_H) + b_H(\mathbf{v}_H, \boldsymbol{\lambda}_H) = f(\mathbf{v}_H)\,, \qquad \mathbf{v}_H \in X_H\,, \tag{39a}$$

$$b_H^{\mathrm{n}}(\mathbf{u}_H, \boldsymbol{\mu}_H - \boldsymbol{\lambda}_H) \leq \langle d, (\boldsymbol{\mu}_H - \boldsymbol{\lambda}_H)\,\mathbf{n}\rangle_{\Gamma_{\mathrm{C}}^{\mathrm{s}}}\,, \quad \boldsymbol{\mu}_H \in M_H^+\,, \tag{39b}$$

$$b_H^{\mathrm{t}}(\mathbf{u}_H, \boldsymbol{\nu}_H - \boldsymbol{\lambda}_H) \leq 0\,, \qquad \boldsymbol{\nu}_H \in \Lambda_H(\mathbf{u}_H)\,. \tag{39c}$$

In (39), the bilinear form $a_H(\cdot,\cdot)$ and the space X_H are obtained from (2)–(4) with corresponding boundary conditions, whereas the bilinear form $b_H(\cdot,\cdot) = b_H^{\mathrm{n}}(\cdot,\cdot)+b_H^{\mathrm{t}}(\cdot,\cdot)$ and the spaces M_H^+, Λ_H are responsible for incorporating the contact conditions. A close account and discussion of these contact conditions is given in the report of the project B8 within this collection. There, we also show how the advantages of the dual approach can be fully exploited for the efficient solution of (39).

In order to improve the solution $\mathbf{u}_H$ of (39), we solve on the patch ω problem (39) with H replaced by h, where the elements of the solution space X_h respect the boundary condition $\mathbf{u}_h = \Pi_h \mathbf{u}_H$ with Π_h denoting the mortar projection onto Γ. In Fig. 16, we demonstrate the effect of our approach, (a) visualizes a reference solution obtained on a quite fine global grid, (b) shows the solution $\mathbf{u}_H$ calculated on a coarse global grid. Comparing with the reference solution, we observe that $\mathbf{u}_H$ does not approximate the solution very well. In Fig. 16(c), the improved solution $\mathbf{u}_h$ is plotted on the patch ω. It is obvious that $\mathbf{u}_h$ resolves the characteristics of the reference solution much better.

7.4 Heat Transfer

We consider the time-dependent problem of a body sliding against another body causing heat generation on the interface due to friction, see also [35, 41, 49]. Neglecting the mechanical part, we assume that the pressure on the contact interface is a known function. A complete thermo-mechanical model

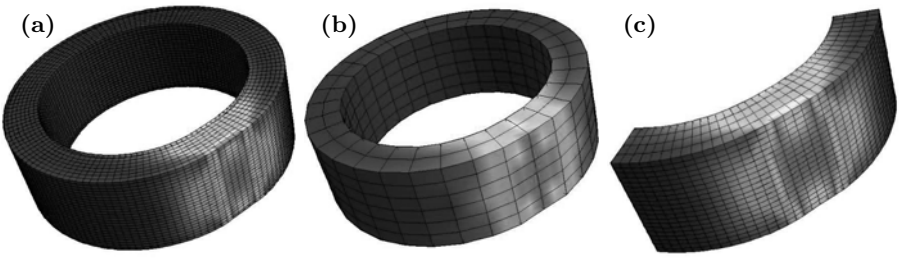


Fig. 16. **(a)** reference solution, **(b)** solution $\mathbf{u}_H$, **(c)** solution $\mathbf{u}_h$

can be found in [35, 49]. As before, we consider the domain Ω, decomposed into two non-overlapping subdomains Ω^{m} and Ω^{s} with the common interior interface Γ. The heat conduction equation for both bodies can be written as

$$\rho_i c_i \dot{u}_i - \operatorname{div} \alpha_i \operatorname{grad} u_i = f_i \ \text{ in } \ \Omega_i \times (0,T)\,, \quad i = \mathrm{m}, \mathrm{s}\,, \tag{40}$$

