Lecture Notes in Applied and Computational Mechanics 81

Kerstin Weinberg
Anna Pandolfi *Editors*

Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems

In Honor of Michael Ortiz's 60th Birthday

Springer

Lecture Notes in Applied and Computational Mechanics

Volume 81

About this Series

This series aims to report new developments in applied and computational mechanics—quickly, informally and at a high level. This includes the fields of fluid, solid and structural mechanics, dynamics and control, and related disciplines. The applied methods can be of analytical, numerical and computational nature.

More information about this series at http://www.springer.com/series/4623

Kerstin Weinberg · Anna Pandolfi
Editors

Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems

In Honor of Michael Ortiz's 60th Birthday

Springer

Editors

Kerstin Weinberg
Universität Siegen
Siegen
Germany

Anna Pandolfi
Politecnico di Milano
Milan
Italy

ISSN 1613-7736 ISSN 1860-0816 (electronic)
Lecture Notes in Applied and Computational Mechanics
ISBN 978-3-319-39021-5 ISBN 978-3-319-39022-2 (eBook)
DOI 10.1007/978-3-319-39022-2

Library of Congress Control Number: 2016940007

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

During the past two decades research in the field of computational mechanics has progressed remarkably, mainly because of the development of a sound mathematical background and the introduction of new efficient computational strategies. Beyond the classical finite element method, several innovative techniques and novel approaches for the analysis of microstructural evolution, growth, damage, and structural failure in multi-field and multi-scale problems have emerged vigorously.

With the aim to discuss different computational strategies for multi-field and multi-scale problems, a remarkable group of scientists gathered in September 2014 to the IUTAM symposium "Innovative numerical approaches for materials and structures in multi-field and multi-scale problems". Hosted by the University of Siegen, the venue of the symposium was the Castle Burg Schnellenberg, a mighty fortress located in the green heart of Westphalia, Germany. There we discussed the new horizons and perspectives of multi-field applied mechanics. The symposium covered a large domain of recent research, from computational materials modeling, crystal plasticity, micro-structured materials, and biomaterials to multi-scale simulations of multi-physics phenomena. The pioneering discretization methods for the solution of coupled nonlinear problems at different length scales were particularly emphasized.

The special occasion that motivated the organization of the symposium was the 60th birthday of Professor Michael Ortiz. Along his exceptional career, Michael Ortiz has been at the forefront of computational mechanics, his work being a constant source of inspiration for many. All participants of this symposium are grateful to Michael Ortiz for being such an enthusiastic collaborator, a reliable colleague, an illuminating scientist, and a valuable friend.

The friendship and fellow-feeling felt during the symposium inspired the idea to collect the presentations of some of the convened researchers in a special book. Our choice was to organize a book as part of the series 'Lecture Notes in Applied and Computational Mechanics' (LNACM), which aims to document new high-level developments in applied and computational mechanics. We are happy to present here 13 high-quality contributions of current and past collaborators of Michael

Ortiz. All contributions have undergone full peer review, and we take the occasion to thank the reviewers for their valuable comments.

It is our hope that the present volume will give the reader an insight into the exciting new developments of computational solid mechanics which is still wide open to discovery. The book attempts to provide a flavor of this challenging field and to contribute to its popularity within the mechanics and physics communities.

Siegen
February 2016

Kerstin Weinberg
Anna Pandolfi

Contents

Robust Numerical Schemes for an Efficient Implementation of Tangent Matrices: Application to Hyperelasticity, Inelastic Standard Dissipative Materials and Thermo-Mechanics at Finite Strains

Masato Tanaka, Daniel Balzani and Jörg Schröder

Abstract In this contribution robust numerical schemes for an efficient implementation of tangent matrices in finite strain problems are presented and their performance is investigated through the analysis of hyperelastic materials, inelastic standard dissipative materials in the context of incremental variational formulations, and thermo-mechanics. The schemes are based on highly accurate and robust numerical differentiation approaches which use non-real numbers, i.e., complex variables and hyper-dual numbers. The main advantage of these approaches are that, contrary to the classical finite difference scheme, no round-off errors in the perturbations due to floating-point arithmetics exist within the calculation of the tangent matrices. This results in a method which is independent of perturbation values (in case of complex step derivative approximations if sufficiently small perturbations are chosen). An efficient algorithmic treatment is presented which enables a straightforward implementation of the method in any standard finite-element program. By means of hyperelastic, finite strain elastoplastic, and thermo-elastoplastic boundary value problems, the performance of the proposed approaches is analyzed.

1 Introduction

Many materials utilized in industry are characterized by micro-heterogeneous properties and a number of direct micro-macro transition methods have been proposed in order to estimate the nonlinear stress-strain relationship at every macroscopic point

M. Tanaka (✉)
Toyota Central R&D Labs., Inc., Nagakute, Japan
e-mail: tanamasa@mosk.tytlabs.co.jp

D. Balzani
Institute of Mechanics and Shell Structures, TU Dresden, Dresden, Germany
e-mail: daniel.balzani@tu-dresden.de

J. Schröder
Institute of Mechanics, University Duisburg-Essen, Essen, Germany
e-mail: j.schroeder@uni-due.de

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_1

by detailed modeling of the microstructure, i.e., the representative volume element (RVE), see [2, 8, 19, 25, 30]. Especially when using the implicit finite element method, the constitutive equations and consistent tangent moduli need to be explicitly specified and precisely implemented in computer programs. Whereas the exact calculation of the stresses determines the accuracy of the numerical simulation, the consistent algorithmic tangent moduli are required to achieve quadratic convergence in a Newton iteration scheme as well as to detect material instabilities in localization analysis, cf. e.g., [26]. However, for some complex material models their analytic derivatives can be extremely elaborate and error-prone to be implemented. In those cases numerical differentiation may be a useful alternative in particular to decrease scientific development time, provided that robust methods are available.

Different robust numerical approaches have been proposed in the literature based on either complex-step derivative approximation (CSDA) or hyper-dual-step derivatives (HDSD). The CSDA scheme goes back to [14] and has been used since many years in a different context than followed here, e.g. for the calculation of sensitivities in structural optimization problems [16]. The concept of hyper-dual numbers has been introduced in Fike [7] and also used for structural optimization in [6]. For interested readers, C++ or Matlab implementation packages can be downloaded from the web site by J.A. Fike (http://adl.stanford.edu/hyperdual/#Code), see also a source code example for the implementation of HDNs using operator overloading in Fortran 90/95 in [32]. In a different context, i.e. for the implementation of material models in solid mechanics at finite strains, numerical methods based on such robust derivative approximations have been proposed, cf. [3, 9–11, 23, 24, 31, 32].

These approaches mainly extend Miehe's implementation technique [17] based on finite differences (FD) to obtain more robust approximations of tangent matrices, either for material tangent moduli or for tangent stiffness matrices in finite element formulations. In this contribution we review some numerical schemes which are suitable for the implementation of finite strain constitutive models, and present a new framework for incremental variational formulations based on hyper-dual numbers.

It is organized as follows. Section 2 recapitulates some numerical differentiation techniques, i.e., FD, CSDA and HDSD. Section 3 shows a formulation to numerically compute tangent moduli from stress equations using the CSDA scheme and Sect. 4 shows a formulation to automatically compute both, stress and tangent moduli from a strain energy function of hyperelastic material models using the HDSD scheme. Section 5 extends the HDSD-based implementation method to an inelastic standard dissipative material model by applying it to an incremental variational formulation. In Sect. 6 the numerical calculation of tangent stiffness matrices in the context of nonlinear thermo-mechanical finite element problems is shown.

2 Numerical Differentiation Techniques

For a summary of the individual numerical differentiation methods simple scalar-valued functions of one single scalar quantity are considered. Given the function $f(x)$, most methods are based on the Taylor series expansion

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!} f''(x) + \frac{h^3}{3!} f'''(x) + \cdots . \tag{1}$$

The often-used classical FD schemes consider a real-valued small perturbation h and neglect the higher order terms $\mathscr{O}(h^2)$ and thus end up in the approximation

$$f'_{\text{FD}}(x) \approx \frac{f(x+h) - f(x)}{h}, \tag{2}$$

wherein, $\mathscr{O}(h^2)$ denotes Landau's symbol to describe the asymptotic behavior of the higher-order terms. This approach, however, is highly sensitive with respect to the values of the perturbation h and of reduced accuracy even in the optimal range of values which is typically small. For small h the explicit calculation of $f(x+h)$ leads to roundoff errors and for large h it is not acceptable to disregard higher order terms in (1). To overcome this difficulty of the FD schemes, Lyness [14] devised the CSDA scheme. In order to avoid the direct addition of perturbations along the same (real) axis, the CSDA method uses perturbations along the imaginary axis of complex numbers by replacing h by ih in (1) and by taking the imaginary part one obtains the approximation

$$f'_{\text{CSDA}}(x) \approx \frac{\Im[f(x+ih)]}{h}. \tag{3}$$

$\Im$ denotes the operation of taking the imaginary part of complex functions. This expression provides a high accuracy for very small values of h being remarkably close to the analytic ones. This is due to the fact that no roundoff errors occur in the application of the perturbation itself and technically perturbation values up to e.g. $h = 10^{-99}$ are possible. The main drawback of this approach is that higher-order derivatives can not be directly computed and combinations of FD and CSDA have to be considered, see [13]. This however still suffers from the problems of the FD method. Automatic differentiation (AD) techniques, which are typically derived based on repeated application of the chain rule of differentiation, see e.g. [15], are alternatives to compute highly accurate derivatives. However, these schemes are difficult to be derived and formulated, in particular for tensorial derivatives. An equivalent with a forward type AD, but more practicable approach can be accomplished by using dual numbers, which have been originally introduced by Clifford [5]. Fike [6] developed a higher dimensional version of the dual numbers, the hyper-dual numbers (HDNs), which are mainly characterized by the consideration of more than one imaginary axis. The HDNs of second-order have two non-real unit numbers ε_1 and ε_2 with the properties

$$\varepsilon_1^2 := 0, \quad \varepsilon_2^2 := 0, \quad \varepsilon_1\varepsilon_2 := \varepsilon_2\varepsilon_1, \tag{4}$$

which implies that each of the numbers squared is 0 and products of them are commutative. In order to better understand calculations with hyper-dual numbers we shortly recapitulate some basic characteristics and operations. Given the two HDNs $a = (a_1 + a_2\varepsilon_1 + a_3\varepsilon_2 + a_4\varepsilon_1\varepsilon_2)$ and $b = (b_1 + b_2\varepsilon_1 + b_3\varepsilon_2 + b_4\varepsilon_1\varepsilon_2)$, a and b are equal if and only if all of their real and non-real parts are equal, i.e., $a_1 = b_1, a_2 = b_2$, $a_3 = b_3, a_4 = b_4$. We define the symbols $\Im_{\varepsilon_1}$, $\Im_{\varepsilon_2}$ and $\Im_{\varepsilon_1\varepsilon_2}$ which denote the operations of taking the imaginary parts of HDNs such that $\Im_{\varepsilon_1}[a] = a_2$, $\Im_{\varepsilon_2}[a] = a_3$ and $\Im_{\varepsilon_1\varepsilon_2}[a] = a_4$. As proposed by Fike [6] the HDSD for a scalar function $f := f(x)$ of a single scalar-valued argument x is obtained by inserting $h\varepsilon_1 + h\varepsilon_2$ into h in (1) with exactly vanishing higher order terms $\mathscr{O}(h^3)$. The first derivative $f'(x)$ can then be obtained by taking the coefficient of ε_1 or ε_2, i.e.

$$f'_{\text{HDSD}}(x) = \frac{\Im_{\varepsilon_1}[f(x + h\varepsilon_1 + h\varepsilon_2)]}{h} = \frac{\Im_{\varepsilon_2}[f(x + h\varepsilon_1 + h\varepsilon_2)]}{h}. \tag{5}$$

The second derivative $f''(x)$ can be obtained by taking the coefficient, i.e.

$$f''_{\text{HDSD}}(x) = \frac{\Im_{\varepsilon_1\varepsilon_2}[f(x + h\varepsilon_1 + h\varepsilon_2)]}{h^2}. \tag{6}$$

Note that since expressions (5) and (6) do not have any subtraction operation the roundoff errors do not arise. Furthermore, since $\mathscr{O}(h^3)$ is exactly 0, the expressions do not have any truncation errors, too. This means that the value of h could be completely arbitrary and $h = 1$ could be considered for simpler formulae.

3 Approximation of Material Tangents in Hyperelasticity

In this section, different methods for the numerical approximation of the derivatives of stress tensors with respect to deformation tensors, known as tangent moduli, are compared. For this purpose numerical differentiation schemes pointed out in the previous section are extended to directional derivatives of tensor fields.

3.1 Numerical Approximation of Directional Derivatives

Let $\boldsymbol{A}$ be an arbitrary second-order tensor on $\mathbb{R}^2$ or $\mathbb{R}^3$. Then, the directional derivative of a second-order tensor function $\boldsymbol{Z}(\boldsymbol{X})$ of a second-order tensor argument $\boldsymbol{X}$ in direction $\boldsymbol{A}$ is given by

$$\mathrm{D}\boldsymbol{Z}(\boldsymbol{X})[\boldsymbol{A}] := \lim_{h \to 0} \frac{\boldsymbol{Z}(\boldsymbol{X} + h\boldsymbol{A}) - \boldsymbol{Z}(\boldsymbol{X})}{h} = \frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}} : \boldsymbol{A}. \tag{7}$$

Herein, $\partial\boldsymbol{Z}/\partial\boldsymbol{X}$ is a fourth-order tensor that implies a gradient of $\boldsymbol{Z}$ with respect to $\boldsymbol{X}$. This expression directly yields the formula for the finite difference (FD) approach by skipping the limit and treating h as the perturbation value:

$$\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}} : \boldsymbol{A} \approx \frac{\boldsymbol{Z}(\boldsymbol{X} + h\boldsymbol{A}) - \boldsymbol{Z}(\boldsymbol{X})}{h}, \tag{8}$$

Second-order derivatives can be directly calculated by applying this formula also to the second derivatives and using first-order FD for the individual first-order derivative function evaluations. By replacing the FD term by the complex step derivative approximation, one obtains

$$\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}} : \boldsymbol{A} \approx \frac{\Im[\boldsymbol{Z}(\boldsymbol{X} + ih\boldsymbol{A})]}{h}. \tag{9}$$

Unfortunately, the direct application of CSDA also to the second derivative yields an expression which is not free from associated round-off errors and thus it may also combined with the FD for the second derivative. The only approach which enables second-order derivative approximation of high accuracy independent of the choice of the perturbation values is based on hyper dual numbers. Such second-order derivatives may be important in the context of hyperelasticity when considering the strain energy function whose first derivative yields the stress and the second derivative the tangent moduli. Thus, first and second directional derivatives of a scalar function with respect to deformation tensors are considered now. Let $\boldsymbol{A}, \boldsymbol{B}$ be arbitrary second-order tensors on $\mathbb{R}^2$ or $\mathbb{R}^3$. Then, the directional derivative of a scalar function $z(\boldsymbol{X})$ of a second-order tensor argument $\boldsymbol{X}$ in direction $\boldsymbol{A}$ is given by

$$\mathrm{D}z(\boldsymbol{X})[\boldsymbol{A}] := \lim_{h\to 0} \frac{z(\boldsymbol{X} + h\boldsymbol{A}) - z(\boldsymbol{X})}{h} = \frac{\partial z}{\partial \boldsymbol{X}} : \boldsymbol{A}, \tag{10}$$

with $\partial z / \partial \boldsymbol{X}$ denoting a second-order tensor. The second order directional derivative of a scalar function $z(\boldsymbol{X})$ of a second-order argument $\boldsymbol{X}$ in the two directions $\boldsymbol{A}$, $\boldsymbol{B}$ is

$$\begin{aligned} \mathrm{D}^2 z(\boldsymbol{X})[\boldsymbol{A}, \boldsymbol{B}] &:= \lim_{h,k\to 0} \frac{z(\boldsymbol{X} + h\boldsymbol{A} + k\boldsymbol{B}) - z(\boldsymbol{X} + h\boldsymbol{A}) - z(\boldsymbol{X} + k\boldsymbol{B}) + z(\boldsymbol{X})}{hk} \\ &= \boldsymbol{A} : \frac{\partial^2 z}{\partial \boldsymbol{X} \partial \boldsymbol{X}} : \boldsymbol{B}, \end{aligned} \tag{11}$$

wherein $\partial^2 z / \partial \boldsymbol{X} \partial \boldsymbol{X}$ is a fourth-order tensor that implies a Hessian of z with respect to $\boldsymbol{X}$. This results in the formulae for the HDSD scheme

$$\begin{aligned} \frac{\partial z}{\partial \boldsymbol{X}} : \boldsymbol{A} &= \frac{\Im_{\varepsilon_1}[z(\boldsymbol{X} + h\varepsilon_1 \boldsymbol{A})]}{h} \\ \boldsymbol{A} : \frac{\partial^2 z}{\partial \boldsymbol{X} \partial \boldsymbol{X}} : \boldsymbol{B} &= \frac{\Im_{\varepsilon_1 \varepsilon_2}[z(\boldsymbol{X} + h\varepsilon_1 \boldsymbol{A} + k\varepsilon_2 \boldsymbol{B})]}{hk}, \end{aligned} \tag{12}$$

with small perturbation values h, k. Provided that appropriate energy, stress, strain and directional tensors are substituted in z, $\boldsymbol{Z}$, $\boldsymbol{X}$ and $\boldsymbol{A}$, respectively, the required tangent moduli can be automatically derived as shown in the following.

3.2 Application to Hyperelastic Constitutive Equations

Here the representations in oblique-angled coordinate systems are provided in order to highlight the mappings required for a consistent perturbation calculation. However, for the examples we focus on Cartesian coordinates and then we identify $\boldsymbol{G}^{\mathrm{I}} = \boldsymbol{E}^{\mathrm{I}}$, $\boldsymbol{G}_{\mathrm{I}} = \boldsymbol{E}_{\mathrm{I}}, \boldsymbol{g}^{\mathrm{i}} = \boldsymbol{e}^{\mathrm{i}}, \boldsymbol{g}_{\mathrm{i}} = \boldsymbol{e}_{\mathrm{i}}$. Herein, $\boldsymbol{G}_{\mathrm{I}}$ and $\boldsymbol{G}^{\mathrm{I}}$ denote the co- and contravariant general base vectors in the reference configuration and $\boldsymbol{g}_{\mathrm{i}}$ and $\boldsymbol{g}^{\mathrm{i}}$ in the actual configuration; the Cartesian base vectors in the reference and in the actual configuration are $\boldsymbol{E}^{\mathrm{I}} = \boldsymbol{E}_{\mathrm{I}}$ and $\boldsymbol{e}^{\mathrm{i}} = \boldsymbol{e}_{\mathrm{i}}$. Note that italic and non-italic indices represent components and "name" indices, respectively. For an abbreviated representation here we describe the constitutive equations in a material setting. Then the existence of a strain energy function $\psi := \psi(\boldsymbol{C})$ or $\psi := \psi(\boldsymbol{E})$ with the right Cauchy-Green deformation tensor $\boldsymbol{C}$ and the Green-Lagrange strain tensor $\boldsymbol{E}$ is postulated. The constitutive relation for the second Piola-Kirchhoff stress tensor $\boldsymbol{S}$ then reads

$$\boldsymbol{S} = \frac{\partial \psi}{\partial \boldsymbol{E}} = 2\frac{\partial \psi}{\partial \boldsymbol{C}}. \tag{13}$$

The material tangent moduli $\mathbb{C}$ are defined as the derivative of $\boldsymbol{S}$ with respect to $\boldsymbol{E}$:

$$\mathbb{C} = \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}} = \frac{\partial^2 \psi}{\partial \boldsymbol{E} \partial \boldsymbol{E}}, \quad \text{or} \quad \mathbb{C} = 2\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{C}} = 4\frac{\partial^2 \psi}{\partial \boldsymbol{C} \partial \boldsymbol{C}}. \tag{14}$$

Now, substitute $\boldsymbol{S}$ in $\boldsymbol{Z}$, $\boldsymbol{C}$ in $\boldsymbol{X}$, and $\overset{*}{\boldsymbol{C}}{}^{\mathrm{KL}}$ in $\boldsymbol{A}$ in (8) and (9), where $\overset{*}{\boldsymbol{C}}{}^{\mathrm{KL}}$ is defined as

$$\overset{*}{\boldsymbol{C}}{}^{\mathrm{KL}} = \frac{1}{2}(\boldsymbol{G}^{\mathrm{K}} \otimes \boldsymbol{G}^{\mathrm{L}} + \boldsymbol{G}^{\mathrm{L}} \otimes \boldsymbol{G}^{\mathrm{K}}). \tag{15}$$

Then, one obtains for the finite difference and the complex-step-derivative approximation scheme

$$\mathbb{C}_{\mathrm{FD}}^{IJ\mathrm{KL}} = \frac{S^{IJ}(\boldsymbol{C} + h\,\overset{*}{\boldsymbol{C}}{}^{\mathrm{KL}}) - S^{IJ}(\boldsymbol{C})}{h}, \tag{16}$$

$$\mathbb{C}_{\mathrm{CSDA}}^{IJ\mathrm{KL}} = 2\frac{\Im\left[S^{IJ}(\boldsymbol{C} + ih\,\overset{*}{\boldsymbol{C}}{}^{\mathrm{KL}})\right]}{h}. \tag{17}$$

As mentioned above, the function evaluations for the second Piola-Kirchhoff stresses could be replaced by numerical approximations in terms of the FD or the CSDA approach. However neither of them would lead to a method without round-off errors. This can be achieved by using hyper dual numbers. For this purpose consider Eq. (12) and substitute z by ψ, $\boldsymbol{X}$ by $\boldsymbol{C}$, $\boldsymbol{A}$ by $\overset{*}{\boldsymbol{C}}{}^{\text{IJ}}$ and $\boldsymbol{B}$ by $\overset{*}{\boldsymbol{C}}{}^{\text{KL}}$, where $\overset{*}{\boldsymbol{C}}{}^{\text{IJ}}$ and $\overset{*}{\boldsymbol{C}}{}^{\text{KL}}$ are defined in (15), then one obtains

$$S^{IJ}_{\text{HDSD}} = 2\frac{\Im_{\varepsilon_1}\left[\psi(\boldsymbol{C} + h\varepsilon_1\, \overset{*}{\boldsymbol{C}}{}^{\text{IJ}})\right]}{h}, \tag{18}$$

$$\mathbb{C}^{IJKL}_{\text{HDSD}} = 4\frac{\Im_{\varepsilon_1\varepsilon_2}\left[\psi(\boldsymbol{C} + h\varepsilon_1\, \overset{*}{\boldsymbol{C}}{}^{\text{IJ}} + k\varepsilon_2\, \overset{*}{\boldsymbol{C}}{}^{\text{KL}})\right]}{hk}. \tag{19}$$

3.3 Numerical Examples

The numerical calculation of stresses and tangent moduli of a hyperelastic material model is analyzed. For that purpose the FD, CSDA and HDSD scheme are compared with the implementation of the analytic stresses and moduli. In this study one of the anisotropic polyconvex models proposed in [1] for fiber-reinforced materials is considered where an isotropic part describes the ground substance and an anisotropic part represents the embedded fibers as

$$\psi = \alpha_1(I_1\, I_3^{-1/3} - 3) + \alpha_2(I_3^{\alpha_3} + I_3^{-\alpha_3} - 2) + \sum_{a=1}^{n_\text{f}} \left[\beta_1 \left\langle I_1 J_4^{(a)} - J_5^{(a)} - 2\right\rangle^{\beta_2}\right]. \tag{20}$$

Herein, $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_3 > 0$, $\beta_1 > 0$ and $\beta_2 > 2$ are material parameters, n_f is the number of fiber families and $\langle\bullet\rangle$ denotes the Macaulay bracket. The invariants are given by $I_1 = \text{tr}[\boldsymbol{C}]$, $I_3 = \det[\boldsymbol{C}]$, $J_4^{(a)} = \text{tr}[\boldsymbol{C}\boldsymbol{M}_{(a)}]$ and $J_5^{(a)} = \text{tr}[\boldsymbol{C}^2\boldsymbol{M}_{(a)}]$ with the structural tensor $\boldsymbol{M}_{(a)} = \boldsymbol{a}_{0(a)} \otimes \boldsymbol{a}_{0(a)}$, where $\boldsymbol{a}_{0(a)}$ denotes the fiber direction.

3.3.1 Homogeneous Problem

As a first investigation, a homogeneous test is considered where a specific deformation gradient $\boldsymbol{F}$ is applied. This deformation gradient includes rotations $\boldsymbol{Q}$ as well as deformations $\boldsymbol{F}_0$ and is given by

$$\boldsymbol{F} = \boldsymbol{Q}\boldsymbol{F}_0 \quad \text{with} \quad \boldsymbol{Q} = \boldsymbol{R}_{\theta_1}\boldsymbol{R}_{\theta_2}\boldsymbol{R}_{\theta_3}. \tag{21}$$

Herein, $\boldsymbol{R}_{\theta_1}$, $\boldsymbol{R}_{\theta_2}$, $\boldsymbol{R}_{\theta_3}$ are the individual rotation tensors and the deformations $\boldsymbol{F}_0$ contain dilatation and shear deformation. They are chosen as

$$\boldsymbol{R}_{\theta_1} = \begin{bmatrix} \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} & 0 \\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \boldsymbol{R}_{\theta_2} = \begin{bmatrix} \cos\frac{\pi}{3} & 0 & \sin\frac{\pi}{3} \\ 0 & 1 & 0 \\ -\sin\frac{\pi}{3} & & \cos\frac{\pi}{3} \end{bmatrix},$$
$$\boldsymbol{R}_{\theta_3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ 0 & \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{bmatrix}, \ \boldsymbol{F}_0 = \begin{bmatrix} 1.1 & \gamma & \gamma \\ 0 & 0.9535 & \gamma \\ 0 & 0 & 0.9535 \end{bmatrix}, \quad (22)$$

γ is the amplitude of the shear deformation. The material parameters are chosen as $\alpha_1 = 1.0$, $\alpha_2 = 1.0$, $\alpha_3 = 0.1$, $\beta_1 = 1.0$ and $\beta_2 = 3.0$. Two preferred directions are considered, i.e. $n_\mathrm{f} = 2$, which are defined as $\boldsymbol{a}_{0(1)} = 1/3\left(1\ 2\ 2\right)^T$ and $\boldsymbol{a}_{0(2)} = 1/\sqrt{5}\left(2\ 1\ 0\right)^T$. In this example we focus on the numerical performance of computing the tangent moduli by using different numerical schemes, i.e., the FD approach (16) and the CSDA scheme (17) starting from the stress tensor, and the HDSD scheme (19) starting from the strain energy function (20). We compare their results with the ones of an analytically derived tangent modulus (see [31, 32] for the detailed expressions of stress and tangent moduli of the model (20)). In order to compare the results the relative error e_C is defined as

$$e_C = \left[\sum_{I,J,K,L}\left((\mathbb{C}_\mathrm{analyt})^{IJKL} - (\mathbb{C}_\mathrm{approx})^{IJKL}\right)^2\right]^{\frac{1}{2}} \Big/ \left[\sum_{I,J,K,L}\left((\mathbb{C}_\mathrm{analyt})^{IJKL}\right)^2\right]^{\frac{1}{2}}, \quad (23)$$

where $(\mathbb{C}_\mathrm{analyt})^{IJKL}$ denotes the coefficients of the analytic material tangent modulus tensor in the Cartesian coordinate system, $(\mathbb{C}_\mathrm{approx})^{IJKL}$ denotes the values of the approximated counterparts. This relative error e_C is depicted for each numerical approximation scheme in Fig. 1. As can be seen, the FD approach shows a quite sensitive behavior having its optimal accuracy of $e_C \approx 10^{-7}$ at perturbation value of

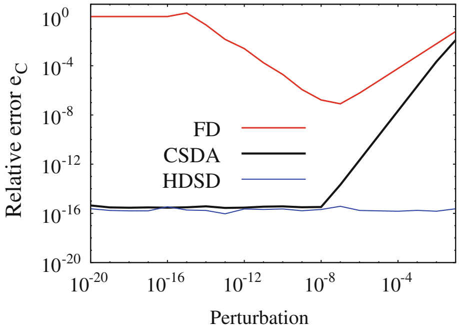


Fig. 1 Relative errors e_C of approximated tangent moduli by using the FD, the CSDA and the HDSD scheme. Perturbation values are varied from 1.0×10^{-20} to 1.0×10^{-1}. In the case of the HDSD, we fix $h = 1.0$ and instead vary the perturbation k in Eq. (19)

$\approx 10^{-7}$. For increasing and decreasing perturbation values the error increases. The CSDA approach already reaches an accuracy of $e_C \approx 10^{-7}$ for perturbations smaller than 10^{-4} and reaches an error at computer accuracy for perturbation values smaller than approximately 10^{-9}. The HDSD is independent on the perturbation value and always yields an error of $e_C \approx 10^{-16}$ which is computer accuracy.

3.3.2 Cook-Type Problem

In order to analyze the influence of the numerical differentiation schemes onto the convergence behavior of Newton-Raphson iterations required for the solution of nonlinear boundary value problems, the formulations described above are implemented in the general-purpose finite-element program FEAP, developed by R.L. Taylor (http://www.ce.berkeley.edu/projects/feap/). In this example, we analyze a Cook-type cantilever beam which is schematically illustrated in Fig. 2a. Now, only one fiber family ($n_\mathrm{f} = 1$) is considered which is oriented in direction $\boldsymbol{a}_{(1)} = 1/\sqrt{3}\,\big(1\ 1\ 1\big)^T$, in order to induce some torsion around the x-axis accompanied by some out of x-y-bending leading to a pronounced combined bending/torsion deformation. The set of material parameters is chosen as $\alpha_1 = 6.0$, $\alpha_2 = 100.0$, $\alpha_3 = 5.0$, $\beta_1 = 100.0$ and $\beta_2 = 2.5$. The maximum load p_0 is increased in equidistant load steps until the ultimate maximum load $p_0 = 5.0$ is reached. In this example, we compare different versions of numerical approximations. Starting from the strain energy function (20) we apply the HDSD scheme (18), (19), in order to directly compute both, stresses and moduli. Then, the FD scheme is applied twice to also

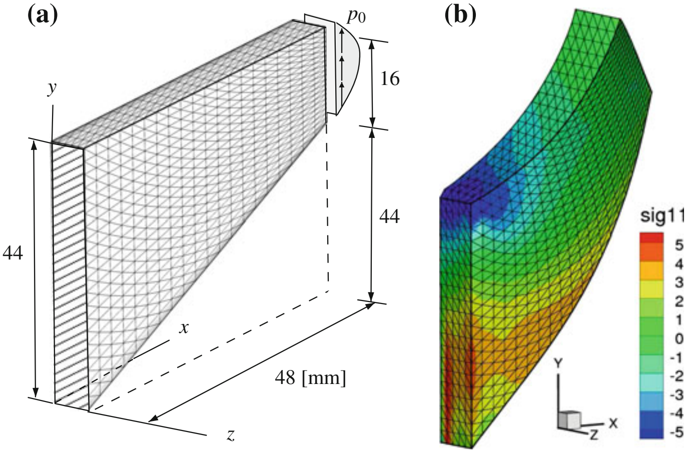


Fig. 2 **a** Geometry with boundary conditions. The beam is fixed in all directions at $x = 0$ (*dashed area*) and a distributed load is applied at the opposite side in vertical direction leading to a bending-dominated deformation. **b** Deformed configuration with distribution of Kirchhoff stresses τ^{11} in x-direction

compute stresses and moduli numerically based on the strain energy function. As a third option we implement the analytic expression for the first derivative of the strain energy and compute then numerically the tangent moduli using the CSDA scheme (17). The Kirchhoff stress distributions τ_{xx} in the most deformed configuration are depicted in Fig. 2b. For any numerical differentiation scheme, the stress distribution is very similar to the analytic implementation of stress and moduli. The Euclidean norm of the residual vector of the discretized weak form of equilibrium for the calculations based on the analytic tangent, the FD and HDSD approach is plotted in Table 1, where $h_S = h_C = 1$ is considered for the HDSD. Even the values of the residuals are almost identical for the HDSD method compared with the ones resulting from an analytic implementation of stress and moduli. The values for the FD method differ by orders of magnitudes and thus, no quadratic convergence is observed. Next, the computing time of the different approaches to calculate the stresses and tangent moduli is compared in Fig. 3. As expected, the computing time of the FD method depends on the perturbation values since a different number of iterations results therefrom. When considering the best performing FD scheme with the perturbation values $h_S = 10^{-4}$ and $h_S = 10^{-6}$, the lowest computing time is obtained for coarse meshes; the HDSD however, is only slightly slower. This behavior changes for an increasing number of elements and at a reasonable discretization the HDSD becomes even faster than the FD method with optimal values. This is due to the fact that with an increasing number of elements the amount of time required for the solution of the system of equations becomes dominant compared to the calculation of the stiffness matrix, which is influenced by the numerical approximation of stress and moduli. Then, for finer meshes the number of iterations and thus the advantage of the HDSD and CSDA method becomes more important. It is emphasized that the optimal choice of perturbation values is of course not known in general and thus, the FD method

Table 1 Euclidean norm of the residual of the Cook-type problem based on the FD and HDSD scheme; red colors indicate that no quadratic convergence could be obtained

Analytical	FD with $h_S = 10^{-4}$			HDSD
	$h_C = 10^{-4}$	$h_C = 10^{-6}$	$h_C = 10^{-8}$	
5.4504×10^{-1}	$6.3167 \times 10^{+1}$	$6.3167 \times 10^{+1}$	$6.3167 \times 10^{+1}$	5.4504×10^{-1}
$1.8568 \times 10^{+2}$	$9.3803 \times 10^{+1}$	$9.0099 \times 10^{+1}$	$9.0116 \times 10^{+1}$	$1.8568 \times 10^{+2}$
$1.5341 \times 10^{+0}$	$2.0612 \times 10^{+0}$	$2.4092 \times 10^{+0}$	$2.4096 \times 10^{+0}$	$1.5341 \times 10^{+0}$
$3.4458 \times 10^{+0}$	$3.2625 \times 10^{+0}$	$3.2628 \times 10^{+0}$	$3.2630 \times 10^{+0}$	$3.4458 \times 10^{+0}$
8.4932×10^{-3}	6.5687×10^{-3}	1.4598×10^{-2}	1.4692×10^{-2}	8.4932×10^{-3}
2.0186×10^{-4}	4.2600×10^{-4}	5.3153×10^{-4}	5.7587×10^{-4}	2.0187×10^{-4}
5.0665×10^{-8}	2.3998×10^{-5}	2.0946×10^{-6}	5.5211×10^{-6}	5.0161×10^{-8}
–	2.3392×10^{-6}	3.3563×10^{-7}	3.3765×10^{-7}	–
–	4.2348×10^{-7}	–	–	–
–	3.3564×10^{-7}	–	–	–

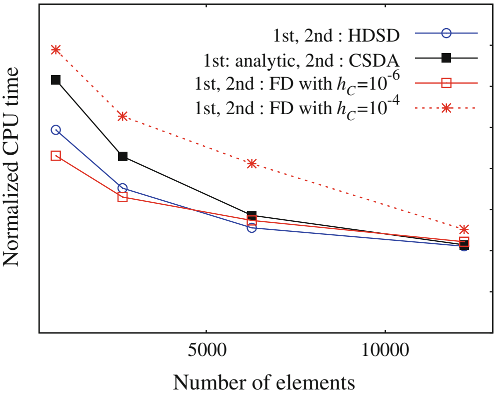


Fig. 3 Comparison of CPU time for Cook-type problem, for the FD scheme the perturbation value $h_S = 10^{-4}$ is considered, for the HDSD scheme we use $h_S = h_C = 1$ and for the CSDA scheme we consider $h = 10^{-30}$. The CPU time is normalized with respect to the time required for the calculation resulting from the implementation of analytic stresses and moduli

can typically be expected to be slower since not the optimal perturbation values will be used. This is shown by the computing time resulting from the values $h_S = 10^{-4}$, $h_C = 10^{-4}$ in Fig. 3 leading to a computing time being significantly higher as for the HDSD and CSDA. Moreover, the HDSD method is still faster than the CSDA scheme, although the first and second derivatives are calculated numerically in the case of HDSD, while the CSDA scheme requires the analytic expressions for the stresses.

4 HDSD Scheme for Incremental Variational Formulations

In this section a framework for the fully automatic calculation of internal variables, stresses and consistent tangent moduli for dissipative materials is proposed. It mainly consists of applying the HDSD scheme to incremental variational formulations (IVFs). Such formulations, see e.g. [18, 22], recast inelasticity theory as an equivalent optimization problem where the incremental stress potential within a discrete time interval is minimized in order to obtain the values of internal variables. The IVFs provide a general framework for a broad range of standard dissipative constitutive models in incremental stress potential regimes. Throughout this section, isothermal conditions are considered.

4.1 Incremental Variational Formulation

Ortiz and Stainier [22] proposed an effective incremental stress potential W^{eff} within a discrete time interval $\Delta t := t_{n+1} - t_n$ as

$$W^{\text{eff}} := \inf_{\boldsymbol{q}_{n+1}} W(\boldsymbol{F}_{n+1}, \boldsymbol{q}_{n+1}). \tag{24}$$

The deformation gradient tensor and the generalized vector of a collection of internal variables at time t_{n+1} are denoted by $\boldsymbol{F}_{n+1}$ and $\boldsymbol{q}_{n+1}$, respectively. The incremental stress potential W comprises a Helmholtz free energy function ψ and a dissipation potential ϕ by

$$W(\boldsymbol{F}_{n+1}, \boldsymbol{q}_{n+1}) = \psi(\boldsymbol{F}_{n+1}, \boldsymbol{q}_{n+1}) - \psi(\boldsymbol{F}_n, \boldsymbol{q}_n) + \Delta t \phi(\boldsymbol{q}_{n+1}, \frac{\Delta \boldsymbol{q}}{\Delta t}); \tag{25}$$

$\Delta \boldsymbol{q} := \boldsymbol{q}_{n+1} - \boldsymbol{q}_n$. In many cases the minimization problem (24) is solved by using a Newton-Raphson iteration and $\boldsymbol{q}_{n+1}$ is determined using the first and second derivatives $\partial_{\boldsymbol{q}} W$, $\partial^2_{\boldsymbol{qq}} W$ by updating

$$\boldsymbol{q}^{(k+1)}_{n+1} = \boldsymbol{q}^{(k)}_{n+1} - \left(\partial^2_{\boldsymbol{qq}} W^{(k)}\right)^{-1} \cdot \partial_{\boldsymbol{q}} W^{(k)}, \quad k = 0, 1, 2, \ldots, \tag{26}$$

until it converges; (k) denotes the iteration number. The stress response $\boldsymbol{P}_{n+1}$ at time t_{n+1} is given via the effective incremental potential function W^{eff} as

$$\boldsymbol{P}_{n+1} = \partial_{\boldsymbol{F}} W^{\text{eff}}(\boldsymbol{F}_{n+1}, \boldsymbol{q}_{n+1}(\boldsymbol{F}_{n+1})), \tag{27}$$

and the consistent tangent modulus $\mathbb{A}_{n+1}$ at time t_{n+1} is given by

$$\mathbb{A}_{n+1} = \partial^2_{\boldsymbol{FF}} W^{\text{eff}}(\boldsymbol{F}_{n+1}, \boldsymbol{q}_{n+1}(\boldsymbol{F}_{n+1})), \tag{28}$$

where $\boldsymbol{P}$ is the first Piola-Kirchhoff stress tensor and $\mathbb{A}$ is the corresponding nominal tangent modulus tensor. Note that $\boldsymbol{q}_{n+1}$ is dependent on the current deformation gradient $\boldsymbol{F}_{n+1}$.