where ρ_i is the density, c_i is the specific heat, α_i is the thermal conductivity, and f_i is the heat source, $i = \mathrm{m}, \mathrm{s}$. The initial temperature is prescribed as $u_i(0,x) = u_i^0(x)$, $i = 1,2$. The transmission conditions on the interface are given in terms of the heat fluxes $\alpha_1 \operatorname{grad} u^{\mathrm{m}} \cdot \mathbf{n}^{\mathrm{m}}$ and $\alpha_2 \operatorname{grad} u^{\mathrm{s}} \cdot \mathbf{n}^{\mathrm{s}}$ across Γ from the first and the second body, respectively with

$$\alpha_1 \operatorname{grad} u^{\mathrm{m}} \cdot \mathbf{n}^{\mathrm{m}} = c_D \beta v p - \hat{a}[u]\,, \text{ and } \alpha_2 \operatorname{grad} u^{\mathrm{s}} \cdot \mathbf{n}^{\mathrm{s}} = c_D(1-\beta) v p + \hat{a}[u]\,,$$

where c_D is the frictional constant, $\hat{a}$ is the heat transfer parameter, and $\beta = \frac{\alpha_1}{\alpha_1+\alpha_2}$. The functions v and p are the relative velocity and the pressure at the real contact interface at the point x and at time t, respectively. We assume that the heat transfer parameter $\hat{a}$ is directly proportional to the contact pressure p on the contact interface so that $\hat{a} = \bar{\gamma}_c p$, where $\bar{\gamma}_c$ is the heat transfer coefficient, see [41, 35]. We consider the boundary conditions of Robin type on $\Gamma_i \times (0,T)$, $\Gamma_i := \partial\Omega_i \backslash \Gamma$ for $i = 1,2$, with $\alpha_i \operatorname{grad} u_i \cdot \mathbf{n}_i = -\tilde{a}_i u_i$, where $\tilde{a}_i$ is the coefficient of convective heat transfer for Ω_i. We point out that here the Neumann jump of the solution g_N is given by $g_N = c_D v p$, and the jump of the solution is coupled with the heat flux on the interface Γ. Introducing the heat flux on Γ with $\lambda = \alpha_2 \nabla u^{\mathrm{s}}(t,x) \cdot \mathbf{n}^{\mathrm{s}}$ being the Lagrange multiplier, the mortar formulation is attained by writing the weak form of

$$\lambda = c_D(1-\beta) v p + \bar{\gamma}_c p\,[u]\,. \tag{41}$$

In contrast to the previous sections, the weak formulation cannot be achieved by multiplying (41) by a dual test function. Since $\lambda \in (H^{1/2}(\Gamma))'$, we have to multiply (41) with a more regular test function $\phi \in H^{1/2}(\Gamma)$ to get the weak form. This leads to a Petrov–Galerkin mortar formulation. Defining $g_D = -c_D(1-\beta) v p$, this formulation in the discrete setting can be written as: find $(u_h(t), \lambda_h(t)) \in L^2(0,T;\, X_h \times M_h)$ so that

$$\begin{aligned}(\rho_k c_k \dot{u}_h(t), v_h) + a(u_h(t), v_h) \;+\; b_1(v_h, \lambda_h(t)) &= f(v_h)\,, \quad v_h \in X_h\,,\\ b_2(u_h(t), \phi_h) \,-\, (\lambda_h(t),\, \phi_h)_\Gamma &= g(\phi_h)\,, \quad \phi_h \in W_h^{\mathrm{s}}\,,\end{aligned} \tag{42}$$

where the forms $a(\cdot,\cdot)$, $f(\cdot)$, $g(\cdot)$, and the sets X_h and M_h are defined in the obvious way, whereas the coupling is realized by

$$b_1(v_h, \lambda_h) = (\lambda_h,\, [v_h])_\Gamma \quad \text{and} \quad b_2(u_h, \phi_h) = (\bar{\gamma}_c p[u_h],\, \phi_h)_\Gamma\,.$$

As before, W_h^{s} denotes the trace of the finite element space from the slave side of the interface Γ. Apart from the mass term containing $\dot{u}_h$ in (42), we now face a generalized saddle point problem. The conditions for the wellposedness of such problems can be found in [7]. Similar to the situation in Sect. 7.1, these conditions have to be incorporated into the variational framework of evolutionary problems, [8].