4.2 Implementation Using HDSD Scheme

At first, find a solution $\boldsymbol{q}_{n+1}$ to the minimization problem (24) by using the HDNs ε_1 and ε_2. Here, an asterisk $(*)$ is used as a superscript denoting the perturbed values by means of the HDNs. The internal variables at k-th iteration $\boldsymbol{q}^{(k)}_{n+1}$ are perturbed by using the HDN units ε_1 and ε_2 as

$$\overset{*}{\boldsymbol{q}}{}^{(k)}_{n+1} = \boldsymbol{q}^{(k)}_{n+1} + \varepsilon_1 \boldsymbol{i}_{\text{r}} + \varepsilon_2 \boldsymbol{i}_{\text{s}}, \tag{29}$$

wherein $\boldsymbol{i}_{\text{r}}$ denotes a unit vector of a set of internal variables $\boldsymbol{q}_{n+1}$ such that the j-th component of $\boldsymbol{i}_{\text{r}}$ is defined as $i_{\text{r}j} = 1$ if $j = \text{r}$ and 0 else. Then, the r-th components of $\partial_{\boldsymbol{q}} W^{(k)}$ and accordingly of $\partial^2_{\boldsymbol{qq}} W^{(k)}$ are obtained by taking the coefficients with respect to ε_1 and $\varepsilon_1 \varepsilon_2$ such as

$$\left(\partial_{\boldsymbol{q}} W^{(k)}\right)_r = \Im_{\varepsilon_1}\left[\overset{*}{W}{}^{(k)}\right], \quad \left(\partial^2_{\boldsymbol{qq}} W^{(k)}\right)_{rs} = \Im_{\varepsilon_1 \varepsilon_2}\left[\overset{*}{W}{}^{(k)}\right], \tag{30}$$

and $\boldsymbol{q}_{n+1}$ is updated by Eq. (26). As a second step, the stresses $\boldsymbol{P}$ and tangent moduli $\mathbb{A}$ are obtained by differentiating the minimized effective stress potential W^{eff} with respect to the deformation gradient $\boldsymbol{F}_{n+1}$ as

$$\boldsymbol{P}_{n+1} = \partial_{\boldsymbol{F}} W^{\text{eff}} + \partial_{\boldsymbol{q}} W^{\text{eff}} \cdot \partial_{\boldsymbol{F}} \boldsymbol{q}_{n+1}, \tag{31}$$

$$\begin{aligned}\mathbb{A}_{n+1} &= \partial^2_{\boldsymbol{F}\boldsymbol{F}} W^{\text{eff}} + \partial^2_{\boldsymbol{F}\boldsymbol{q}} W^{\text{eff}} \cdot \partial_{\boldsymbol{F}} \boldsymbol{q}_{n+1} \\ &\quad + \partial_{\boldsymbol{F}} \boldsymbol{q}_{n+1} \cdot \left(\partial^2_{\boldsymbol{q}\boldsymbol{F}} W^{\text{eff}} + \partial^2_{\boldsymbol{q}\boldsymbol{q}} W^{\text{eff}} \cdot \partial_{\boldsymbol{F}} \boldsymbol{q}_{n+1}\right) + \partial_{\boldsymbol{q}} W^{\text{eff}} \cdot \partial^2_{\boldsymbol{F}\boldsymbol{F}} \boldsymbol{q}_{n+1}.\end{aligned} \tag{32}$$

Note that Eqs. (31) and (32) are indeterminate since $\partial_{\boldsymbol{F}} \boldsymbol{q}_{n+1}$ and $\partial^2_{\boldsymbol{F}\boldsymbol{F}} \boldsymbol{q}_{n+1}$ cannot be computed using only 2nd-order HDNs. In order to avoid the calculations of $\partial_{\boldsymbol{F}} \boldsymbol{q}_{n+1}$ and $\partial^2_{\boldsymbol{F}\boldsymbol{F}} \boldsymbol{q}_{n+1}$, we assume the strict stationary condition

$$\partial_{\boldsymbol{q}} W^{\text{eff}} = \boldsymbol{0}, \tag{33}$$

and also

$$D_{\boldsymbol{F}}(\partial_{\boldsymbol{q}} W^{\text{eff}}) = \boldsymbol{0} \quad \text{or equivalently} \quad \partial^2_{\boldsymbol{q}\boldsymbol{F}} W^{\text{eff}} + \partial^2_{\boldsymbol{q}\boldsymbol{q}} W^{\text{eff}} \cdot \partial_{\boldsymbol{F}} \boldsymbol{q}_{n+1} = \boldsymbol{0}. \tag{34}$$

Inserting (33) and (34) into (31) and (32), the stresses and tangent moduli in this formulation are obtained by

$$\boldsymbol{P}_{n+1} = \partial_{\boldsymbol{F}} W^{\text{eff}}, \tag{35}$$

$$\mathbb{A}_{n+1} = \partial^2_{\boldsymbol{F}\boldsymbol{F}} W^{\text{eff}} - \partial^2_{\boldsymbol{F}\boldsymbol{q}} W^{\text{eff}} \cdot \left(\partial^2_{\boldsymbol{q}\boldsymbol{q}} W^{\text{eff}}\right)^{-1} \cdot \partial^2_{\boldsymbol{q}\boldsymbol{F}} W^{\text{eff}}. \tag{36}$$

The derivatives $\partial_{\boldsymbol{F}} W^{\text{eff}}$, $\partial^2_{\boldsymbol{F}\boldsymbol{F}} W^{\text{eff}}$, $\partial^2_{\boldsymbol{F}\boldsymbol{q}} W^{\text{eff}}$ and $\partial^2_{\boldsymbol{q}\boldsymbol{F}} W^{\text{eff}}$ can be computed using only ε_1 and ε_2 in an analogous way as for the derivatives with respect to the internal variables. Note that the main advantage of this proposed framework is that the users are only required to implement the scalar energy functions ψ and ϕ.

4.3 Numerical Example: Elastoplastic Microstructure

In this numerical example, we analyze a finite strain elastoplastic model using a von Mises yield function including exponential isotropic hardening, cf. [4, 20], see also [27, 29]. The deformation gradient is multiplicatively decomposed into elastic and isochoric plastic parts as $\boldsymbol{F} = \boldsymbol{F}^{\text{e}} \cdot \boldsymbol{F}^{\text{p}}$ with $\det \boldsymbol{F}^{\text{p}} = 1$ and the Helmholtz free energy is additively decomposed into elastic and plastic parts as $\psi = \psi^{\text{e}} + \psi^{\text{p}}$. In this model, the elastic response is captured by

$$\psi^{\text{e}} = \frac{\lambda}{2}\left[b^{\text{e}}_1 + b^{\text{e}}_2 + b^{\text{e}}_3\right]^2 + \mu\left[(b^{\text{e}}_1)^2 + (b^{\text{e}}_2)^2 + (b^{\text{e}}_3)^2\right]. \tag{37}$$

λ and μ are material parameters and b^{e}_A $(A = 1, 2, 3)$ is the logarithm of each eigenvalue λ^{e}_A of the elastic left Cauchy-Green deformation tensor $\boldsymbol{b}^{\mathrm{e}} = \boldsymbol{F}^{\mathrm{e}}\boldsymbol{F}^{\mathrm{e}T}$ as $b^{\mathrm{e}}_A = \log(\lambda^{\mathrm{e}}_A)$. The plastic response ψ^{p} is captured by the exponential hardening

$$\psi^{\mathrm{p}} = y_\infty\alpha - \frac{1}{\eta}(y_0 - y_\infty)\exp(-\eta\alpha) + \frac{1}{2}h\alpha^2, \tag{38}$$

with α being a strain-like isotropic hardening variable and the material parameters y_∞, y_0, η and h representing the initial yield strength, the plastic yield strength at the transition from exponential to linear hardening, the degree of exponential hardening, and the slope of superimposed linear hardening, respectively. In our formulation, we identify the internal variable $\boldsymbol{q}$ as

$$\boldsymbol{q} = \left[\boldsymbol{F}^{\mathrm{p}},\, \alpha\right]^T, \tag{39}$$

wherein the square brackets denote an appropriate arrangement as column matrix. The generalized internal force vector $\boldsymbol{y}$ dual to $\boldsymbol{q}$ is determined as

$$\boldsymbol{y} := -\frac{\partial\psi}{\partial\boldsymbol{q}} = \left[-\frac{\partial\psi^{\mathrm{e}}}{\partial\boldsymbol{F}^{\mathrm{p}}},\, -\frac{\partial\psi^{\mathrm{p}}}{\partial\alpha}\right]^T = \left[\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{p}-T},\, \beta\right]^T, \tag{40}$$

with $\boldsymbol{\Sigma}$ denoting the Mandel stress tensor and β denoting the thermodynamic conjugate force to α. The dissipation potential ϕ is obtained by the maximum-dissipation principle as

$$\phi = \sup_{(\boldsymbol{\Sigma}_{n+1},\beta_{n+1})\in\mathbb{E}} \left[\boldsymbol{\Sigma}_{n+1} : \Delta\boldsymbol{L}^{\mathrm{p}} + \beta_{n+1}\cdot\Delta\alpha\right], \tag{41}$$

Herein, $\Delta\boldsymbol{L}^{\mathrm{p}}$ is a constant plastic velocity gradient in the current time increment and $\Delta\alpha := \alpha_{n+1} - \alpha_n$ is a constant increment of hardening variable. The elastic domain $\mathbb{E}$ is defined by a yield function f as

$$\mathbb{E} := \left\{(\boldsymbol{\Sigma}, \beta)\,\middle|\, f(\boldsymbol{\Sigma}, \beta) = \|\mathrm{dev}\boldsymbol{\Sigma}\| - \sqrt{2/3}\beta \le 0\right\}; \tag{42}$$

$\mathrm{dev}(\bullet)$ is the deviatoric operator. Through the use of Karush-Kuhn-Tucker conditions, the inf-sup problem (24) and (41) can be recast by one parameter minimization problem using the Lagrange multiplier $\Delta\gamma$ [21] as

$$W^{\mathrm{eff}} = \inf_{\Delta\gamma\ge 0}\left[\psi(\boldsymbol{F}_{n+1}, \Delta\gamma) - \psi(\boldsymbol{F}_n) + \Delta\gamma f\right] \quad \text{with} \quad \Delta\gamma f = 0, \quad f \le 0, \tag{43}$$

with the updated variables as

$$\boldsymbol{F}^{\mathrm{p}}_{n+1} = \exp\left(\Delta\gamma\frac{\partial f}{\partial\boldsymbol{\Sigma}}\right)\cdot\boldsymbol{F}^{\mathrm{p}}_n \quad \text{and} \quad \alpha_{n+1} = \alpha_n + \Delta\gamma\frac{\partial f}{\partial\beta}. \tag{44}$$

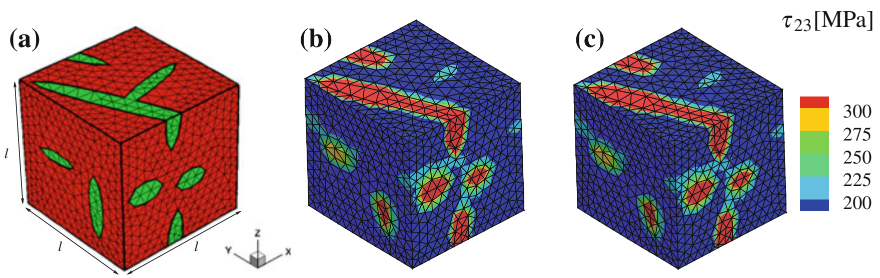


Fig. 4 **a** Initial configuration of SSRVE (the total number of elements is 10,889 and the number of degrees of freedom is 43,578), **b** distribution of Kirchhoff stresses τ_{yz} in the deformed configuration resulting from the classical standard return mapping method and **c** IVF with HDSD scheme

Table 2 Material parameters for individual phases of DP steel microstructure

	λ (MPa)	μ (MPa)	y_0 (MPa)	y_∞ (MPa)	η (–)	h (MPa)
Matrix (Ferrite)	118846.2	79230.77	260.0	580.0	9.0	70.0
Inclusion (Martensite)	118846.2	79230.77	1000.0	2750.0	35.0	10.0

The performance of the proposed implementation scheme for finite strain plasticity is investigated by the finite element simulation of a statistically similar representative volume element (SSRVE) of a dual-phase (DP) steel microstructure. The SSRVE shown in Fig. 4a is obtained by constructing a simplified microstructure in terms of statistical measures as similar as possible to the real random microstructure, cf. [2]. To somehow represent a typical DP steel microstructure consisting of a ferritic matrix phase with an embedded martensitic inclusion phase the material parameters in Table 2 are used. A macroscopic deformation gradient $\bar{\boldsymbol{F}}$ including the shear $\bar{F}_{yz}$ is applied by prescribing the homogeneous deformation field $\boldsymbol{x} = \bar{\boldsymbol{F}}\boldsymbol{X} + \tilde{\boldsymbol{w}}$ to the SSRVE and periodic deformation fluctuations $\tilde{\boldsymbol{w}}$ and anti-periodic tractions

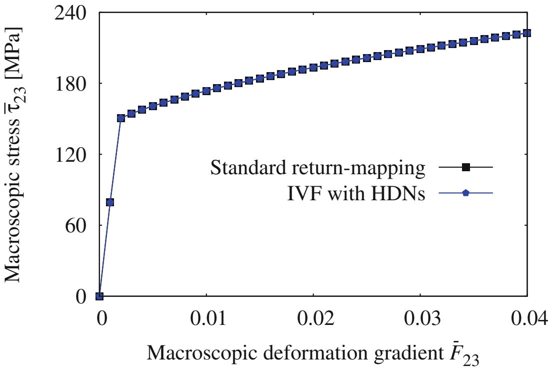


Fig. 5 Macroscopic stress-strain diagram for simple shear deformation of statistically similar RVE of DP steel microstructure. The diagram compares the results of the HDSD scheme with the results of the classical standard return mapping scheme

at the boundary of the SSRVE. The resulting macroscopic stress-strain response is calculated as shown in Fig. 5. Figure 4b, c show deformed configurations of the SSRVE with stress distributions τ_{yz} as a result of two different schemes, i.e., the proposed HDSD based implementation scheme and the classical standard return mapping scheme following the implementation given in [12]. According to those results, a good agreement is observed at both, the microscopic as well as the macroscopic scale. However, it is worth pointing out that the proposed scheme has the important advantage of a straightforward implementation of any other complicated constitutive model in the context of incremental variational formulations since users are only required to modify scalar-valued quantities such as the functions ψ, ϕ and the yield function f.

5 Robust Approach to Compute Tangent Stiffness Matrix in Thermo-Mechanical Problems Based on CSDA

In this section a CSDA-based robust approximation scheme for the numerical calculation of tangent stiffness matrices is presented in the context of nonlinear thermo-mechanical finite element problems and its performance is analyzed.

5.1 Formulations

The thermo-mechanical framework at large strains relies on the governing equations, namely the balance of linear momentum, and the balance of energy, i.e.

$$-\operatorname{Div}\boldsymbol{F}\boldsymbol{S} - \boldsymbol{f} = 0, \tag{45}$$

$$\boldsymbol{S}\cdot\frac{1}{2}\dot{\boldsymbol{C}} + \rho_0 r - \operatorname{Div}\boldsymbol{q}_0 - \rho_0(\dot{\psi} + \dot{\overline{\theta\eta}}) = 0. \tag{46}$$

Herein, the Legendre transform $\psi = U - \theta\eta$ has been performed, where ψ, U, θ and η denote the Helmholtz free energy, the specific internal energy, temperature and the specific entropy. The internal dissipation is considered to consist of two additive parts, i.e. $\mathscr{D}_{\text{int}} = \mathscr{D}_{\text{mech}} + \mathscr{D}_{\text{therm}}$, with the thermal part $\mathscr{D}_{\text{therm}} = \rho_0\theta\dot{\eta}_{\text{p}}$ and a mechanical part $\mathscr{D}_{\text{mech}}$. $\boldsymbol{q}_0$ is the heat flux through the body in the reference configuration, which is related to the Cauchy heat flux $\boldsymbol{q} = -k_\theta \operatorname{grad}\theta$ by $\boldsymbol{q}_0 = J\boldsymbol{F}^{-1}\boldsymbol{q}$. Herein, k_θ is the heat conduction coefficient and J is the determinant of $\boldsymbol{F}$. Note that grad$(\bullet)$ denotes the gradient with respect to coordinates in the reference configuration, respectively, and Div$(\bullet)$ denotes the divergence with respect to coordinates in the reference configuration. Also, $\boldsymbol{f}$, r and ρ_0 are the body force vector, internal heat source and the reference density of the body, respectively. For the solution of boundary value problems the standard Galerkin method is typically applied which requires the weak form of the balance equations. In this context see e.g. [33] or [29] for further details. Multiplying the balance Eqs. (45), (46) with test functions and integrating over the physical domain leads to the weak forms $G_u := G_u^{\text{int}} - G_u^{\text{ext}} = 0$ and $G_\theta = G_\theta^{\text{int}} - G_\theta^{\text{ext}} = 0$, wherein the internal and external parts of the weak forms are given by

$$G_u^{\text{int}} := \int_{\mathscr{B}_0^e} \boldsymbol{S} \cdot \frac{1}{2} \delta \boldsymbol{C} \, \mathrm{d}V, \tag{47}$$

$$\begin{aligned} G_\theta^{\text{int}} := \int_{\mathscr{B}_0^e} (& \boldsymbol{q}_0 \cdot \mathrm{Grad} \delta\theta + \rho_0 \, \theta \, \partial^2_{\theta\theta} \psi \, \dot{\theta} \, \delta\theta + \rho_0 \, \theta \, \partial^2_{\theta\alpha} \psi \, \dot{\alpha} \, \delta\theta \\ & + \rho_0 \, \theta \, \partial^2_{\theta \boldsymbol{b}^{\mathrm{e}}} \psi \cdot \dot{\boldsymbol{b}}^{\mathrm{e}} \, \delta\theta + \mathscr{D}_{\text{mech}} \delta\theta \,) \, \mathrm{d}V, \end{aligned} \tag{48}$$

$$G_u^{\text{ext}} := \int_{\partial \mathscr{B}_0^e} \boldsymbol{t}_0 \cdot \delta \boldsymbol{u} \, \mathrm{d}A + \int_{\mathscr{B}_0^e} \boldsymbol{f} \cdot \delta \boldsymbol{u} \, \mathrm{d}V, \tag{49}$$

$$G_\theta^{\text{ext}} := \int_{\partial \mathscr{B}_0^e} \mathscr{Q} \cdot \boldsymbol{N} \delta\theta \, \mathrm{d}A. \tag{50}$$

Herein, $\mathscr{B}_0^e$ is the domain of the reference configuration, $\delta \boldsymbol{u}$ and $\delta\theta$ denote the test functions associated with the displacement and temperature fields, respectively. Also note that $\boldsymbol{t}_0$ and $\mathscr{Q}$ are the surface traction and the heat flux vectors, respectively, which are assumed here to be independent of the displacements or temperatures. Exploiting the principle of maximum dissipation and applying the Karush-Kuhn-Tucker optimality conditions, c.f. [28], the explicit form for the mechanical dissipation for isotropic hardening is $\mathscr{D}_{mech} = \lambda \sqrt{\frac{2}{3}} \, y(\theta)$, with the consistency parameter λ and the initial yield stress $y(\theta)$. Here, for the finite element implementation the discretized weak forms of the balance equations using matrix notation read

$$G_{\text{u}}^{\text{int}} \approx G_{\text{u}}^{\text{int},h} = \sum_{e=1}^{n_{\text{ele}}} (\delta \boldsymbol{d}_{\text{u}}^e)^{\text{T}} \boldsymbol{r}_{\text{u}}^{int,e} \quad \text{and} \quad G_\theta^{int} \approx G_\theta^{\text{int},h} = \sum_{e=1}^{n_{\text{ele}}} (\delta \boldsymbol{d}_\theta^e)^{\text{T}} \boldsymbol{r}_\theta^{\text{int},e}, \tag{51}$$

with n_{ele} denoting the number of finite elements. The external parts of the discretized weak forms $G_{\text{u}}^{\text{ext},h}$ and $G_\theta^{\text{ext},h}$ are reformulated accordingly. The element vector of mechanical degrees of freedom is denoted by $\boldsymbol{d}_{\text{u}}^e$ which can be represented for a three-dimensional setting by $\boldsymbol{d}_{\text{u}}^e = [(\boldsymbol{d}_{\text{u}}^{I=1})^{\text{T}} \; (\boldsymbol{d}_{\text{u}}^{I=2})^{\text{T}} \; \ldots \; (\boldsymbol{d}_{\text{u}}^{nen})^{\text{T}}]^{\text{T}}$ with nen being the number of nodes per element and $\boldsymbol{d}_u^I = [d_{\text{u}x} \, d_{\text{u}y} \, d_{\text{u}z}]^{\text{T}}$ denoting the displacements at a particular node I.

In a three-dimensional physical space the number of degrees of freedom associated with the displacements is $ndf_{\text{u}} = 3$, i.e. d_{u_x}, d_{u_y}, and d_{u_z}. Then the total number of mechanical degrees of freedom for one element is $tdof_{\text{u}} = nen \times ndf_{\text{u}}$. Analogously, the number of thermal degrees of freedom is $ndf_\theta = 1$, i.e. d_θ, such that the total number of thermal degrees of freedom per element is $tdof_\theta = nen \times ndf_\theta$.

After inserting standard approximations for the nodal displacements and temperatures and shifting to the Voigt notation to represent tensorial quantities of second order as vectors, the discretized internal element residual vectors are

$$\boldsymbol{r}_{\mathrm{u}}^{e,\mathrm{int}} = \sum_{I=1}^{nen} \int_{\mathscr{B}_0^e} (\boldsymbol{B}_{\mathrm{u}}^I)^{\mathrm{T}} \boldsymbol{S} \, \mathrm{d}V, \tag{52}$$

$$\begin{aligned}\boldsymbol{r}_{\theta}^{e,\mathrm{int}} = &\sum_{I=1}^{nen} \int_{\mathscr{B}_0^e} \big((\boldsymbol{B}_{\theta}^I)^{\mathrm{T}} \boldsymbol{q}_0 + N^I \rho_0 \theta \partial^2_{\theta\theta} \psi \dot{\theta} + N^I \rho_0 \theta \partial^2_{\theta\theta} \psi \, \dot{\alpha} \\ &+ N^I \rho_0 \theta \partial^2_{\theta \boldsymbol{b}^{\mathrm{e}}} \psi \cdot \dot{\boldsymbol{b}}^{\mathrm{e}} + N^I \lambda \sqrt{2/3} y(\theta)) \, \mathrm{d}V.\end{aligned} \tag{53}$$

Herein, $\mathscr{B}_0^e$ denotes the domain of the initial configuration of finite element e. The external residual vectors $\boldsymbol{r}_{\mathrm{u}}^{e,\mathrm{ext}}$ and $\boldsymbol{r}_{\theta}^{e,\mathrm{ext}}$ are obtained analogously. Here, the same nodal shape functions N^I are used for the displacements and the temperature; $\boldsymbol{B}_{\mathrm{u}}^I$ are the standard mechanical B-matrices associated with node I, c.f. [33], which consist of the gradients of the nodal shape functions. The thermal B-matrix is given by $\boldsymbol{B}_{\theta}^I = [N_{,1}^I \; N_{,2}^I \; N_{,3}^I]^T$, where $N_{,i}^I$ represents the derivative of the shape function with respect to physical coordinate X_i. Due to the material nonlinearities these equations have to be solved numerically. For that purpose their linearizations $\mathrm{Lin}G_{\mathrm{u}}^{\mathrm{h}} = G_{\mathrm{u}}^{\mathrm{h}} + \Delta G_{\mathrm{u}}^{\mathrm{h}}$ and $\mathrm{Lin}G_{\theta}^{\mathrm{h}} = G_{\theta}^{\mathrm{h}} + \Delta G_{\theta}^{\mathrm{h}}$ are required. Here $\Delta G_{\mathrm{u}}^{\mathrm{h}}$ and $\Delta G_{\theta}^{\mathrm{h}}$ are the linear increments of the weak forms obtained as

$$\Delta G_{\mathrm{u}}^{\mathrm{int,h}} = \sum_{e=1}^{n_{\mathrm{ele}}} (\delta \boldsymbol{d}_{\mathrm{u}}^e)^{\mathrm{T}} \left(\boldsymbol{k}_{\mathrm{uu}}^e \, \Delta \boldsymbol{d}_{\mathrm{u}}^e + \boldsymbol{k}_{\mathrm{u}}^e \theta \, \Delta \boldsymbol{d}_{\theta}^e \right), \tag{54}$$

$$\Delta G_{\theta}^{\mathrm{int,h}} = \sum_{e=1}^{n_{\mathrm{ele}}} (\delta \boldsymbol{d}_{\theta}^e)^{\mathrm{T}} \left(\boldsymbol{k}_{\theta\mathrm{u}}^e \, \Delta \boldsymbol{d}_{\mathrm{u}}^e + \boldsymbol{k}_{\theta\theta}^e \, \Delta \boldsymbol{d}_{\theta}^e \right), \tag{55}$$

with the individual derivatives of the residual vectors given by

$$\boldsymbol{k}_{\mathrm{uu}}^e = \frac{\partial \boldsymbol{r}_{\mathrm{u}}^{e,\mathrm{int}}}{\partial \boldsymbol{d}_{\mathrm{u}}^e}, \quad \boldsymbol{k}_{\mathrm{u}\theta}^e = \frac{\partial \boldsymbol{r}_{\mathrm{u}}^{e,\mathrm{int}}}{\partial \boldsymbol{d}_{\theta}^e}, \quad \boldsymbol{k}_{\theta\mathrm{u}}^e = \frac{\partial \boldsymbol{r}_{\theta}^{e,\mathrm{int}}}{\partial \boldsymbol{d}_{\mathrm{u}}^e}, \quad \boldsymbol{k}_{\theta\theta}^e = \frac{\partial \boldsymbol{r}_{\theta}^{e,\mathrm{int}}}{\partial \boldsymbol{d}_{\theta}^e}. \tag{56}$$

Instead of implementing the analytic expressions for these derivatives the CSDA scheme can be applied. Remember that the mechanical and thermal residual vectors depend on all (mechanical and thermal) degrees of freedom, i.e. $\boldsymbol{r}_{\mathrm{u}}^e := \boldsymbol{r}_{\mathrm{u}}^e(\boldsymbol{d}_{\mathrm{u}}^e, \boldsymbol{d}_{\theta}^e)$ and $\boldsymbol{r}_{\theta}^e := \boldsymbol{r}_{\theta}^e(\boldsymbol{d}_{\mathrm{u}}^e, \boldsymbol{d}_{\theta}^e)$. Then the approximations of the k-th column vectors $\tilde{\boldsymbol{k}}_{\mathrm{uu}(k)}^e$ and $\tilde{\boldsymbol{k}}_{\theta\mathrm{u}(k)}^e$ in $\boldsymbol{k}_{\mathrm{uu}}^e$ and $\boldsymbol{k}_{\theta\mathrm{u}}^e$, respectively, and of the j-th column vectors $\tilde{\boldsymbol{k}}_{\mathrm{u}\theta(j)}^e$ and $\tilde{\boldsymbol{k}}_{\theta\theta(j)}^e$ in $\boldsymbol{k}_{\mathrm{u}\theta}^e$ and $\boldsymbol{k}_{\theta\theta}^e$, respectively, can alternatively be approximated by

$$\tilde{\boldsymbol{k}}_{\mathrm{uu}(k)}^e := \frac{\partial \boldsymbol{r}_{\mathrm{u}}^{e,\mathrm{int}}}{\partial \left\{ d_{\mathrm{u}}^e \right\}_k} \approx \frac{\Im \left[\boldsymbol{r}_{\mathrm{u}}^e(\boldsymbol{d}_{\mathrm{u}}^e + ih\tilde{\boldsymbol{d}}_{\mathrm{u}(k)}^e, \boldsymbol{d}_{\theta}^e) \right]}{h}, \tag{57}$$

$$\tilde{\boldsymbol{k}}_{\mathrm{u}\theta(j)}^e := \frac{\partial \boldsymbol{r}_{\mathrm{u}}^{e,\mathrm{int}}}{\partial \left\{ d_{\theta}^e \right\}_j} \approx \frac{\Im \left[\boldsymbol{r}_{\mathrm{u}}^e(\boldsymbol{d}_{\mathrm{u}}^e, \boldsymbol{d}_{\theta}^e + ih\tilde{\boldsymbol{d}}_{\theta(j)}^e) \right]}{h}, \tag{58}$$

$$\tilde{\boldsymbol{k}}^e_{\theta\mathrm{u}(k)} := \frac{\partial \boldsymbol{r}^{e,\mathrm{int}}_\theta}{\partial \left\{d^e_\mathrm{u}\right\}_k} \approx \frac{\Im\left[\boldsymbol{r}^e_\theta(\boldsymbol{d}^e_\mathrm{u} + ih\tilde{\boldsymbol{d}}^e_{\mathrm{u}(k)}, \boldsymbol{d}^e_\theta)\right]}{h}, \tag{59}$$

$$\tilde{\boldsymbol{k}}^e_{\theta\theta(j)} := \frac{\partial \boldsymbol{r}^{e,\mathrm{int}}_\theta}{\partial \left\{d^e_\theta\right\}_j} \approx \frac{\Im\left[\boldsymbol{r}^e_\theta(\boldsymbol{d}^e_\mathrm{u}, \boldsymbol{d}^e_\theta + ih\tilde{\boldsymbol{d}}^e_{\theta(j)})\right]}{h}, \tag{60}$$

where the indices $k \in [1, tdof_\mathrm{u}]$ and $j \in [1, tdof_\theta]$ on the left hand side of the equations represent the column index. On the right hand side these indices correspond to the individual perturbation vectors whose components with indices $m \in [1, tdof_\mathrm{u}]$ and $q \in [1, tdof_\theta]$, respectively, are defined as

$$\{\tilde{d}^e_{\mathrm{u}(k)}\}_m = \delta_{(k)m} \quad \text{and} \quad \{\tilde{d}^e_{\theta(j)}\}_q = \delta_{(j)q}. \tag{61}$$

The Kronecker symbol is defined as $\delta_{ab} = 1$ for $a = b$ and $\delta_{ab} = 0$ otherwise.

5.2 Numerical Example: Thermo-Elastoplastic Microstructure

Now the performance of the proposed approximation scheme is investigated in a thermo-elastoplastic problem. A simplified RVE of DP steel is considered which consists of a spherical inclusion embedded in a cubic matrix, cf. Fig. 6a. The displacements at all surfaces are linked in normal direction such that the outer surfaces remain planar. The surfaces at $X = 0$, $Y = 0$ and $Z = 0$ are fixed in the respective normal direction and a compressive normal stress is applied to the top surface along the Z-direction. First, compression is increased up to 650 MPa and then it is

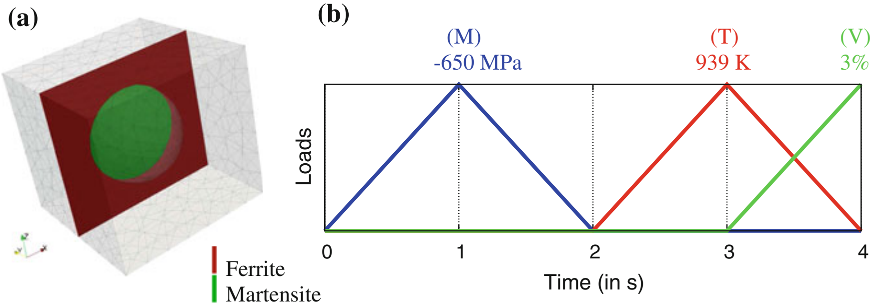


Fig. 6 **a** RVE, discretized with 3623 quadratic tetrahedral elements, and **b** loading protocol; M, T and V indicate the mechanical compressive load, the thermal change and the volumetric expansion (of the inclusion), respectively

Table 3 Material parameters of the phases

	E (MPa)	ν (–)	α_t (1/K)	c (mm^2/s^{2}K)	k (NK/s)	H (MPa)	ω (MPa/K)	y_0 (MPa)
Ferrite	206,000	0.3	1×10^{-5}	0.46×10^{-9}	49	5000	−0.295	260
Martensite	206,000	0.3	1×10^{-5}	0.46×10^{-9}	43	25000	−0.586	1000

unloaded again. Second, the temperature is prescribed over the entire microstructure, increasing first from 293 K up to 939 K and then decreasing back to 293 K, to reflect some heat treatment. Third, during the cooling procedure additionally a volumetric change is applied in the inclusion phase to characterize the mechanical fields resulting from a 3 % phase transformation volume change. This loading protocol is depicted in Fig. 6b. It is remarked that this procedure is only an idealization of the production process of DP steel and serves here only as a numerical example showing the applicability of the approximation scheme. For the calculation the strain energy function as proposed in [29] is used along with a linear hardening law. The material parameters chosen for the two phases are listed in Table 3, wherein the hardening modulus is denoted by H. For the initial yield stress the temperature dependency $y = \langle \omega(\theta - \theta_0) + y_0 - \tilde{y}_0 \rangle + \tilde{y}_0$ is taken into account, where y_0 is the initial yield stress at room temperature $\theta_0 = 293$ K. The parameters ω and y_0 for the two phases are also given in Table 3. The resulting accumulation of plastic strains in the microstructure is visualized in Fig. 7. At time $t = 1$ s, after the application of compression, we see a development of plastic strains in the ferrite, which is however, lower than at time $t = 4$ s. To demonstrate the performance of the CSDA scheme the convergence patterns are analyzed at particular times of interest during the entire simulation. Thus, the times $t = 1$ s where the total compressive load is active and $t = 4$ s at the end of the volume jump and cooling process are considered. The norms of the residual vectors versus the corresponding Newton iteration, obtained at each of these times, are plotted in Fig. 8 for the perturbation values of $h = 10^{-5}$, $h = 10^{-8}$,

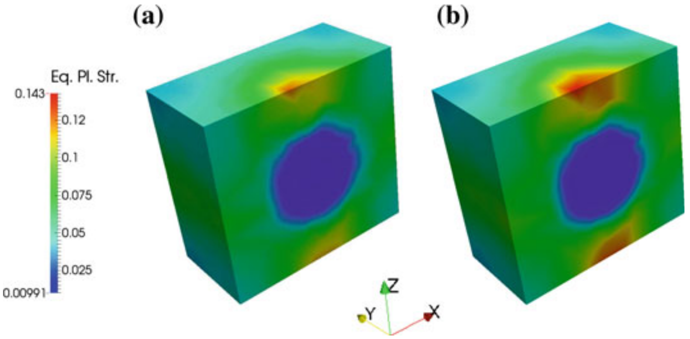


Fig. 7 Values of internal variable α in the microstructure at (**a**) $t = 1$ s, after the application of compressive loads, and at (**b**) $t = 4$ s, after the application of the volumetric expansion

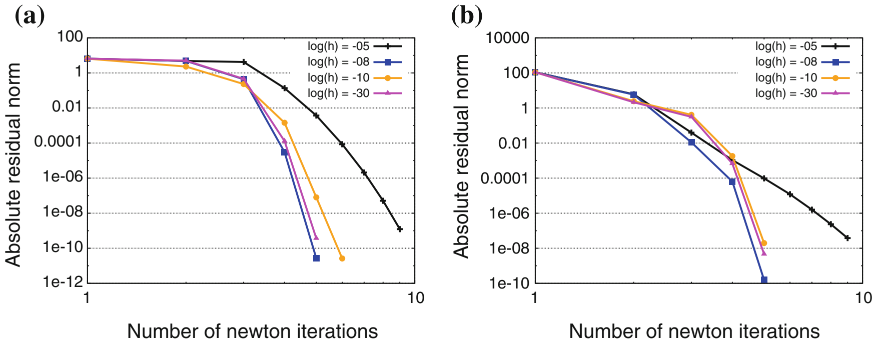


Fig. 8 Absolute values of the residual norms versus corresponding Newton iteration at times (**a**) $t = 1$ s, (**b**) $t = 4$ s. Both axes are depicted in logarithmic scale

$h = 10^{-10}$ and $h = 10^{-30}$. As can be seen, the perturbation $h = 10^{-5}$ leads to too inaccurate approximations and therefore gives unsatisfactory results, but smaller perturbation values lead to quadratically converging Newton iterations. Summarizing, a stable convergence of the CSDA approach for the tangent stiffness matrix is demonstrated in this example for a thermo-elastoplastic problem with a variety of loading conditions.

6 Conclusion

This contribution presented robust numerical schemes for an efficient implementation of tangent matrices in finite strain problems. The schemes were based on highly accurate numerical differentiation approaches which use non-real numbers, i.e., complex-step derivative approximation and hyper-dual-step derivatives. Their excellent performance was confirmed by analyzing different numerical problems, a hyperelastic material, an inelastic standard dissipative material in the context of incremental variational formulations and a thermo-mechanical calculation. The simplicity of the algorithms enabled an uncomplicated implementation such that the user is only required to program the scalar energy function for the approximation of stresses and moduli for hyperelastic and standard dissipative materials, or the residual vector for an approximation of the tangent stiffness matrix in finite element environments. This advantage of saving time during the development process while still attaining accurate approximations could be exploited for implementing complex element formulations or material models for which the derivation and programming of analytical derivatives would be difficult. In the context of incremental variational formulations it has additionally the advantage of enabling a fully automatic formulation which just requires the definition of strain energy and dissipation potential.

Acknowledgments Financial funding by the DFG Priority Program 1648 (SPPEXA "Software for Exascale Computing"), projects BA 2823/8-1 and SCHR 570/19-1 is greatly acknowledged by D. Balzani and J. Schröder. Furthermore, assistance for the thermo-mechanical calculations by Ashutosh Gandhi is appreciated.

References

1. Balzani, D., Neff, P., Schröder, J., & Holzapfel, G. A. (2006). A polyconvex framework for soft biological tissues. Adjustment to experimental data. *International Journal of Solids and Structures*, *43*(20), 6052–6070.
2. Balzani, D., Scheunemann, L., Brands, D., & Schröder, J. (2014). Construction of two- and three-dimensional statistically similar RVEs for coupled micro-macro simulations. *Computational Mechanics*, *54*, 1269–1284.
3. Balzani, D., Gandhi, A., Tanaka, M., & Schröder, J. (2015). Numerical calculation of thermomechanical problems at large strains based on robust approximations of tangent stiffness matrices. *Computational Mechanics*, *55*, 861–871.
4. Bleier, N., & Mosler, J. (2012). Efficient variational constitutive updates by means of a novel parameterization of the flow rule. *International Journal for Numerical Methods in Engineering*, *89*, 1120–1143.
5. Clifford, W. K. (1873). Preliminary sketch of biquaternions. *Proceedings of the London Mathematical Society*, *4*(64), 381–395.
6. Fike, J. A. (2013). Multi-objective optimization using hyper-dual numbers. Ph.D. thesis, Stanford university.
7. Fike, J. A., & Alonso, J. J. (2011). The development of hyper-dual numbers for exact second-derivative calculations. In *49th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*.
8. Golanski, D., Terada, K., & Kikuchi, N. (1997). Macro and micro scale modeling of thermal residual stresses in metal matrix composite surface layers by the homogenization method. *Computational Mechanics*, *19*, 188–201.
9. Kim, S., Ryu, J., & Cho, M. (2011). Numerically generated tangent stiffness matrices using the complex variable derivative method for nonlinear structural analysis. *Computer Methods in Applied Mechanics and Engineering*, *200*, 403–413.
10. Kiran, R., & Khandelwal, K. (2015). Automatic implementation of finite strain anisotropic hyperelastic models using hyper-dual numbers. *Computational Mechanics*, *55*, 229–248.
11. Kiran, R., & Khandelwal, K. (2014). Complex step derivative approximation for numerical evaluation of tangent moduli. *Computers and Structures*, *140*, 1–13.
12. Klinkel, S. (2000). Theorie und Numerik eines Volumen-Schalen-Elementes bei finiten elastischen und plastischen Verzerrungen. *Dissertation thesis*, Institut für Baustatik, Universität Karlsruhe.
13. Lai, K.-L., & Crassidis, J. L. (2008). Extensions of the first and second complex-step derivative approximations. *Journal of Computational and Applied Mathematics*, *219*, 276–293.
14. Lyness, J. N. (1968). Differentiation formulas for analytic functions. *Mathematics of Computation*, 352–362.
15. Martins, J. R. R. A., & Hwang, J. T. (2013). Review and unification of discrete methods for computing derivatives of single- and multi-disciplinary computational models. *AIAA Journal*, *51*(11), 2582–2599.
16. Martins, J. R. R. A., Sturdza, P., & Alonso, J. J. (2003). The complex-step derivative approximation. *ACM Transactions on Mathematical Software*, *29*, 245–262.
17. Miehe, C. (1996). Numerical computation of algorithmic (consistent) tangent moduli in large-strain computational inelasticity. *Computer Methods in Applied Mechanics and Engineering*, *134*, 223–240.

18. Miehe, C., & Lambrecht, M. (2003). Analysis of microstructure development in shearbands by energy relaxation of incremental stress potentials: Large-strain theory for standard dissipative solids. *International Journal for Numerical Methods in Engineering*, *58*, 1–41.
19. Miehe, C., Schotte, J., & Schröder, J. (1999). Computational micro-macro-transitions and overall moduli in the analysis of polycrystals at large strains. *Computational Materials Science*, *16*, 372–382.
20. Mosler, J., & Bruhns, O. T. (2009). Towards variational constitutive updates for non-associative plasticity models at finite strain: Models based on a volumetric-deviatoric split. *International Journal of Solids and Structures*, *46*, 1676–1684.
21. Mosler, J., & Bruhns, O. T. (2010). On the implementation of rate-independent standard dissipative solids at finite strain—variational constitutive updates. *Computer Methods in Applied Mechanics and Engineering*, *199*, 417–429.
22. Ortiz, M., & Stainier, L. (1999). The variational formulation of viscoplastic constitutive updates. *Computer Methods in Applied Mechanics and Engineering*, *171*, 419–444.
23. Pérez-Foguet, A., Rodríguez-Ferran, A., & Huerta, A. (2000). Numerical differentiation for local and global tangent operators in computational plasticity. *Computer Methods in Applied Mechanics and Engineering*, *189*, 277–296.
24. Pérez-Foguet, A., Rodríguez-Ferran, A., & Huerta, A. (2000). Numerical differentiation for non-trivial consistent tangent matrices: An application to the mrs-lade model. *International Journal for Numerical Methods in Engineering*, *48*, 159–184.
25. Schröder, J. 2013. A numerical two-scale homogenization scheme: the FE^2-method. In *Plasticity and beyond—microstructures, chrystal-plasticity and phase transitions (CISM Lecture Notes)*. Vienna: Springer.
26. Schröder, J., Neff, P., & Balzani, D. (2005). A variational approach for materially stable anisotropic hyperelasticity. *International Journal of Solids and Structures*, *42*(15), 4352–4371.
27. Simo, J. C. (1988). A framework for finite strain elastoplasticity based on maximum plastic dissipation and the multiplicative decomposition: Part I. Continuum formulation. *Computer Methods in Applied Mechanics and Engineering*, *66*, 199–219.
28. Simo, J., & Hughes, T. J. R. (1998). *Computational inelasticity*. Berlin: Springer.
29. Simo, J., & Miehe, C. (1992). Associative coupled thermoplasticity at finite strains: Formulation, numerical analysis and implementation. *Computer Methods in Applied Mechanics and Engineering*, *98*, 41–104.
30. Smit, R. J. M., Brekelmans, W. A. M., & Meijer, H. E. H. (1998). Prediction of the mechanical behavior of nonlinear heterogeneous systems by multi-level finite element modeling. *Computer Methods in Applied Mechanics and Engineering*, *155*, 181–192.
31. Tanaka, M., Fujikawa, M., Balzani, D., & Schröder, J. (2014). Robust numerical calculation of tangent moduli at finite strains based on complex-step derivative approximation and its application to localization analysis. *Computer Methods in Applied Mechanics and Engineering*, *269*, 454–470.
32. Tanaka, M., Sasagawa, T., Omote, R., Fujikawa, M., Balzani, D., & Schröder, J. (2015). A highly accurate 1st- and 2nd-order differentiation scheme for hyperelastic material models based on hyper-dual numbers. *Computer Methods in Applied Mechanics and Engineering*, *283*, 22–45.
33. Zienkiewicz, O. C., & Taylor, R. L. (1967). *The finite element method for solid and structural mechanics*. Oxford: Butterworth-Heinemann.

Folding Patterns in Partially Delaminated Thin Films

David Bourne, Sergio Conti and Stefan Müller

Abstract Michael Ortiz and Gustavo Gioia showed in the 90s that the complex patterns arising in compressed elastic films can be analyzed within the context of the calculus of variations. Their initial work focused on films partially debonded from the substrate, subject to isotropic compression arising from the difference in thermal expansion coefficients between film and substrate. In the following two decades different geometries have been studied, as for example anisotropic compression. We review recent mathematical progress in this area, focusing on the rich phase diagram of partially debonded films with a lateral boundary condition.

1 Introduction

Elastic films deposited on a substrate are often subject, after thermal expansion, to compressive strains which are released by debonding and buckling, generating a variety of microstructures. The work of Michael Ortiz and Gustavo Gioia in the 90s [1, 2] opened the way for the use of the tools of calculus of variations in the study of these structures. Their starting point was the Föppl-von Kármán plate theory, as given in (4) below. One of their insights was that the key nonconvexity which gives rise to the microstructure can be understood in terms of the out-of-plane displacement alone, leading after some rescalings to the Eikonal functional, as given in (1) below. This functional contains a term of the form $(|Dw|^2 - 1)^2$, where w is the normal

D. Bourne
Department of Mathematical Sciences, Durham University, Durham, UK
e-mail: david.bourne@durham.ac.uk

S. Conti (✉) · S. Müller
Institut Für Angewandte Mathematik, Universität Bonn, Bonn, Germany
e-mail: sergio.conti@uni-bonn.de

S. Müller
e-mail: stefan.mueller@hcm.uni-bonn.de

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_2

displacement, which favours deformations with the property that the gradient of w is approximately a unit vector, independently of the orientation. Since the film is still bound to the substrate at the boundary of the debonded region, the appropriate boundary condition is $w = 0$, which prescribes that the average over Ω of the gradient of w vanishes. Therefore the resulting low-energy deformations have gradient Dw oscillating between different values. As in many nonconvex variational problems, oscillations on very small scales may be energetically convenient, see [3, 4]. Correspondingly, the variational problem $\int_\Omega (|Dw|^2 - 1)^2 dx$ is not lower semicontinuous, and - depending on the boundary data and forcing - does not have minimizers. However, the curvature term $\sigma^2 |D^2 w|^2$ penalizes oscillations on an exceedingly fine scale and thereby ensures existence of minimizers. The solutions then have oscillations on an intermediate scale, which is determined by the competition between the two terms. The analysis of the specific functional proposed by Ortiz and Gioia is reviewed in Sect. 2 below.

The approach of Ortiz an Gioia was later extended to the full vectorial Föppl-von Kármán energy, and also to three-dimensional elasticity. These refinements explained the appearance of oscillations on two different length scales, with coarse oscillations in a direction normal to the boundary, and fine oscillations in the direction tangential to the boundary, as discussed in Sect. 3 below.

Recently interest has been directed to controlling the microstructures by designing the geometry of the debonded region appropriately [5, 6]. The key idea is to introduce a sacrificial layer between the film and the substrate, and then to selectively etch away a part of it, so that the boundary of the debonded region is straight. The film then partially rebonds to the surface, leading to complex patterns of tunnels. A study of these patterns within the Ortiz-Gioia framework, with a variational functional containing the Föppl-von Kármán energy and a fracture term proportional to the debonded area, is presented in Sect. 4. The mathematical analysis leading to the upper bounds of Theorem 6 suggests the presence of different types of patterns in different parameter ranges. The picture is rather easy in the two extreme cases in which the bonding energy per unit area is very small or very large. Indeed, in the first one the patterns observed for completely debonded films give the optimal energy scaling, in the second one the optimal state corresponds to the film completely bound to the substrate. In the intermediate regime we expect a richer picture, with bonded areas separated by thin debonded tunnels. For a certain regime, depending on the relation between the bonding energy per unit area, the film thickness and the compression ratio, a construction in which the tunnels branch and refine close to the boundary has a lower energy than the one with straight tunnels, see discussion in Sect. 4 below. The microstructure formation in thin films can be understood at a qualitative level as a form of Euler buckling instability. The relevant experiments, however, are well beyond the stability threshold, as discussed in Sect. 5 below.

2 Scalar Modeling of Compressed Thin Films

Ortiz and Gioia showed that, if tangential displacements are neglected, the energy of a compressed thin film can be characterized by the functional

$$I_\sigma[w] = \int_\Omega \left(\left(|Dw|^2 - 1\right)^2 + \sigma^2 |D^2 w|^2 \right) dx, \tag{1}$$

subject to $w = 0$ on $\partial\Omega$ and $w \geq 0$ in Ω. Here $\Omega \subset \mathbb{R}^2$ represents the debonded region, $w : \Omega \to [0, \infty)$ the rescaled normal displacement, and σ is a small parameter related to the thickness of the film. This functional arises also naturally in the study of liquid crystal configurations [7] and of magnetic structures in thin films [8]. Despite a large mathematical effort [7, 9–15] the problem (1) is not yet completely understood; it has been shown that the minimal energy is proportional to σ but the Γ-limit of $\sigma^{-1} I_\sigma[w]$ has only been partially identified. The natural candidate is

$$I_0[w] = \frac{1}{3} \int_{J_{Dw}} |[Dw]|^3 d\mathscr{H}^1 \tag{2}$$

restricted to functions $w : \Omega \to \mathbb{R}$ which solve the Eikonal equation $|Dw| = 1$ and are sufficiently regular. Here, J_{Dw} denotes the set of points (typically, a curve) where the gradient Dw is not continuous, $[Dw]$ denotes its jump across the interface, and $d\mathscr{H}^1$ the line integral along the interface. In particular, under the additional assumption that Dw is a function of bounded variation, it has been shown that for $\sigma \to 0$ the scaled functionals $\sigma^{-1} I_\sigma$ converge, in the sense of Γ-convergence, to I_0, see [11–13] for the lower bound and [14–16] for the upper bound. However, it is also clear that finiteness of the energy does not imply that Dw has bounded variation, but only that w belongs to a larger space, called $AG(\Omega)$, see [10, 11]. Therefore the result is still incomplete.

The Eikonal equation $|Dw| = 1$ with the boundary data $w = 0$ on $\partial\Omega$ is solved by the distance to the boundary, $w_0(x) = \text{dist}(x, \partial\Omega)$. For example, if Ω is a square this leads to the tent-shaped deformation illustrated in Fig. 1. The function w_0 is however only Lipschitz continuous, not twice differentiable, and makes the curvature term

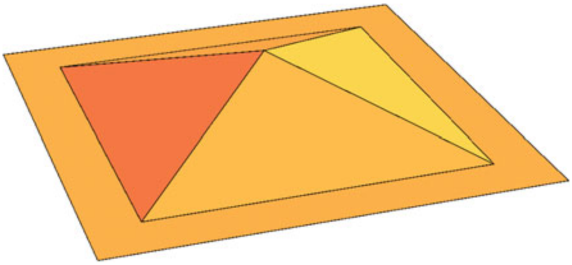


Fig. 1 Sketch of a deformation achieving the optimal energy in (1). Here the debonded region $\Omega = (0, 1)^2$ is a *square*, and the distance to the boundary gives a "tent"-form. The convolution in (3) makes the folds have smooth transitions on a small scale

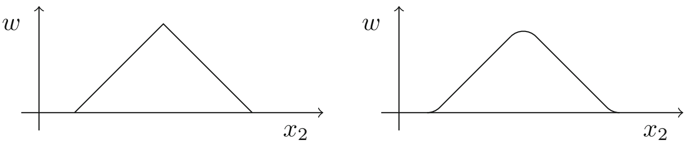


Fig. 2 Sketch of the effect of the mollification in (3) in a direction orthogonal to the fold. *Left panel*: the distance from the boundary $\mathrm{dist}(x, \partial\Omega)$ is a function with slope ± 1 and sharp kinks. *Right panel*: the mollification defined in (3) still has slope ± 1 on large parts of the domain, but has smooth transitions from one value to the other over a length of the order of σ, see Fig. 3

$\int_\Omega \sigma^2 |D^2 w|^2 dx$ infinite. Therefore Ortiz and Gioia [1, 2] proposed to use a smoothed version of the distance function,

$$w_\sigma(x) = \int_\Omega \mathrm{dist}(y, \partial\Omega)\varphi_\sigma(x-y)dy \tag{3}$$

where φ_σ is a mollifier on the scale σ, i.e., $\varphi_\sigma \in C_c^\infty(B_\sigma)$ with $\int_{\mathbb{R}^2} \varphi_\sigma\, dx = 1$ and $|D\varphi_\sigma| \le c/\sigma^3$. Then the regularized gradient Dw_σ has length close to 1 on most of the domain Ω, but at boundaries between regions where Dw_0 has different orientations Dw_σ changes smoothly over a length scale σ from one value to the other. The bending energy is correspondingly localized in a stripe of thickness 2σ around the interfaces, see Fig. 2. The prediction that minimizers of (1) are well represented by w_σ is in good agreement, at least for some geometries, with experimental observations [1, 2].

The work of Ortiz and Gioia was then extended to related problems, showing for example that under anisotropic compression branching-type microstructures appear close to the boundary [17, 18], or that in certain regimes telephone-cord blisters develop [19–21] thanks to the interaction between the elastic deformation and the fracture problem that determines the boundary of the debonded region.

3 Pattern Formation in Debonded Thin Films

A finer analysis of the nonlinear elasticity model that had led to (1) showed that, in the case of isotropic compression, also the in-plane components exhibit fine-scale oscillations which refine close to the boundary [22–25]. This analysis was based on the Föppl-von Kármán model, which includes the tangential components of the displacement u as well. After rescaling the energy takes the form (in the case of zero Poisson's ratio for simplicity)

$$E_\sigma[u, w] = \int_\Omega \left(|Du + Du^T + Dw \otimes Dw - \mathrm{Id}|^2 + \sigma^2 |D^2 w|^2\right) dx. \tag{4}$$

Here $\Omega \subset \mathbb{R}^2$ is, as above, the debonded region, and the displacements u and w vanish at the boundary of Ω, corresponding to the fact that the rest of the film is still bound to the substrate. The isotropic compressive strain has been scaled to 1, and one can check that $E_\sigma[0, w] = 1 + I_\sigma[w]$. The key result from [22, 23] was that the minimum energy scales proportional to σ:

Theorem 1 (From [22, 23]) *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with piecewise smooth boundary. Then there are two constants $c_L, c_U > 0$ such that*

$$c_L\sigma \leq \min\{E_\sigma[u, w] : u = 0, w = 0 \text{ } on \text{ } \partial\Omega\} \leq c_U\sigma. \tag{5}$$

The argument used for proving the lower bound also proves that a finite fraction of the energy is localized in a thin strip close to the boundary.

Similar statements hold if the plate theory in (4) is replaced by a fully three-dimensional nonlinear elastic model. For $v : \Omega \times (0, h) \to \mathbb{R}^3$, $h > 0$, we define

$$E_h^{3D}[v] = \frac{1}{h} \int_{\Omega\times(0,h)} W(Dv)dx \tag{6}$$

where $W : \mathbb{R}^{3\times3} \to [0, \infty)$ is the elastic stored energy density, which vanishes on the set of proper rotations SO(3) and has quadratic growth, in the sense that

$$c \operatorname{dist}^2(F, \mathrm{SO}(3)) \leq W(F) \leq c' \operatorname{dist}^2(F, \mathrm{SO}(3)) \tag{7}$$

for some positive constants c and c'. The factor $1/h$ is included explicitly in (6) to obtain an energy per unit thickness, corresponding to (4).

In the nonlinear case the thickness h of the film and the compression δ enter the problem separately, however to leading order and after scaling the optimal energy only depends on the combination $\sigma = h/\delta^{1/2}$. In order to understand this expression it is instructive to recall the relation between the three-dimensional problem E_h^{3D} and its two-dimensional counterpart E_σ. In particular, a given pair (u, w) in (4) corresponds to a three-dimensional deformation v_δ of the form

$$v_\delta(x_1, x_2, x_3) = (1 - \delta) [\psi(x_1, x_2) + x_3 n(x_1, x_2)] \tag{8}$$

where

$$\psi(x_1, x_2) = \begin{pmatrix} x_1 + 2\delta u_1(x_1, x_2) \\ x_2 + 2\delta u_2(x_1, x_2) \\ (2\delta)^{1/2} w(x_1, x_2) \end{pmatrix} \tag{9}$$

represents the deformation of the $x_3 = 0$ layer and

$$n(x_1, x_2) = \begin{pmatrix} -(2\delta)^{1/2}\partial_1 w(x_1, x_2) \\ -(2\delta)^{1/2}\partial_2 w(x_1, x_2) \\ 1 \end{pmatrix} \tag{10}$$

is, to leading order, the normal to the surface described by ψ and gives the out-of-plane component of the strain. An expansion of $E_h^{3D}[v_\delta]$ for small δ shows that the leading order contribution is proportional to $\delta^2 E_\sigma[u, w]$ if the Poisson's ratio of the material vanishes. See for example [22, App. A and App. B] for a more detailed discussion of this point. A rigorous relation between E_σ and E_h^{3D} was derived in [26, 27] by means of Γ-convergence, these results however are appropriate for a different regime, with much smaller energy, and therefore do not apply directly to the situation of interest here.

Theorem 2 (From [25]) *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with piecewise smooth boundary, $\delta \in (0, 1)$, $h \in (0, \delta^{1/2})$. Then there are two constants $c_L, c_U > 0$ such that*

$$c_L\sigma \leq \min\{\frac{1}{\delta^2}E_h^{3D}[u] : u(x) = (1-\delta)x \textit{ for } (x_1, x_2) \in \partial\Omega\} \leq c_U\sigma \qquad (11)$$

where $\sigma = h/\delta^{1/2}$.

The significance of Theorems 1 and 2 is best understood by considering the key ideas in the proofs. The upper bound in (5) and (11) is proven by explicitly constructing a suitable deformation field (u, w). This is done in several steps. The first step is the Ortiz-Gioia construction given in (3), which correctly describes the large-scale behavior of the film and relaxes the compression in direction normal to the boundary, as in Fig. 1. In the second step one adds fine-scale oscillations in the orthogonal direction, as illustrated in Fig. 3. This microstructure does not change the average shape significantly but relaxes the strain component tangential to the boundary. Finally, one realizes that optimal deformations have oscillations on a very fine scale close to the boundary, to adequately match the boundary data, but much coarser oscillations in

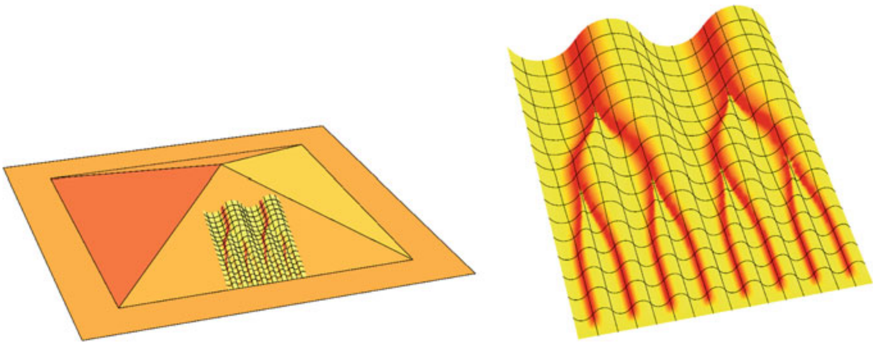


Fig. 3 Sketch of a deformation achieving the optimal upper bound in (5). As in Fig. 1, the debonded region $\Omega = (0, 1)^2$ is a square. The starting point, at a coarse scale, is the "tent"-form illustrated in Fig. 1. At a finer scale, folds orthogonal to the boundary relax the tangential compression (*left panel*, folds are only drawn in a small region). The period of the folds is of order h close to the boundary, and via a sequence of period-doubling steps becomes coarser in the inside (*right*, blow-up of the folds from the *left panel*)

the interior, to minimize the bending energy. Therefore a number of period-doubling steps are inserted, as illustrated in Fig. 3. Analogous self-similar branched patterns had previously appeared in the study of microstructures in shape-memory alloys [28, 29], where for a simplified model it had been possible to show that minimizers are indeed asymptotically self-similar [30]. The scaling in the presence of finite elasticity, both of the martensite and in the surrounding austenite, was then studied in [31, 32]; vectorial variants of the model were considered in [33, 34]. A similar approach has been useful also for a variety of other problems, ranging from magnetic patterns in ferromagnets [35–37] to field penetration in superconductors [38, 39], dislocation structures in crystal plasticity [40] and coarsening in thin film growth [41].

This variational approach to microstructure formation in thin elastic sheets is much more general, and indeed it can be applied to a number of related problems. One example is paper crumpling [42, 43] in which a thin plate, completely detached from the substrate, is confined to a small volume. In this case it has been possible to construct deformations with much smaller energy per unit volume. In particular one can obtain an energy per unit thickness proportional to $h^{5/3}$ [44, 45], and one can approximate any compressive deformation with this energy.

Theorem 3 (From [45]) *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain, $r > 0$. Then there is a map $v : \Omega \times (0, h) \to B_r(0)$ such that*

$$E_h^{3D}[v] \leq ch^{5/3}\,. \tag{12}$$

The constant c may depend on Ω and r but not on h. Further, if $v_0 : \Omega \to \mathbb{R}^3$ is a short map, i.e., a map which obeys $|v_0(x) - v_0(y)| \leq |x - y|$ for all $x, y \in \Omega$, then there is a sequence v_h, converging to v_0, such that

$$\lim_{h\to 0} \frac{1}{h^\alpha} E_h^{3D}[v_h] = 0 \tag{13}$$

for any $\alpha < 5/3$. Convergence of v_h is understood as uniform convergence of the vertical averages.

The proof of this is based on the combination of three ingredients. The first one is an approximation of short maps with Origami maps:

Theorem 4 (From [45]) *Let $v_0 : \Omega \to \mathbb{R}^3$ be a short map, i.e., a map which obeys $|v_0(x) - v_0(y)| \leq |x - y|$ for all $x, y \in \Omega$. Then there is a sequence v_j of Origami maps converging uniformly to v_0.*

Here we say that a map $v : \mathbb{R}^2 \to \mathbb{R}^3$ is an Origami map if it is continuous and piecewise isometric, i.e., if the domain can be subdivided into pieces such that v is a linear isometry (a translation plus a rotation) in each piece. The number of pieces is allowed to diverge only at infinity, in the sense that only finitely many pieces are allowed in any bounded subset of $\mathbb{R}^2$.

The second step is to approximate any Origami maps with low-energy maps:

Theorem 5 (From [45]) *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain, $v_0 : \Omega \times (0, h) \to B_r(0)$ be an Origami map. Then for any Origami map v_0 there is a sequence of maps $v_h : \Omega \times (0, h) \to \mathbb{R}^3$, converging to v_0, such that*

$$E_h^{3D}[v_h] \le Ch^{5/3}\,. \tag{14}$$

The constant may depend on Ω and v_0 but not on h.

This is proven by an explicit construction around each fold.

Another related problem of high current interest is the study of wrinkling patterns in graphene sheets [46, 47]. This has been addressed by a similar model, in which the boundary conditions are replaced by a viscous term describing the interaction with a substrate [48–50]. It would be interesting to see if the methods discussed here can be useful also for this variant of the problem.

4 Pattern Formation in Rebonded Thin Films

The microstructures spontaneously developed by compressed thin films can be controlled if the geometry of the debonded region is designed appropriately [5, 6]. One possibility is to introduce a sacrificial layer between the film and the substrate, and then to selectively etch away a part of it, so that the boundary of the debonded region is straight, see sketch in Fig. 4. The film then partially rebonds to the surface, leading to complex patterns of tunnels, which in some cases refine close to the boundary, see Fig. 5.

These patterns can be studied by coupling the von-Kármán energy with a fracture term proportional to the debonded area,

$$E_{\sigma,\gamma}[u, w] = \int_\Omega \left(|Du + Du^T + Dw \otimes Dw - \mathrm{Id}|^2 + \sigma^2|D^2w|^2\right) dx + \gamma|\{w > 0\}|. \tag{15}$$

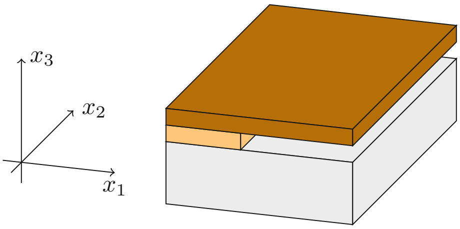


Fig. 4 Geometry of the partially delaminated film. The intermediate sacrificial layer is removed chemically only for $x_1 > 0$. The free-standing film is subject to compression at the Dirichlet boundary $x_1 = 0$ and may rebond to the substrate

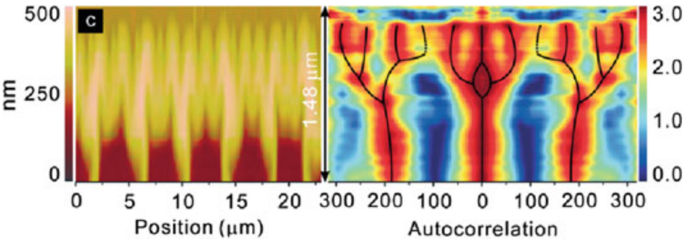


Fig. 5 Experimental picture of tube branching in $Si_{1-x}Ge_x$ film on a thick SiO_2 substrate. *Left* AFM image of the network near the etching front. *Right* autocorrelation pattern. Reprinted from [5, Fig. 2] with permission from Wiley

The three terms represent stretching, bending and bonding energies respectively. Here $u : \Omega \to \mathbb{R}^2$ are the (scaled) tangential displacements and $\gamma > 0$ is the bonding energy per unit area (related to Griffith's fracture energy), $|\{w > 0\}|$ represents the area of the set where the vertical displacement w is nonzero. Equivalently one could take the debonded state as reference and consider a negative term proportional to the rebonded area, $-\gamma|\{w = 0\}|$; the two energies only differ by an additive constant. The appropriate boundary conditions correspond to the film being bound to a substrate on one side of the domain; for simplicity we shall focus on $\Omega = (0, 1)^2$ with $u = 0$ and $w = 0$ on the $\{x_1 = 0\}$ side of Ω. As above, we assume $w \geq 0$ everywhere.

The mathematical analysis of the energy (15) leads to the rich phase diagram sketched in Fig. 6, which contains four different regimes [51] that we now illustrate.

For large specific bonding energy γ the film is completely bound to the substrate. In particular the film is flat, so that there is no bending energy, but the stretching energy is not released. The total energy is then proportional to the area of Ω, and one obtains $E_{\sigma,\gamma}[0, 0] = 2$. This is regime A in Fig. 6 and Theorem 6.

The opposite case of very small bonding energy γ is also easy to understand after the foregoing discussion: here the bonding term plays no significant role and

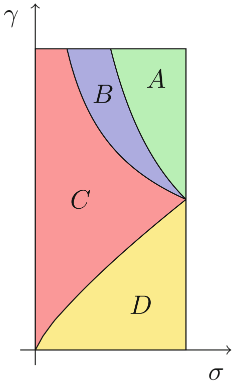


Fig. 6 Phase diagram for $E_{\sigma,\gamma}[u, w]$ in the (σ, γ) plane

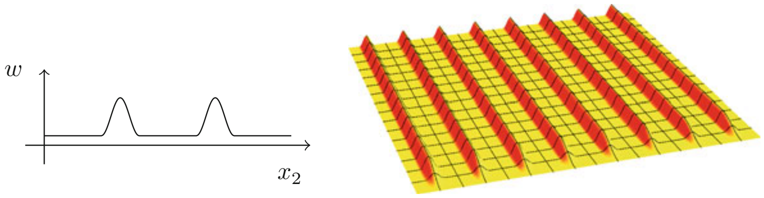


Fig. 7 Sketch of the laminate regime (B)

the film is completely detached from the substrate. One recovers the result of the blistering problem of Theorem 1, $E_{\sigma,\gamma}[u,w] \simeq E_\sigma[u,w] \le c\sigma$. The corresponding deformations are those illustrated in Fig. 3. This is regime D in Fig. 6 and Theorem 6.

For intermediate values of γ the situation is more complex, in particular debonded channels are formed, which separate wider bonded regions. In regime B the pattern is periodic and, away from the Dirichlet boundary, depends only on the tangential variable x_2. A large part of the film is bonded to the substrate, but bonded regions are separated by thin tubes, see Fig. 7. Denoting by h the period of the oscillations, and by δ the width of a tube, the total volume fraction of the tubes is δ/h, therefore the bonding energy is proportional to $\gamma\delta/h$. Each tube has to release a compression of h over a width δ, therefore the term $|Dw|^2$ is of order h/δ inside the tubes (the stretching energy is then completely relaxed). This gives $|Dw| \sim (h/\delta)^{1/2}$ in the tubes, and hence $|D^2w| \sim (h/\delta)^{1/2}/\delta$. Therefore the total energy can be estimated by

$$\gamma\frac{\delta}{h} + \sigma^2\frac{\delta}{h}\left(\frac{h^{1/2}/\delta^{1/2}}{\delta}\right)^2 = \gamma\frac{\delta}{h} + \frac{\sigma^2}{\delta^2}\,. \tag{16}$$

Optimizing in δ we obtain $\delta \sim \sigma^{2/3}h^{1/3}\gamma^{-1/3}$ (this is clearly only admissible if $\delta \le h \le 1$). The period h is fixed by the energetic cost of the interpolation region close to the boundary. In this part of the domain there is no stretch-free construction, and indeed an interpolation over a boundary layer of thickness ε results in a total stretching energy of $\varepsilon(1 + h^2/\varepsilon^2)$. Optimizing over ε we obtain $\varepsilon \sim h$, and therefore the total energy for the laminate construction is

$$h + \gamma\frac{\delta}{h} + \frac{\sigma^2}{\delta^2}. \tag{17}$$

Inserting the value of δ obtained above and minimizing in h we conclude that h and E are proportional to $(\sigma\gamma)^{2/5}$. The width of each tube δ is then proportional to $\sigma^{4/5}\gamma^{-1/5}$. This is regime B in Fig. 6; a precise version of this construction proves the second bound in Theorem 6.

If the bending term becomes more important, it is convenient to insert period-doubling steps, just like in the discussion of the functional (4). The resulting pattern is shown in Fig. 8. In comparison to the pattern of Fig. 3 the key difference is that the bending is localized to a small region, whereas large parts of the film are bond to the

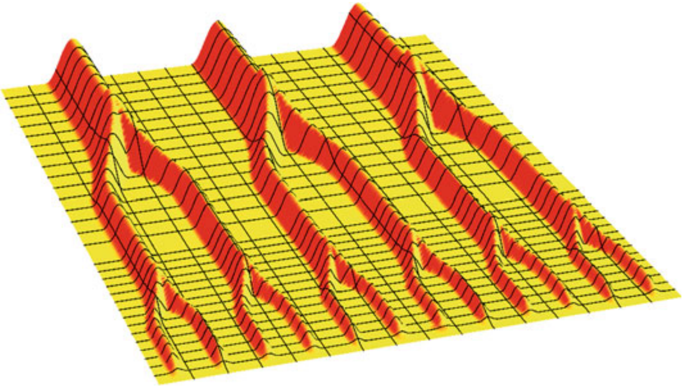


Fig. 8 Sketch of the tube branching regime (C)

substrate. The period-doubling steps are only possible at the expense of stretching energy; balancing the different terms one finds [51] that the resulting energy is proportional to $\sigma^{1/2}\gamma^{5/8}$. The result of the construction is summarized in the following statement.

Theorem 6 (From [51]) *Let* $\gamma > 0$*,* $\sigma \in (0, 1)$*. There are u, w which obey the stated boundary conditions and*

$$E_{\sigma,\gamma}[u,w] \le c \begin{cases} 1 & \textit{if } \sigma\gamma > 1 & \textit{(regime A)}, \\ (\sigma\gamma)^{2/5} & \textit{if } \sigma^{-4/9} \le \gamma \le \sigma^{-1} & \textit{(regime B)}, \\ \sigma^{1/2}\gamma^{5/8} & \textit{if } \sigma^{4/5} \le \gamma \le \sigma^{-4/9} & \textit{(regime C)}, \\ \sigma & \textit{if } \gamma < \sigma^{4/5} & \textit{(regime D)}. \end{cases} \tag{18}$$

The proof is based on making the constructions sketched above precise, details are given in [22] for regime D and in [51] for regimes B and C. Regime A, as discussed above, is immediate.

Optimality of the phase diagram just discussed can be at least partially proven by providing matching lower bounds on the energy. In particular, one can show the following.

Theorem 7 (From [51]) *Let* $\gamma > 0$*,* $\sigma \in (0, 1)$*. For any u, w which obey the stated boundary conditions one has*

$$E_{\gamma,\sigma}[u,w] \ge c \begin{cases} 1 & \textit{if } \sigma\gamma > 1 & \textit{(regime A)}, \\ (\sigma\gamma)^{2/3} & \textit{if } \sigma^{1/2} \le \gamma \le \sigma^{-1} & \textit{(regime B}'\textit{)}, \\ \sigma & \textit{if } \gamma < \sigma^{1/2} & \textit{(regime D}'\textit{)}. \end{cases} \tag{19}$$

Whereas the statement in regime D′ follows from [22], the other two bounds are proven in [51] using the Korn-Poincaré inequality for *SBD*^{2} functions obtained in [52].

Theorem 7 proves optimality in phases A and D. The bound in the intermediate region does not, however, match the upper bounds stated in Theorem 6. Therefore it is at this stage not clear if the branching patterns illustrated in Fig. 8 are optimal.

5 Linear Stability Analysis

The general form of the linearized Föppl-von Kármán plate theory under isotropic compression is [53, 54]

$$E_{\mathrm{FvK}}[u,w]=\frac{1}{2}Yh\int_{\Omega}\left[(1-\nu)|\varepsilon|^2+\nu(\mathrm{Tr}\varepsilon)^2+\frac{h^2}{12}\left[(1-\nu)|D^2w|^2+\nu(\Delta w)^2\right]\right]dx\,, \tag{20}$$

see also [2, 22] for a discussion in the present context and [27] for a rigorous mathematical derivation. Here $\nu\in[-1,1/2]$ is the Poisson ratio, Y Young's modulus, h the film thickness, and the strain ε is defined by

$$\varepsilon=Du+(Du)^T+Dw\otimes Dw-2\delta\mathrm{Id}\,, \tag{21}$$

where δ is the eigenstrain (i.e., the compression enforced by the substrate). We recall that we use $|M|^2=\mathrm{Tr}M^TM$ for the matrix norm. For $\nu=0$, after a rescaling (20) reduces to (4). We recall that in [22, App. B] it was shown that the scaling behavior of the functional E_{FvK} is the same for all $\nu\in(-1,1/2]$, hence our results hold also for generic values of the Poisson ratio. Of course, the regime $\nu\geq 0$ is the most relevant.

For small δ one can linearize around the state $u=0$, $w=0$. After straightforward computations this leads to

$$\begin{aligned}E^{\mathrm{lin}}_{\mathrm{FvK}}[u,w]=\frac{1}{2}Yh\int_{\Omega}\Big[&(1-\nu)|Du+Du^T-2\delta\mathrm{Id}|^2-4(1-\nu)\delta|Dw|^2\\&+\nu(2\mathrm{div}u-4\delta)^2-8\delta\nu|Dw|^2+\frac{h^2}{12}\left[(1-\nu)|D^2w|^2+\nu(\Delta w)^2\right]\Big]dx\,.\end{aligned}$$

In this linearized functional u and w are decoupled. The dependence on u is convex, hence $u=0$ is the minimizer with the given boundary data. The dependence on w is however not necessarily convex. Working for concreteness in a circle of radius R, we can assume w to be radial, $w(x)=\varphi(|x|)$, subject to $\varphi(R)=0$, so that

$$Dw(x)=\varphi'(|x|)\frac{x}{|x|}$$

and

$$D^2w(x)=\varphi''(|x|)\frac{x}{|x|}\otimes\frac{x}{|x|}+\varphi'(|x|)\left(\frac{\mathrm{Id}}{|x|}-\frac{x\otimes x}{|x|^3}\right)\,.$$

Inserting into the energy leads to the one-dimensional variational problem

$$\frac{1}{2}Yh\int_0^R \Bigg[-4\delta(1+\nu)(\varphi'(r))^2 + \frac{h^2}{12}\Bigg[(\varphi''(r))^2 + \left(\frac{\varphi'(r)}{r}\right)^2 + 2\nu\frac{\varphi'(r)\varphi''(r)}{r}\Bigg]\Bigg] r dr\,.$$

This is positive definite if the first term, of order δ, is not larger then the second term, of order h^2/R^2. Therefore the loss of stability, which corresponds to Euler buckling, occurs at strains $\delta \sim h^2/R^2$. Inserting the experimental data from [5], namely, $h \sim 20$ nm, $R \sim 10\,\mu$m, $\nu \sim 0.277$, leads to $\delta_{\text{crit}} \sim 4 \times 10^{-6}$, which corresponds to a strain of 0.0004 %. This is over three orders of magnitude smaller than the experimentally applied strain $\delta_{\text{Exp}} \sim 0.011 = 1.1\,\%$. Therefore the experiments we discussed take place well beyond the loss of linear stability, and a buckling-postbuckling analysis does not seem appropriate to understand the deformations. Our variational approach is instead constructed to deal with deformations and microstructures that appear in the deeply nonlinear regime and is therefore more suitable to study the mentioned experiments.

References

1. Ortiz, M., & Gioia, G. (1994). Dynamically propagating shear bands in impact-loaded prenotched plates II. Numerical simulations. *Journal of the Mechanics and Physics of Solids*, *42*, 531.
2. Gioia, G., & Ortiz, M. (1997). Delamination of compressed thin films. *Advances in Applied Mechanics*, *33*, 119.
3. Dacorogna, B. (1989). Direct methods in the calculus of variations. *Applied Mathematical Sciences*, *78*. Springer.
4. Müller, S.: F. Bethuel et al. (Ed.) *Calculus of variations and geometric evolution problems*. Lecture Notes in Math (1713, pp. 85–210). Springer.
5. Mei, Y., Thurmer, D. J., Cavallo, F., Kiravittaya, S., & Schmidt, O. G. (2007). Semiconductor sub-micro-/nanochannel networks by deterministic layer wrinkling. *Advanced Materials*, *19*, 2124.
6. Cendula, P., Kiravittaya, S., Mei, Y. F., Deneke, C., & Schmidt, O. G. (2009). Bending and wrinkling as competing relaxation pathways for strained free-hanging films. *Physical Review B*, *79*, 085429.
7. Aviles, P., & Giga, Y. (1987). A mathematical problem related to the physical theory of liquid crystal configurations. *Proceedings of the Centre for Mathematical Analysis, Australian National University*, *12*, 1.
8. DeSimone, A., Kohn, R. V., Müller, S., & Otto, F. (2000). Magnetic microstructures : a paradigm of multiscale problems. In J.M. Ball (ed.) *ICIAM 99: proceedings of the Fourth International Congress on Industrial and Applied Mathematics* (pp. 175–190). Oxford: Oxford University Press.
9. Aviles, P., & Giga, Y. (1996). The distance function and energy. *Proceedings of the Royal Society of Edinburgh Section A*, *126*, 923.

10. Aviles, P., & Giga, Y. (1999). On lower semicontinuity of a defect energy obtained by a singular limit of the Ginzburg-Landau type energy for gradient fields. *Proceedings of the Royal Society of Edinburgh Section A*, *129*, 1.
11. Ambrosio, L., De Lellis, C., & Mantegazza, C. (1999). Line energies for gradient vector fields in the plane. *Calculus of Variations and Partial Differential Equations*, *9*, 327.
12. Jin, W., Kohn, R.V. (2000). Singular perturbation and the energy of folds. *Journal of Nonlinear Science*, *10*, 355.
13. DeSimone, A., Kohn, R. V., Müller, S., & Otto, F. (2001). A compactness result in the gradient theory of phase transitions. *Proceedings of the Royal Society of Edinburgh Section A*, *131*, 833.
14. Poliakovsky, A. (2007). Upper bounds for singular perturbation problems involving gradient fields. *Journal of the European Mathematical Society*, *9*(1), 1.
15. Conti, S., & De Lellis, C. (2007). Sharp upper bounds for a variational problem with singular perturbation. *Mathematische Annalen*, *338*, 119.
16. Poliakovsky, A. (2005). A method for establishing upper bounds for singular perturbation problems. *Comptes Rendus Mathematique Academy of Science Paris*, *341*, 97.
17. Gioia, G., DeSimone, A., Ortiz, M., & Cuitino, A. M. (2002). Folding energetics in thin-film diaphragms. *Proceedings of the Royal Society of London. Series A*, *458*, 1223.
18. Conti, S., DeSimone, A., & Müller, S. (2005). Self-similar folding patterns and energy scaling in compressed elastic sheets. *Computer Methods in Applied Mechanics and Engineering*, *194*, 2534.
19. Gioia, G., & Ortiz, M. (1998). Determination of thin-film debonding parameters from telephone-cord measurements. *Acta Materialia*, *46*(1), 169.
20. Audoly, B. (1999). Stability of straight delamination blisters. *Physical Review Letters*, *83*, 4124.
21. Audoly, B. (2000). Mode-dependent toughness and the delamination of compressed thin films. *Journal of the Mechanics and Physics of Solids*, *48*, 2315.
22. Ben Belgacem, H., Conti, S., DeSimone, A., Müller, S. (2000). Rigorous bounds for the Föppl-von Kármán theory of isotropically compressed plates. *Journal of Nonlinear Science*, *10*, 661.
23. Jin, W., & Sternberg, P. (2001). Energy estimates for the von Kármán model of thin-film blistering. *Journal of Mathematical Physics*, *42*, 192.
24. Jin, W., & Sternberg, P. (2002). In-plane displacements in thin-film blistering. *Proceedings of the Royal Society of Edinburgh Section A*, *132*, 911.
25. Ben Belgacem, H., Conti, S., DeSimone, A., Müller, S. (2002). Energy scaling of compressed elastic films. *Archive for Rational Mechanics and Analysis*, *164*, 1.
26. Friesecke, G., James, R. D., & Müller, S. (2002). A theorem on geometric rigidity and the derivation of nonlinear plate theory from 3d elasticity. *Communications on Pure and Applied Mathematics*, *55*, 1461.
27. Friesecke, G., James, R. D., & Müller, S. (2006). A hierarchy of plate models derived from nonlinear elasticity by Gamma-convergence. *Archive for Rational Mechanics and Analysis*, *180*, 183.
28. Kohn, R. V., & Müller, S. (1992). Branching of twins near an austenite-twinned-martensite interface. *Philosophical Magazine A*, *66*, 697.
29. Kohn, R. V., & Müller, S. (1994). Surface energy and microstructure in coherent phase transitions. *Communications on Pure and Applied Mathematics*, *47*, 405.
30. Conti, S. (2000). Branched microstructures: scaling and asymptotic self-similarity. *Communications on Pure and Applied Mathematics*, *53*, 1448.
31. Conti, S. (2006). A lower bound for a variational model for pattern formation in shape-memory alloys. *Continuum Mechanics and Thermodynamics*, *17*, 469.
32. Zwicknagl, B. (2014). Microstructures in low-hysteresis shape memory alloys: scaling regimes and optimal needle shapes. *Archive for Rational Mechanics and Analysis*, *213*, 355.
33. Diermeier, J. (2013). Master's thesis, Universität Bonn.
34. Chan, A., & Conti, S. (2015). Energy scaling and branched microstructures in a model for shape-memory alloys with SO (2) invariance. *Mathematical Models and Methods in Applied Sciences*, *25*, 1091.