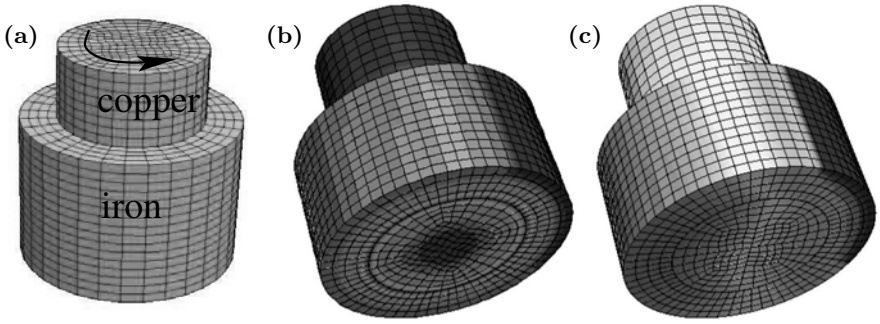


Fig. 17. **(a)** problem setting and grid, **(b)** solution after 1 time step, **(c)** solution after 200 time steps

As a numerical example, we consider the situation illustrated in Fig. 17(a). A cylinder Ω^{m} of radius 0.25 cm and height 0.8 cm made of copper, ρ_{m}=8960 kg/m^3, $c_{\mathrm{m}} = 385\,\mathrm{J/(kg\,K)}$, $\alpha_{\mathrm{m}} = 386\,\mathrm{W/(m\,K)}$, rotates inside a cylinder ring Ω^{s} of thickness 0.15 cm made of iron, $\rho_{\mathrm{s}} = 7860\,\mathrm{kg/m^3}$, $c_{\mathrm{s}} = 444\,\mathrm{J/(kg\,K)}$, $\alpha_{\mathrm{s}} = 80.2\,\mathrm{W/(m\,K)}$. In Fig. 17(b) and (c), the initial heat distribution and the one after 200 time steps is visualized, respectively. We observe that the position of the grid on Ω^{m} with respect to the grid on Ω^{s} changes considerably between the two pictures. In the context of sliding meshes, the advantages of non-matching grids become very obvious. Especially, no complicated re-meshing process is necessary from one time step to the next. In each time step, the original grid can be used for calculation.

8 Conclusions

Since its start in May 2002, the project C12 has helped to accomplish many important contributions to the numerically efficient treatment of coupled field problems. In the following, we summarize the achieved results by listing our research papers which can be seen in direct connection to the project.

One main focus is the extension of the concept of dual Lagrange multipliers from low order and straight interfaces towards higher order and curvilinear interfaces. Concerning the extension to higher order, we would like to mention [30] concentrating on quadratic dual multipliers, [38] dealing with the influence of quadrature formulas, and [47] introducing continuous dual multipliers, [32] dealing with Serendipity elements, [29] providing a general and rigorous mathematical framework for the 3D case, and [33] giving a construction principle for arbitrary order in the 2D case based on Gauß–Lobatto nodes. For the treatment of curvilinear interfaces, we develop in [15] a general framework for the 2D case, in [16] a sometimes necessary modification for 2D elasticity, and in [19] an extension to arbitrary quadrilateral surface meshes in 3D.

Another main topic of the project is the investigation of overlapping domain decompositions with nested domains. In [17], a priori estimates for a one-directionally coupled model problem are derived, and in [14] a mathematical framework in terms of generalized saddle point problems is established. An application within the elasticity setting for domains with holes is investigated in [40]. A major application is the treatment of eddy current problems involving moving conductors. These contributions are concerned with the investigation of the wellposedness for the stationary case in [37] and for the dynamic case in [39], as well as with the presentation of numerical examples in [12] and [13].

Additionally, the project deals with the development and analysis of fast iterative solvers for coupled problems. Our contributions can be grouped into the two categories of Dirichlet–Neumann algorithms and multigrid schemes. For the first category, we added with [27] a new approach for the treatment of multi-body contact problems with and without Coulomb friction, for which we proved in [10] uniform convergence rates. Concerning multigrid methods, we present an abstract framework for mortar finite elements in [45], as well as an additive variant in [9]. Moreover, the proof of level-independent convergence rates for a $\mathcal{V}$-cycle approach is given in [48].