35. Choksi, R., & Kohn, R. V. (1998). Bounds on the micromagnetic energy of a uniaxial ferromagnet. *Communications on Pure and Applied Mathematics*, *51*, 259.
36. Choksi, R., Kohn, R. V., & Otto, F. (1999). Domain branching in uniaxial ferromagnets: a scaling law for the minimum energy. *Communications on Pure and Applied Mathematics*, *201*, 61.
37. Viehmann, T. (2009). Uniaxial ferromagnets. Ph.D. thesis, Universität Bonn.
38. Choksi, R., Kohn, R. V., & Otto, F. (2004). Energy minimization and flux domain structure in the intermediate state of a type-I superconductor. *Journal of Nonlinear Science*, *14*, 119.
39. Choksi, R., Conti, S., Kohn, R. V., & Otto, F. (2008). Ground state energy scaling laws during the onset and destruction of the intermediate state in a type-I superconductor. *Communications on Pure and Applied Mathematics*, *61*, 595.
40. Conti, S., & Ortiz, M. (2005). Dislocation microstructures and the effective behavior of single crystals. *Archive for Rational Mechanics and Analysis*, *176*, 103.
41. Conti, S., & Ortiz, M. (2008). Minimum principles for the trajectories of systems governed by rate problems. *Journal of the Mechanics and Physics of Solids*, *56*, 1885.
42. Lobkovsky, A. E., Gentges, S., Li, H., Morse, D., & Witten, T. A. (1995). Boundary layer analysis of the ridge singularity in a thin plate. *Science*, *270*, 1482.
43. Witten, T. A. (2007). Stress focusing in elastic sheets. *Reviews of Modern Physics*, *79*, 643.
44. Venkataramani, S. C. (2004). Lower bounds for the energy in a crumpled elastic sheet a minimal ridge. *Nonlinearity*, *17*, 301.
45. Conti, S., & Maggi, F. (2008). Confining thin elastic sheets and folding paper. *Archive for Rational Mechanics and Analysis*, *187*, 1.
46. Li, X., Cai, W., An, J., Kim, S. et al. (2009). Large-area synthesis of high-quality and uniform graphene films on copper foils. *Science*, *324*(5932), 1312.
47. Zang, J., Ryu, S., Pugno, N., Wang, Q., Tu, Q., Buehler, M. J., et al. (2013). Multifunctionality and control of the crumpling and unfolding of large-area graphene. *Nature Materials*, *12*, 321.
48. Zhang, K., & Arroyo, M. (2013). Adhesion and friction control localized folding in supported graphene. *Journal of Applied Physics*, *113*, 193501.
49. Zhang, K., & Arroyo, M. (2014). Understanding and strain-engineering wrinkle networks in supported graphene through simulations. *Journal of Mechanics and Physics of Solids, 72*, 61.
50. Arroyo, M., Zhang, K., & Rahimi, M. (2014). Mechanics of confined solid and fluid thin films: Graphene and lipid bilayers. *IUTAM Symposium on innovative numerical approaches for materials and structures in multi-field and multi-scale problems*.
51. Bourne, D., Conti, S., & Müller, S. (2015). Energy bounds for a compressed elastic film on a substrate. *Preprint* arXiv:1512.07416
52. Chambolle, A., Conti, S., & Francfort, G. (2014). Korn-Poincaré inequalities for functions with a small jump set. *To appear in Indiana University Mathematics Journal*. *Preprint* hal-01091710v1.
53. Antman, S. S. (1995). Applied Mathematical Sciences. In *Nonlinear problems in elasticity*, *107*. Springer.
54. Ciarlet, P. G. (1997). Theory of plates, Mathematical elasticity, vol. II. Elsevier.

Thermo-mechanical Behavior of Confined Granular Systems

Gülşad Küçük, Marcial Gonzalez and Alberto M. Cuitiño

Abstract We present a mathematical formulation that integrates thermal contact and Hertzian deformation models to understand the thermo-mechanical behavior of consolidated granular systems. The model assumes quasi-static equilibrium and quasi-steady heat conduction conditions that are appropriate for many thermally-assisted manufacturing processes. We perform a parametric study that explores the effect of applied thermal and mechanical loads, and of particles' thermal expansion. The nonlinearity of the multi-physics problem reveals that thermo-mechanical coupling enhances the effective thermal conductivity and mechanical stiffness by directly impacting the interrelation between contact conductance and overlapping between the particles. Alterations in temperature profiles and displacements of particles are significant for materials with higher thermal expansion coefficients. In this regards, it is worth noting that the results of the proposed thermo-mechanical model depart from those of conventional compaction models based on a continuum mechanics description.

1 Introduction

Understanding the fundamental multi-physics behind the thermo-mechanically coupled deformation of granular systems and its projections in macroscopic scale provides the essentials to fabricate particulate assemblies with specific functionalities.

G. Küçük (✉) · A.M. Cuitiño
Department of Mechanical and Aerospace Engineering, Rutgers University,
Piscataway, NJ 08854, USA
e-mail: gulsad@gmail.com

A.M. Cuitiño
e-mail: alberto.cuitino@rutgers.edu

M. Gonzalez
School of Mechanical Engineering, Purdue University, 585 Purdue Mall,
West Lafayette, IN 47907-2088, USA
e-mail: marcial-gonzalez@purdue.edu

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_3

A proper estimate of the mechanical strength, and of the thermal and electrical conductivity of a compacted solid is contingent upon the knowledge of microstructure formation during the deformation stage of the compression. Since thermally assisted compaction of granular matter is of great importance for a wide set of manufacturing processes, theoretical modeling and numerical simulations serve as significant tools to forecast the macroscopic behavior of materials, essentially when experimental techniques are also unfeasible.

At present, one of the most implemented methodologies to elucidate the collective behavior of particulate materials is the continuum mechanics approach, in which the granular material is assumed to be statistically homogenous [1]. This is achieved by treating the system as units of ordered groups, simulating disordered arrangements by statistical correlation functions or using empirical correlations. The statistical averaging technique provides homogenized solutions of the highly heterogeneous granular media at the cost of imposing two assumptions: (i) affine motion approximation, namely the motion of each grain follows the macroscopic strain, and (ii) well-bonded structure, contact number and positioning do not change under the applied load. Despite the fact that the effective medium theory particularly estimates the effective elastic moduli of packed bed of spherical particles to a large extent, the discrepancy between numerical and experimental results is remarkable. Makse and co-workers questioned the relevance of force laws defined at single contact level, where they pointed out that the simplification done in effective medium theory is the misleading element in the formulation [2, 3]. Affine motion assumption demolishes the ability of the approach to account for the relaxation and rearrangement of particles that are under shear deformation. Moreover concerning the variety of boundary conditions and geometrical effects, experimentation techniques become insufficient in providing sufficient information about the microstructure to feed empirical correlations.

The second most adopted approach treats the particles as individual bodies. Originating from particle-particle interactions based on constitutive relations of contact mechanics [4–6], the discrete element method has been widely used in the field of particle scale research [7]. Pioneers of this approach, Cundall and Strack introduced an explicit numerical scheme to practice the granular dynamics by defining particles' interactions over the contact networks and solving for particles' motion under the state of force balance equilibrium [8]. The integration of particle motion and energy to the macroscopic behavior of the assembly, provides the required understanding of overall behavior of the confined material [9]. The main advantage of this methodology is the capability of presenting broad information about the micro-structural arrangement of the granular media. Although there exists computational challenges to model a large number of particles system with discrete elements methods, advances in simulation techniques enhance the implementation of this approach into the field of multi-physics problems of granular systems.

Recently researchers also focus on multi-scale approaches to describe the macroscopic behavior of granular systems. Zheng and Cuitino implemented a quasi-continuum approach to bridge the gap between micro and meso scale description by using a discrete-continuum formulation of elastic-inelastic deformations occurring in the post-rearrangement regime of consolidation of inhomogeneous granular beds [10]. Since this approach provides the flexibility of storing individual particle interactions in a FEM scheme, it provides the overall behavior of the entire body without loosing critical information specific to microstructure. Koynov et al. presented a notable adaptation of this approach on the topic of powder compactions for pharmaceutical purposes [11]. In this study we present a new methodology to explore the family of multi-physics problems such as thermo-mechanical coupling. The method is an extension of discrete element method that accounts for the effective modeling of heat conduction, and similar in spirit to early studies of Vargas and McCarthy [12] and Feng et al. [9].

Current study incorporates early mathematical models that are developed for conforming thermal contact of elastic, spherical surfaces [13–15]. These theoretical models are validated though experimental studies [16–18]. Also there exists studies that aim to relax some of the assumptions by focusing on elasto-plastic contacts [19], or rough surfaces of non-conforming contact [20–22]. Recently the field of granular matter gained importance in the light of understanding the correlation between geometry, loading conditions and anisotropic microstructural arrangements that determine the macroscopic behavior of compacted particulate system [23]. Gonzalez and Cuitino introduced a new formulation that account for the interplay of nonlocal mesoscopic deformations characteristic of confined granular systems. In the absence of the classical restriction of independent contacts of Hertz law, the extended theory of nonlocal contact formulation provides predictive models at moderate levels of deformation and high confinement [24]. In their study on effects of packing grains by thermal cycling, Chen et al. [25] showed that thermal expansion, due to the imposed thermal gradient, has significant effect on the rearrangement of particle bed. Vargas and McCarthy focused on the problem of how the forces supporting the grains are distributed under the effect of thermal expansion [26].

It is the purpose of this study to suggest that insight into the nature of thermo-mechanical behavior of confined granular materials. We aim to discover the effects of thermal and mechanical coupling at the particle level and implement the required amendments to continuum level models. We present the system of governing equations, which define prescribed state of the assembly under steady state conditions, in terms of heat and force transfer between the contacting particle pairs. Owing to the fact that the nature of the problem leads to highly non-linear coupled equations, regular packing simplifies the problem and makes it mathematically traceable. Moreover we consider the analogous problem from the perspective of the conventional continuum mechanics approach. Practicing a thermo-elastic continuum model to simulate the system, we focus on the effective mechanical and transport properties to account for the unique characteristics of granular materials.

2 Particle Mechanics Approach

Our point of departure for the particle-scale description of thermo-elastic contact of spherical smooth particles is to integrate the well-known theory of Hertzian deformation, [4], and heat conduction through the common interface of deformed particles in contact, [13, 14]. Under steady state conditions, the total heat transferred to individual particle m from neighboring particles n and the total of forces acting on particle m are zero,

$$Q^m = \sum_{n \varepsilon \mathcal{N}_m} Q^{mn} = 0 \tag{1}$$

$$\mathbf{F}^m = \sum_{n \varepsilon \mathcal{N}_m} F^{mn} \mathbf{n}^{mn} = 0 \tag{2}$$

$$\mathbf{n}^{mn} = \frac{\mathbf{x}^m - \mathbf{x}^n}{\|\mathbf{x}^m - \mathbf{x}^n\|} . \tag{3}$$

where $\mathbf{n}^{mn}$ is the unit normal vector defined from centers of particle n to particle m. $\mathbf{x^m}$ and $\mathbf{x^n}$ are the position of the particles.

Johnson identifies the elastic deformation of locally spherical particles that are subject to a compression load by contact mechanics considerations in his book [27]. Small-strain deformation of conforming surfaces results in a flat circle of contact area. Collinear contact force at this elastic contact of the particles m and n is defined through Young's moduli, E^m and E^n; Poisson's ratios, ν^m and ν^n; particle radii, R^m and R^n of particle m and n; and overlap, γ^{mn}, between these particles,

$$F^{mn} = \frac{4}{3} E^{mn} (R^{mn})^{1/2} (\gamma^{mn})^{3/2} \tag{4}$$

where

$$R^{mn} = \left[\frac{1}{R^m} + \frac{1}{R^n} \right]^{-1} \tag{5}$$

$$E^{mn} = \left[\frac{1 - (\nu^m)^2}{E_m} + \frac{1 - (\nu^n)^2}{E_n} \right]^{-1} \tag{6}$$

$$\gamma^{mn} = R^m + R^n - \|\mathbf{x}^m - \mathbf{x}^n\| . \tag{7}$$

One particular effect of applied thermal load on the system of particles is the change in radii due to thermal expansion. Similar to previous studies in the literature [26, 28], in the present study, linear thermal expansion formulation is taken into consideration.

$$R^m = R^m_{ref} \left[1 + \alpha^m \left(T^m - T^m_{ref} \right) \right] \tag{8}$$

Here α^m is the thermal expansion coefficient, T_{ref} is the reference temperature and R^m_{ref} is the radius of particle at the reference temperature. Due to the dependence of contact geometry on the nature of thermo-mechanically coupled problem, it is expected to capture a distribution of contact area formation throughout the compacted medium.

There has been considerable research on thermal-contact models. The major heat transfer mechanisms in compacted particle beds consist of conduction through solid, conduction through the contact area between two touching particles, conduction to/from interstitial fluid, heat transfer via convection, radiation between particle surfaces, radiation between neighboring voids [12]. For a system of granular media where the thermal conductivity of the solid particles is much larger than the interstitial medium, the driving mechanisms for the heat transfer are the first two. Concerning the problem of thermally-assisted compaction of spherical particles in vacuum, we focus on the thermal contact models that consider the conduction through solid particle and through the contact area between two touching particles.

Analytical solution of the heat conduction through the solid phase of ordered spherical particles has been proposed by Chan and Tien [14] and Kaganer [15]. Moreover the problem of heat transfer regarding the compaction of particles that are in or nearly in contact is deeply investigated by Batchelor and O'Brien [13]. In an attempt to find the approximate effective thermal conductivity of ordered and randomly packed granular beds, Batchelor and O'Brien discussed the heat flux across the flat circle of contact between smooth, conforming, and elastic particles. In this study we adopt Batchelor and O'Brien's model for predicting the heat conductance, which is the ability of two touching surfaces to transmit heat through their mutual interface. Heat flux across the contact area of two spherical smooth particles is given as

$$Q^{mn} = 2a^{mn}k^{mn}(T^m - T^n) \tag{9}$$

where k^{mn} is the arithmetic mean of the thermal conductivities of two conforming particles, and a^{mn} is the Hertzian contact area.

$$k^{mn} = \frac{1}{2}\left[\frac{1}{k^m} + \frac{1}{k^n}\right]^{-1} \tag{10}$$

$$a^{mn} = \sqrt{\gamma^{mn}R^{mn}} \tag{11}$$

The total heat flow to an individual particle, Eq. (1), is calculated by adding the heat flow at each contact of the particle between its neighboring particles Eq. 9. As discussed by the thermal contact models introduced in the literature [13, 14], Eq. (1) requires that at each contact of the individual particle, the temperature is equal to the temperature calculated at the center of the particle. In other words, the temperature does not very significantly within the particle, which also imposes that the contact conductance at the interface of conforming particles is relatively smaller than the heat conductance within the particle.

$$\frac{2k^{mn}a^{mn}}{k^{mn}A/R^{m}} \ll 1 \tag{12}$$

where A is the cross sectional area, $A = \pi(R^m)^2$ and Eq. (12) defines the state of Biot number much less than 1. This assertion is applied by several authors in earlier studies [12, 29]. The condition of $a^{mn} \ll R^{mn}$ is also enforced by the assumption of small-strain deformation of elastic bodies in contact.

2.1 Simulation Configuration

Referring to the previous experimental studies on regular and random packings of granular media, Walton points out that although the regular packing models are founded on extreme assumptions, they are capable of capturing vast majority of the characteristics of a real granular media [30]. In the present study we consider a simple cubic packing of identical elastic spheres, which are constrained between parallel planes of infinite extent. Compression load, temperature gradient are applied along the major and finite direction. Stress and heat flux are defined to only depend on externally applied thermal and mechanical loads, and weight of the particles is neglected. For such regular packings each layer of arrangement is isothermal normal to the direction of applied load. Also, since these transversely oriented particles are, at most, point contact, for each individual particle there is only one pair of contact area aligned with the direction of applied thermal and mechanical load. Due to the symmetry of the problem, it is sufficient to consider a single column of square cross-section containing the longitudinally compressed spheres together. This concept is similar, in spirit, to the work of Chan and Tien [14], who proposed the effective thermal resistance, and to the work of Walton [30] who presented a method to calculate the effective elastic moduli of such packings. The above description for the specified granular media, which is under thermally-assisted compaction, can be modeled as a chain of elastic particles, seen in Fig. 1.

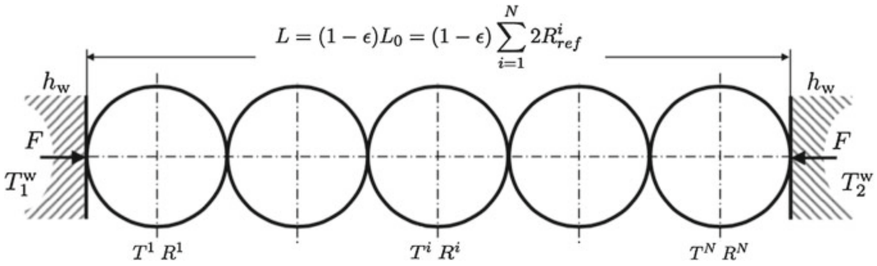

$L = (1-\epsilon)L_0 = (1-\epsilon)\sum_{i=1}^{N} 2R^i_{ref}$
h_w
F
T_1^w
h_w
F
T_2^w
$T^1\ R^1$
$T^i\ R^i$
$T^N\ R^N$

Fig. 1 Sketch of initial configuration

2.2 Wall-Particle Interaction

Given such a setting that the chain of particles is compressed between two parallel walls, which are maintained at different temperatures, the wall-particle interaction is one of the important factors of this problem. In this study, analogous to ghost-cell method, the contact between boundary particle and adjacent wall is simulated as the contact between a boundary particle and a ghost particle. Based on the rigid wall assumption, ghost particle and boundary particle are set to have the same material properties and radius. The temperature difference between the ghost particle and the wall surface is the same as the temperature deviation between the wall surface and the boundary particle. The boundary wall is assumed to be located in the midst of these symmetrically deformed particles. The temperature difference, overlap and contact area are formulated in Eqs. (13)–(15), where T^{ws} and T^{w} refer to the temperature at the wall surface and wall temperature, respectively. Subscript g is used to indicate the ghost particle.

$$\Delta T^{mg} = 2(T^{m} - T^{ws}) \tag{13}$$

$$\gamma^{mg} = 2(R^{m} - ||\mathbf{x^{m}} - \mathbf{x^{ws}}||) \tag{14}$$

$$a^{mg} = (\gamma^{mg}\frac{R^{m}}{2})^{1/2} \tag{15}$$

At the boundary surfaces, heat transfer between the boundary particle to the wall can be expressed by two main heat transfer mechanisms, (i) heat is conducted over flat circle of contact between the particle and adjacent wall surface; (ii) convective heat transfer, which is dependent on walls' convection coefficient of h_w, takes place between the wall surface and the wall.

$$Q^{m-ws} = k^{m}a^{mg}\Delta T^{mg} \tag{16}$$

$$Q^{ws-w} = -h_{w}\pi(a^{mg})^{2}(T^{ws} - T^{w}) \tag{17}$$

The temperature at the wall surface can be obtained for the equilibrium of Eqs. (16) and (17). The final set of equations, which define the wall-particle interaction, are the following:

$$Q = 4ka^{mg}\big(T^{m} - \frac{4k^{m}T^{m} + h_{w}\pi a^{mg}T^{w}}{4k + h_{w}\pi a^{mg}}\big) \tag{18}$$

$$F = \frac{E^{m}}{3(1-(\nu^{m})^{2})}R^{m}\gamma^{mg}\,. \tag{19}$$

3 Conventional Continuum Mechanics Approach

While the particle mechanics approach aims to elucidate the formation and the evolution of the microstructure of the granular media at particle-level, there has been considerable research directed towards understanding macroscopic behavior of compacted materials. Some of the early work on theoretical modeling of transport properties are devoted to the estimation of thermal, and electrical conductivity, elastic, plastic mechanical properties of ordered and disordered arrangements. Originating from the pair interactions between particles, the macroscopic properties are obtained using various homogenization techniques and postulating continuum constitutive laws [31]. In this study, we consider a continuum system that mimics the particle level description for small strain thermoelasticity, which incorporates the proposed effective mechanical and thermal properties for granular beds under compaction. Governing field equations of motion and energy are the following

$$\operatorname{div}(\boldsymbol{\sigma}) = \mathbf{0} \tag{20}$$

$$\operatorname{div}\left[k \operatorname{grad}(T)\right] = 0 \tag{21}$$

where the Cauchy's stress, $\boldsymbol{\sigma}$, is formulated as combination of classical linear elasticity theory and simple linear thermal expansion.

$$\boldsymbol{\sigma} = -\lambda \operatorname{tr}(\boldsymbol{\varepsilon})\mathbf{I} - 2\mu\boldsymbol{\varepsilon} + (3\lambda + 2\mu)\alpha(T - T_{ref})\mathbf{I} \tag{22}$$

For the basic problem of one dimensional steady state thermoelasticity of continuum media, where the body forces are neglected, the solution depends linearly on elastic constants (λ, μ), thermal expansion and conduction coefficients, compaction strain and thermal gradient. Since $\varepsilon_{22} = \varepsilon_{33} = 0$ holds, ε_{11} is referred as ε. Equations of motion and energy Eqs. (20) and (21) can be rewritten as

$$\sigma = -(\lambda + \mu)\varepsilon + \alpha(3\lambda + 2\mu)(T - T_{ref}) \,, \tag{23}$$

$$q = k\frac{\partial T}{\partial x} \,. \tag{24}$$

Effective mechanical properties of granular beds are of heavy interest in many theoretical studies. Some of these include: calculation of the principal elastic modulus of vertical compression of spherical particles without any lateral extension [30]; derivation of the finite and incremental elasticity of random packing of identical particles using energy methods [32]; enhancement of the derived formulas based on the pressure dependence of the elastic moduli of granular packings [2, 3].

In this study we extend the effective medium theory with the thermal contact model principles by incorporating the particle interactions to account for the local field effects. We re-formulate the effective elastic properties and effective thermal conductivity accordingly, and implement these parameters in continuum mechanics model. The effective mechanical properties can be expressed in terms of the applied stress, σ, and bulk material properties [32],

$$C_n = 4\frac{\mu}{1-\nu} = 4\frac{E}{2(1+\nu)}\frac{1}{1-\nu} = \frac{2E}{1-\nu^2} \tag{25}$$

$$\tilde{\lambda} + 2\tilde{\mu} = \frac{3}{20\pi}C_n(\phi_s Z)^{2/3}\big(\frac{6\pi\sigma}{C_n}\big)^{1/3} \tag{26}$$

$$3\tilde{\lambda} + 2\tilde{\mu} = \frac{1}{4\pi}C_n(\phi_s Z)^{2/3}\big(\frac{6\pi\sigma}{C_n}\big)^{1/3} \tag{27}$$

where C_n is named as stiffness of the system, ϕ_s is the packing fraction, and Z is the coordination number. Effective mechanical properties, $\tilde{\lambda}$, $\tilde{\mu}$, and effective thermal conductivity, $\tilde{k}$, are implemented in continuum mechanics model and listed as conventional continuum solution. As a simple application of this theory, we consider a case of particles' chain that is compacted by a ratio of ε, under the effect of a thermal gradient of $T_2^w - T_1^w$. The expression that defines the stress evaluation through the chain is found as

$$\sigma = \phi_s Z C_n\big(\frac{3}{32\pi^2}\big)^{1/2}\big(|-\varepsilon\frac{3}{5} - \alpha(\frac{T_2^w + T_1^w}{2} - T_{ref})|\big)^{3/2}. \tag{28}$$

Effective thermal conductivity, $\tilde{k}$, of the granular bed is substantially sensitive to the thermal and the elastic properties of individual particle. Regarding ordered cubic packing configuration, it is known that thermal contact models provide accurate results in estimating steady and average temperature profiles [33]. Three major analytical solutions in literature, by Batchelor and O'Brien [13], Chan and Tien [14], Kaganer [15] and Siu and Lee [34], are proposed to determine the effective thermal conductivity. Since Batchelor and O'Brien's [13] solution stays in remarkable agreement with the particle mechanics results in terms of heat transferred through the chain, we adopted this solution for effective thermal conductivity coefficient in our continuum mechanics approach. This comparison is shown in Fig. 2, where PMA and CMA refer to particle mechanics approach, and conventional continuum mechanics approach, respectively.

$$\begin{aligned}
\tilde{k}^{B\&O} &= k\big(\frac{6\sigma}{C_n}\big)^{1/3} \\
\tilde{k}^{C\&T} &= 0.9454k\big(\frac{6\sigma}{C_n}\big)^{1/3} \\
\tilde{k}^{S\&L} &= 0.8278k\big(\frac{6\sigma}{C_n}\big)^{1/3}
\end{aligned}$$

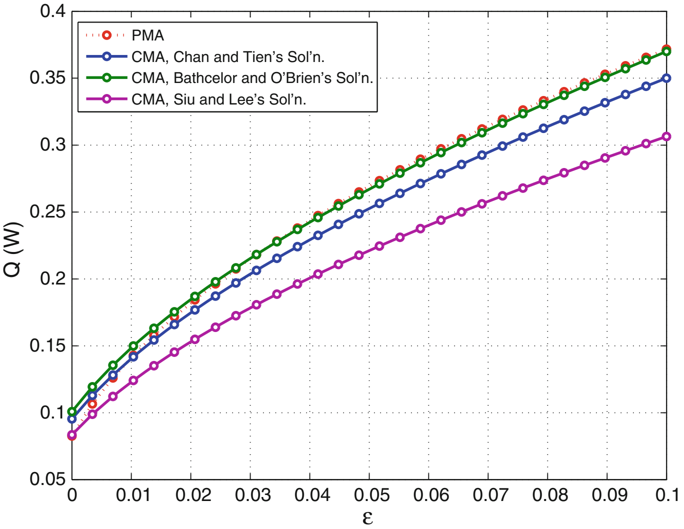


Fig. 2 Comparison of continuum solutions adopting different thermal contact models with respect to particle mechanics solution. Heat versus compaction strain, ε, is evaluated at $T_2^w - T_1^w = 600$ K

4 Results

According to the Hertz theory [35], the collinear contact force between the elastically compressed particles is a nonlinear function of the overlap, which is generated under the effect of the external load, between the particles. For the case of thermally-assisted compaction of granular system of particles, this dependency is altered under the effect of applied thermal gradient. Figure 3 shows the ratio of force, needed to compress the system, in particle mechanics approach to the force in conventional continuum mechanics approach. CMA significantly overestimates the thermal stress within the chains system, particularly for the range of high thermal gradient and low mechanical load. Moreover concerning the highly compacted systems CMA underestimates PMA solution for the system of particles by 10 %.

Similar to compaction force comparison, conventional continuum solution predicts higher heat transferred values for analogous particles system. It is also shown in Fig. 4 that as the packing density of the deformed particles system is increased, conventional continuum mechanics solution becomes more effective in estimating the particle-level solution.

Under three difference compaction strains, 2.5, 5, 10 %, the effect of wall-particle interaction is examined through the chain of particles by imposing a thermal gradient of 300 K between the two boundary walls of the system. For each case, wall heat transfer coefficient, h_w, is ranged from 1 to 10^7 W/m^2 K. Figure 5 indicates the two limiting cases of perfect insulating and perfect conducting walls.

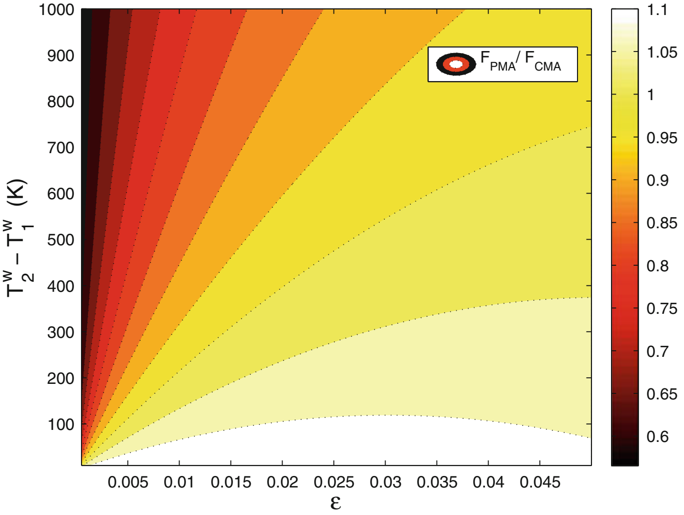


Fig. 3 Comparison of force calculated in PMA and CMA under varying thermal and mechanical loading conditions

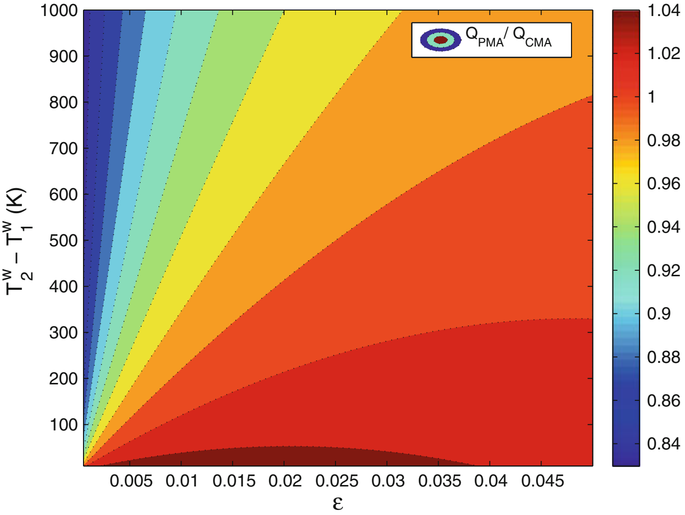


Fig. 4 Comparison of heat calculated in PMA and CMA under varying thermal and mechanical loading conditions

4.1 Role of Thermal Expansion

Systems of granular materials with different thermal expansion properties respond in various re-arrangements to a particular thermal and mechanical load. The following

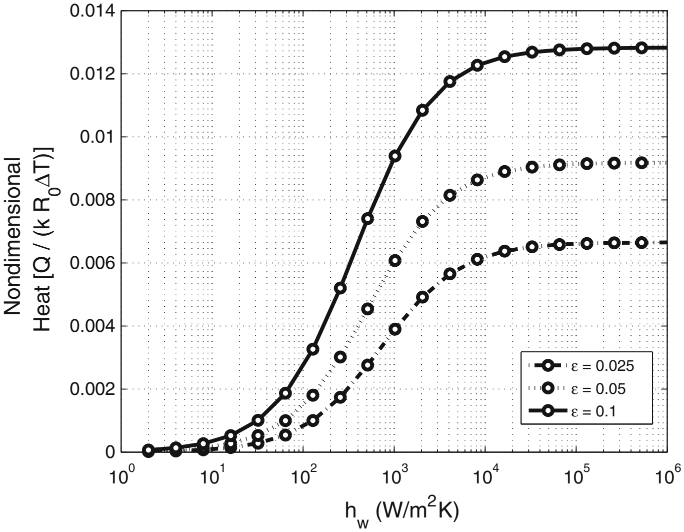


Fig. 5 Correlation between heat and wall conductance at different compaction strains

numerical experiment compiles the results of three different homogeneous system of particles: SS304, Aluminum, and Teflon with different thermal expansion values, 17.3 10^{-6} 1/K, 23.6 10^{-6} 1/K, 250 10^{-6} 1/K, respectively. In the above-mentioned cases of comparison, the chain is compacted to 2.5 % of the initial length and a total thermal gradient of 300 K is applied between the two boundary walls. Alterations in displacement of each particle due to increase of thermal stress can be traced to unveil the effect of thermal expansion coefficient on the system of particles under thermally-assisted compaction. In Fig. 6 the displacement of each particle is divided by the total mechanical deformation applied on the system. The non-dimensional displacement of the particle in contact with the fixed boundary is listed as 0, whereas the one in contact with the heated moving boundary wall is 1.

While reaching to equilibrium the two dominant mechanisms, thermal and mechanical stresses, induce a nonlinear distribution of displacements, which is a unique characteristic of particulate systems. This deviation from linear continuum solution is enhanced for systems with high thermal expansion property.

4.2 Role of Applied Mechanical Load

Under the effect of a modest thermal gradient of 300 K, three different mechanical loading conditions, 1, 2, 5–10 %, are compared in Fig. 7. In order to compare the coupled effect of thermal gradient and mechanical deformation, two extreme cases for wall-particle interactions are also considered. In case of perfect insulating walls, h_w is

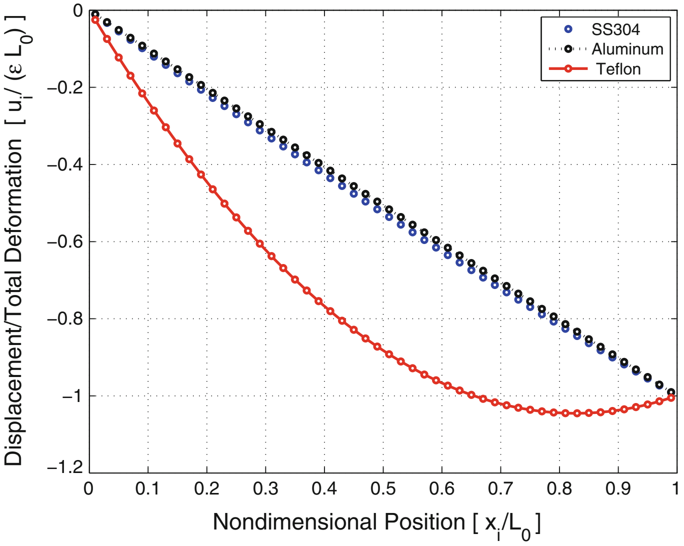


Fig. 6 Relative displacement of each particle within the chain at $\varepsilon = 0.025$ and $T_2^w - T_1^w = 300\,\mathrm{K}$

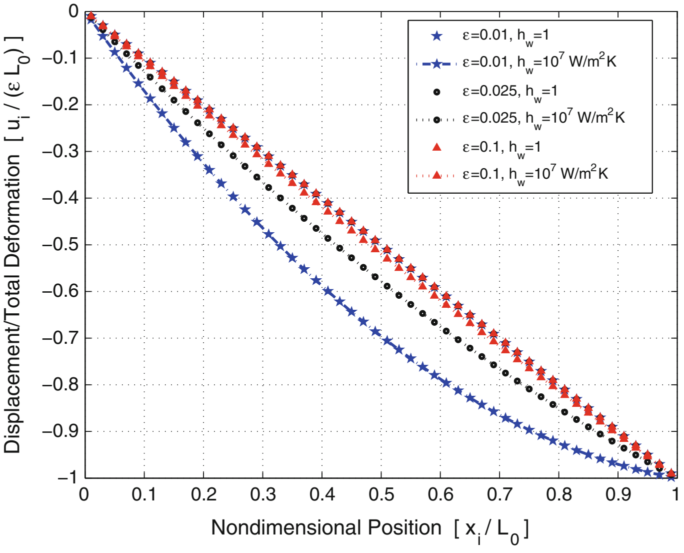


Fig. 7 Non-dimensional displacement versus initial position of the particles under different mechanical loads, with varying heat convection coefficients of the boundary walls

assumed to be 1 $\mathrm{W/m^2\,K}$. This particular condition is a simulation of pure mechanical loading where we expect to have linear distribution of non-dimensionalized positions' of particles within the chain system. On the other hand the case of $h_w = 10^7\,\mathrm{W/m^2\,K}$ simulates the condition of perfectly conducting walls.

Regarding the system of SS304 spherical particles' chain Fig. 7 indicates that nonlinearity in distribution of displacements is more dominant for low mechanical loading cases.

4.3 Role of Thermal Gradient

A recent experimental study on silos of spherical glass particles showed that thermal cycling, and the difference in thermal expansion properties of the granular material with respect to its container, significantly affect the packing fractions of granular materials in the absence of mechanical compaction [25]. In the current study we focus on the active interval where thermal gradient acts as a dominant mechanism compared to mechanical deformation. A chain of spherical particles is gradually consolidated up to a compaction strain of 5 % of their initial length, while thermal gradient between the two boundary walls is increased to 1000 K.

The ratio of the displacements calculated in PMA to CMA indicates a discrepancy between these two approaches. In Fig. 8 the maximum difference between particle mechanics approach and conventional continuum mechanics approach is traced. Under the effect of low mechanical deformation, such as $\varepsilon < 0.02$, and high thermal gradient conditions the continuum solution overestimates the actual position of particles up to 40 % of the solution provided by particle mechanics approach. The difference between these two solutions diminishes as the packing density of the granular system increases.

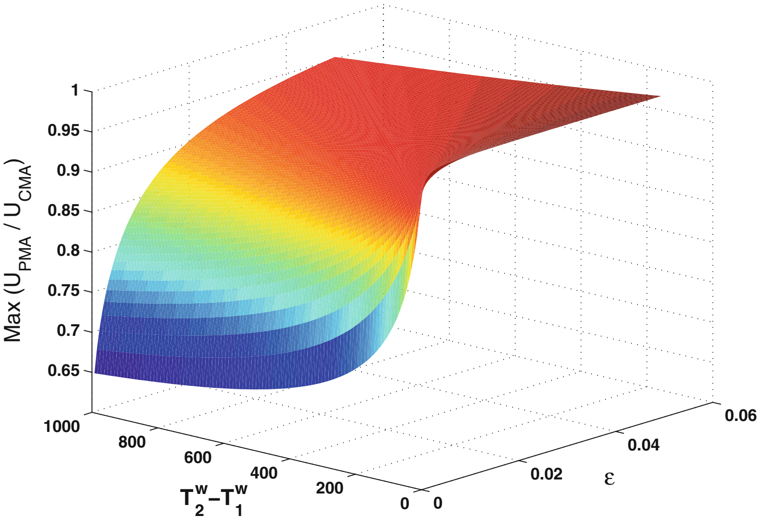


Fig. 8 Maximum difference between the calculated displacements in particle mechanics and in conventional continuum mechanics approaches

5 Conclusion

In this study we present a numerical model to describe the thermo-mechanical behavior of a confined granular system by adopting a detailed description at the particle level. We integrate thermal-contact and Hertzian deformation models to simulate the temperature and displacement of consolidated granular medium. One-dimensional model provides an opportunity to unveil the relation between two dominant mechanisms affecting the thermal and mechanical equilibrium of the particulate systems. In order to capture the actual physical conditions, we consider wall-particle interactions ranging from perfect insulating to perfect conducting walls.

The numerical results indicate that integration of thermal deformation with the elastic contact models induces the incongruity seen in mechanical deformation based compaction models. The coupled phenomena introduce highly nonlinear system of equations, and it imposes variation in contact areas and nonlinear temperature distribution within the particulate material. This effect is enhanced for particles with larger thermal expansion coefficient. It appears that the critical regime, where the nonlinearity due to thermo-mechanical coupling becomes more dominant, is low mechanical load and high thermal gradient conditions.

As a multi-physics problem, thermally-assisted compaction shows a significant dependence on the thermal expansion of the particles. Discrete solution based on the particle mechanics approach that adopts the thermal contact model, carries out this dependence and the nonlinearity enhanced by thermal strains, successively. Despite the fact that effective medium theory improves the continuum solution to a large extend, it fails to capture the characteristics of multi-physics of the problem, particularly for the cases of low thermal gradient coupled with high mechanical load.

Looking toward to future, we are now in a position to address a variety of important questions, such as; (i) what can be a further improvement in effective medium theory that also account for an effective thermal expansion coefficient depending not only on the bulk properties but also loading conditions of the compacted granular assembly? (ii) what is the role of uneven distribution of contact areas and nonlinear temperature distribution on formations of heterogeneous force and heat networks within the concept of the micro-structural arrangement of granular system to macroscopic behavior of the thermally-assisted compacted end product?

Acknowledgments This work has been partially supported by U.S. Army ARDEC grant under the project titled as: Multifunctional Nanomaterials: Processing, Properties, and Applications. The authors would also like to acknowledge the support provided by the National Science Foundation Engineering Research Center for Structured Organic Particle Systems (C-SOPS).

References

1. Vargas-Escobar, W. L. (2002). *Discrete modeling of heat conduction in granular media*. Ph.D. thesis, University of Pittsburgh.
2. Makse, H. A., Gland, N., Johnson, D. L., & Schwartz, L. M. (1999). Why effective medium theory fails in granular materials. *Physical Review Letters*, *83*(24), 5070–5073.
3. Makse, H. A., Gland, N., Johnson, D. L., & Schwartz, L. (2001). The apparent failure of effective medium theory in granular materials. *Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy*, *26*(1), 107–111.
4. Hertz, H. (1881). On the contact of elastic solids. *Journal für die reine und angewandte Mathematik*, *92*(156–171), 110.
5. Mindlin, R. D. (1949). Compliance of elastic bodies in contact. *Journal of Applied Mechanics*, *16*.
6. Mindlin, R. D., & Deresiewicz, H. (1953). Elastic spheres in contact under varying oblique forces. *Journal of Applied Mechanics*, *20*.
7. Zhu, H. P., Zhou, Z. Y., Yang, R. Y., & Yu, A. B. (2007). Discrete particle simulation of particulate systems: Theoretical developments. *Chemical Engineering Science*, *62*(13), 3378–3396.
8. Cundall, P. A., & Strack, O. D. L. (1979). A discrete numerical model for granular assemblies. *Geotechnique*, *29*(1), 47–65.
9. Feng, Y. T., Han, K., Li, C. F., & Owen, D. R. J. (2008). Discrete thermal element modelling of heat conduction in particle systems: Basic formulations. *Journal of Computational Physics*, *227*(10), 5072–5089.
10. Zheng, S., & Cuitino, A. M. (2002). Consolidation behavior of inhomogeneous granular beds of ductile particles using a mixed discrete-continuum approach. *Kona*, *20*, 168–177.
11. Koynov, A., Akseli, I., & Cuitiño, A. M. (2011). Modeling and simulation of compact strength due to particle bonding using a hybrid discrete-continuum approach. *International Journal of Pharmaceutics*, *418*(2), 273–285.
12. Vargas, W. L., & McCarthy, J. J. (2001). Heat conduction in granular materials. *AIChE Journal*, *47*(5), 1052–1059.
13. Batchelor, G. K., & O'Brien, R. W. (1977). Thermal or electrical conduction through a granular material. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, *355*(1682), 313–333.
14. Chan, C. K., & Tien, C. L. (1973). Conductance of packed spheres in vacuum. *Journal of Heat Transfer (United States)*, *95*(3).
15. Kaganer, M. G. (1966). Contact heat transfer in granular material under vacuum. *Journal of Engineering Physics*, *11*(1), 19–22.
16. Hadley, G. R. (1986). Thermal conductivity of packed metal powders. *International Journal of Heat and Mass Transfer*, *29*(6), 909–920.
17. Nozad, I., Carbonell, R. G., & Whitaker, S. (1985). Heat conduction in multiphase systems—II: Experimental method and results for three-phase systems. *Chemical Engineering Science*, *40*(5), 857–863.
18. Shonnard, D. R., & Whitaker, S. (1989). The effective thermal conductivity for a pointcontact porous medium: An experimental study. *International Journal of Heat and Mass Transfer*, *32*(3), 503–512.
19. Sridhar, M. R., & Yovanovich, M. M. (1996). Elastoplastic contact conductance model for isotropic conforming rough surfaces and comparison with experiments. *Transactions-American Society of Mechanical Engineers Journal of Heat Transfer*, *118*, 3–9.
20. Bahrami, M., Yovanovich, M. M., & Culham J. R., et al. (2005). A compact model for spherical rough contacts. *Transactions-American Society of Mechanical Engineers Journal of Tribology*, *127*(4), 884.
21. Fletcher, L. S. (1988). Recent developments in contact conductance heat transfer. *ASME Transactions Journal of Heat Transfer*, *110*, 1059–1070.

22. Majumdar, A., & Tien, C. L. (1991). Fractal network model for contact conductance. *Journal of Heat Transfer (Transactions of the ASME (American Society of Mechanical Engineers), Series C;(United States)*, *113*(3).
23. Majmudar, T. S., & Behringer, R. P. (2005). Contact force measurements and stress-induced anisotropy in granular materials. *Nature*, *435*(7045), 1079–1082.
24. Gonzalez, M., & Cuitiño, A. M. (2012). A nonlocal contact formulation for confined granular systems. *Journal of the Mechanics and Physics of Solids*, *60*(2), 333–350.
25. Chen, K., Cole, J., Conger, C., Draskovic, J., Lohr, M., Klein, K., et al. (2006). Granular materials: Packing grains by thermal cycling. *Nature*, *442*(7100), 257–257.
26. Vargas, W. L., & McCarthy J. J. (2007). Thermal expansion effects and heat conduction in granular materials. *Physical Review E*, *76*(4), 041301.
27. Johnson, K. L. (1987). *Contact mechanics*. Cambridge University press.
28. Lu, Z., Abdou, M., & Ying, A. (2001). 3d micromechanical modeling of packed beds. *Journal of nuclear materials*, *299*(2), 101–110.
29. Siu, W. W. M., & Lee, S. H.-K. (2004). Transient temperature computation of spheres in three-dimensional random packings. *International Journal of Heat and Mass Transfer*, *47*(5), 887–898.
30. Walton, K. (1975). The effective elastic moduli of model sediments. *Geophysical Journal of the Royal Astronomical Society*, *43*(2), 293–306.
31. Markov, K. Z. (2000). Elementary micromechanics of heterogeneous media. In *Heterogeneous media* (pp. 1–162). Springer.
32. Norris, A. N., & Johnson, D. L. (1997). Nonlinear elasticity of granular media. *Transactions-American Society of Mechanical Engineers Journal of Applied Mechanics*, *64*, 39–49.
33. Vargas, W. L., & McCarthy, J. J. (2002). Stress effects on the conductivity of particulate beds. *Chemical Engineering Science*, *57*(15), 3119–3131.
34. Siu, W. W. M., & Lee, S. H.-K. (2000). Effective conductivity computation of a packed bed using constriction resistance and contact angle effects. *International Journal of Heat and Mass Transfer*, *43*(21), 3917–3924.
35. Landau, L. D., & Lifshitz, E. M. (1959). *Theory of elasticity*. Pergamon Press.

Elastomeric Gels: A Model and First Results

Mariarita de Luca and Antonio DeSimone

Abstract An elastomeric gel is a cross–linked polymer network swollen by a solvent. Computational models of gels need to resolve the strong coupling between the diffusion of the solvent and the deformation of the elastic network. We present here a continuum mechanics model to describe the gel deformation and the coupled fluid permeation in the polymeric network, and the first results we have obtained with it. These consist of numerical simulations of two basic experiments: the free swelling deformation of a dry specimen and an indentation test performed on a swollen sample.

1 Introduction

An elastomeric gel is a cross–linked polymer network swollen by a solvent. Computational models of gels need to resolve the strong coupling between the diffusion of the solvent and the deformation of the elastic network.

In the literature there are many approaches to study the evolution of a system composed by a polymeric structure plus solvent. Hong and coworkers [9] consider the gel as a bulk soft material characterized by the free energy function proposed by Flory and Rehner in 1943 [8]. Such free energy, derived on the basis of statistical mechanics, takes into account the entropy change induced by stretching the polymer network, and the enthalpy of mixing of polymer and solvent. Hong and others also introduced a dynamic model for the diffusion of solvent molecules inside the gel based on the assumption that the solvent molecules self diffuse inside the gel and there is no macroscopic flow [9].

Following the same theory, Kang and Huang [12] first developed a finite element model to compute the equilibrium swelling deformation of a gel in contact with a

M. de Luca · A. DeSimone (✉)
SISSA, Via Bonomea 265, 34136 Trieste, Italy
e-mail: desimone@sissa.com

M. de Luca
e-mail: mariarita.deluca@gmail.com

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_4

solvent constrained to a fixed support. Subsequently, Zhang et al. [17] proposed a full three dimensional model to simulate the diffusion of the solvent inside the gel and the corresponding deformation.

Another approach is to consider the gel composed by a porous elastic material and a fluid (solvent) flowing through its pores. This model combines the elasticity of the porous structure with a mass transport model in the porous medium (Darcy Law), obtaining the Biot model [2]. In its original formulation, the Biot model considers only the pore pressure in a material element, defined as the pressure in an hypothetical reservoir which is in equilibrium with this element, as the mechanism regulating the flux of fluid between the reservoir and the material element [16]. For such reason it is suitable to describe quasi-static processes, but it is not adequate for swelling gels, then it needs to be extended in order to be able to correctly capture all the phenomena that are interesting for gel applications. Following this idea, Murad et al. [14] have extended the Biot model in the framework of hybrid mixture theory to develop a theory for swelling porous media, in which the variation of the chemical potential is the mechanism regulating the swelling and draining processes.

In this work we start from the idea proposed in [3], and we use a continuum mechanics approach to describe the gel deformation and the coupled fluid permeation in the polymeric network. We present here the first preliminary results obtained with our model, which has been implemented in the software package Comsol Multiphysics®, and used to simulate two basic experiments. These are the free swelling deformation of a dry specimen and an indentation test performed on a swollen sample.

2 Governing Equations

In the following sections we present the governing equations of the model, the balance of mass and the balance of linear momentum and we recall the kinematics needed to describe the model.

2.1 Kinematics

Let $\Omega_0 \subset \mathbb{R}^3$ be the initial placement, at time $t = 0$, of the body. Points $\mathbf{X} \in \Omega_0$ correspond to both fluid and solid particles and it is not possible to distinguish among them, hence the polymer-solvent mixture is treated as a single homogeneus continuum. The vector function $\mathbf{f} : \Omega_0 \times [0, T] \to \mathbb{R}^3$ is the motion of the body, mapping points $\mathbf{X} \in \Omega_0$ at time $t \in [0, T]$ to points $\mathbf{x} \in \Omega_t \subset \mathbb{R}^3$, where $\Omega_t = \mathbf{f}(\Omega_0, t)$. We write $\mathbf{x} = \mathbf{f}(\mathbf{X}, t)$, $\mathbf{F}(\mathbf{X}, t) = \operatorname{Grad} \mathbf{f}(\mathbf{X}, t) = \partial \mathbf{f}(\mathbf{X}, t)/\partial \mathbf{X}$ for the deformation gradient, subject to the orientation constraint $J = \det \mathbf{F} > 0$, $\mathbf{v}(\mathbf{X}, t) = \dot{\mathbf{f}}(\mathbf{X}, t) = \partial \mathbf{f}(\mathbf{X}, t)/\partial t$ the velocity, and $\mathbf{L} = \operatorname{grad} \mathbf{v} = \partial \mathbf{v}(\mathbf{f}^{-1}(\mathbf{x}, t), t)/\partial \mathbf{x} = \dot{\mathbf{F}}\mathbf{F}^{-1}$ the spatial gradient of the velocity.

As in [3], we consider a multiplicative decomposition of the deformation gradient into a “swelling” and an “elastic” part:

$$\mathbf{F} = \mathbf{F}^e\,\mathbf{F}^s, \quad \text{with}\,\mathbf{F}^s = \lambda^s\mathbf{I}, \quad \lambda^s > 0, \tag{1}$$

where the swelling part $\mathbf{F}^s$ is assumed isotropic, with λ^s the swelling stretch. We refer te reader to [5, 6] for further examples of using the multiplicative decomposition of the deformation gradient in nonlinear models of polymers and of phase-changing elastomers.

At a fixed time t^*, $\mathbf{F}^s(\mathbf{X}, t^*)$ represents the distortion of the body only due to swelling, while $\mathbf{F}^e(\mathbf{X}, t^*)$ is the subsequent stretching and rotation of the corresponding swollen network structure, representing the mechanical elastic distortion.

Then $J^e = \det\mathbf{F}^e$, $J^s = \det\mathbf{F}^s$, and $J = J^e\,J^s$. The right Cauchy–Green strain tensor is

$$\mathbf{C} = \mathbf{F}^T\,\mathbf{F} = (\mathbf{F}^e\,\mathbf{F}^s)^T\,\mathbf{F}^e\,\mathbf{F}^s = \mathbf{F}^{sT}\,\mathbf{F}^{eT}\,\mathbf{F}^e\,\mathbf{F}^s = \mathbf{F}^{sT}\,\mathbf{C}^e\,\mathbf{F}^s, \tag{2}$$

where

$$\mathbf{C}^e = \mathbf{F}^{eT}\,\mathbf{F}^e \tag{3}$$

is the elastic part.

The velocity gradient

$$\begin{aligned}\mathbf{L} = \dot{\mathbf{F}}\,\mathbf{F}^{-1} &= (\mathbf{F}^e\dot{}\mathbf{F}^s)\,(\mathbf{F}^e\,\mathbf{F}^s)^{-1} = (\dot{\mathbf{F}}^e\,\mathbf{F}^s + \mathbf{F}^e\,\dot{\mathbf{F}}^s)\,\mathbf{F}^{s-1}\,\mathbf{F}^{e-1}\\ &= \dot{\mathbf{F}}^e\,\mathbf{F}^{e-1} + \mathbf{F}^e\,\dot{\mathbf{F}}^s\,\mathbf{F}^{s-1}\,\mathbf{F}^{e-1} = \mathbf{L}^e + \mathbf{F}^e\,\mathbf{L}^s\,\mathbf{F}^{e-1},\end{aligned} \tag{4}$$

where

$$\mathbf{L}^e = \dot{\mathbf{F}}^e\,\mathbf{F}^{e-1} \tag{5}$$

$$\mathbf{L}^s = \dot{\mathbf{F}}^s\,\mathbf{F}^{s-1}. \tag{6}$$

It is convenient to compute the time derivative of the determinants

$$\dot{J} = J\,\mathbf{F}^{-T} : \dot{\mathbf{F}} = J\,\dot{\mathbf{F}}\,\mathbf{F}^{-1} : \mathbf{I} = J\,\mathrm{tr}\mathbf{L} \tag{7}$$

$$\dot{J^s} = J^s\,\mathbf{F}^{s-T} : \dot{\mathbf{F}^s} = J^s\,\dot{\mathbf{F}^s}\,\mathbf{F}^{s-1} : \mathbf{I} = J^s\,\mathrm{tr}\mathbf{L}^s, \tag{8}$$

then

$$\mathbf{L}^s = \frac{1}{3}\,\dot{J^s}\,J^{s-1}\,\mathbf{I}. \tag{9}$$

With reference to Fig. 1, $\mathbf{F}$ describes the deformation from the reference state to the current state, while $\mathbf{F}_d$ is the deformation gradient from the dry state (when there is no solvent) to the current state, $\mathbf{F}$ and $\mathbf{F}_d$ relate through $\mathbf{F}_0$ which is the spontaneous

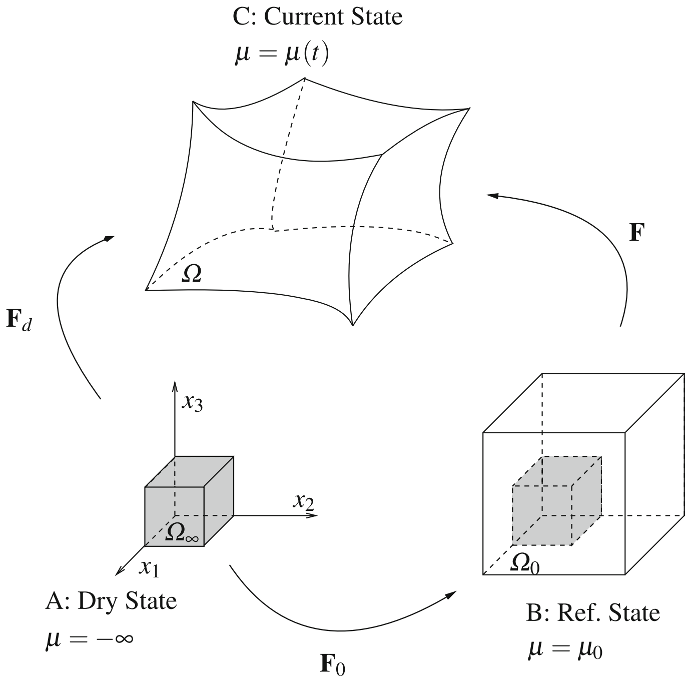


Fig. 1 Cartoon of dry, reference and current configurations

deformation that a dry body undergoes when immersed in a solvent (free swelling deformation) [13]

$$\mathbf{F}_d = \mathbf{F}\,\mathbf{F}_0, \tag{10}$$

where

$$\mathbf{F}_0 = \lambda_0\,\mathbf{I}. \tag{11}$$

The free swelling deformation is isotropic, characterized by the free swelling stretch λ_0. Then

$$\mathbf{C}_d = \mathbf{F}_d{}^T\,\mathbf{F}_d = (\mathbf{F}\,\mathbf{F}_0)^T\,\mathbf{F}\,\mathbf{F}_0 = \mathbf{F}_0{}^T\,\mathbf{F}^T\,\mathbf{F}\,\mathbf{F}_0 = \mathbf{F}_0{}^T\,\mathbf{C}\,\mathbf{F}_0. \tag{12}$$

2.2 Conservation of Mass

We call current concentration

$$c_c(\mathbf{x}, t) = \frac{n(\mathbf{x}, t)}{V}, \tag{13}$$

the number of solvent moles $n(\mathbf{x}, t)$ absorbed by the solid network per unit current volume V. The concentration field is a scalar function $c_c : \Omega_t \times [0, T] \to \mathbb{R}$. The mass of solvent in a region $R_t \subset \Omega_t$ is

$$M(t) = \int_{R_t} c_c(\mathbf{x}, t)\, dV. \tag{14}$$

The referential concentration, namely, the number of solvent moles absorbed by the solid network per unit reference volume V_0 is

$$c(\mathbf{X}, t) = \frac{n(\mathbf{X}, t)}{V_0}, \tag{15}$$

where $\mathbf{X} = \mathbf{f}^{-1}(\mathbf{x}, t)$ and $c : \Omega_0 \times [0, T] \to \mathbb{R}$.

If we consider a fixed region $R_0 \subset \Omega_0$ the conservation of mass implies

$$M(t) = \int_{R_0} c(\mathbf{X}, t)\, dV_0 = \int_{R_t} c_c(\mathbf{x}, t)\, dV, \tag{16}$$

using the volume transformation $dV = J\, dV_0$, then

$$\int_{R_0} c(\mathbf{X}, t)\, dV_0 = \int_{R_0} c_c(\mathbf{x}, t)\, J\, dV_0, \tag{17}$$

and

$$c = J\, c_c. \tag{18}$$

Similarly,

$$c_d = \frac{n}{V_d} \tag{19}$$

is the concentration referred to the dry volume of the body and

$$c_d = J_0\, J\, c_c = J_0\, c, \tag{20}$$

where $dV_0 = J_0\, dV_d$.

If n_0 is the number of solvent moles absorbed by the polymer network during the spontaneous deformation (free swelling),

$$c_0 = \frac{n_0}{V_0} \tag{21}$$

is the concentration in the free swollen reference state per unit reference volume, while

$$c_{d0} = \frac{n_0}{V_d} \tag{22}$$

is the concentration in the free swollen reference state per unit dry volume.

In our model, we impose the constraint of "molecular incompressibility". This means that the increase in volume of the body is only due to the solvent that enters the polymer network. The increase in volume associated with the spontaneous free swelling deformation described by $\mathbf{F}_0$ is

$$J_0 = 1 + \nu\, c_{d0}, \tag{23}$$

where ν is the molar volume of the solvent.

Any deformation from the reference configuration induces a change in volume described by $J = J^e J^s$, where J^s accounts for the swelling part of the deformation. Then the "molecular incompressibility" constraint reads

$$J^s = J_0^{-1} + \nu c, \quad \text{or } J^s J_0 = 1 + \nu\, J_0\, c, \tag{24}$$

because $V_0 = V_d + \nu n_0,\;\; V = V_0 + \nu(n - n_0) = V_d + \nu n$, and dividing all the terms by V_d we obtain (24).

The global balance of liquid mass over a convected region R_t with boundary measure dA in $\mathbb{R}^3$ is:

$$\frac{d}{dt}\int_{R_t} c_c(\mathbf{x}, t)\, dV = -\int_{\partial R_t} \mathbf{j}_c(\mathbf{x}, t) \cdot \mathbf{n}\, dA, \tag{25}$$

where $\mathbf{j}_c$ is the relative solvent flux through the boundary of R_t and $\mathbf{n}$ is the outward unit normal to ∂R_t.

Using the formulas for surface and volume change

$$dV = J\, dV_0, \quad \mathbf{n}\, dA = J\mathbf{F}^{-T}\mathbf{n}_0\, dA_0, \tag{26}$$

we obtain

$$\frac{d}{dt}\int_{R_0} J\, c_c(\mathbf{x}, t)\, dV_0 = -\int_{\partial R_0} \mathbf{j}_c(\mathbf{x}, t) \cdot J\, \mathbf{F}^{-T}\, \mathbf{n}_0\, dA_0, \tag{27}$$

and

$$\frac{d}{dt}\int_{R_0} J\, c_c(\mathbf{x}, t)\, dV_0 = -\int_{\partial R_0} J\mathbf{F}^{-1}\mathbf{j}_c(\mathbf{x}, t) \cdot \mathbf{n}_0\, dA_0. \tag{28}$$

Then, applying the divergence theorem and switching the derivative on the left hand side with the integral sign we have:

$$\int_{R_0} \frac{d}{dt} c(\mathbf{X}, t)\, dV_0 = -\int_{R_0} \mathrm{Div}\, \mathbf{j}(\mathbf{X}, t)\, dV_0, \tag{29}$$

where

$$\mathbf{j}(\mathbf{X}, t) = J\mathbf{F}^{-1}\mathbf{j}_c(\mathbf{x}, t) \tag{30}$$

is the Piola transformation of the relative solvent flux.

Finally, due to the arbitrariness of R_0, the local form of the balance of solvent mass in the reference region Ω_0 is

$$\frac{dc}{dt} = -\mathrm{Div}\,\mathbf{j}. \tag{31}$$

2.2.1 Polymer Fraction

We introduce the variable ϕ, the volume fraction occupied by the solid network in the current configuration, defined as

$$\phi = \frac{V_d}{V}, \tag{32}$$

then

$$\phi^{-1} = \frac{V}{V_d} = \frac{V_d + \nu n}{V_d} = 1 + \nu\,\frac{n}{V_d} = 1 + \nu c_d = 1 + \nu\,J_0\,c = J_0\,J^s. \tag{33}$$

In the reference configuration Ω_0, the corresponding polymer volume fraction is

$$\phi_0 = \frac{V_d}{V_0} = J_0{}^{-1} \tag{34}$$

and combining Eqs. (33) and (34) we have

$$\phi = J^{s^{-1}}\phi_0. \tag{35}$$

From (33) we can derive the concentration

$$c = \frac{1}{\nu\,J_0}(\phi^{-1} - 1), \tag{36}$$

and its time derivative

$$\dot{c} = -\frac{1}{\nu\,J_0}\frac{\dot{\phi}}{\phi^2} = \frac{\phi_0}{\nu}\frac{\dot{\phi}}{\phi^2} \tag{37}$$

Inserting relation (37) in Eq. (31) we have the mass balance with the polymer fraction as independent variable:

$$-\frac{\phi_0}{\nu}\frac{\dot{\phi}}{\phi^2} = -\mathrm{Div}\,\mathbf{j}. \tag{38}$$

It is useful to define the fluid volume fraction to be able to compare our model with the standard poro-elastic one

$$\phi^f = 1 - \phi = \frac{V^f}{V}, \tag{39}$$

where $V^f = V - V_d$ is the solvent volume. In the reference configuration Ω_0

$$\phi_0^f = 1 - \phi_0 = \frac{V_0^f}{V_0}, \tag{40}$$

where $V_0^f = V_0 - V_d$ is the fluid volume enclosed by the solid elastic network in the reference configuration, then thanks to Eqs. (35), (39) and (40) we have

$$\phi^f = 1 - J^{s-1}(1 - \phi_0^f), \tag{41}$$

and

$$\Phi^f = J^s - 1 + \phi_0^f, \tag{42}$$

where $\Phi^f = J^s \phi^f = V^f / V_0$ is the fluid fraction referred to the reference volume. We remark that in classical poroelastic models, ϕ^f is the "porosity" and it varies with the deformation following the law (41) as described in [7], and

$$e = \frac{\phi^f}{1 - \phi^f} = \frac{\phi^f}{\phi} \tag{43}$$

is the "void ratio".

2.3 Balance of Linear Momentum

The balance of linear and angular momentum are unchanged with respect to the usual expressions. Time scales associated with fluid diffusion are ususally considerably larger then those associated with wave propagation, hence inertial effects can be neglected. The body is subjected to volume forces **b** and surface forces **t**, where, denoting by **n** the outward unit normal to $\partial\Omega_t$, we have

$$\mathbf{t}(\mathbf{n}) = \mathbf{T}\,\mathbf{n}, \tag{44}$$

with **T** the Cauchy stress tensor.

The balance of linear momentum in a convected region $R_t \subset \Omega_t$ reads

$$\int_{\partial R_t} \mathbf{t}(\mathbf{n})\,\mathrm{d}A + \int_{R_t} \mathbf{b}\,\mathrm{d}V = 0, \tag{45}$$

or, using (44),

$$\int_{\partial R_t} \mathbf{T}\,\mathbf{n}\,\mathrm{d}A + \int_{R_t} \mathbf{b}\,\mathrm{d}V = 0. \tag{46}$$

By using the area and volume transformation formulas (26) we obtain the balance of linear momentum in a fixed region $R_0 \subset \Omega_0$

$$\int_{\partial R_0} J\,\mathbf{T}\,\mathbf{F}^{-T}\,\mathbf{n}_0\,dA_0 + \int_{R_0} J\,\mathbf{b}\,dV_0 = 0. \tag{47}$$

Thanks to the divergence theorem this becomes

$$\int_{R_0} \mathrm{Div}(J\,\mathbf{T}\,\mathbf{F}^{-T})\,dV_0 + \int_{R_0} J\,\mathbf{b}\,\mathrm{d}V_0 = 0, \tag{48}$$

where

$$\mathbf{P} := J\,\mathbf{T}\,\mathbf{F}^{-T} \tag{49}$$

is the First Piola—Kirchhoff stress tensor and $\mathbf{b}_0 = J\,\mathbf{b}$ are the reference body forces, so that

$$\int_{R_0} (\mathrm{Div}\mathbf{P} + \mathbf{b}_0)\;dV_0 = 0. \tag{50}$$

Due to arbitrariness of R_0, we obtain the local form of the balance of linear momentum in the reference configuration, which reads

$$\mathrm{Div}\mathbf{P} + \mathbf{b}_0 = 0 \tag{51}$$

The balance of angular momentum implies

$$\mathbf{T} = \mathbf{T}^T, \tag{52}$$

and, in terms of the first Piola-Kirchhoff stress tensor,

$$\mathbf{P}\,\mathbf{F}^T = \mathbf{F}\,\mathbf{P}^T. \tag{53}$$

We can also introduce the second Piola-Kirchhoff stress tensor

$$\mathbf{S} := \mathbf{F}^{-1}\,\mathbf{P} = J\,\mathbf{F}^{-1}\,\mathbf{T}\,\mathbf{F}^{-T}, \tag{54}$$

which is a symmetric second order tensor, i.e. $\mathbf{S} = \mathbf{S}^T$.

3 Thermodynamics

In this section we derive a consistent thermodynamic theory which involves chemical, thermal and mechanical processes. To do so, we recall the following quantities defined per unit reference volume: ε_0 is the internal energy, η_0 is the entropy density, Q_0 is the external heat supply, and ρ_0 is the mass density. Moreover, we consider the heat flux per unit reference area $\mathbf{q}_0$, the chemical potential μ, and the absolute temperature θ.

We note that the mass density satisfies $\rho = \rho^s \phi + \rho^f \phi^f$, where ρ^s and ρ^f are the current mass densities (mass densities per unit current volume) corresponding respectively to the solid and fluid part.

Given a region $R_0 \subset \Omega_0$ with outward unit normal $\mathbf{n}_0$ we can define the corresponding global energy terms

$$\mathscr{I}(R_0) = \int_{R_0} \mathbf{P} : \dot{\mathbf{F}}\, dV_0 : \quad \text{power of internal forces,} \tag{55}$$

$$\mathscr{W}(R_0) = \int_{\partial R_0} \mathbf{P}\,\mathbf{n}_0 \cdot \dot{\mathbf{x}}\, dA_0 + \int_{R_0} \mathbf{b}_0 \cdot \dot{\mathbf{x}}\, dV_0 : \quad \text{power of external forces,} \tag{56}$$

$$\mathscr{E}(R_0) = \int_{R_0} \varepsilon_0\, dV_0 : \quad \text{internal energy,} \tag{57}$$

$$\mathscr{K}(R_0) = \int_{R_0} \frac{1}{2}\rho_0\, |\dot{\mathbf{x}}|^2\, dV_0 : \quad \text{kinetic energy,} \tag{58}$$

$$\mathscr{Q}(R_0) = -\int_{\partial R_0} \mathbf{q}_0 \cdot \mathbf{n}_0\, dA_0 + \int_{R_0} Q_0\, dV_0 : \quad \text{heat flow,} \tag{59}$$

$$\mathscr{T}(R_0) = -\int_{\partial R_0} \mu\, \mathbf{j} \cdot \mathbf{n}_0\, dA_0 : \quad \text{energy flow due to fluid diffusion,} \tag{60}$$

$$\mathscr{S}(R_0) = \int_{R_0} \eta_0\, dV_0 : \quad \text{entropy,} \tag{61}$$

$$\mathscr{J}(R_0) = -\int_{\partial R_0} \frac{1}{\theta}\, \mathbf{q}_0 \cdot \mathbf{n}_0\, dA_0 + \int_{R_0} \frac{Q_0}{\theta}\, dV_0 : \quad \text{entropy flow.} \tag{62}$$

The conventional power balance is

$$\mathscr{W}(R_0) = \dot{\mathscr{K}}(R_0) + \mathscr{I}(R_0), \tag{63}$$

and the first law of thermodynamics reads

$$\dot{\mathscr{E}}(R_0) + \dot{\mathscr{K}}(R_0) = \mathscr{W}(R_0) + \mathscr{Q}(R_0) + \mathscr{T}(R_0) \tag{64}$$

$$= \dot{\mathscr{K}}(R_0) + \mathscr{I}(R_0) + \mathscr{Q}(R_0) + \mathscr{T}(R_0). \tag{65}$$

Thus,

$$\dot{\mathscr{E}}(R_0) = \mathscr{I}(R_0) + \mathscr{Q}(R_0) + \mathscr{T}(R_0) \tag{66}$$

By inserting Eqs. (57), (55), (59), and (60) in Eq. (66) we have

$$\int_{R_0} \dot{\varepsilon}_0 \, dV_0 = \int_{R_0} \mathbf{P} : \dot{\mathbf{F}} \, dV_0 - \int_{\partial R_0} \mathbf{q}_0 \cdot \mathbf{n}_0 \, dA_0 + \int_{R_0} Q_0 \, dV_0 - \int_{\partial R_0} \mu \, \mathbf{j} \cdot \mathbf{n}_0 \, dA_0. \tag{67}$$

Then, by applying the divergence theorem

$$\int_{R_0} \dot{\varepsilon}_0 \, dV_0 = \int_{R_0} \big(\mathbf{P} : \dot{\mathbf{F}} - \mathrm{Div}\, \mathbf{q}_0 + Q_0 - \mathrm{Div}\,(\mu \, \mathbf{j}) \big) \, dV_0, \tag{68}$$

and finally, since the equation holds for each subset R_0,

$$\dot{\varepsilon}_0 = \mathbf{P} : \dot{\mathbf{F}} - \mathrm{Div}\, \mathbf{q}_0 + Q_0 - \mu \, \mathrm{Div}\, \mathbf{j} - \nabla \mu \cdot \mathbf{j}. \tag{69}$$

The second law of thermodynamics states that the net entropy production should always be non negative:

$$\dot{\mathscr{S}}(R_0) - \mathscr{J}(R_0) \geq 0. \tag{70}$$

By inserting Eqs. (61) and (62) in Eq. (70) we have

$$\int_{R_0} \dot{\eta}_0 \, dV_0 \geq - \int_{\partial R_0} \frac{1}{\theta} \mathbf{q}_0 \cdot \mathbf{n}_0 \, dA_0 + \int_{R_0} \frac{Q_0}{\theta} \, dV_0. \tag{71}$$

Then, thanks to the divergence theorem

$$\int_{R_0} \dot{\eta}_0 \, dV_0 \geq - \int_{R_0} \left(\mathrm{Div} \left(\frac{1}{\theta} \mathbf{q}_0 \right) + \frac{Q_0}{\theta} \right) dV_0 \tag{72}$$

and finally, since each integral holds $\forall R_0 \subset \Omega_0$

$$\dot{\eta}_0 \geq \frac{1}{\theta} \left(Q_0 - \mathrm{Div}\, \mathbf{q}_0 + \frac{1}{\theta} \mathbf{q}_0 \cdot \nabla \theta \right). \tag{73}$$

From Eq. (69) we can derive $Q_0 - \mathrm{Div}\, \mathbf{q}_0$ and inserting it in Eq. (73) to obtain

$$\dot{\eta}_0 \geq \frac{1}{\theta} \left(\dot{\varepsilon}_0 - \mathbf{P} : \dot{\mathbf{F}} + \mu \, \mathrm{Div}\, \mathbf{j} + \nabla \mu \cdot \mathbf{j} + \frac{1}{\theta} \mathbf{q}_0 \cdot \nabla \theta \right). \tag{74}$$

Then, multiplying by θ each term and bringing all the terms on the right we have

$$\dot{\varepsilon}_0 - \theta \, \dot{\eta}_0 - \mathbf{P} : \dot{\mathbf{F}} + \mu \, \mathrm{Div}\, \mathbf{j} + \nabla \mu \cdot \mathbf{j} + \frac{1}{\theta} \mathbf{q}_0 \cdot \nabla \theta \leq 0. \tag{75}$$

We define the Helmotz free energy ψ_0 as

$$\psi_0 = \varepsilon_0 - \theta \, \eta_0, \tag{76}$$

then its time derivative is

$$\dot{\psi}_0 = \dot{\varepsilon}_0 - \dot{\theta}\,\eta_0 - \theta\,\dot{\eta}_0. \tag{77}$$

We can replace $\dot{\varepsilon}_0 - \theta\,\dot{\eta}_0 = \dot{\psi}_0 + \dot{\theta}\,\eta_0$ in Eq. (75) to have

$$\dot{\psi}_0 + \dot{\theta}\eta_0 - \mathbf{P} : \dot{\mathbf{F}} + \mu\,\mathrm{Div}\,\mathbf{j} + \nabla\mu \cdot \mathbf{j} + \frac{1}{\theta}\mathbf{q}_0 \cdot \nabla\theta \leq 0, \tag{78}$$

that in isothermal condition reduces to

$$\dot{\psi}_0 - \mathbf{P} : \dot{\mathbf{F}} + \mu\,\mathrm{Div}\,\mathbf{j} + \nabla\mu \cdot \mathbf{j} \leq 0. \tag{79}$$

Thanks to the replacement of the mass balance (31) in Eq. (79) we finally have the thermodynamic constraint

$$\dot{\psi}_0 - \mathbf{P} : \dot{\mathbf{F}} - \mu\,\dot{c} + \nabla\mu \cdot \mathbf{j} \leq 0. \tag{80}$$

3.1 Stress Power

In this section we compute the $\mathbf{P} : \dot{\mathbf{F}}$ term in the thermodynamic inequality (80). Given the kinematic decomposition of the deformation gradient $\mathbf{F}$ in Eq. (1), the time derivative of $\mathbf{F}$ reads

$$\dot{\mathbf{F}} = \dot{\mathbf{F}}^e\,\mathbf{F}^s + \mathbf{F}^e\,\dot{\mathbf{F}}^s, \tag{81}$$

then

$$\begin{aligned}
\mathbf{P} : \dot{\mathbf{F}} &= \mathbf{P} : \dot{\mathbf{F}}^e\,\mathbf{F}^s + \mathbf{P} : \mathbf{F}^e\,\dot{\mathbf{F}}^s && (82)\\
&= \mathbf{P}\mathbf{F}^{sT} : \dot{\mathbf{F}}^e + \mathbf{F}^{eT}\mathbf{P} : \dot{\mathbf{F}}^s && (83)\\
&= J\,\mathbf{T}\,\mathbf{F}^{-T}\,\mathbf{F}^{sT} : \dot{\mathbf{F}}^e + J\,\mathbf{F}^{eT}\,\mathbf{T}\,\mathbf{F}^{-T} : \dot{\mathbf{F}}^s && (84)\\
&= J\,\mathbf{T}\,(\mathbf{F}^e\,\mathbf{F}^s)^{-T}\,\mathbf{F}^{sT} : \dot{\mathbf{F}}^e + J\,\mathbf{F}^{eT}\,\mathbf{T}\,(\mathbf{F}^e\,\mathbf{F}^s)^{-T} : \dot{\mathbf{F}}^s && (85)\\
&= J\,\mathbf{T}\,\mathbf{F}^{e-T}\,\mathbf{F}^{s-T}\,\mathbf{F}^{sT} : \dot{\mathbf{F}}^e + J\,\mathbf{F}^{eT}\,\mathbf{T}\,\mathbf{F}^{e-T}\,\mathbf{F}^{s-T} : \dot{\mathbf{F}}^s && (86)\\
&= J\,\mathbf{T}\,\mathbf{F}^{e-T} : \dot{\mathbf{F}}^e + J\,\mathbf{F}^{eT}\,\mathbf{T}\,\mathbf{F}^{e-T} : \dot{\mathbf{F}}^s\,\mathbf{F}^{s-1} && (87)\\
&= J\,\mathbf{T}\,\mathbf{F}^{e-T} : \dot{\mathbf{F}}^e + J\,\mathbf{F}^{eT}\,\mathbf{T}\,\mathbf{F}^{e-T} : \mathbf{L}^s && (88)\\
&= \mathbf{P}^e : \dot{\mathbf{F}}^e + \mathbf{M}^e : \mathbf{L}^s, && (89)
\end{aligned}$$

in which we have used relations (49) and (6), and we have defined two new measures of stress

$$\begin{aligned}
\mathbf{P}^e &= J\,\mathbf{T}\,\mathbf{F}^{e-T} && (90)\\
\mathbf{M}^e &= J\,\mathbf{F}^{eT}\,\mathbf{T}\,\mathbf{F}^{e-T}. && (91)
\end{aligned}$$

Recalling (9), and replacing it in Eq. (89) we obtain

$$\mathbf{P}:\dot{\mathbf{F}} = \mathbf{P}^e:\dot{\mathbf{F}}^e + \mathbf{M}^e:\frac{1}{3}J^{s-1}\dot{J}^s\,\mathbf{I} \tag{92}$$

$$= \mathbf{P}^e:\dot{\mathbf{F}}^e + \frac{1}{3}J^{s-1}\dot{J}^s\,\mathrm{tr}\mathbf{M}^e \tag{93}$$

$$= \mathbf{P}^e:\dot{\mathbf{F}}^e - \bar{p}\,\dot{J}^s, \tag{94}$$

where $\bar{p}$ is the mean pressure

$$\bar{p} = -\frac{1}{3}J^{s-1}\,\mathrm{tr}\mathbf{M}^e. \tag{95}$$

We define another symmetric measure of stress $\mathbf{S}^e$ as

$$\mathbf{S}^e = J\,\mathbf{F}^{e-1}\,\mathbf{T}\,\mathbf{F}^{e-T}, \tag{96}$$

and we observe that

$$\mathbf{P}^e = \mathbf{F}^e\,\mathbf{S}^e, \tag{97}$$

Thanks to Eqs. (91) and (96) we have

$$\bar{p} = -\frac{1}{3}J^e\,\mathrm{tr}\mathbf{T} = -\frac{1}{3}J^{s-1}\,\mathbf{S}^e:\mathbf{C}^e. \tag{98}$$

Then, recalling relation (3) we find

$$\mathbf{S}^e:\dot{\mathbf{C}}^e = \mathbf{S}^e:(\dot{\mathbf{F}}^{e^T}\,\mathbf{F}^e + \mathbf{F}^{eT}\,\dot{\mathbf{F}}^e) \tag{99}$$

$$= \mathbf{S}^e:\dot{\mathbf{F}}^{e^T}\,\mathbf{F}^e + \mathbf{S}^e:\mathbf{F}^{eT}\,\dot{\mathbf{F}}^e \tag{100}$$

$$= \mathbf{S}^e\,\mathbf{F}^{eT}:\dot{\mathbf{F}}^{e^T} + \mathbf{F}^e\,\mathbf{S}^e:\dot{\mathbf{F}}^e \tag{101}$$

$$= \mathbf{F}^e\,\mathbf{S}^e:\dot{\mathbf{F}}^e + \mathbf{F}^e\,\mathbf{S}^e:\dot{\mathbf{F}}^e \tag{102}$$

$$= 2\,\mathbf{P}^e\;:\dot{\mathbf{F}}^e. \tag{103}$$

3.2 Elastic Incompressibility Constraint

We introduce in the model the incompressibility constraint for the elastic polymer network by imposing

$$J^e = 1. \tag{104}$$

Then its time derivative is

$$\dot{J}^e = J^e\,\dot{\mathbf{F}}^e\,\mathbf{F}^{e-1} : \mathbf{I} = J^e\,\dot{\mathbf{F}}^e : \mathbf{F}^{e-T} = 0, \tag{105}$$

that implies orthogonality between $\mathbf{F}^{e-T}$ and $\dot{\mathbf{F}}^e$. It follows that the introduction of the incompressibility constraint in the form (105) does not have any effect if added to Eq. (94) and, we can write

$$\mathbf{P} : \dot{\mathbf{F}} = \mathbf{P}^e : \dot{\mathbf{F}}^e - \bar{p}\,\dot{J}^s + p\,\mathbf{F}^{e-T} : \dot{\mathbf{F}}^e, \tag{106}$$

where p is a Lagrange multiplier associated with the imposed constraint.

Using Eq. (3), we can compute

$$\begin{aligned}
\mathbf{F}^{e-T} : \dot{\mathbf{F}}^e &= \mathbf{F}^{e-T} : \left(\mathbf{F}^{e-T}\,\dot{\mathbf{C}}^e - \mathbf{F}^{e-T}\,\dot{\mathbf{F}}^{e^T}\,\mathbf{F}^e\right) && (107)\\
&= \mathbf{F}^{e-T} : \mathbf{F}^{e-T}\,\dot{\mathbf{C}}^e - \mathbf{F}^{e-T} : \mathbf{F}^{e-T}\,\dot{\mathbf{F}}^{e^T}\,\mathbf{F}^e && (108)\\
&= \mathbf{F}^{e-1}\,\mathbf{F}^{e-T} : \dot{\mathbf{C}}^e - \mathbf{F}^{e-1}\,\mathbf{F}^{e-T}\,\mathbf{F}^{eT} : \dot{\mathbf{F}}^{e^T} && (109)\\
&= \mathbf{C}^{e-1} : \dot{\mathbf{C}}^e - \mathbf{F}^{e-T} : \dot{\mathbf{F}}^e, && (110)
\end{aligned}$$

then

$$\mathbf{C}^{e-1} : \dot{\mathbf{C}}^e = 2\,\mathbf{F}^{e-T} : \dot{\mathbf{F}}^e. \tag{111}$$

Introducing Eqs. (103) and (111) in Eq. (106) we have

$$\mathbf{P} : \dot{\mathbf{F}} = \frac{1}{2}\left(\mathbf{S}^e + p\,\mathbf{C}^{e-1}\right) : \dot{\mathbf{C}}^e - \bar{p}\,\dot{J}^s. \tag{112}$$

To conclude our computation we observe that

$$\dot{J}^s = \nu\,\dot{c} \tag{113}$$

thank to Eqs. (24), (112) becomes

$$\mathbf{P} : \dot{\mathbf{F}} = \frac{1}{2}\left(\mathbf{S}^e + p\,\mathbf{C}^{e-1}\right) : \dot{\mathbf{C}}^e - \nu\,\bar{p}\,\dot{c}. \tag{114}$$

3.3 Dissipation Inequality

The introduction of Eq. (114) in the free energy inequality (80) gives:

$$\dot{\psi}_0 - \frac{1}{2}\left(\mathbf{S}^e + p\,\mathbf{C}^{e-1}\right) : \dot{\mathbf{C}}^e - (\mu - \nu\,\bar{p})\,\dot{c} + \nabla\mu \cdot \mathbf{j} \leq 0. \tag{115}$$

We define

$$\mathbf{S}^e_{\text{act}} = \mathbf{S}^e + p\,\mathbf{C}^{e-1}, \tag{116}$$

$$\mu_{\text{act}} = \mu - \nu\,\bar{p}, \tag{117}$$

where $\mathbf{S}^e_{\text{act}}$ is called “active stress”, and μ_{act} is called “active chemical potential”. Introducing Eqs. (116) and (117) in Eq. (115), the free energy inequality becomes

$$\dot{\psi}_0 - \frac{1}{2}\,\mathbf{S}^e_{\text{act}} : \dot{\mathbf{C}}^e - \mu_{\text{act}}\,\dot{c} + \nabla\mu \cdot \mathbf{j} \leq 0 \tag{118}$$

The independent variables of our model are the elastic deformation represented by the right Cauchy—Green strain tensor $\mathbf{C}^e$, and the concentration c of the solvent. Hence we write

$$\psi_0 = \hat{\psi}_0(\mathbf{C}^e, c), \tag{119}$$

$$\mathbf{S}^e_{\text{act}} = \hat{\mathbf{S}}^e_{\text{act}}(\mathbf{C}^e, c), \tag{120}$$

$$\mu_{\text{act}} = \hat{\mu}_{\text{act}}(\mathbf{C}^e, c), \tag{121}$$

$$\mathbf{j} = \hat{\mathbf{j}}(\mathbf{C}^e, c, \nabla\mu), \tag{122}$$

with $\mathbf{C}^e$ constrained to satisfy $\sqrt{\det \mathbf{C}^e} = J^e = 1$.