Besides the applications for overlapping methods mentioned above, many applications for non-overlapping decompositions are considered. It is worth to mention [31] investigating interface problems, [34] presenting various applications, [18] investigating coupled problems in nonlinear elasticity, and [11] dealing with applications to computational acoustics. We also would like to mention the close relation to the work undertaken within the project B8.

Although many questions have been answered due to the research work achieved within this project, there are still many tasks to be accomplished for the proper mathematical treatment of nonconforming coupling problems. Now, the main challenge for the nearest future is to evolve from the successful

treatment of application-oriented, yet still academic examples, towards an efficient and robust numerical simulation of real-life problems. The achieved results provide a sound and most promising basis for this goal.

References

1. D. N. Arnold, D. Boffi, and R. S. Falk. Approximation by quadrilateral finite elements. *Math. Comp.*, 71(239):909–922, 2002.
2. A. Bamberger, R. Glowinski, and Q. H. Tran. A domain decomposition method for the acoustic wave equation with discontinuous coefficients and grid change. *SIAM J. Numer. Anal.*, 34(2):603–639, 1997.
3. F. Ben Belgacem. The mortar finite element method with Lagrange multipliers. *Numer. Math.*, 84(2):173–197, 1999.
4. C. Bernardi, Y. Maday, and A. T. Patera. A new nonconforming approach to domain decomposition: the mortar element method. In *Nonlinear partial differential equations and their applications. Collège de France Seminar, Vol. XI (Paris, 1989–1991)*, volume 299 of *Pitman Res. Notes Math. Ser.*, pages 13–51. Longman Sci. Tech., Harlow, 1994.
5. F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*. Springer-Verlag, New York, 1991.
6. K. S. Chavan, B. P. Lamichhane, and B. I. Wohlmuth. Locking-free finite element methods for linear and nonlinear elasticity in 2D and 3D. Technical Report 13, University of Stuttgart, SFB 404, 2005. To appear in Comp. Meth. Appl. Mech. Engrg.
7. P. Ciarlet, Jr., J. Huang, and J. Zou. Some observations on generalized saddle-point problems. *SIAM J. Matrix Anal. Appl.*, 25(1):224–236, 2003.
8. R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 5: Evolution problems I.* Springer-Verlag, Berlin, 1992.
9. M. Dryja, A. Gantner, O. B. Widlund, and B. I. Wohlmuth. Multilevel additive Schwarz preconditioner for nonconforming mortar finite element methods. *J. Numer. Math.*, 12(1):23–38, 2004.
10. C. Eck and B. Wohlmuth. Convergence of a contact-Neumann iteration for the solution of two-body contact problems. *Math. Models Methods Appl. Sci.*, 13(8):1103–1118, 2003.
11. B. Flemisch, M. Kaltenbacher, and B. I. Wohlmuth. Elasto-acoustic and acoustic-acoustic coupling on nonmatching grids. Technical Report 10, University of Stuttgart, SFB 404, 2005. To appear in Internat. J. Numer. Methods Engrg.
12. B. Flemisch, Y. Maday, F. Rapetti, and B. Wohlmuth. Coupling scalar and vector potentials on nonmatching grids for eddy currents in a moving conductor. *J. Comput. Appl. Math.*, 168(1-2):191–205, 2004.
13. B. Flemisch, Y. Maday, F. Rapetti, and B. Wohlmuth. Scalar and vector potentials' coupling on nonmatching grids for the simulation of an electromagnetic brake. *COMPEL*, 24(3):1061–1070, 2005.
14. B. Flemisch, M. Mair, and B. I. Wohlmuth. Nonconforming discretization techniques for overlapping domain decompositions. In *M. Feistauer et al., editors, Numerical mathematics and advanced applications. Proceedings of Enumath 2003, Prague, Czech Republic, August 18–22, 2003*, pages 316–325. Springer, Berlin, 2004.

15. B. Flemisch, J. M. Melenk, and B. I. Wohlmuth. Mortar methods with curved interfaces. *Appl. Numer. Math.*, 54(3-4):339–361, 2005.
16. B. Flemisch, M. A. Puso, and B. I. Wohlmuth. A new dual mortar method for curved interfaces: 2D elasticity. *Internat. J. Numer. Methods Engrg.*, 63(6):813–832, 2005.
17. B. Flemisch and B. I. Wohlmuth. A domain decomposition method on nested domains and nonmatching grids. *Numer. Methods Partial Differential Equations*, 20(3):374–387, 2004.
18. B. Flemisch and B. I. Wohlmuth. Nonconforming methods for nonlinear elasticity problems. Technical Report 03, University of Stuttgart, SFB 404, 2005. To appear in the Proceedings of the 16th International Conference on Domain Decomposition Methods.
19. B. Flemisch and B. I. Wohlmuth. Stable Lagrange multipliers for quadrilateral meshes of curved interfaces in 3D, IANS preprint 2005/005. Technical report, University of Stuttgart, 2005. To appear in Comp. Meth. Appl. Mech. Engrg.
20. R. Glowinski, J. He, A. Lozinski, J. Rappaz, and J. Wagner. Finite element approximation of multi-scale elliptic problems using patches of elements. *Numer. Math.*, 101(4):663–687, 2005.
21. J. Gopalakrishnan. *On the mortar finite element method.* PhD thesis, Texas A&M University, 1999.
22. P. Hauret. *Numerical methods for the dynamic analysis of twoscale incompressible nonlinear structures.* PhD thesis, Ecole Polytechnique, Paris, 2004.
23. P. Hauret and P. L. Tallec. Dirichlet–Neumann preconditioners for elliptic problems with small disjoint geometric refinements on the boundary. Technical Report 552, CMAP - Ecole Polytechnique, 2004.
24. S. Hüeber and B. I. Wohlmuth. A primal-dual active set strategy for non-linear multibody contact problems. *Comput. Methods Appl. Mech. Engrg.*, 194:3147–3166, 2005.
25. T. Hughes. *The Finite Element Method.* Prentice-Hall, New Jersey, 1987.
26. C. Kim, R. D. Lazarov, J. E. Pasciak, and P. S. Vassilevski. Multiplier spaces for the mortar finite element method in three dimensions. *SIAM J. Numer. Anal.*, 39(2):519–538, 2001.
27. R. H. Krause and B. I. Wohlmuth. A Dirichlet-Neumann type algorithm for contact problems with friction. *Comput. Vis. Sci.*, 5(3):139–148, 2002.
28. B. P. Lamichhane. *Higher order mortar finite elements with dual Lagrange multiplier spaces and applications.* PhD thesis, University of Stuttgart, 2006.
29. B. P. Lamichhane, R. P. Stevenson, and B. I. Wohlmuth. Higher order mortar finite element methods in 3D with dual Lagrange multiplier bases. *Numer. Math.*, 102(1):93–121, 2005.
30. B. P. Lamichhane and B. I. Wohlmuth. Higher order dual Lagrange multiplier spaces for mortar finite element discretizations. *Calcolo*, 39(4):219–237, 2002.
31. B. P. Lamichhane and B. I. Wohlmuth. Mortar finite elements for interface problems. *Computing*, 72(3-4):333–348, 2004.
32. B. P. Lamichhane and B. I. Wohlmuth. A quasi-dual Lagrange multiplier space for serendipity mortar finite elements in 3D. *M2AN Math. Model. Numer. Anal.*, 38(1):73–92, 2004.
33. B. P. Lamichhane and B. I. Wohlmuth. Biorthogonal bases with local support and approximation properties. Technical Report 02, University of Stuttgart, SFB 404, 2005. To appear in Math. Comp.