3.4 Thermodynamic Restrictions

In order to determine the thermodynamic restriction imposed by the free energy inequality (118) we compute the time derivative of the free energy which, thanks to Eq. (119), reads

$$\dot{\psi}_0 = \frac{\partial\hat{\psi}_0}{\partial\mathbf{C}^e} : \dot{\mathbf{C}}^e + \frac{\partial\hat{\psi}_0}{\partial c}\,\dot{c}. \tag{123}$$

Substituting in Eq. (118) gives

$$\left(\frac{\partial\hat{\psi}_0}{\partial\mathbf{C}^e} - \frac{1}{2}\,\mathbf{S}^e_{\text{act}}\right) : \dot{\mathbf{C}}^e + \left(\frac{\partial\hat{\psi}_0}{\partial c} - \mu_{\text{act}}\right)\dot{c} + \hat{\mathbf{j}}(\mathbf{C}^e, c, \nabla\mu) \cdot \nabla\,\mu \leq 0. \tag{124}$$

Inequality (124) must hold for all values of $\mathbf{C}^e$, c, and $\nabla\,\mu$. Since $\mathbf{C}^e$ and c appear linearly, the only possibility is that their coefficients vanish. Then inequality (124) implies

$$\mathbf{S}^e_{\text{act}} = 2\,\frac{\partial\hat{\psi}_0}{\partial\mathbf{C}^e}, \tag{125}$$

$$\mu_{\text{act}} = \frac{\partial\hat{\psi}_0}{\partial c}, \tag{126}$$

and the dissipation inequality reduces to

$$\hat{\mathbf{j}}(\mathbf{C}^e, c, \nabla\mu) \cdot \nabla\,\mu \leq 0. \tag{127}$$

Recalling definitions (116) and (117), we finally have the expressions for the stress tensor and the chemical potential:

$$\mathbf{S}^e = 2\,\frac{\partial\hat{\psi}_0}{\partial\mathbf{C}^e} - p\,\mathbf{C}^{e-1}, \tag{128}$$

$$\mu = \frac{\partial\hat{\psi}_0}{\partial c} + \nu\,\bar{p}. \tag{129}$$

Combining Eqs. (128) and (98) we obtain

$$p = J^s\,\bar{p} - \frac{2}{3}\,\frac{\partial\hat{\psi}_0}{\partial\mathbf{C}^e} : \mathbf{C}^e, \tag{130}$$

and the expression of the stress tensor $\mathbf{S}^e$ depending on the mean (pore) pressure is:

$$\mathbf{S}^e = \frac{8}{3}\,\frac{\partial\hat{\psi}_0}{\partial\mathbf{C}^e} - J^s\,\bar{p}\,\mathbf{C}^{e-1}. \tag{131}$$

4 Constitutive Theory

The constitutive theory and the thermodynamic restriction that we propose are based on ψ_0, the free energy density defined in the reference, free swollen, configuration. In the literature the free energy density for polymeric swelling material is usually defined per unit volume in the dry configuration, ψ_d, and is composed by a term due to the mixing of the polymer and the solvent, $\psi_{d\,\mathrm{mix}}$, and a term that accounts for the energy change due to the stretching of the polymer molecules, $\psi_{d\,\mathrm{mech}}$:

$$\psi_d(\mathbf{C}_d, c_d) = \mu^0\,c_d + \psi_{d\,\mathrm{mix}}(\mathbf{C}_d, c_d) + \psi_{d\,\mathrm{mech}}(\mathbf{C}_d, c_d), \tag{132}$$

where μ^0 is the value of the chemical potential of pure solvent, and

$$\psi_{d\,\mathrm{mix}}(\mathbf{C}_d, c_d) = R\,\theta\,c_d\left(\log\left(\frac{\nu\,c_d}{1+\nu\,c_d}\right) + \chi\,\left(\frac{1}{1+\nu\,c_d}\right)\right), \tag{133}$$

$$\psi_{d\,\mathrm{mech}}(\mathbf{C}_d, c_d) = \frac{G}{2}\,(\mathrm{tr}\mathbf{C}_d - 3). \tag{134}$$

In Eq. (133), $R = k_B N_A$ is the ideal gas constant, that is equal to the Boltzmann's constant k_B times the Avogadro's number N_A, and θ is the absolute temperature, while G in Eq. (134) is the shear modulus.

4.1 Free Energy Density per Unit Reference Volume

Recalling Eqs. (12) and (20) we obtain the free energy density per unit reference volume as

$$\psi_0(\mathbf{C}, c) = \frac{1}{J_0}\, \psi_d(\mathbf{F}_0{}^T\, \mathbf{C}\, \mathbf{F}_0, J_0\, c), \tag{135}$$

where

$$\psi_{0\text{mix}}(\mathbf{C}, c) = R\,\theta\, c\left(\log\left(\frac{\nu\, J_0\, c}{1+\nu\, J_0\, c}\right) + \chi\, \left(\frac{1}{1+\nu\, J_0\, c}\right)\right), \tag{136}$$

$$\psi_{0\text{mech}}(\mathbf{C}, c) = \frac{G}{2\, J_0}\, (\text{tr}\left(\mathbf{F}_0{}^T\, \mathbf{C}\, \mathbf{F}_0\right) - 3). \tag{137}$$

In the constitutive theory we chose as independent variables the concentration c and the elastic deformation $\mathbf{C}^e$, and since there is no dependence on the elastic deformation

$$\psi_{0\text{mix}}(\mathbf{C}^e, c) = R\,\theta\, c\left(\log\left(\frac{\nu\, J_0\, c}{1+\nu\, J_0\, c}\right) + \chi\, \left(\frac{1}{1+\nu\, J_0\, c}\right)\right). \tag{138}$$

To compute the mechanical part, thanks to Eq. (2), we have

$$\text{tr}\left(\mathbf{F}^T{}_0\, \mathbf{C}\, \mathbf{F}_0\right) = \text{tr}\left(\mathbf{F}_0{}^T\, \mathbf{F}^{sT}\, \mathbf{C}^e\, \mathbf{F}^s\, \mathbf{F}_0\right), \tag{139}$$

where

$$\mathbf{F}^s = \lambda_s\, \mathbf{I} = (J^s)^{1/3}\, \mathbf{I} = (J_0)^{-1/3}\, (1+\nu\, J_0\, c)^{1/3}\, \mathbf{I}, \tag{140}$$

then

$$\psi_{0\text{mech}}(\mathbf{C}^e, c) = \frac{G}{2\, J_0}\, ((J_0)^{-2/3}\, (1+\nu\, J_0\, c)^{2/3}\text{tr}\left(\mathbf{F}_0{}^T\, \mathbf{C}^e\, \mathbf{F}_0\right) - 3). \tag{141}$$

Having in mind future applications to biological tissues and intervertebral discs, we want to also allow the presence of fibres in the gel, and we consider two families of fibres that are stress-free in the swollen configuration. The energy density that accounts for the fibre contribution is defined per unit reference volume, and it vanishes when there is no elastic deformation $\mathbf{C}^e = \mathbf{I}$:

$$\psi_{0\text{fibre}}(\mathbf{C}^e, c) = \Sigma_{A_\alpha} \frac{k_1}{2\, k_2}\left[\exp\left(k_2\, (\mathbf{C}^e : \mathbf{A}_\alpha - 1)^2\right) - 1\right], \tag{142}$$

where $\alpha = 1, 2$ identifies the first and the second family of fibres.

$$\mathbf{A}_1 = \mathbf{a}_1 \otimes \mathbf{a}_1, \text{ and } \mathbf{A}_2 = \mathbf{a}_2 \otimes \mathbf{a}_2, \tag{143}$$

are the structural tensor that account for the material anisotropy and $\mathbf{a}_1$ and $\mathbf{a}_2$ are the material unit vectors defining the direction of the two family of fibres.

To conclude, accounting for all contributions, the energy density for our material reads

$$\psi_0(\mathbf{C}^e, c) = \mu^0\, c + \psi_{0\text{mix}}(\mathbf{C}^e, c) + \psi_{0\text{mech}}(\mathbf{C}^e, c) + \psi_{0\text{fibre}}(\mathbf{C}^e, c). \tag{144}$$

4.2 Partial Derivatives of the Free Energy ψ_0

In this section we compute the partial derivatives of the energy density ψ_0 that are necessary to compute the stress and the chemical potential. We start with

$$\frac{\partial \psi_{0\text{mix}}(\mathbf{C}^e, c)}{\partial \mathbf{C}^e} = 0, \tag{145}$$

$$\frac{\partial \psi_{0\text{mech}}(\mathbf{C}^e, c)}{\partial \mathbf{C}^e} = \frac{G}{2 J_0} (J_0)^{-2/3} (1 + \nu J_0\, c)^{2/3}\, \mathbf{F}_0\, \mathbf{F}_0{}^T, \tag{146}$$

$$\begin{aligned}\frac{\partial \psi_{0\text{fibre}}(\mathbf{C}^e, c)}{\partial \mathbf{C}^e} =& k_1\, (I^e_{A_1} - 1) \exp\left[k_2 \left(I^e_{A_1} - 1\right)^2\right] \mathbf{A}_1 + \\ & k_1\, (I^e_{A_2} - 1) \exp\left[k_2 \left(I^e_{A_2} - 1\right)^2\right] \mathbf{A}_2,\end{aligned} \tag{147}$$

where

$$I^e_{A_\alpha} = \mathbf{C}^e : \mathbf{A}_\alpha, \quad \text{with} \quad \alpha = 1, 2. \tag{148}$$

Then

$$\frac{\partial \psi_0(\mathbf{C}^e, c)}{\partial \mathbf{C}^e} = \frac{\partial \psi_{0\text{mix}}(\mathbf{C}^e, c)}{\partial \mathbf{C}^e} + \frac{\partial \psi_{0\text{mech}}(\mathbf{C}^e, c)}{\partial \mathbf{C}^e} + \frac{\partial \psi_{0\text{fibre}}(\mathbf{C}^e, c)}{\partial \mathbf{C}^e} \tag{149}$$

We compute the partial derivatives of ψ_0 with respect to the concentration c:

$$\frac{\partial \psi_{0\text{mix}}(\mathbf{C}^e, c)}{\partial c} = R\,\theta \left[\log\left(\frac{\nu J_0\, c}{1 + \nu J_0\, c}\right) + \frac{1}{1 + \nu J_0\, c} + \frac{\chi}{(1 + \nu J_0\, c)^2}\right], \tag{150}$$

$$\frac{\partial \psi_{0\text{mech}}(\mathbf{C}^e, c)}{\partial c} = \frac{1}{3} \frac{\nu G}{J_0} \left(J_0{}^{1/3} (1 + \nu J_0\, c)^{-1/3} \operatorname{tr}\left(\mathbf{F}_0{}^T\, \mathbf{C}^e\, \mathbf{F}_0\right)\right), \tag{151}$$

and

$$\frac{\partial \psi_{0\text{fibre}}(\mathbf{C}^e, c)}{\partial c} = 0, \tag{152}$$

then

$$\frac{\partial \psi_0(\mathbf{C}^e, c)}{\partial \mathbf{C}^e} = \mu^0 + \frac{\partial \psi_{0\text{mix}}(\mathbf{C}^e, c)}{\partial c} + \frac{\partial \psi_{0\text{mech}}(\mathbf{C}^e, c)}{\partial c} + \frac{\partial \psi_{0\text{fibre}}(\mathbf{C}^e, c)}{\partial c}. \tag{153}$$

4.3 Stress Tensor

The insertion of Eq. (149) in Eq. (128) finally gives the expression of the elastic part of the second Piola-Kirchhoff stress tensor

$$\begin{aligned}\mathbf{S}^e(\mathbf{C}^e, c) = & \frac{G}{J_0} J^{s\,2/3}\, \mathbf{F_0}\, \mathbf{F_0}^T - p\, \mathbf{C}^{e-1} \\ & + 2\, k_1\, (I^e_{A_1} - 1) \exp\left[k_2 \left(I^e_{A_1} - 1\right)^2\right] \mathbf{A}_1 \\ & + 2\, k_1\, (I^e_{A_2} - 1) \exp\left[k_2 \left(I^e_{A_2} - 1\right)^2\right] \mathbf{A}_2, \end{aligned} \tag{154}$$

where

$$J^s = J_0{}^{-1}\ (1 + \nu\, J_0\, c)\,. \tag{155}$$

Moreover, if we want to express the stress tensor $\mathbf{S}^e$ as function of the deformation and the polymer fraction, we have:

$$\begin{aligned}\mathbf{S}^e(\mathbf{C}^e, \phi) = & \frac{G}{J_0}\, \phi^{-2/3}\, \phi_0{}^{2/3}\, \mathbf{F_0}\, \mathbf{F_0}^T - p\, \mathbf{C}^{e-1} \\ & + 2\, k_1\, (I^e_{A_1} - 1) \exp\left[k_2 \left(I^e_{A_1} - 1\right)^2\right] \mathbf{A}_1 \\ & + 2\, k_1\, (I^e_{A_2} - 1) \exp\left[k_2 \left(I^e_{A_2} - 1\right)^2\right] \mathbf{A}_2, \end{aligned} \tag{156}$$

since

$$J^s = \phi^{-1}\, \phi_0. \tag{157}$$

Using Eqs. (96) and (54) we recover the expression of the second Piola-Kirchhoff stress tensor

$$\mathbf{S} = \mathbf{F}^{s-1}\, \mathbf{S}^e\, \mathbf{F}^{s-T} = J^{s-2/3}\, \mathbf{S}^e = \phi^{2/3}\, \phi_0{}^{-2/3}\, \mathbf{S}^e, \tag{158}$$

and, thanks to Eq. (156) we have

$$\begin{aligned}\mathbf{S}(\mathbf{C}^e, \phi) = & \frac{G}{J_0}\, \mathbf{F_0}\, \mathbf{F_0}^T - p\, \mathbf{C}^{-1} \\ & + 2\, \phi^{2/3}\, \phi_0{}^{-2/3}\, k_1\, (I^e_{A_1} - 1) \exp\left[k_2 \left(I^e_{A_1} - 1\right)^2\right] \mathbf{A}_1 \\ & + 2\, \phi^{2/3}\, \phi_0{}^{-2/3}\, k_1\, (I^e_{A_2} - 1) \exp\left[k_2 \left(I^e_{A_2} - 1\right)^2\right] \mathbf{A}_2, \end{aligned} \tag{159}$$

where

$$\mathbf{C}^{-1} = \phi^{2/3}\,{\phi_0}^{-2/3}\,\mathbf{C}^{e-1}. \tag{160}$$

4.4 Chemical Potential

The final expression of the chemical potential can be obtained by replacing Eq. (153) in Eq. (129) to obtain

$$\begin{aligned}\mu(\mathbf{C}^e, c) = &\mu^0 + R\,\theta\left[\log\left(\frac{\nu\,J_0\,c}{1+\nu\,J_0\,c}\right) + \frac{1}{1+\nu\,J_0\,c} + \frac{\chi}{(1+\nu\,J_0\,c)^2}\right]\\ &+ \frac{1}{3}\,\frac{\nu\,G}{J_0}\,\left({J_0}^{1/3}\,(1+\nu\,J_0\,c)^{-1/3}\,\mathrm{tr}\left({\mathbf{F}_0}^T\,\mathbf{C}^e\,\mathbf{F}_0\right)\right) + \nu\,\bar{p},\end{aligned} \tag{161}$$

where $\bar{p}$, defined in Eq. (98), can be explicitly computed once the expression of the stress tensor $\mathbf{S}^e$ is known, as in Eq. (154). In terms of the polymer fraction ϕ, the chemical potential becomes

$$\begin{aligned}\mu(\mathbf{C}^e, \phi) = &\mu^0 + R\,\theta\,\left[\log(1-\phi) + \phi + \chi\,\phi^2\right]\\ &+ \frac{1}{3}\,\frac{\nu\,G}{J_0}\,\left({\phi_0}^{-1/3}\,\phi^{1/3}\,\mathrm{tr}\left({\mathbf{F}_0}^T\,\mathbf{C}^e\,\mathbf{F}_0\right)\right) + \nu\,\bar{p}.\end{aligned} \tag{162}$$

4.5 Solvent Flux

In the constitutive theory, we made the assumption that the solvent flux $\mathbf{j}$ depends on the deformation, the concentration, and the gradient of the chemical potential, see Eq. (122). Following [3] we suppose that the fluid flux obeys a Darcy-tipe relation, hence it depends linearly on the gradient of the chemical potential

$$\mathbf{j} = -\mathbf{M}\,\nabla\mu, \tag{163}$$

where $\mathbf{M}$ is the mobility tensor. To respect the constitutive restriction (127) the mobility $\mathbf{M}$ must be a positive definite tensor.

The fluid flux $\mathbf{j}$ is the amount of solvent that flows through the boundary of the body per unit surface and per unit time, then it is measured in mol/(s m^2). If we introduce Eq. (129) in Eq. (163) we have

$$\mathbf{j} = -\mathbf{M}\,\nabla\left(\frac{\partial\psi_0}{\partial c} + \nu\,\bar{p}\right), \tag{164}$$

and, in turn,

$$\nu\,\mathbf{j} = -\nu^2 \mathbf{M}\,\nabla\left(\frac{\partial \psi_0}{\partial \nu c} + \bar{p}\right). \tag{165}$$

We observe that $\nu\,\mathbf{j}$ has the dimension of $\mathrm{m}^3/(\mathrm{m}^2\,\mathrm{s})$,since the molar volume $[\nu] = \mathrm{m}^3/\mathrm{mol}$, that still represents the amount of solvent flowing through the boundary per unit time and per unit surface, but simplifying we obtain that it has the dimension of a velocity, i.e. m/s.

Recalling definition (34), (24), and (40), we have

$$\Phi^f = \nu c, \tag{166}$$

and Eq. (165) becomes

$$\mathbf{j}\,\frac{\Phi^f}{c} = -\nu^2 \mathbf{M}\,\nabla\left(\frac{\partial \psi_0}{\partial \Phi^f} + \bar{p}\right), \tag{167}$$

where $\mathbf{j}/c$ represents the relative velocity $\mathbf{v}^{fs}$ of the fluid with respect to the solid network that usually appears in the generalized Darcy's Law

$$\Phi^f\,\mathbf{v}^{fs} = -\mathbf{K}\,\nabla\left(\frac{\partial \psi_0}{\partial \Phi^f} + \bar{p}\right), \tag{168}$$

where $\mathbf{K}$ is the permeability tensor corresponding to

$$\mathbf{K} = \nu^2\,\mathbf{M}. \tag{169}$$

We remark that in Terzaghi's consolidation theory the term $\frac{\partial \psi_0}{\partial \Phi^f}$, which accounts for adsorption and capillarity, is neglected. In our case this term is not negligible due to the importance of swelling phenomena.

For isotropic materials the mobility, and also the permeability, can be represented by scalar quantities, so that the corresponding tensors are

$$\mathbf{M} = m\,\mathbf{I}, \quad \mathbf{K} = k\,\mathbf{I}, \tag{170}$$

and thanks to Eq. (169) we have a scalar relation between mobility and permeability:

$$k = \nu^2\,m. \tag{171}$$

In [4] a linear dependence of the mobility upon the polymer fraction is proposed as

$$m(\phi) = \frac{D}{\nu R \theta}\,\gamma(\phi), \tag{172}$$

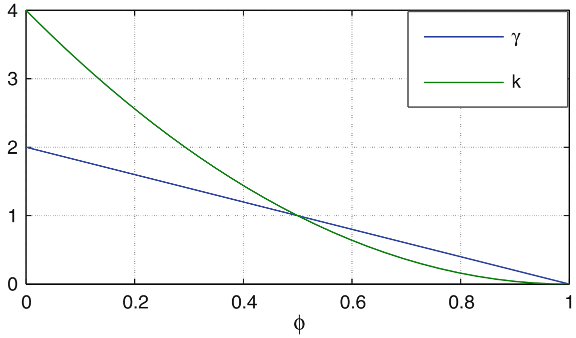


Fig. 2 Comparison between the two laws proposed to model the dependence of the permeability upon the polymer fraction ϕ. The linear law is the function γ in Eq. (173), and the exponential law is the function k in Eq. (175)

where D is a diffusion coefficient and $\gamma(\phi)$ is a positive function of ϕ

$$\gamma(\phi) = (1-\gamma_s)\,\phi\ \phi_0^{-1} + \gamma_s, \quad \gamma_s > 1, \tag{173}$$

such that in the reference state ($\phi = \phi_0$), $\gamma = 1$, and $m = D/(\nu R\theta)$.

Having in mind a specific application to intervertebral discs, in [1] an exponential law is proposed for the permeability

$$k = k_0 \left[\frac{e\,(1+e_0)}{e_0\,(1+e)}\right]^2 \exp\left[M\left(\frac{1+e}{1-e} - 1\right)\right], \tag{174}$$

where k_0 is the permeability and e_0 the void ratio in the reference state, and M is an empirical positive coefficient. Using Eqs. (43) and (39) we obtain the relation between the permeability and ϕ:

$$k(\phi) = k_0 \left[\frac{1-\phi}{1-\phi_0}\right]^2 \exp\left[M\left(\frac{\phi}{\phi_0} - 1\right)\right]. \tag{175}$$

In the reference state $\phi = \phi_0$, and the permeability k assumes its reference value k_0.

In Fig. 2, a comparison between the two laws (173) and (175) is shown. Only for this purpose arbitrary values of the constant are chosen, namely $M = 10^{-6}$, $\gamma_s = 2$, $\phi_0 = 0.5$, $k_0 = 1$. The important common feature of the proposed laws is the increase of permeability with swelling, which causes a decrease in the polymer fraction.

5 Summary of the Equations of the Model

In this section we summarize the system of equations needed to solve a swelling-deformation problem. We have a system of four equations, whose independent variables are the displacement $\mathbf{u}$, the chemical potential μ, the polymer fraction ϕ and the pressure p:

$$\begin{cases} \mathrm{Div}\mathbf{P} + \mathbf{b}_0 = 0 \\ -\frac{\phi_0}{\nu}\,\frac{\dot{\phi}}{\phi^2} = -\mathrm{Div}\,(\mathbf{j}) \\ \mu = \mu^0 + R\,\theta\,\left[\log\left(1-\phi\right) + \phi + \chi\,\phi^2\right] + \nu\,p\,\phi\,\phi_0^{-1} \\ J^e = 1 \end{cases} \tag{176}$$

The first equation is the balance of forces, the second the balance of fluid mass, the third is the relation between the chemical potential and the polymer fraction and the last represents the constraint of elastic incompressibility. To close the problem we need to consider the boundary condition for the displacement and the fluid mass equations.

6 Spontaneous Swelling Deformations

The spontaneous swelling deformation is achieved when a dry sample is fully immersed in solvent and undergoes free swelling. The computation of the spontaneous deformation is fundamental when we choose a stress free configuration as reference configuration and we want to study and simulate a particular phenomenon starting from that configuration.

The free energy function of the material in the dry configuration is given by Eq. (132), and the spontaneous deformation is the deformation that minimizes such energy. The free energy function $\psi_d(\mathbf{C}_d, c_d)$, in Eq. (132), is composed by the elastic part $\psi_{d_{mech}}$, in Eq. (134), and the mixing part $\psi_{d_{mix}}$, in Eq. (133). For the computation of the spontaneous deformation the free energy contribution due to the fibres is not considered. We consider only the case of isotropic deformation, since free swelling deformations are isotropic, such that

$$\mathbf{F}_d = \lambda_d\,\mathbf{I}, \tag{177}$$

and the incompressibility constraint (105) becomes

$$J_d = 1 + \nu\,c_d, \quad \text{and} \quad c_d = \frac{J_d - 1}{\nu}, \tag{178}$$

with $J_d = \lambda_d^3$. When in Eq. (132) we replace $\mathbf{F}_d$ and c_d with (177) and (178), we obtain

$$\psi_d(\lambda_d) = \frac{R\,\theta}{\nu}\left(\left(\lambda_d^3 - 1\right)\,\log\left(1 - \frac{1}{\lambda_d^3}\right) + \chi\,\left(1 - \frac{1}{\lambda_d^3}\right)\right) + \frac{3}{2}\,G_0\,\left(\lambda^2 - 1\right). \tag{179}$$

The minimum of free energy (179) is numerically computed using Newton's method, for which expressions of its first and second derivatives with respect to λ_d are needed:

$$\frac{\mathrm{d}\psi_d(\lambda_d)}{\mathrm{d}\lambda_d} = \frac{R\theta}{\nu}\left(3\,\lambda_d^2 l \log\left(1-\frac{1}{\lambda_d^3}\right)+\frac{3}{\lambda_d}+\frac{3\,\chi}{\lambda_d^4}\right)+3\,G_0\,\lambda_d, \tag{180}$$

and

$$\frac{\mathrm{d}^2\psi_d(\lambda_d)}{\mathrm{d}\lambda_d^2} = \frac{R\,\theta}{\nu}\left(6\,\lambda_d\,\log\left(1-\frac{1}{\lambda_d^3}\right)+\frac{9\,\lambda_d}{\lambda_d^3-1}-\frac{3}{\lambda_d^2}-\frac{12\,\chi}{\lambda_d^5}\right)+3\,G_0. \tag{181}$$

Using numerical data $\chi = 0.1, \theta = 298\,\mathrm{K}, R = 8.314472\,\mathrm{J\,K/mol}, G_0 = 1e6\,\mathrm{Pa}$, we find the value $\lambda^* = 1.6475$ for which energy $\psi_d(\lambda^*)$ is minimal, as shown in Fig. 3.

Recalling Eq. (11), we set $\lambda_0 = \lambda^*$ as the spontaneous (free swelling) deformation, such that in the reference state Ω_0 (in Fig. 1), the free energy is minimum and the body is stress free. Hence, thanks to Eq. (10), since $\mathbf{F}_d = \mathbf{F}_0$, in the reference free swollen state the deformation gradient $\mathbf{F} = \mathbf{I}$.

Once the spontaneous deformation is determined, thank to Eq. (34), and recalling that $J_0 = \lambda_0^3$, we can compute the polymer fraction ϕ_0 in the reference state. The smaller ϕ_0 is, the grater is the amount of solvent that entered the body. For $\lambda_0 = 1.6475$ we have $\phi_0 = 0.2236$.

A key role in the determination of the spontaneous deformation is played by the χ parameter. In Flory's theory [8], χ is called *enthalpy of mixing* and it represents the energetic contribution due to the mixing of polymer and solvent. Values of $\chi \leq 0.5$ promote swelling, while greater values prevent swelling due to some repulsion between the polymer and the solvent. Looking at Fig. 4, $\psi_{d\,mix}$ has an horizontal asymptote, as the deformation increases. While for $\chi \leq 0.5$, $\psi_{d\,mix}$ strictly decreases for deformations greater than 1, for $\chi > 0.5$ the free energy $\psi_{d\,mix}$ shows a change of concavity and the presence of a minimum that becomes closer to the identity deformation as χ increases. This can be attributed to growing repulsion between the polymer and the solvent, until no solvent is allowed to enter the polymer. As an exam-

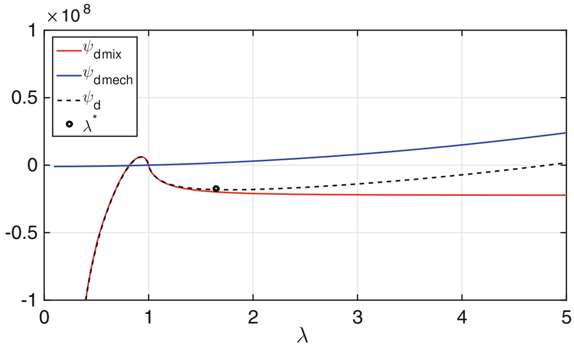


Fig. 3 The free energy ψ_d in the dry configuration, composed by a mechanical part $\psi_{d\,mech}$ and a mixing contribution $\psi_{d\,mix}$. The value λ^* corresponds to the deformation, greater than 1, that minimizes ψ_d. The value $\chi = 0.1$ has been used for the plot

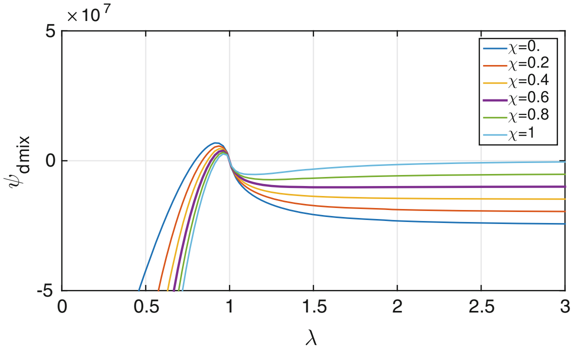


Fig. 4 Dependence of the mixing free energy $\psi_{d\,mix}$ upon the entalphy of mixing χ

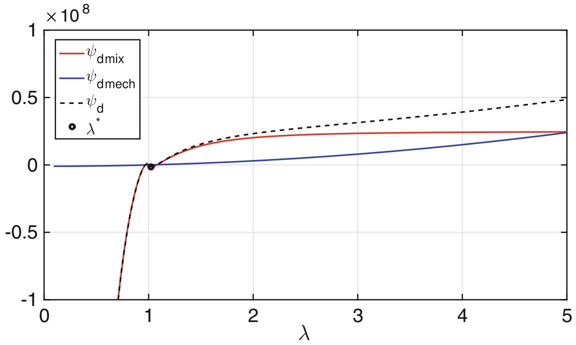


Fig. 5 The free energy ψ_d in the dry configuration, composed by a mechanical part $\psi_{d\,mech}$ and a mixing contribution $\psi_{d\,mix}$. The value λ^* corresponds to the deformation, greater that 1, that minimizes ψ_d. The value $\chi = 2$ has been used in the plot

ple, in Fig. 5, the free energy ψ_d is computed using $\chi = 2$, and the corresponding minimizing spontaneous deformation is $\lambda_0 = \lambda^* = 1.023$. The corresponding polymer fraction is $\phi_0 = 0.934$, i.e., a very small amount of solvent entered the polymer network and the material stays almost dry.

7 Numerical Experiments

In this section we describe a series of numerical experiments that can be performed, using the presented model, to simulate the physical behaviour of a hydrogel.

7.1 Free Swelling Deformation

A free swelling deformation, as previously pointed out, is an isotropic deformation a body undergoes when, fully immersed in the solvent, it is free to swell without any constraint. As described in the previous section, a freely swollen configuration corresponds to a state in which the free energy is minimum and the body is stress free. In this section, we want to study the dynamics of the free swelling deformation, by simulating the hydrogel that, starting from the dry configuration swells until it reaches the free swelling state. Then, in this case, the initial configuration will be the dry, not stress free, configuration Ω_∞ in Fig. 1. The presented model can be used to simulate such problem by setting $\mathbf{F}_0 = \mathbf{I}$, so that $\mathbf{F} = \mathbf{F}_d$ and $\psi_d = \psi_0$, and the system of equations we are going to solve is the following:

$$\begin{cases} \mathrm{Div}\mathbf{P} + \mathbf{b}_0 = 0 \\ -\frac{1}{\nu}\frac{\dot{\phi}}{\phi^2} = -\mathrm{Div}\,(\mathbf{j}) \\ \mu = \mu^0 + R\theta\left[\log(1-\phi) + \phi + \chi\,\phi^2\right] + \nu\, p\, \phi \\ J^e = 1 \end{cases} \tag{182}$$

Figure 6 shows the initial dry configuration of the body. It is a cube, whose edge is 0.1 cm long. For symmetry reasons only one eighth of the cube is shown. The boundary conditions for the problem are $\mu = \mu^0$ for all the boundaries of the body, because we suppose that the body is fully immersed in the pure solvent.

In Fig. 8 the body is fully swollen, the solvent inside the body is in equilibrium with the solvent outside (they have the same chemical potential), and the polymer fraction $\phi = 0.2236$ is uniform throughout the body. The corresponding free swelling deformations in the three directions are the same, confirming that the deformation is isotropic and uniform, with values $\lambda_x = \lambda_y = \lambda_z = 1.6475$. The computed values for the polymer fraction and the deformations coincide with the values for the spontaneous deformation computed in the previous section through the minimization of the free energy.

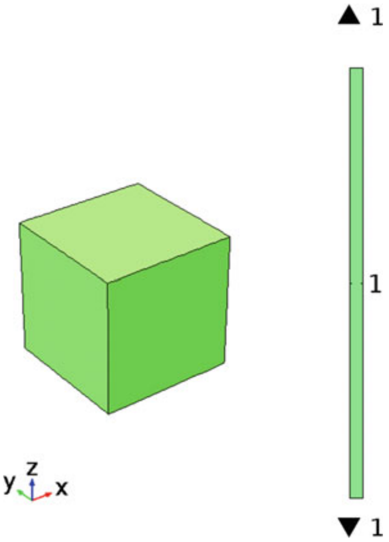


Fig. 6 Initial dry configuration for the free swelling deformation. In the figure the variable ϕ is plotted and the legend shows the assumed values throughout the body. The value $\phi = 1$ corresponds to the total absence of solvent

The numerical simulation of the swelling process allows us to observe the dynamics of the solvent flowing inside the body. The intermediate states between the initial and the final configuration are not characterized by an isotropic uniform deformation as shown by Fig. 7. The solvent flows across the boundaries and migrates toward the inner part of the body due to the gradient of the chemical potential. The diffusion coefficient $D = 5e - 12\,\text{m}^2/\text{s}$ determines how fast the diffusion process happens. After 36 h the dynamics of the solvent is over and equilibrium is reached (Fig. 8).

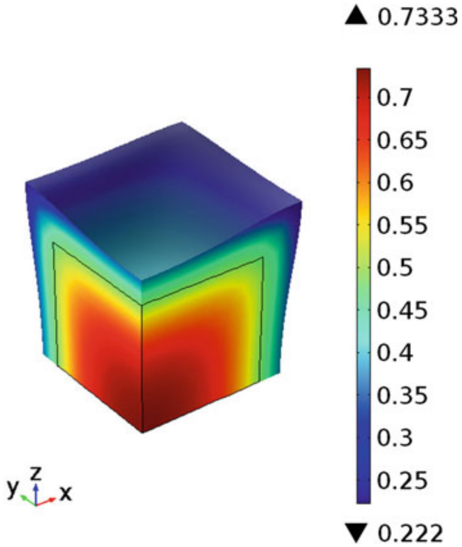


Fig. 7 Intermediate state for the free swelling deformation: the material is swelling. The legend shows the values of ϕ throughout the body: at this stage the polymer fraction is not uniform in the volume. The edges of the *cube* swell more than the rest of the body, due to a larger concentration of solvent near the boundaries

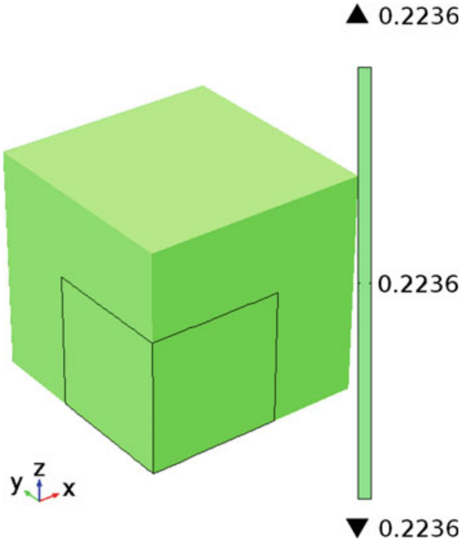


Fig. 8 Final state for the free swelling deformation. The material is fully swollen, the dymanics of the solvent flowing through the body is over and ϕ is again uniform throughout the body

7.2 Indentation Experiment

Indentation is one of the most common techniques for the mechanical characterization of materials. The test can be realized by pressing an indenter with different possible shapes (spherical, conical, cylindrical flat-ended) on the upper surface of a block of material. A control parameter can be the indentation depth or the pressure exerted by the indenter and, from the material response it is possible to compute material parameters which as the shear modulus and the Poisson coefficient.

When using sharp or spherical indenters the contact area varies during the indentation, while with a flat-ended cylindrical indenter the contact area remains constant [15]. In the following we will only consider the latter type of indenter, as shown in Fig. 9.

If the deformation induced by the punch is sufficiently small, then the linear theory of elasticity is applicable, and provided that the dimensions of the bodies are large compared with the dimension of the contact area, the stresses in this region are independent upon the shape of the bodies and the way they are supported far from the contact area [11].

The displacement variable is u, the small strain tensor is

$$\varepsilon = \frac{1}{2}\left(\nabla u + \nabla u^T\right), \tag{183}$$

and the stress tensor is

$$\sigma = 2G\left(\varepsilon + \frac{\nu}{1-2\nu}\mathrm{tr}(\varepsilon)I\right), \tag{184}$$

where G is the shear modulus, and ν is the Poisson coefficient.

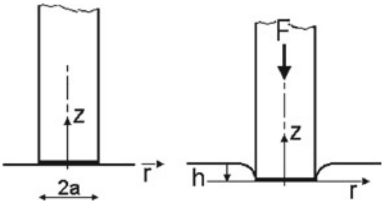
Using the reference system depicted in Fig. 9, we consider the z axis along the axis of the indenter and, due to the symmetry of the problem, a radial axis on the surface of the material. The radius of the flat-ended indenter is a. The equilibrium problem $\mathrm{div}\sigma = 0$ with the boundary conditions

$$\sigma_{zz}(r,0) = 0,\ r > a \tag{185}$$

$$\sigma_{rz}(r,0) = 0,\ 0 \le r \le a \tag{186}$$

$$u_z(r,0) = h,\ 0 \le r \le a \tag{187}$$

Fig. 9 Flat-ended punch indentation scheme, adapted from [15]



is solved in [10], and the results are reported in the following.

With reference to Fig. 9, corresponding to a penetration depth h, the distribution of pressure σ_{zz} under the punch is

$$\sigma_{zz}(r,0) = -\frac{2Gh}{\pi(1-\nu)\sqrt{a^2-r^2}}. \tag{188}$$

When the material is incompressible the distribution of pressure under the punch is obtained by replacing $\nu = 1/2$ in Eq. (188):

$$\sigma_{zz}^{inc}(r,0) = -\frac{4Gh}{\pi\sqrt{a^2-r^2}}. \tag{189}$$

The contact force exerted by the punch on the material can be obtained by integrating the distribution of pressure over the area of the punch:

$$F = -\int_0^{2\pi}\int_0^a \sigma_{zz}\, r\, dr\, d\alpha, \quad \text{and} \quad F^{inc} = -\int_0^{2\pi}\int_0^a \sigma_{zz}^{inc} r\, dr\, d\alpha, \tag{190}$$

obtaining

$$F = \frac{4Gah}{1-\nu}, \quad \text{and} \quad F^{inc} = 8Gah, \tag{191}$$

which give the reaction forces that a material exerts against the indenter pressed with indentation depth h. In Figure 11 we refer to F as F^{comp}, because it is the value for a compressible material.

To perform the numerical simulation of the indentation test we start from a cylinder of hydro-gel fully swollen, in its stress-free reference state. The shear modulus we are using in the numerical computation is $G = 10^5$ Pa, and the corresponding initial polymer fraction $\phi_0 = 0.0615$. We suppose that the material is immersed in pure solvent and it is in equilibrium, so that the chemical potential through all the body is zero. From this condition a cylindrical flat indenter is pressed on the top center of the cylinder, as shown in Fig. 10. The radius of the indenter is $a = 0.1$ cm, the indentation depth is $h = 0.03$ cm.

When the indenter is pressed into the hydro-gel, the latter immediately deforms elastically. The movement of the polymeric network causes a variation in the chemical potential, that in order to find new equilibrium induces the discharge of the solvent from the hydro-gel. Figure 10 shows a detail of the loaded area and the flux of the solvent that is depicted by the arrows, which have the direction of the chemical potential gradient. The solvent cannot flow through the contact area because the indenter is considered impermeable. Figure 11 shows the graph of the reaction force that the hydro-gel exerts against the indenter. It initially responds as an incompressible elastic material and then the forces relax due to the migration of the solvent until the new equilibrium is reached. The corresponding F^{comp} is computed using the first of Eq. (191), with a Poisson coefficient $\nu = 0.3$.

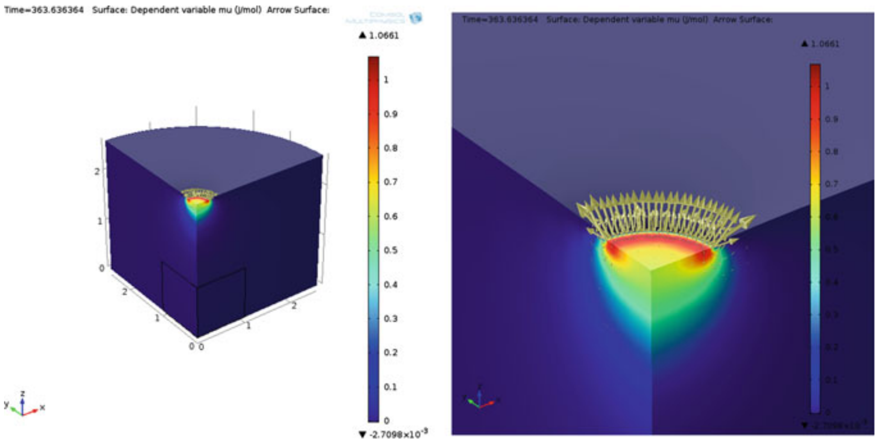

Time=363.636364 Surface: Dependent variable mu (J/mol) Arrow Surface:
▲ 1.0661
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
▼ -2.7098×10⁻³
2
1
0
2
1
0 0
1
2
y
z
x

Fig. 10 Fully swollen hydro-gel loaded by a cylindrical flat indenter. The indentation causes a change in the chemical potential, shown by the *color bar*, that induces a discharge of some solvent from the body. A detail of the indentation zone is shown, where the *arrows* have the direction of the chemical potential gradient. The tip of the indenter is considered impermeable

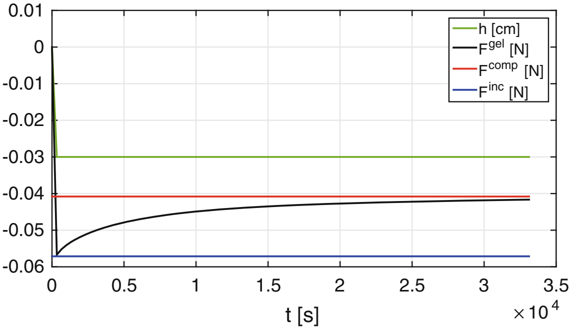


Fig. 11 Comparison of the reaction force $\mathsf{F}^{\mathrm{gel}}$ exerted by the hydro-gel against the indenter with the analytical values of the reaction forces for a compressible ($\mathsf{F}^{\mathrm{comp}}$) and an incompressible ($\mathsf{F}^{\mathrm{inc}}$) material. As expected, when the indentation starts the hydro-gel responds as an elastic (incompressible) material, then after some time the material relaxes due to the discharge of some solvent. h is the indentation depth

References

1. Argoubi, M., & Shirazi-Adl, A. (1996). Poroelastic creep response analysis of a lumbar motion segment in compression. *Journal of Biomechanics*, *29*(10), 1331–1339.
2. Biot, M. A. (1972). Theory of finite deformation of porous solids. *Indiana University Mathematics Journal*, *21*(7).
3. Chester, S. A., & Anand, L. (November 2010). A coupled theory of fluid permeation and large deformations for elastomeric materials. *Journal of the Mechanics and Physics of Solids*, *58*(11), 1879–1906.

4. Chester, S. A., & Anand, L. (2011). A thermo-mechanically coupled theory for fluid permeation in elastomeric materials: Application to thermally responsive gels. *Journal of the Mechanics and Physics of Solids*, *59*(10), 1978–2006.
5. de Luca, M., & DeSimone, A. (2012). Mathematical and numerical modeling of liquid crystal elastomer phase transition and deformation. *In MRS Proceedings*, *1403*, 2012. doi:10.1557/opl.2012.249. Copyright Materials Research Society 2012, Published online by Cambridge University Press: 2012.
6. de Luca, M., DeSimone, A., Petelin, A., & Čopič, M. (2013). Sub-stripe pattern formation in liquid crystal elastomers: Experimental observations and numerical simulations. *Journal of the Mechanics and Physics of Solids*, *61*(11), 2161–2177.
7. Ferguson, S. J., Ito, K., & Nolte, L.-P. (2004). Fluid flow and convective transport of solutes within the intervertebral disc. *Journal of Biomechanics*, *37*(2), 213–221.
8. Flory, P. J., & Rehner, J, Jr. (1943). Statistical mechanics of cross-linked polymer networks i. rubberlike elasticity. *The Journal of Chemical Physics*, *11*, 512.
9. Hong, W., Zhao, X., & Suo, Z. (2009). Formation of creases on the surfaces of elastomers and gels. *In Applied Physics Letters*, *95*
10. Ian, N. (1965). Sneddon. The relation between load and penetration in the axisymmetric boussinesq problem for a punch of arbitrary profile. *International Journal of Engineering Science*, *3*(1), 47–57.
11. Johnson, K. L. (1987). *Contact mechanics*. Cambridge: Cambridge university press.
12. Kang, M. K., & Huang, R. (2010). A variational approach and finite element implementation for swelling of polymeric hydrogels under geometric constraints. *Journal of Applied Mechanics*, *77*(6), 61004.
13. Lucantonio, A., Nardinocchi, P., & Teresi, L. (2012). Transient analysis of swelling-induced large deformations in polymer gels. *Journal of the Mechanics and Physics of Solids*.
14. Murad, M. A., Bennethum, L. S., & Cushman, J. H. (1995). A multi-scale theory of swelling porous media: I. application to one-dimensional consolidation. *Transport in Porous Media*, *19*(2), 93–122.
15. Riccardi, B., & Montanari, R. (2004). Indentation of metals by a flat-ended cylindrical punch. *Materials Science and Engineering: A*, *381*(1), 281–291.
16. Rice, J. R., & Cleary, M. P. (1976). Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. *Reviews of Geophysics and Space Physics*, *14*(2), 227–241.
17. Zhang, J., Zhao, X., Suo, Z., & Jiang, H. (2009). A finite element method for transient analysis of concurrent large deformation and mass transport in gels. *Journal of Applied Physics*, *105*(9), 093522–093522.

A Tensegrity Paradigm for Minimal Mass Design of Roofs and Bridges

Gerardo Carpentieri, Fernando Fraternali and Robert E. Skelton

Abstract This work presents a parametric design approach to simply-supported structures, exhibiting minimal mass *tensegrity* architectures (axially-loaded prestressible configurations of axially-loaded members) in two-dimensions. This provides minimal mass bridge structures in the plane. The mass minimization problem considers a distributed loading condition, under buckling and yielding constraints. The minimal mass structure is proved to be a tensegrity system with an optimal complexity. This optimal complexity (number of structural elements) depends only on material properties and the magnitude of the external load. The fact that the minimal mass structure is a Class 1 Tensegrity *substructure* has significant economic advantage. Class 1 structures are less expensive to construct, and substructures are easily deployable, offering portable applications for small spans. They can be easily assembled for prefabricated component parts for large spans. This minimal mass theory is then used to design a support structure for a solar panel cover of water canals, stopping evaporative losses and generating power without requiring additional land.

1 Introduction

Tensegrity structures are very efficient, and tend to provide minimal mass solutions to structure design under certain conditions. Some tensegrity papers have shown minimal mass for tensile structures, subject to a stiffness constraint [1]. Some have

G. Carpentieri (✉) · F. Fraternali
Department of Civil Engineering, University of Salerno, Via Giovanni Paolo II, 132,
84084 Fisciano, SA, Italy
e-mail: gcarpentieri@unisa.it

F. Fraternali
e-mail: f.fraternali@unisa.it

R.E. Skelton
Mechanical and Aerospace Engineering, University of California San Diego,
9500 Gilman Drive MC 0411, La Jolla, CA 92093-0411, USA
e-mail: bobskelton@ucsd.edu

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_5

shown minimal mass for: compressive loads [2], cantilevered bending loads [3, 4], torsional loads [5], simply-supported bending loads [6], and distributed loads on simply-supported spans, where significant structure is not allowed below the roadway, [7]. The present work formulates a parametric design approach to simply-supported (bridge-like) structures, which produces minimal mass shapes among all possible tensegrity topologies (configurations of members).

The subject of form-finding of tensegrity structures continues to be an active research area [8–13]. Particularly interesting is the use of fractal geometry as a form-finding method for tensegrity structures, which is well described in [2, 3, 5, 14]. Such an optimization strategy exploits the use of fractal geometry to design tensegrity structures, through a finite or infinite number of self-similar subdivisions of basic modules. The strategy looks for the optimal number of self-similar iterations to achieve minimal mass or other design criteria. This number is called the optimal *complexity*, since this number fixes the total number of parts in the structure. The 'fractal' approach to tensegrity form-finding paves the way to an effective implementation of the tensegrity paradigm in *parametric architectural design* [9, 10, 15, 16]. Discrete to continuum approaches to trusses and tensegrity structures are available in [17].

This paper finds the minimum mass design of tensegrity structures carrying simply supported and distributed bending loads. In [7] numerical solutions where found for a specified topology, without any theoretical guarantees that those topologies produced minimal mass. This paper provides more fundamental proofs that provide necessary and sufficient conditions for minimal mass.

The remainder of the paper is organized as follows. Section 2 describes the topology of the tensegrity bridge under examination. For a simply-supported structure of the simplest complexity, Sect. 3 describes the minimal mass bridge when the admissible topology allows *substructure* or *superstructure* (that is, respectively, structure below and above the roadbed). Section 4 defines deck mass and provides closed-form solutions to the minimal mass bridge designs when only sub- or super-structure is allowed. This finalizes the proof that the minimal mass bridge is indeed the substructure bridge. Section 4 also adds joint mass and shows that the optimal complexity is finite. In Sect. 5, we describe an application of the above theory to the design of a tensegirty bridge to be used for a solar panel covering of water canals. Conclusions are offered at the end.

2 Planar Topologies of the Tensegrity Bridges Under Study

The tensegrity structures in this paper will be composed of rigid compressive members called *bars*, and elastic tensile members called *cables*. We will assume that a tensile member obeys Hooke's law,

$$t_s = k(s - s_0), \tag{1}$$

where k is cable stiffness, t_s is tension in the cable, s is the length of the cable, and $s_0 < s$ is the rest length of the cable. The tension members cannot support compressive loads. For our purposes, a compressive member is a solid cylinder, called a bar. All results herein are trivially modified to accommodate pipes, tubes of any material, but the concepts are more easily demonstrated and the presentation is simplified by using the solid bar in our derivations. The minimal mass of a cable with loaded length s, yielding strength σ_s, mass density ρ_s, and maximal tension t_s is

$$m_s = \frac{\rho_s}{\sigma_s} t_s s. \tag{2}$$

The minimal mass to avoid yielding or buckling of a bar of length b, yielding strength σ_b, mass density ρ_b, modulus of elasticity E_b and with compression force f_b are, respectively

$$m_{b,Y} = \frac{\rho_b}{\sigma_b} f_b b, \quad m_{b,B} = 2\rho_b b^2 \sqrt{\frac{f_b}{\pi E_b}}. \tag{3}$$

The planar bridge topology is considered here to elucidate the fundamental properties that are important in the vertical plane. We use the following nomenclature, referring to Fig. 1: (i) a *superstructure* bridge has no structure below the deck level; (ii) a *substructure* bridge has no structure above the deck level; (iii) a *nominal* bridge contains both *substructure* and *superstructure*; (iv) Y means the design was constrained against yielding for both cables and bars; (v) B means the design was constrained against yielding for cables and buckling for bars; (vi) n means the number of self-similar iterations involved in the design ($n = 1$ in Fig. 1); (vii) p means the complexity of each iteration in the substructure ($p = 1$ in Fig. 1c); (viii) q means the complexity of each iteration in the superstructure ($q = 1$ in Fig. 1b); (ix) α is the aspect angle of the *superstructure* measured from the horizontal; (x) β is the aspect angle of the *substructure* measured from the horizontal.

We define the *superstructure* bridge of complexity $(n, p = 0, q)$ where the *substructure* below is deleted. We define the *substructure* bridge of complexity $(n, p, q = 0)$ where the *superstructure* above is deleted. It will be convenient to define the following constants:

$$\rho = \frac{\rho_b/\sigma_b}{\rho_s/\sigma_s}, \quad \eta = \frac{\rho_b L}{(\rho_s/\sigma_s)\sqrt{\pi E_b F}}. \tag{4}$$

Define a normalization of the system mass m by the dimensionless quantity μ:

$$\mu = \frac{m}{(\rho_s/\sigma_s)\, FL}, \tag{5}$$

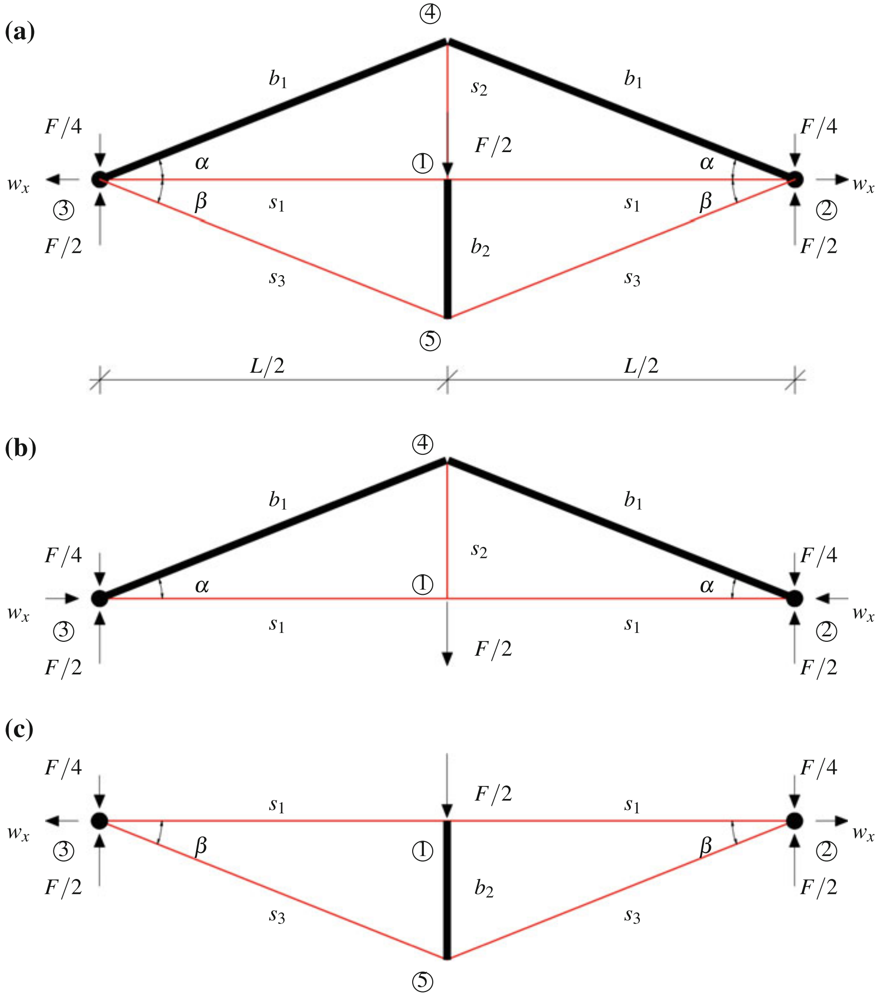


Fig. 1 Basic modules of the tensegrity bridge with: **a** nominal bridge: $n = q = p = 1$; **b** superstructure: $n = q = 1$; **c** substructure: $n = p = 1$. Compressive members (bars) are heavy *black lines*, tensile members (cables) are thin *red lines*

where the mass m at the yielding condition is:

$$m = \frac{\rho_b}{\sigma_b} \sum_{i=1}^{n_b} f_i b_i + \frac{\rho_s}{\sigma_s} \sum_{i=1}^{n_s} t_i s_i , \tag{6}$$

where (b_i, s_i) is respectively the length of the ith bar or ith cable, and respectively (f_i, t_i) is the force in the ith bar or cable.

3 Mass of Bridges of Complexity $(n, p, q) = (1, p, q)$, Under Yielding and Buckling Constraints

Now we consider structures by increasing p, q. This section finds the minimal mass of *substructure*, and *superstructure* bridges with complexity $(n, p, q) = (1, p, q)$, for any p and q greater then 1.

3.1 Substructure *Bridge with Complexity* $(n, p, q) = (1, p > 1, 0)$

Refer to Fig. 2 for the notation. The angle between the bars is:

$$\gamma = \frac{2\beta}{p-1}. \tag{7}$$

The lengths of the bars and cables are:

$$s_0 = \frac{L}{2}, \quad s_1 = \frac{L}{2}\cos\beta, \quad s_2 = L\sin\beta\sin\left(\frac{\beta}{p-1}\right), \quad b_1 = b_2 = \frac{L}{2}\sin\beta. \tag{8}$$

From the equilibrium equations, we obtain the following relations for the forces:

$$f_1 = \frac{F}{4\left[\cos\beta + \sin\left(\frac{\beta(p-2)}{p-1}\right) / \sin\left(\frac{\beta}{p-1}\right)\right]}, \quad f_2 = 2f_1, \tag{9}$$

$$t_2 = \frac{f_2}{2\sin\left(\frac{\beta}{p-1}\right)}, \quad t_1 = t_2\cos\left(\frac{\beta}{p-1}\right). \tag{10}$$

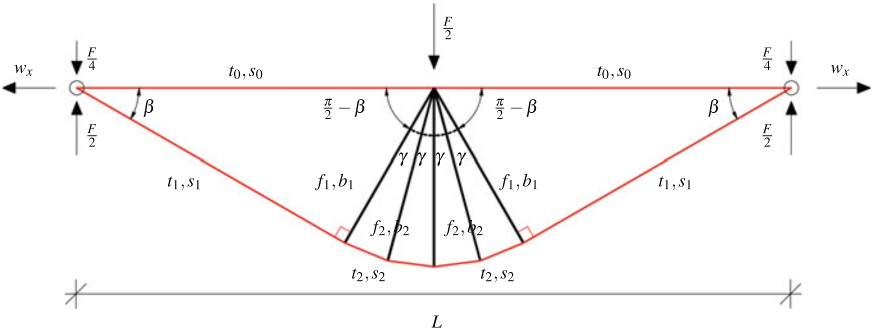


Fig. 2 Notations for forces and lengths of bars and cables for a *substructure* with complexity $n = 1$ and $p > 1$

The following theorems and corollaries can be obtained by minimizing the total masses of the bridge in Fig. 2 (refer to [6] for an extended proof of the above theorem and following theorems of this section).

Theorem 1 *Consider a* substructure *bridge with topology described by (8), with complexity* $(n, p, q) = (1, p, 0)$ *(Fig. 2). At the yielding condition the dimensionless total mass is:*

$$\mu_Y(\beta, p) = \frac{t_0}{F} + \frac{1}{4}\left[\frac{(p-1)\sin\beta\sin\left(\frac{\beta}{p-1}\right) + \cos\beta\cos\left(\frac{\beta}{p-1}\right)}{\cos\beta\sin\left(\frac{\beta}{p-1}\right) + \sin\left(\frac{\beta(p-2)}{p-1}\right)}\right] + \rho\frac{(p-1)\sin\beta}{4\left[\cos\beta + \sin\left(\frac{\beta(p-2)}{p-1}\right)/\sin\left(\frac{\beta}{p-1}\right)\right]}. \tag{11}$$

Note that t_0 is the force in deck cables s_0 (see Fig. 2).

Corollary 1 *The minimal mass in (11) is achieved at infinite complexity* $p \to \infty$ *and* $t_0 = 0$*. The minimal mass at yielding for a* substructure *bridge is:*

$$\mu_Y^*(\beta_Y^*, p^*) = \frac{1}{4}\left[\sqrt{\rho} + (1+\rho)\arctan\frac{1}{\sqrt{\rho}}\right], \tag{12}$$

where $p^* \to \infty$ *and the optimal angle* β_Y^* *is:*

$$\beta_Y^* = \arctan\left(\frac{1}{\sqrt{\rho}}\right). \tag{13}$$

Theorem 2 *Consider a* substructure *bridge with topology defined by (8), with complexity* $(n, p, q) = (1, p, 0)$*, See Fig. 2. At the buckling condition the dimensionless total mass is minimized at* $p = 2$ *and* $t_0 = 0$*, where:*

$$\mu_B(\beta, p=2) = \frac{1+\tan^2\beta}{4\tan\beta} + \frac{\eta}{2}\frac{\tan^2\beta}{\left(1+\tan^2\beta\right)^{3/4}}. \tag{14}$$

Corollary 2 *The minimal mass* substructure *is achieved for* $p = 1$.

3.2 Superstructure *Bridge with Complexity* $(n, p, q) = (1, 0, q > 1)$

Refer to Fig. 3 for the notation. The angle between the bars is:

$$\gamma = \frac{2\alpha}{q-1}. \tag{15}$$

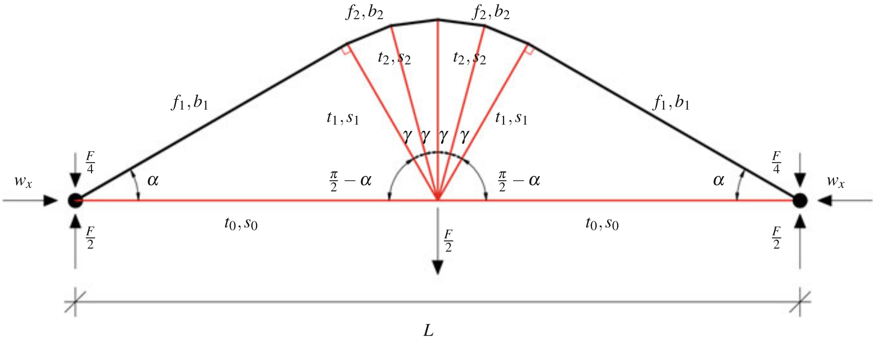


Fig. 3 Notations for forces and lengths of bars and cables for a *superstructure* with complexity $n = 1$ and $q > 1$

The lengths of the bars and cables are:

$$s_0 = \frac{L}{2}, \quad s_1 = s_2 = \frac{L}{2}\sin\alpha, \quad b_1 = \frac{L}{2}\cos\alpha, \quad b_2 = L\sin\alpha\sin\left(\frac{\alpha}{q-1}\right). \quad (16)$$

From the equilibrium equations, we obtain the following relations for the forces:

$$t_2 = \frac{F}{2\left[\cos\alpha + \sin\left(\frac{\alpha(q-2)}{q-1}\right)/\sin\left(\frac{\alpha}{q-1}\right)\right]}, \quad t_1 = \frac{t_2}{2}, \quad (17)$$

$$f_2 = \frac{t_2}{2\sin\left(\frac{\alpha}{q-1}\right)}, \quad f_1 = f_2\cos\left(\frac{\alpha}{q-1}\right). \quad (18)$$

The following theorems and corollaries can be obtained by minimizing the total masses of the bridge in Fig. 3 (refer to [6] for an extended proof of the above theorem and following theorems of this section).

Theorem 3 *Consider a* superstructure *bridge, of total span L, topology defined by (16), with complexity ($n = 1, q > 1$), Fig. 3. At the yielding condition under a vertical load F the dimensionless total mass is:*

$$\mu_Y(\alpha, q) = \frac{t_0}{F} + \frac{(q-1)\sin\alpha}{4\left[\cos\alpha + \sin\left(\frac{\alpha(q-2)}{q-1}\right)/\sin\left(\frac{\alpha}{q-1}\right)\right]}$$
$$+ \frac{\rho}{4}\frac{(q-1)\sin\alpha\sin\left(\frac{\alpha}{q-1}\right) + \cos\alpha\cos\left(\frac{\alpha}{q-1}\right)}{\sin\left(\frac{\alpha}{q-1}\right)\cos\alpha + \sin\left(\frac{\alpha(q-2)}{q-1}\right)}. \quad (19)$$

Corollary 3 *The minimal mass in (19) is achieved at infinite complexity $q \to \infty$ and $t_0 = 0$. Then the minimal mass at yielding for a* superstructure *bridge is:*

$$\mu_Y^*(\alpha_Y^*, q^*) = \frac{1}{4}\left[(1+\rho)\arctan\sqrt{\rho} + \sqrt{\rho}\right], \tag{20}$$

where $q^* \to \infty$ and the optimal angle α_Y^* is:

$$\alpha_Y^* = \arctan\sqrt{\rho}. \tag{21}$$

The left side of Fig. 4 illustrates *superstructure* bridges as $q \to \infty$, where masses are given for any q by (19).

Theorem 4 *Consider a* superstructure *bridge with topology (16), and complexity $(n, p, q) = (1, 0, q > 1)$, see Fig. 3. At the buckling condition the dimensionless total mass is:*

$$\mu_B(\alpha, q) = \frac{t_0}{F} + \frac{(q-1)\sin\alpha}{4\left[\cos\alpha + \sin\left(\frac{\alpha(q-2)}{q-1}\right)/\sin\left(\frac{\alpha}{q-1}\right)\right]} + \eta\left[\frac{\cos^2\alpha\sqrt{\cos\left(\frac{\alpha}{q-1}\right)} + 2(q-1)\sin^2\alpha\sin^2\left(\frac{\alpha}{q-1}\right)}{2\sqrt{\sin\left(\frac{\alpha}{p-1}\right)\cos\alpha + \sin\left(\frac{\alpha(q-2)}{q-1}\right)}}\right]. \tag{22}$$

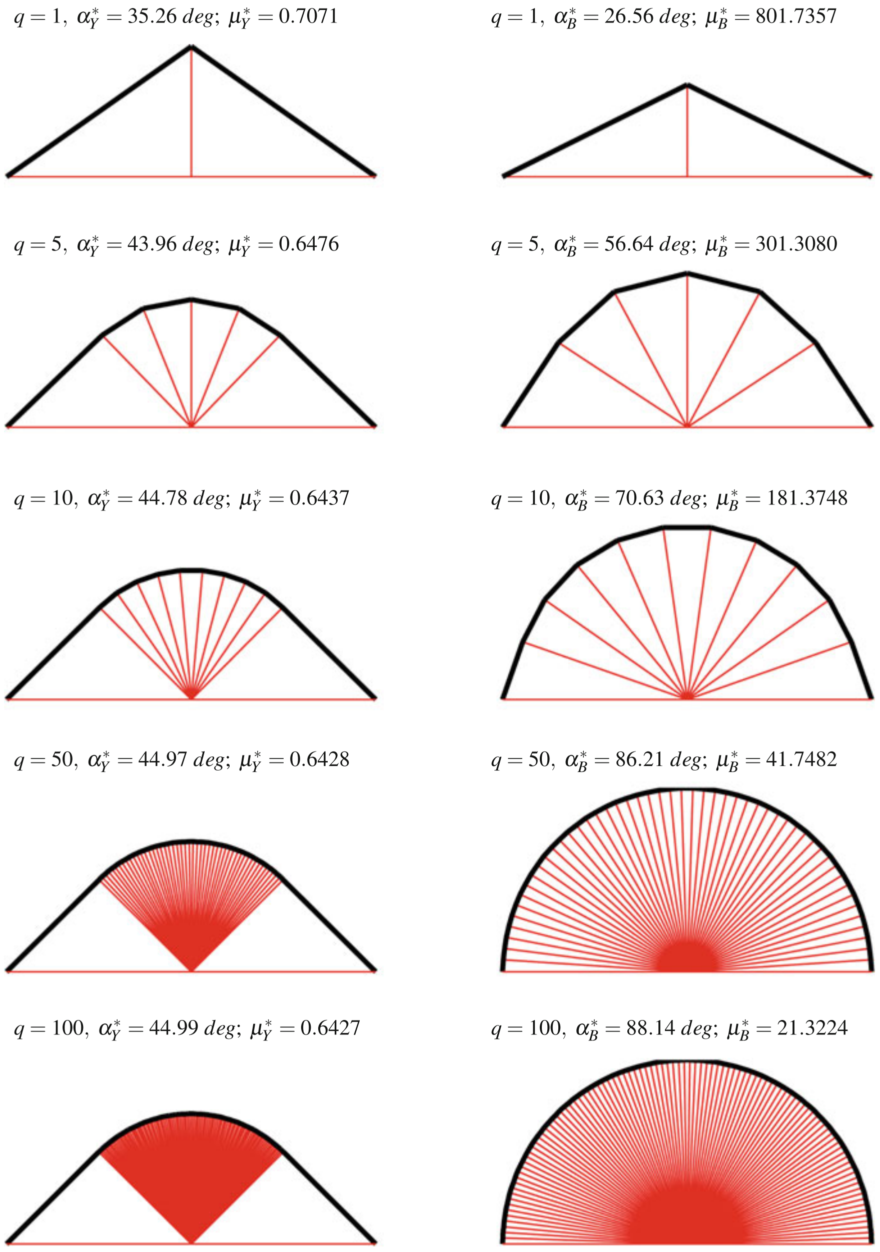
Corollary 4 *The minimal mass* superstructure *is achieved for $q \to \infty$ and $t_0 = 0$, leading to the following mass:*

$$\mu_B(\alpha, q \to \infty) = \frac{\alpha}{4} + \frac{\eta\cos^2\alpha}{2\sqrt{\sin\alpha}}. \tag{23}$$

It is important to consider that, for the solution $q \to \infty$, buckling is not the mode of failure since the lengths of the bars approaches zero. Also note that at $\alpha = 90°$, $\mu_B = \pi/8$.

The left side of Fig. 4 shows a sequence of *superstructures* under yielding constraints, as q increases. From (19) the mass is minimized at $q \to \infty$ and $\alpha_Y^* = 45°$ ($\rho = 1$). The right side of Fig. 4 shows a sequence of *superstructures* under buckling constraints, as q increases. The mass is minimized at $\alpha = 90°$ for $q = \infty$ ($\eta = 857.71$, same steel/steel material as above).

Moreover, the left side of Fig. 5 shows a sequence of *substructures* under yielding constraints, as p increases. From (11) the mass is minimized at $p \to \infty$ and $\beta_Y^* = 45°$ ($\rho = 1$). The right side of Fig. 5 shows a sequence of *substructures* under buckling constraints, as p increases. The mass is minimized at $\beta = 90°$ for $p = 1$ ($\eta = 857.71$, same steel/steel material as above).



$q = 1,\ \alpha_Y^* = 35.26\, deg;\ \mu_Y^* = 0.7071$

$q = 1,\ \alpha_B^* = 26.56\, deg;\ \mu_B^* = 801.7357$

$q = 5,\ \alpha_Y^* = 43.96\, deg;\ \mu_Y^* = 0.6476$

$q = 5,\ \alpha_B^* = 56.64\, deg;\ \mu_B^* = 301.3080$

$q = 10,\ \alpha_Y^* = 44.78\, deg;\ \mu_Y^* = 0.6437$

$q = 10,\ \alpha_B^* = 70.63\, deg;\ \mu_B^* = 181.3748$

$q = 50,\ \alpha_Y^* = 44.97\, deg;\ \mu_Y^* = 0.6428$

$q = 50,\ \alpha_B^* = 86.21\, deg;\ \mu_B^* = 41.7482$

$q = 100,\ \alpha_Y^* = 44.99\, deg;\ \mu_Y^* = 0.6427$

$q = 100,\ \alpha_B^* = 88.14\, deg;\ \mu_B^* = 21.3224$

Fig. 4 Optimal topologies of *superstructure* bridges with complexity $(n, p, q) = (1, 0, q \rightarrow \infty)$ under yielding constraints (*left*) and buckling constraints (*right*) for different q, (steel for bars and cables, $F = 1$ N, $L = 1$ m)

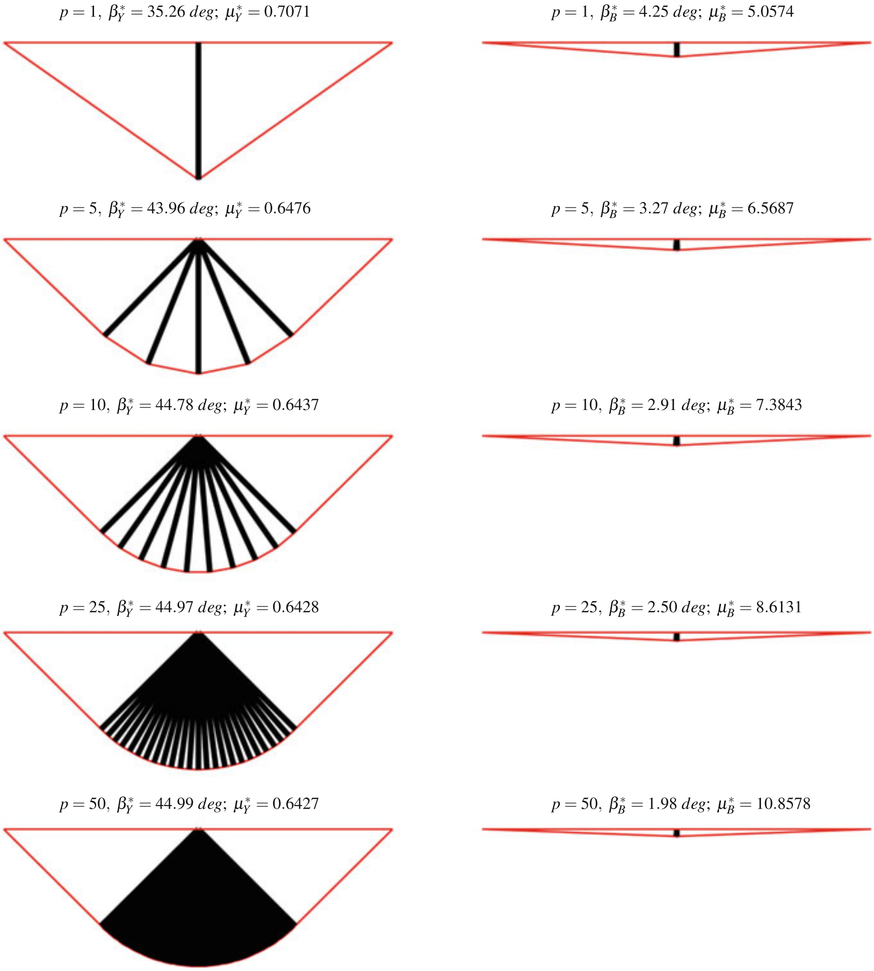

$p = 1,\ \beta_Y^* = 35.26\ deg;\ \mu_Y^* = 0.7071$
$p = 1,\ \beta_B^* = 4.25\ deg;\ \mu_B^* = 5.0574$
$p = 5,\ \beta_Y^* = 43.96\ deg;\ \mu_Y^* = 0.6476$
$p = 5,\ \beta_B^* = 3.27\ deg;\ \mu_B^* = 6.5687$
$p = 10,\ \beta_Y^* = 44.78\ deg;\ \mu_Y^* = 0.6437$
$p = 10,\ \beta_B^* = 2.91\ deg;\ \mu_B^* = 7.3843$
$p = 25,\ \beta_Y^* = 44.97\ deg;\ \mu_Y^* = 0.6428$
$p = 25,\ \beta_B^* = 2.50\ deg;\ \mu_B^* = 8.6131$
$p = 50,\ \beta_Y^* = 44.99\ deg;\ \mu_Y^* = 0.6427$
$p = 50,\ \beta_B^* = 1.98\ deg;\ \mu_B^* = 10.8578$

Fig. 5 Optimal topologies of *substructure* bridges with $n = 1$ under yielding constraints (*left*) and buckling constraints (*right*) for different p, (steel for bars and cables, $F = 1$ N, $L = 1$ m)

Theorem 5 *A minimal mass* superstructure *constrained against yielding with hinge/roller boundary conditions, has the same optimal topology as a minimal mass* superstructure *constrained against buckling and hinge/hinge boundary conditions.*

Proof Michell [18] proved that the minimal mass structure constrained against yielding with hinge/roller boundary conditions has the topology of the right side of Fig. 4 as $q \to \infty$ and $\alpha \to 90°$. Theorem 4 provides the same topology for hinge/hinge constraints. □

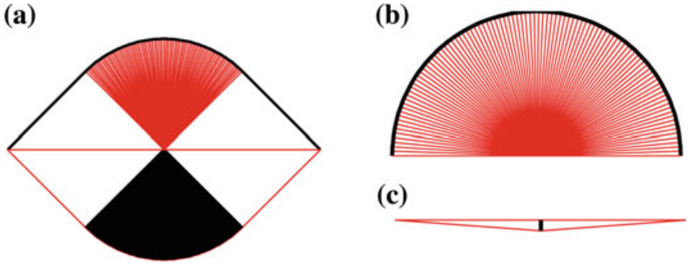


Fig. 6 Minimal mass bridges under, **a** yielding constrained *nominal* bridges, **b** buckling constrained *superstructure* bridge and **c** buckling constrained *substructure* bridge

Theorem 6 *The minimal mass* nominal *bridge constrained against yielding is obtained combining the optimal* superstructure *topology (Fig. 4, left side as $q \to \infty$) with the optimal* substructure *topology (Fig. 4, left side as $p \to \infty$).*

Proof Michell [18] obtained these same results by starting with a continuum and optimizing the shape. □

Figure 6a illustrates the minimal mass *nominal* bridge under yielding constraints (Theorem 6), leading to complexity $(n, p, q) = (1, \infty, \infty)$. Figure 6b illustrates the minimal mass *superstructure* bridge under buckling constraints, leading to complexity $(n, p, q) = (1, 0, q \to \infty)$. Figure 6c illustrates the minimal mass *substructure* bridge under buckling constraints, leading to complexity $(n, p, q) = (1, 1, 0)$.

4 Introducing Deck and Joint Masses

In previous sections, complexity n was restricted to 1. This is appropriate only when the external loads are all applied at the midspan. Real bridges cannot tolerate such an assumption. So in this section we consider a distributed load. Part of the load is the mass of the deck that must span the distance between adjacent support structures (complexity n will add $2^n - 1$ supports). In the Sect. 4.4 we will consider adding mass to make the joints, where high precision joints have less mass then rudely constructed joints.

4.1 Including Deck Mass

The total load that the structure must support includes the mass of the deck, which increases with the distance that must be spanned between support points of the structure design (which is determined by the choice of complexity n). We therefore consider bridges with increasing complexity n. We will show that the smallest $n = 1$

yields smallest structural mass and the largest deck mass. The required deck mass obviously approaches zero as the required deck span approaches zero, which occurs as $n \to \infty$. We will show that the mass of the deck plus the mass of the structure is minimized at a finite value of n.

The deck is composed by 2^n simply supported beams connecting the nodes on the deck.

Let the deck parameters be labeled as: mass m_d, mass density ρ_d, yielding strength σ_d, width w_d. The mass of one deck section is equal to:

$$m_d = \frac{c_1}{2^{3n}} + \frac{c_1}{2^{2n}}\sqrt{c_2 + \frac{1}{2^{2n}}}, \tag{24}$$

where:

$$c_1 = \frac{3\, w_d\, g\, \rho_d^2\, L^3}{8\, \sigma_d}, \quad c_2 = \frac{16\, \sigma_d\, F}{3\, w_d\, g^2\, L^3\, \rho_d^2}. \tag{25}$$

Then, the normalized total mass of the deck structure is:

$$\mu_d^* = \frac{2^n\, m_d}{(\rho_s/\sigma_s)\, FL}. \tag{26}$$

The total force acting on each internal node on the deck is the sum of the force due to the external loads and the force due to the deck:

$$F_{tot} = F + 2^n\, m_d\, g. \tag{27}$$

4.2 *Adding Deck Mass for a* Substructure *Bridge with Complexity* $(n, p, q) = (n, 1, 0)$

In this case, we make use of the notation illustrated in Fig. 7 in which complexity p is fixed to be one. Complexity n is defined to be the number of self-similar iterations of the basic module of Fig. 1c. Each iteration $n = 1, 2, \ldots$ generates different lengths of bars and cables. The lengths at the ith iteration are:

$$b_i = \frac{L}{2^i} \tan\beta, \qquad s_i = \frac{L}{2^i \cos\beta}, \qquad i = 1 - n. \tag{28}$$

Observing the multiscale structure of Fig. 7 it's clear that the number of bars and the number of cables at the ith self-similar iteration are

$$n_{si} = 2^i, \quad n_{bi} = 2^{i-1}. \tag{29}$$

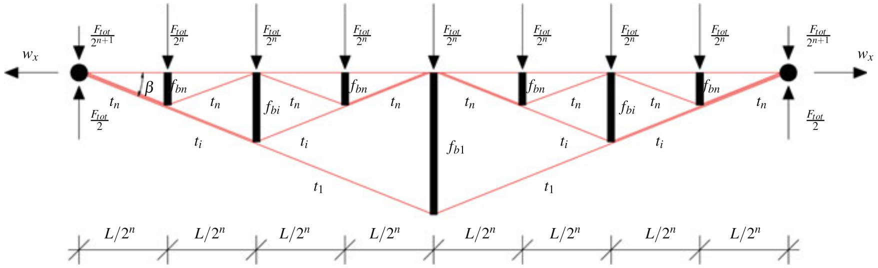


Fig. 7 Adopted notations for forces and lengths of bars and cables for a substructure with generic complexity $(n, p, q) = (n, 1, 0)$

In this case the total force applied to the bridge structure is given by (27) and then the forces in each member become:

$$f_{bi} = \frac{F + 2^n m_d g}{2^i}, \quad t_{si} = \frac{F + 2^n m_d g}{2^{(1+i)} \sin\beta}. \tag{30}$$

Theorem 7 *Consider a* substructure *bridge with deck mass m_d and topology defined by (28), with complexity $(n, p, q) = (n, 1, 0)$, see Fig. 7. The minimal mass design under yielding constraints is given by:*

$$\mu_Y^* = \left(1 - \frac{1}{2^n}\right)\left(1 + 2^n g \frac{m_d}{F}\right)\sqrt{1+\rho}, \tag{31}$$

using the optimal angle:

$$\beta_Y^* = \arctan\left(\frac{1}{\sqrt{1+\rho}}\right). \tag{32}$$

Observe that (31) yields mass $\sqrt{1+\rho}/2$ for complexity $n = 1$ and mass $\sqrt{1+\rho}$ for complexity $n = \infty$. Note from (32), that the optimal angle β_Y^* does not depend upon the choice of n. Indeed, the minimal mass solution under yielding constraints (31) depends on the material choice ρ (4), the complexity parameter n and the deck properties. Note that, since the total external force F is a specified constant, the mass is minimized by the complexity $n = 1$ if $m_d = 0$. However since m_d depends upon n, the total vertical force including deck mass depends upon n, and the optimal complexity will be shown to be $n > 1$ in that case.