34. B. P. Lamichhane and B. I. Wohlmuth. Mortar finite elements with dual Lagrange multipliers: some applications. In *Kornhuber, Ralf (ed.) et al., Domain decomposition methods in science and engineering. Selected papers of the 15th International Conference on Domain Decomposition, Berlin, Germany, July 21-25, 2003*, pages 319–326. Springer, Berlin, 2005.
35. T. A. Laursen. *Computational contact and impact mechanics.* Springer-Verlag, Berlin, 2002.
36. M. J. Lighthill. On sound generated aerodynamically. I. General theory. *Proc. Roy. Soc. London. Ser. A.*, 211:564–587, 1952.
37. Y. Maday, F. Rapetti, and B. I. Wohlmuth. Coupling between scalar and vector potentials by the mortar element method. *C. R. Math. Acad. Sci. Paris*, 334(10):933–938, 2002.
38. Y. Maday, F. Rapetti, and B. I. Wohlmuth. The influence of quadrature formulas in 2D and 3D mortar element methods. In *Recent developments in domain decomposition methods (Zürich, 2001)*, volume 23 of *Lect. Notes Comput. Sci. Eng.*, pages 203–221. Springer, Berlin, 2002.
39. Y. Maday, F. Rapetti, and B. I. Wohlmuth. Mortar element coupling between global scalar and local vector potentials to solve eddy current problems. In *F. Brezzi et al., editors, Numerical mathematics and advanced applications. Proceedings of Enumath 2001, Ischia, July 2001*, pages 847–865. Springer, Berlin, 2003.
40. M. Mair and B. I. Wohlmuth. A domain decomposition method for domains with holes using a complementary decomposition. *Comput. Methods Appl. Mech. Engrg.*, 193(45-47):4961–4978, 2004.
41. T. W. McDevitt and T. A. Laursen. A mortar-finite element formulation for frictional contact problems. *Internat. J. Numer. Methods Engrg.*, 48(10):1525–1547, 2000.
42. P. Oswald and B. I. Wohlmuth. On polynomial reproduction of dual FE bases. In *Domain decomposition methods in science and engineering (Lyon, 2000)*, Theory Eng. Appl. Comput. Methods, pages 85–96. Internat. Center Numer. Methods Eng. (CIMNE), Barcelona, 2002.
43. M. Puso. A 3D mortar method for solid mechanics. *Internat. J. Numer. Methods Engrg.*, 59(3):315–336, 2004.
44. E. Stein and M. Rüter. Finite element methods for elasticity with error-controlled discretization and model adaptivity. In *E. Stein, R. de Borst and T.J.R. Hughes, editors, Encyclopedia of Computational Mechanics*, pages 5–58. Wiley, Chichester, 2004.
45. C. Wieners and B. I. Wohlmuth. Duality estimates and multigrid analysis for saddle point problems arising from mortar discretizations. *SIAM J. Sci. Comput.*, 24(6):2163–2184, 2003.
46. B. I. Wohlmuth. A mortar finite element method using dual spaces for the Lagrange multiplier. *SIAM J. Numer. Anal.*, 38(3):989–1012, 2000.
47. B. I. Wohlmuth. A comparison of dual Lagrange multiplier spaces for mortar finite element discretizations. *M2AN Math. Model. Numer. Anal.*, 36(6):995–1012, 2002.
48. B. I. Wohlmuth. A $\mathcal{V}$-cycle multigrid approach for mortar finite elements. *SIAM J. Numer. Anal.*, 42(6):2476–2495, 2005.
49. P. Wriggers. *Computational contact mechanics.* Wiley, 2002.

Appendix

List of Contributors

Claus G. Bayreuther
P+Z Engineering GmbH
Anton-Ditt-Bogen 3
80939 München, Germany

Martin Becker
Institute of Applied Mechanics
University of Stuttgart
Pfaffenwaldring 7
70569 Stuttgart, Germany

Stefan Berres
Institute for Applied Analysis and
Numerical Simulation
University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany

Raimund Bürger
Departamento de
Ingeniería Matemática
Universidad de Concepción
Casilla 160-C
Concepción, Chile

Holger Class
Institute of Hydraulic Engineering
University of Stuttgart
Pfaffenwaldring 61
70569 Stuttgart, Germany

Michael Dumbser
Institute of Aerodynamics and
Gas Dynamics
University of Stuttgart
Pfaffenwaldring 21
70569 Stuttgart, Germany

Peter Eberhard
Institute of Engineering and
Computational Mechanics
University of Stuttgart
Pfaffenwaldring 9
70569 Stuttgart, Germany

Wolfgang Ehlers
Institute of Applied Mechanics (CE)
University of Stuttgart
Pfaffenwaldring 9
70569 Stuttgart, Germany

Matthias Fischer
Institute of Applied and
Experimental Mechanics
University of Stuttgart
Pfaffenwaldring 9
70569 Stuttgart, Germany

Bernd Flemisch
Institute for Applied Analysis and
Numerical Simulation
University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany

Christiane Förster
Institute of Structural Mechanics
University of Stuttgart
Pfaffenwaldring 7
70569 Stuttgart, Germany

Lothar Gaul
Institute of Applied and
Experimental Mechanics
University of Stuttgart
Pfaffenwaldring 9
70569 Stuttgart, Germany

Steffen Genkinger
Institute of Structural Mechanics
University of Stuttgart
Pfaffenwaldring 7
70569 Stuttgart, Germany

Ercan Gürses
Institute of Applied Mechanics
University of Stuttgart
Pfaffenwaldring 7
70569 Stuttgart, Germany

Jens Harting
Institute for Computational Physics
University of Stuttgart
Pfaffenwaldring 27
70569 Stuttgart, Germany

Martin Hecht
Institute for Computational Physics
University of Stuttgart
Pfaffenwaldring 27
70569 Stuttgart, Germany

Rainer Helmig
Institute of Hydraulic Engineering
University of Stuttgart
Pfaffenwaldring 61
70569 Stuttgart, Germany

Hans J. Herrmann
Institute for Building Materials
ETH Hönggerberg
HIF E 12
8093 Zürich, Switzerland

Stefan Hüeber
Institute for Applied Analysis and
Numerical Simulation
University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany

Yu Jiang
Institute of Engineering and
Computational Mechanics
University of Stuttgart
Pfaffenwaldring 9
70569 Stuttgart, Germany

Dorothee Knees
Weierstrass Institute for
Applied Analysis and Stochastics
Mohrenstraße 39
10117 Berlin, Germany

Andreas Mainik
Institute for Analysis,
Dynamics and Modeling
University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany

Sean McNamara
Institute for Computational Physics
University of Stuttgart
Pfaffenwaldring 27
70569 Stuttgart, Germany

Christian Miehe
Institute of Applied Mechanics
University of Stuttgart
Pfaffenwaldring 7
70569 Stuttgart, Germany

Alexander Mielke
Weierstrass Institute for
Applied Analysis and Stochastics
Mohrenstraße 39
10117 Berlin, Germany

Claus-Dieter Munz
Institute of Aerodynamics and Gas Dynamics
University of Stuttgart
Pfaffenwaldring 21
70569 Stuttgart, Germany

Beate Muth
Institute of Engineering and Computational Mechanics
University of Stuttgart
Pfaffenwaldring 9
70569 Stuttgart, Germany

Malte Neumann
Institute of Structural Mechanics
University of Stuttgart
Pfaffenwaldring 7
70569 Stuttgart, Germany

Jennifer Niessner
Institute of Hydraulic Engineering
University of Stuttgart
Pfaffenwaldring 61
70569 Stuttgart, Germany

Günther Of
Institute for Computational Mathematics
Graz University of Technology
Steyrergasse 30
8010 Graz, Austria

Ulrich Ölmann
Institute of Hydraulic Engineering
University of Stuttgart
Pfaffenwaldring 61
70569 Stuttgart, Germany

Ekkehard Ramm
Institute of Structural Mechanics
University of Stuttgart
Pfaffenwaldring 7
70569 Stuttgart, Germany

Anna-Margarete Sändig
Institute for Applied Analysis and Numerical Simulation
University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany

Bernd Scholz
Institute of Applied Mechanics (CE)
University of Stuttgart
Pfaffenwaldring 9
70569 Stuttgart, Germany

Thomas Schwartzkopff
Institute of Aerodynamics and Gas Dynamics
University of Stuttgart
Pfaffenwaldring 21
70569 Stuttgart, Germany

Olaf Steinbach
Institute for Computational Mathematics
Graz University of Technology
Steyrergasse 30
8010 Graz, Austria

Jens Utzmann
Institute of Aerodynamics and Gas Dynamics
University of Stuttgart
Pfaffenwaldring 21
70569 Stuttgart, Germany

Wolfgang A. Wall
Chair for Computational Mechanics
Technical University of Munich
Boltzmannstraße 15
85747 Garching, Germany

Wolfgang L. Wendland
Institute for Applied Analysis and Numerical Simulation
University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany

Barbara I. Wohlmuth
Institute for Applied Analysis and Numerical Simulation
University of Stuttgart
Pfaffenwaldring 57
70569 Stuttgart, Germany

Index