Theorem 8 *Consider a* substructure *bridge with topology defined by (28), with complexity $(n, p, q) = (n, 1, 0)$. The minimal mass design under yielding and buckling constraints is given by:*

$$\mu_B^* = \beta_1 \frac{(1 + \tan^2\beta_B^*)}{2\tan\beta_B^*} + \eta\beta_2 \tan^2\beta_B^*, \tag{33}$$

using the aspect angle:

$$\beta_B^* = \arctan\left\{\frac{1}{12\beta_2\eta}\left[\beta_3 + \beta_1\left(\frac{\beta_1}{\beta_3} - 1\right)\right]\right\}, \tag{34}$$

where:

$$\beta_1 = \left(1 - \frac{1}{2^n}\right)\left(1 + 2^n g\frac{m_d}{F}\right), \tag{35}$$

$$\beta_2 = \left(\frac{1 + 2\sqrt{2}}{7}\right)\left(1 - \frac{1}{2^{3n/2}}\right)\sqrt{1 + 2^n g\frac{m_d}{F}}, \tag{36}$$

$$\beta_3 = \left(216\beta_1\beta_2^2\eta^2 - \beta_1^3 + 12\sqrt{324\beta_1^2\beta_2^4\eta^4 - 3\beta_1^4\beta_2^2\eta^2}\right)^{1/3}. \tag{37}$$

4.3 *Adding Deck Mass for a* Superstructure *Bridge with Complexity* $(n, p, q) = (n, 0, 1)$

In this case, we make use of the notation illustrated in Fig. 8 in which complexity q is fixed to be one. Complexity n is the number of self-similar iterations of the basic module of Fig. 1b at different scales. After the ith self-similar iterations, the length of the bars and cables for i ranging from 1 to n, are:

$$b_i = \frac{L}{2^i\cos\alpha}, \quad s_i = \frac{L}{2^i}\tan\alpha. \tag{38}$$

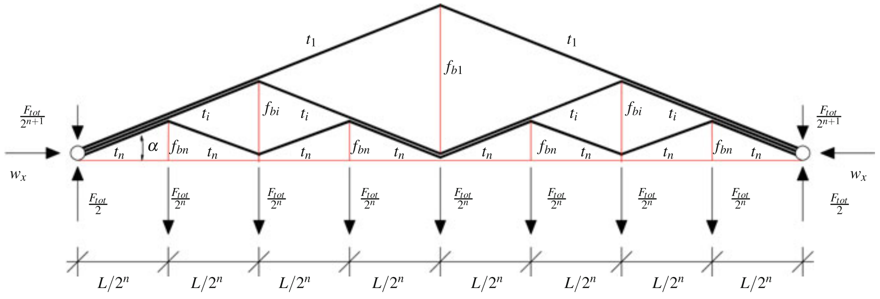


Fig. 8 Adopted notations for forces and lengths of bars and cables for a *superstructure* with complexity $(n, p, q) = (n, 0, 1)$

Observing the multiscale structure of Fig. 8 it's clear that the number of bars and the number of cables after the *i*th self-similar iterations are:

$$n_{si} = 2^{i-1}, \quad n_{bi} = 2^i. \tag{39}$$

In this case the total force applied to the bridge structure is given by (27) and then the forces in each member become:

$$f_{bi} = \frac{F + 2^n m_d g}{2^{(i+1)} \sin \alpha}, \quad t_{si} = \frac{F + 2^n m_d g}{2^i}. \tag{40}$$

Theorem 9 *Consider a* superstructure *bridge with topology defined by (38), with complexity* $(n, p, q) = (n, 0, 1)$*, Fig. 8. Under a given total vertical force (27), the minimal mass design under yielding constraints is given by:*

$$\mu_Y^* = \left(1 - \frac{1}{2^n}\right)\left(1 + 2^n g \frac{m_d}{F}\right)\sqrt{\rho\,(1 + \rho)}, \tag{41}$$

using the aspect angle:

$$\alpha_Y^* = \arctan\left(\sqrt{\frac{\rho}{1+\rho}}\right). \tag{42}$$

Theorem 10 *Consider a* superstructure *bridge with topology defined by (38), and complexity* $(n, p, q) = (n, 0, 1)$*, see Fig. 8. The structure is loaded with a given total vertical force (27) and the minimal bar mass, subject to yielding constraints is given by:*

$$\mu_B^* = \frac{\delta_1}{2} + \eta \delta_2 \frac{5^{5/4}}{4}, \tag{43}$$

using the aspect angle:

$$\alpha_B^* = \arctan \frac{1}{2}, \tag{44}$$

where:

$$\delta_1 = \frac{1}{2}\left(1 + 2^n g \frac{m_d}{F}\right)\left(1 - \frac{1}{2^n}\right), \tag{45}$$

$$\delta_2 = \sqrt{2}\left(\frac{1 + 2\sqrt{2}}{7}\right)\sqrt{1 + 2^n g \frac{m_d}{F}}\left(1 - \frac{1}{2^{3n/2}}\right). \tag{46}$$

The proofs of Theorems 9 and 10 are similar to the proofs of Theorems 7 and 8 in Sect. 4.2 and can be founded in [6].

4.4 Penalizing Complexity with Cost Considerations: Adding Joint Mass

Theorem 7, for $m_d = 0$, leads to an optimal complexity $n = 1$ which corresponds to a minimal mass equal to $\sqrt{1+\rho}/2$. As complexity n approaches infinity, instead, the mass given in (31), for $m_d = 0$, go to a limit equal to $\sqrt{1+\rho}$. However, the addition of the deck mass in Theorem 7 switches the optimal complexity from $n = 1$ to $n = \infty$, so small complexities n are penalized by massive decks. Also in this latter case, the resulting optimal minimal mass is then $\sqrt{1+\rho}$, as can be verified looking the (31) or considering that as n goes to infinity the deck mass given in (24) approaches zero. As a matter of fact, neither $n = 1$ or $n = \infty$ are believable solutions due to practical reasons: the first solution leads only to a single force at the middle of the span, the second solution leads to an infinite number of joints and connections. The minimal masses obtained from (31) with or without deck correspond to perfect massless joints. The addition of the joint masses to a tensegrity structure with n_n nodes, as illustrated in [5], leads to the following total normalized mass:

$$\mu^*_{Y,tot} = \mu^*_Y + \mu^*_d + \Omega n_n. \tag{47}$$

Let $\$_j$ be the cost per kg of making joints and let $\$_b$ be the cost per kg of making bars. Then define $\Omega = \$_b/\$_j$. For perfect joints $\Omega = 0$, for rudely made low cost joints $\$_j$ is small and Ω is larger. Hence Ω is also approximatively the ratio of material cost per joint divided by material cost per structural member being joined.

Consider the minimal masses of the *substructure* bridge (μ^*_Y) constrained against yielding, for the cases with or without deck, see Eq. (31). Assume steel material for cables, bars and deck beams and set $F = 1$ N, $L = w_d = 1$ m. Without deck the optimal aspect angle β^*_Y (32) is 35.26°. For the case with neither deck nor joint mass, the optimum complexity n is 1, which corresponds to an optimal mass $\mu^*_Y = \sqrt{2}/2$. As n approaches infinity the mass tends to a limit equal to $\sqrt{2}$, which is also the optimal mass for the case with deck mass and perfectly manufactured joints, since μ^*_d approaches zero for $n \to \infty$. Note that with the addition of joint masses as illustrated in (47), the optimal complexity n^* can become a finite value. The above procedure can be also used for the design under buckling constraints.

Figures 9 (for yielding) and Fig. 10 (for buckling) show the total minimal masses obtained by using (47). In both Figs. 9 and 10 we also show with red curves the minimal mass of *substructures* or *superstructures* only. In either case, the total mass of the structure with deck (but no joint mass), is shown by black continuous lines in Figs. 9 and 10, reaching minimum for an infinite complexity n. It is worth nothing that, for infinite n, the mass of the deck is zero and the total minimum mass is just the mass of the bridge structure. Then, with the dotted and dashed lines, we show that a finite optimal complexity can be achieved if the joint's masses are considered.

From Fig. 9 note that the minimal mass ($\mu \cong 21$) bridge has complexity $n = 11$ for $\Omega = 0.002$, and has minimal mass $\mu \cong 15$ with complexity $n = 12$ for $\Omega = 0.001$. Economic costs would decide if saving 25 % structural mass is worth the extra cost of improving the joint precision by a factor of 2.

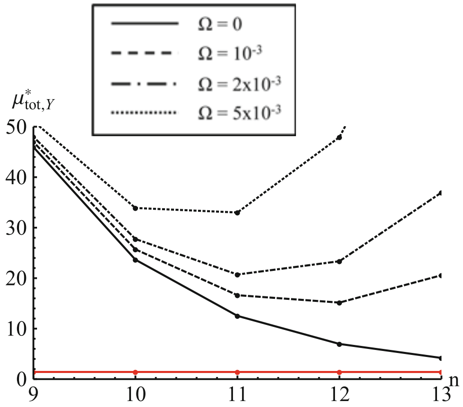


Fig. 9 Optimal masses under yielding of the *substructures* and *superstructure* (*red curve*) and total optimal mass with deck and different joint factors (*dashed* and *dottled curves*) for different values of the complexity n (steel for bars, cables, deck, $F = 1$ N, $L = w_d = 1$ m)

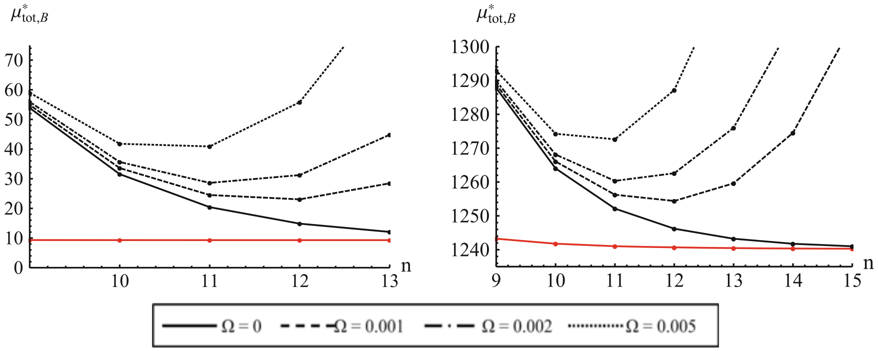


Fig. 10 Optimal masses under buckling of the *substructures* (*left*) and *superstructure* (*right*) (*red curves*) and total optimal masses with deck and different joint factors (*dashed* and *dotted curves*) for different values of the complexity n (steel for bars, cables, deck, $F = 1$ N, $L = w_d = 1$ m)

5 Design of a Deployable Bridge for Solar Energy Harvesting on Water Canals

From open canals that bring water to cities all over the world, water evaporation represents a significant loss of water. This section designs a minimal mass cover for water canals while using solar panels as the cover. This is not a new idea. Since 2012 India has built solar-panel-covered canals (Gujarata, India, 2012, [19]). The truss structure used in Gujarata is massive, threatening the economic survivability of the project. Here we design a minimal mass cover for such canals.

The optimal complexity of a deployable support structure for a solar panel covering of water canals is derived in the following, along with deployable schemes which are useful for construction, repairs, for sun following, and for servicing. It is shown that the minimal structure naturally has deployable features so that extra mass is not needed to add the multifunctional features. The design of bridge structures with tensegrity architecture will show an optimal complexity depending only on material choices and external loads. The minimization problem considers a distributed load (from weight of solar panels and wind loads), subject to buckling and yielding constraints. The result is shown to be a Class 1 tensegrity *substructure* (support structure only below the deck). These structures, composed of axially-loaded members (tension and compressive elements), can be easily deployable and have many portable applications for small spans, or they can be easily assembled for prefabricated component parts for large spans. The focus of this section is an application of these minimal mass tensegrity concepts to design shading devices to prevent or reduce evaporation loss, while generating electric power with solar panels as the cover. While the economics of the proposed designs are far from finalized, we show a technical solution that uses the smallest material resources, and shows the technical feasibility of the concept.

5.1 Description of the Model

In this example, we will assume buckling as a mode of failure of compressive members since it has been shown in [6] that buckling is the mode of failure in most of the practical cases.

In the previous sections we provided a theory to minimize the sum of deck mass, structural mass, and joint mass. The solution is a Class 1 tensegrity structure (compressive members do not make contact) with an optimal *complexity (optimal number of structural members)* that is finite. That is, the optimal structure is not a continuum (in contrast to the Michell truss, [18]) but a discrete structure with an optimal number of elements. This optimal number depends on material choice, the span, and the external load. This optimal bridge has no structure above the horizontal line (we call this a *substructure* bridge). This example assures that the most efficient structure does not extend above horizontal, making it ideal for our proposed solar array surface, since the surface is horizontal, and does not generate any shadows on the solar panels.

For a water canal application, Fig. 11 shows a 3D deployable flat roof made of repetitive 2D *substructure* bridges with multiscale topology defined in Fig. 7. Each planar *substructure* bridge (Fig. 7) is constrained with two fixed hinges at both ends (in practice these hinges might be pulleys that allow roll-up during construction or repair). As illustrated in Fig. 11, this module can be replicated (along the longitudinal direction) to build a deployable three-dimensional structure able to carry vertical loads distributed on the horizontal plane of the solar array. Figure 12 shows a possible application of this module to water canals.

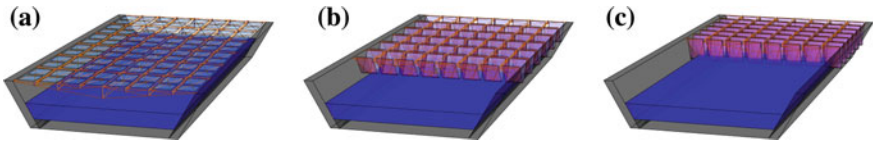


Fig. 11 Different configurations of a deployable solar roof for water canals: **a** open onfiguration, **b** transition between open/closed configurations, **c** closed configuration

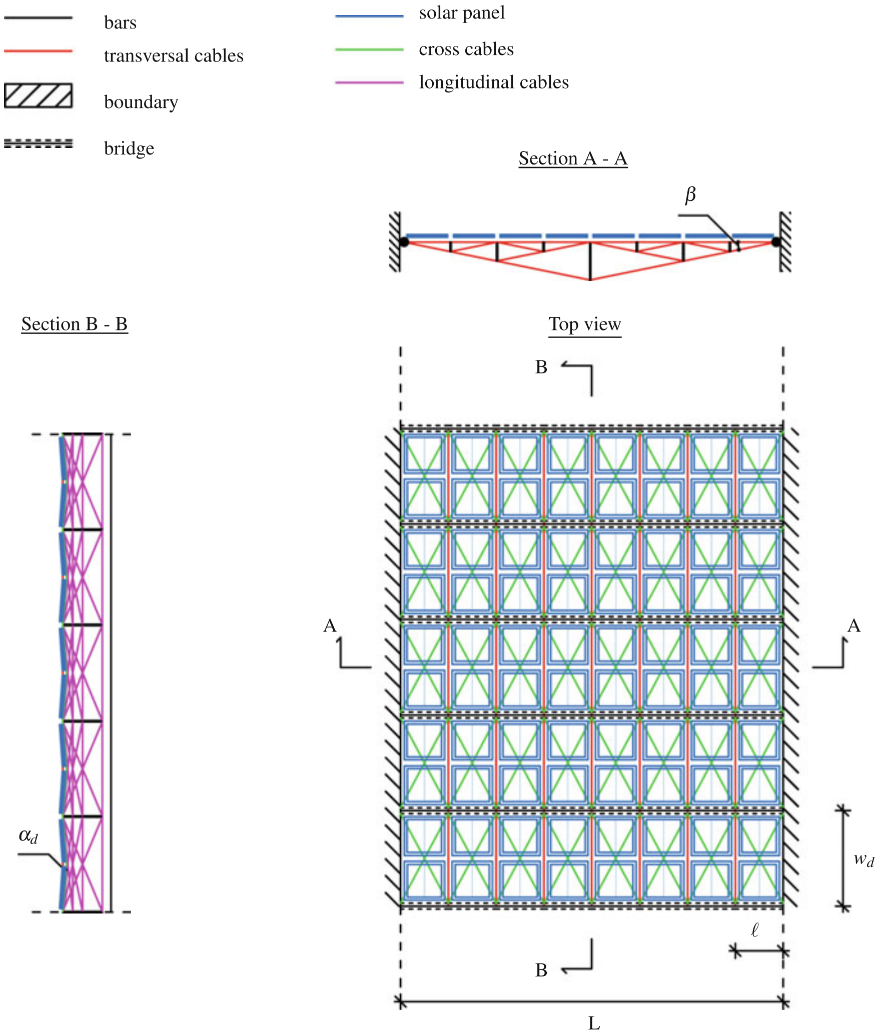


Fig. 12 Schematic of a deployable tensegrity system with solar panel

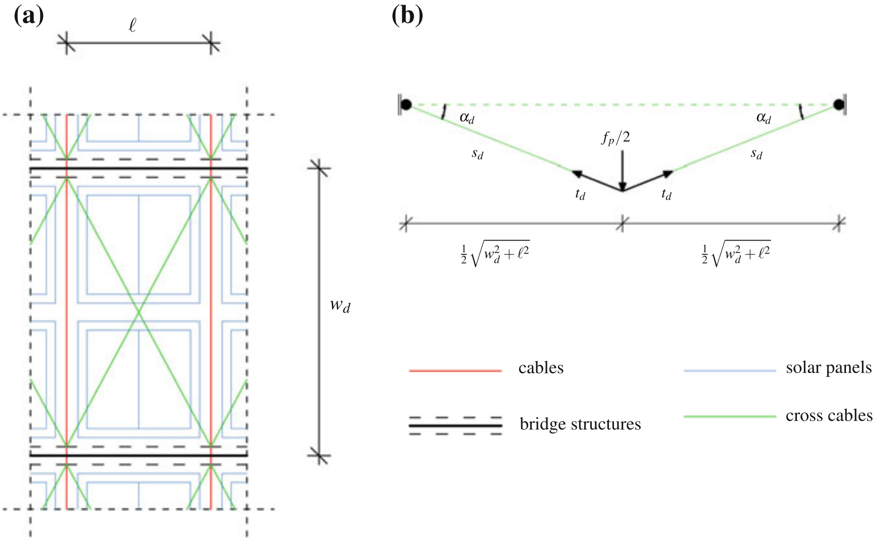


Fig. 13 Details of the canal structure: **a** deck system, **b** deformed shape of the deck cross cables subjected to the solar panel force

The bridge structures are stabilized out of the plane with a set of longitudinal cables as illustrated in Fig. 12. In particular diagonal cables and horizontal longitudinal cables (the magenta element showed in Section B-B of Fig. 12) are used to stabilize out of plane vertical movement.
The deck is composed of different orders of cables (refer to Figs. 12 and 13):

- longitudinal cables: the elements connecting each tensegrity bridge unit along the length of the canal;
- transversal cables: the elements of each tensegrity bridge lying on the transversal direction;
- cross cables: the elements that directly carry the solar panel loads and transfer their weight to the bridge structures.

The total vertical force F_{tot} can be computed designing the deck diagonal cables represented in Fig. 13. These cables directly support two different solar panel modules of sizes ℓ by $w_d/2$ (see Fig. 13). We design these cables assuming that, at the fully-deployed configuration of the structure, the deck diagonal cables are inclined at a fixed angle α_d with respect to the horizontal (Fig. 13). At this configuration the length and the tensile force in each deck diagonal cable are:

$$t_d = \frac{f_p}{4\sin\alpha_d}, \quad s_d = \frac{\sqrt{w_d^2+\ell^2}}{2\cos\alpha_d}. \tag{48}$$

We can compute the total mass of the deck diagonal cables as:

$$m_d = 4\frac{\rho_d}{\sigma_d} t_d s_d = \frac{\rho_d f_p}{\sigma_d} \frac{\sqrt{w_d^2 + \ell^2}}{2 \sin \alpha_d \cos \alpha_d}. \tag{49}$$

The normalized total mass μ_d^* and the total force F_{tot} can be computed with (26) and (27), respectively.

The final total mass to be optimized is then the summation of the mass of the bridge structure (33), the total mass of the deck (26) and the mass of the joints, Ωn_n.

5.2 Numerical Results and Discussion

Let us now focus our attention on numerical results regarding the optimal design of real-life roof structures featuring different complexities n. The examined structures show the following design data: $L = 30.48$ m, $F = 12$ kN, $w_d = 4.88$ m, $\alpha_d = 1°$, and the material properties of steel for bars ($\rho = 7862$ kg/m^3; $\sigma = 6.9 \times 10^8$ N/m^2; $E = 2.06 \times 10^{11}$ N/m^2) and Spectra® for cables ($\rho = 970$ kg/m^3; $\sigma = 2.7 \times 10^9$ N/m^2; $E = 120 \times 10^9$ N/m^2). We investigate on the optimal values of the following parameters: μ_B^*, μ_{tot}^* and β_B^*, which respectively denote the dimensionless minimal mass of a single bridge unit; the dimensionless minimal mass of the overall system formed by the bridge and the deck; and the optimal aspect angle of the bridge structure, under combined yielding and buckling constraints. The optimal angle β_B^* of the examined structures can be computed from Eq. (34), and/or the plots in Fig. 14. It is easy to verify that the minimal mas structure shows a markedly streamlined profile with $\beta_B^* = 2.18°$ (Fig. 14-left). The global minimum of μ_{tot}^* is attained in correspondence with the complexity $n^* = 3$, both for $\Omega = 0$ ($\mu_{tot}^* \cong 23.07$; $m_{tot}^* \cong 3.0318$ kg, cf. Table 1), and for $\Omega > 0$ (Fig. 14-right). The optimal design leads to 0.10 kg mass per cables and 0.02 kg mass per bars per meter of the canal lengthwise span (cf. Table 1).

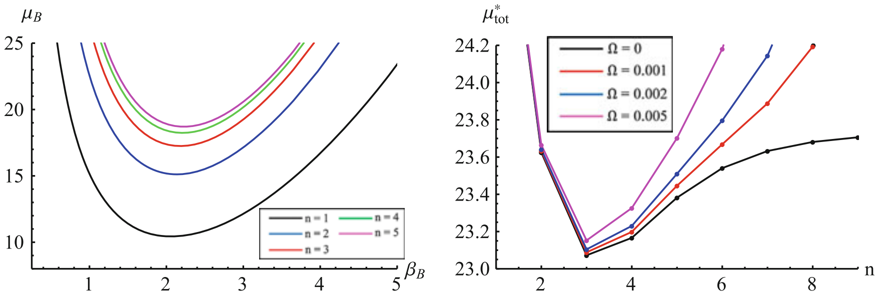


Fig. 14 *Left* dimensionless masses μ_B versus aspect angle β_B; *right* dimensionless total masses μ_{tot} versus complexity n ($\eta = 7569.04$)

Table 1 Optimal masses μ_B^* (33) and μ_{tot}^* and optimal aspect angles β_B^* (34) of *substructure* bridges with steel bars and Spectra®cables, under combined yielding and buckling constraints (B), for different complexities n

n	p	F_{tot} (N)	β_B^* (°)	μ_B^*	μ_{tot}^*	m_{tot}^* (kg)
1	1	12019.39	2.06	10.4357	25.4791	3.3480
2	1	12010.97	2.13	15.1186	23.6251	3.1044
3	1	12007.50	2.18	17.2522	23.0724	3.0318
4	1	12006.35	2.21	18.2414	23.1662	3.0441
5	1	12006.03	2.23	18.7080	23.3822	3.0725

6 Concluding Remarks

This paper provides closed form solutions (analytical expressions) for planar minimal mass tensegrity bridge designs. The forces, locations, and number of members are optimized to minimize mass subject to buckling and yielding constraints for a planar structure with fixed-hinge/fixed-hinge boundary conditions.

We present the optimal complexity of the *substructure* bridge that minimizes the sum of structural mass, deck mass and joint mass. Making better joints (less joint mass) results in higher optimal complexity and less mass. So the economic tradeoff between material cost of the truss structure and costs of making better joints will lead to the proper trade between mass and labor costs.

We also define a 3D deployable tensegrity structure made of repetitive planar *substructure* bridges (spanning the canal in the transversal direction) conveniently stabilized out of plane with a set of cables, in both the transversal and the longitudinal direction of the canal. Each planar structure has a self-similar fractal type of topology generated by the complexity parameter n. The minimal mass solution yields complexity n^* which depends upon material properties. Moreover, the topology of the 3D structure is function of canal width (L), aspect angle (β) of the *substructures* bridges, longitudinal aspect angle (α_d) governing the deploy-ability of the structure, the distance between consecutive repetitive structures in the longitudinal direction (w_d).

The design occupies much less volume and mass than the designs for the most advanced attempts at energy production and shading over water canals (see Gujarata, India, [19]). Formulas are given which will allow economic tradeoffs between material costs of the structure, the labor cost (assuming price per joint is inversely proportional to mass of the joint) of making more refined joints, and the choice of material (steel, Spectra®, or other). Implicit in these tradeoffs, the optimized complexity n^* of the structure is derived to allow economic decisions on the number of components (bars and cables) that will minimize mass for the given choice of material and joint costs.

Numerical and experimental studies on the dynamics of these structures will follow in subsequent work to impose further design constraints on stiffness issues (vibrational frequencies, mode shapes [20, 21], displacements for high winds conditions, etc.), but the capability of all these choices and adjustments are within the free parameters of the designs in this paper. The subsequent dynamics approach will evaluate the value (economics and performance tradeoffs) the use of feedback control for the deployable and service functions, or to adjust the stiffness of the structure (varying the prestress of the cables) to modify stiffness or damping after storm damage.

References

1. Skelton, R. E., & Nagase, K. (2012). Tensile tensegrity structures. *International Journal of Space Structures*, *27*, 131–137.
2. Skelton, R. E., & de Oliveira, M. C. (2010). Optimal complexity of deployable compressive structures. *Journal of the Franklin*, *I*(347), 228–256.
3. Skelton, R. E., & de Oliveira, M. C. (2010). Optimal tensegrity structures in bending: the discrete Michell truss. *Journal of the Franklin*, *I*(347), 257–283.
4. Nagase, K., & Skelton, R. E. (2014). Minimal mass tensegrity structures. *Journal of the International Association for Shell and Spatial Structures*, *55*(1), 37–48.
5. Skelton, R. E., & de Oliveira, M. C. (2010). *Tensegrity systems*. New York: Springer.
6. Carpentieri, G., Skelton, R. E., & Fraternali, F. (2015). On the minimum mass and optimal complexity of planar tensegrity bridges, Internal Report of the University of California, San Diego, Mechanical and Aerospace Engineering, No. 1-2014. www.fernandofraternaliresearch.com/publications/arxiv_tensegrity_bridges_theory_2014.pdf. Accessed 23 June 2015.
7. Skelton, R. E., Fraternali, F., Carpentieri, G., & Micheletti, A. (2014). Minimum mass design of tensegrity bridges with parametric architecture and multiscale complexity. *Mechanics Research Communications*, *58*, 124–132, ISSN 0093-6413, doi:10.1016/j.mechrescom.2013.10.017.
8. Koohestani, K. (2012). Form-finding of tensegrity structures via genetic algorithm. *International Journal of Solids and Structures*, *49*, 739–747.
9. Rhode-Barbarigos, L., Jain, H., Kripakaran, P., & Smith, I. F. C. (2010). Design of tensegrity structures using parametric analysis and stochastic search. *Engineering and Computer*, *26*(2), 193–203.
10. Sakamoto , T., Ferrè, A., & Kubo, M. (Eds.). (2008). From control to design: Parametric/algorithmic architecture. Actar, Barcelona.
11. Sokóf, T., & Rozvany, G. I. N.: New analytical benchmarks for topology optimization and their implications. Part I: bi-symmetric trusses with two point loads between supports. *Structural and Multidisciplinary Optimization*, *46*, 477–486 (2012).
12. Tilbert, A. G., & Pellegrino, S. (2011). Review of form-finding methods for tensegrity structures. *International Journal of Space Structures*, *18*, 209–223.
13. Yamamoto, M., Gan, B. S., Fujita, K., & Kurokawa, J. (2011). A genetic algorithm based form-finding for tensegrity structure. *Procedia Engineering*, *14*, 2949–2956.
14. Fraternali, F., Marino, A., El Sayed, T., & Della Cioppa, A. (2011). On the structural shape optimization through variational methods and evolutionary algorithms. *Mechanics of Advanced Materials and Structures*, *18*, 224–243.
15. Fraternali, F., Farina, I., & Carpentieri, G. (2014). A discrete-to-continuum approach to the curvatures of membrane networks and parametric surfaces. *Mechanics Research Communications*, *56*, 18–25.
16. Phocas, M. C., Kontovourkis, O., & Matheou, M. (2012). Kinetic hybrid structure development and simulation. *International Journal of Architectural Computing*, *10*(1), 67–86.
17. Fraternali, F., & Carpentieri, G. (2013). On the correspondence between 2D force networks and polyhedral stress functions. *International Journal of Space Structures*, *29*(3), 145–159.

18. Michell, A. G. M. (1904). The limits of economy of material in frame-structures. *Philosophical Magazine*, *8*, 589–597.
19. Kahn, M., & Longcore, T. (2014). A Feasibility Analysis of Installing Solar Photovoltaic Panels Over California Water Canals. UCLA Institute of the Environment and Sustainability, Los Angeles, CA. http://www.environment.ucla.edu/perch/resources/files/adeptfinalreport1.pdf. Accessed 22 July 2014.
20. Carpentieri, G., Modano, M., Fabbrocino, F., & Fraternali, F. (2015). Optimal design and dynamics of truss bridges. In *COMPDYN 2015—5th ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering* (pp. 1731–1740).
21. Modano, M., Fabbrocino, F., Gesualdo, A., Matrone, G., Farina, I., & Fraternali, F. (2015). On the forced vibration test by vibrodyne. In *COMPDYN 2015—5th ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering* (pp. 209–217).

Universal Meshes for the Simulation of Brittle Fracture and Moving Boundary Problems

Maurizio M. Chiaramonte, Evan S. Gawlik, Hardik Kabaria and Adrian J. Lew

Abstract Universal meshes have recently appeared in the literature as a computationally efficient and robust paradigm for the generation of conforming simplicial meshes for domains with evolving boundaries. The main idea behind a universal mesh is to immerse the moving boundary in a background mesh (the universal mesh), and to produce a mesh that conforms to the moving boundary at any given time by adjusting a few elements of the background mesh. In this manuscript we present the application of universal meshes to the simulation of brittle fracturing. To this extent, we provide a high level description of a crack propagation algorithm and showcase its capabilities. Alongside universal meshes for the simulation of brittle fracture, we provide other examples for which universal meshes prove to be a powerful tool, namely fluid flow past moving obstacles. Lastly, we conclude the manuscript with some remarks on the current state of universal meshes and future directions.

1 Introduction

Predicting and understanding the behavior of a propagating fracture has applications in a broad spectrum of disciplines. Perhaps the most renowned are applications in civil, mechanical, and aerospace engineering for the safe design of structural and mechanical components. More recently, a new wave of interest in understanding fracture propagation has risen due to the insurgence of hydraulic fracturing for the

M.M. Chiaramonte (✉) · E.S. Gawlik · H. Kabaria · A.J. Lew
Department of Mechanical Engineering and Institute for Computational
and Mathematical Engineering, Stanford University, Stanford, USA
e-mail: mchiaram@stanford.edu

E.S. Gawlik
e-mail: egawlik@stanford.edu

H. Kabaria
e-mail: hardik@stanford.edu

A.J. Lew
e-mail: lewa@stanford.edu

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_6

recovery of shale gas, as well as for engineering geothermal reservoirs. Alongside hydraulic fracturing, the practice of abyssal sequestration [17] for the disposal of radioactive waste also necessitates numerical tools capable of predicting the behavior of fluid driven fractures. Beyond engineering, the modeling of fracturing finds relevance in geophysics, for example for the prediction of ice-sheet separation and its effect on global climate. Due to the pervasive nature of fracture mechanics in many disciplines, there is a need for a deeper understanding of fracture evolution accounting for the three-dimensionality of the fracturing process. These applications motivate the current efforts towards the creation of robust and computationally efficient numerical methods to approximate the solutions of such fracture evolution models.

From the numerical standpoint, one of the crucial challenges faced in this particular class of problems is the approximation of the evolving displacement discontinuity, which is the focus of the work presented here. Several approaches have been proposed in the literature to address this challenge. Albeit a comprehensive literature review is beyond the scope of this manuscript, a very broad classification of the predominant classes of methods capable of handling the evolution of a few cracks can be arguably categorized into basis-enriching methods or mesh-conforming methods. Additionally it is worthwhile mentioning numerical methods to approximate solutions of regularized theories of fracture. These theories, by assigning a finite width to the fracture, circumvent the need to explicitly track the crack geometry. Some examples are phase field methods [6], and Michael Ortiz's own contributions on eigenfracture [37] and eigenerosion [29, 31], to name a few. Also worth mentioning are methods for situations in which massive fragmentation appears, such as the seminal contributions by Michael Ortiz based on cohesive elements [7, 27, 28, 30, 36].

Basis-enriching methods, such as the Extended (XFEM) [4, 24] and Generalized (GFEM) [2, 22] finite element methods, endow the finite dimensional subspace with discontinuous functions. These methods circumvent the need to accommodate the evolving displacement discontinuity in the domain subdivision by implicitly representing it through the discontinuous basis functions. Numerical integration can be rather challenging, and, for problems such as hydraulic fracturing, when coupled governing equations need to be solved on the crack faces, these methods fail to provide a quality subdivision of the crack geometry. An example of the latter is illustrated in Fig. 1.

Alternatively, conforming methods envision generating a subdivision which accommodates exactly the evolving crack geometry. By ensuring that the mesh for the domain always conforms to the crack path, any displacement discontinuity along the crack is easily introduced. While the idea is simple it is nonetheless powerful. The robustness of this class of methods is limited by the generation of a quality conforming subdivision, a process which can be computationally demanding and prone to failure. Some examples of this approach are locally re-meshing methods as encountered in [3, 5, 32] as well as r-adaptive procedures as proposed in [1, 12, 23, 39]. Related to the latter are finite element spaces with embedded discontinuities [18, 26].

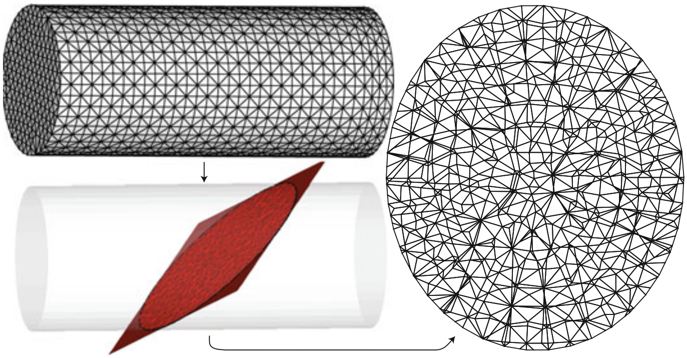


Fig. 1 An arbitrary cut (in *red*) through a quality tetrahedralization, representing an imaginary fracture, yields a poor discretization of the fracture faces

Herein we present a few ideas for the simulation of brittle fracture that fall in the latter category of conforming mesh methods by taking advantage of *universal meshes*. Universal meshes are a paradigm for mesh generation that envision the use of a single "background" mesh (the universal mesh) whose vertices closest to the crack geometry are perturbed to obtain a subdivision conforming to it. An example of such a perturbation is illustrated in Fig. 2. Because the same mesh can be deformed to conform to the geometry of a class of cracks, we say that the mesh is *universal* for such a class. The salient features of the method are its robustness, computational efficiency, and the mesh-independence of the solutions it provides (in fact, convergence).

We provide a description of the algorithm of universal meshes in Sect. 2 followed by the presentation of the algorithm for the simulation of brittle fracture in Sect. 3. Later, in Sect. 4, we highlight some applications of universal meshes beyond brittle fracture. We conclude the manuscript on some recent developments of universal meshes in three dimensions in Sect. 5.

2 Universal Meshes

We introduce the basic algorithmic ideas behind a universal mesh next. For concreteness, we focus the description on the aspects relevant to crack propagation, bearing in mind that similar ideas apply equally well to other classes of evolving domains, such as those encountered in fluid-structure interaction, as discussed later in Sect. 4.

2.1 Algorithm

To illustrate the discretization of an evolving domain with a universal mesh, we consider in this section the problem of triangulating a domain $\Omega(t) \subset \mathbb{R}^2, 0 \leq t \leq T$, which contains an evolving crack. In other words,

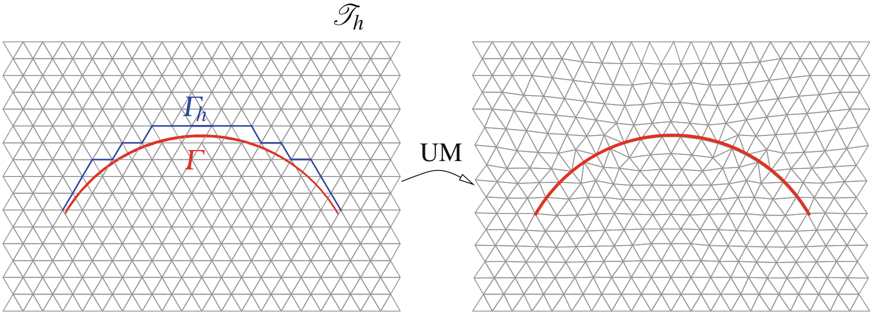


Fig. 2 Using a universal mesh, a triangulation conforming to a crack (*right*) is constructed by immersing the crack in a background triangulation $\mathscr{T}_h$ (*left*) and adjusting a few of its elements. This is accomplished by selecting a set of edges Γ_h in the background triangulation that lie near the crack Γ, mapping them onto Γ via the closest point projection, and relaxing a few nearby vertices to ensure the quality of the resulting triangulation

$$\Omega(t) = \mathscr{D} \setminus \Gamma(t)$$

where $\mathscr{D} \subset \mathbb{R}^2$ is an open, bounded, polygonal domain and $\Gamma(t) \subset \mathscr{D}$ is a simple open rectifiable curve.

Let $\mathscr{T}_h$ be a triangulation of $\mathscr{D}$, hereafter referred to as the *universal mesh*. We use h to denote the maximum diameter of an element of $\mathscr{T}_h$. We do not assume that the universal mesh conforms to $\Gamma(t)$ at any given time; in general, $\Gamma(t)$ may cut through elements of $\mathscr{T}_h$ arbitrarily, as in Fig. 2. Intuition would suggest, however, that a conforming mesh can be constructed by adjusting a few elements of $\mathscr{T}_h$ in a neighborhood of $\Gamma(t)$. This is the basic observation behind universal meshes.

To construct such a conforming mesh from $\mathscr{T}_h$, the following algorithm is adopted. First, a subset of edges in $\mathscr{T}_h$ lying near $\Gamma(t)$ is identified. We denote the union of these edges $\Gamma_h(t)$. Next, these edges are mapped onto $\Gamma(t)$ via the closest point projection $\pi : \mathbb{R}^2 \to \Gamma(t)$, with a suitable modification that places nodes precisely at the crack tips. Finally, the positions of nearby nodes are adjusted via a *relaxation* step that ensures the quality of the resulting triangulation.

The precise choices for the edges constituting $\Gamma_h(t)$ and the nodal adjustments adopted during relaxation are detailed in [35]. Briefly, $\Gamma_h(t)$ consists of *positive edges* of *positively cut* triangles in $\mathscr{T}_h$. To define these notions, one designates an orientation (positive or negative) for points in a neighborhood of $\Gamma(t)$. A triangle in $\mathscr{T}_h$ is called *positively cut* if it has two nodes on the positive side of $\Gamma(t)$ and one on the negative side. An edge is then called a *positive edge* if it belongs to a positively cut triangle and its endpoints both lie on the positive side of $\Gamma(t)$. A minor modification to $\Gamma_h(t)$ is made if a triangle in $\mathscr{T}_h$ has three nodes on $\Gamma_h(t)$; see [35] for details.

A key feature of the algorithm summarized above is its robustness. That is, the algorithm returns a valid mesh, for both the crack and the domain, with the quality of the elements bounded from below independently of the mesh size, provided that three conditions are satisfied: (1) the background mesh is sufficiently refined in a neighborhood of $\Gamma(t)$, (2) all positively cut triangles in $\mathscr{T}_h$ are acute, and (3) the curve $\Gamma(t)$ is sufficiently smooth. This statement was proved for a domain with C^2 boundary (no cracks) in [33, 34]. The numerical examples strongly suggest that this should also be possible for domains with interfaces, such as cracks.

3 Simulating Brittle Fracture with Universal Meshes

The obvious way in which a Universal Mesh is useful for the simulation of a propagating crack is by providing a mesh perfectly conforming to the crack at each step of its evolution. However, there are advantages that are less evident: the conforming mesh enables us to compute stress intensity factors to any order of accuracy, and the few mesh changes from step to step make it possible to retain much of the data structures in the computer implementation. The accuracy in the computation of the stress intensity factors is a determinant factor in observing convergence of the crack evolutions for "reasonable" mesh sizes.

In the following we present a numerical algorithm for the simulation of crack evolution in a restricted set of problems, as introduced in [35]. The presentation of the algorithm is followed by examples, including the formation of oscillatory crack paths in quenched plates, which requires very accurate stress intensity factors to converge.

3.1 Problem Statement

We consider the problem of *an always propagating* crack in an elastic medium, as defined next. We parametrize the crack evolution by the crack length $\ell \in [\ell_0, \ell_{max}]$ and we denote by $\mathscr{C}(\ell)$ the crack tip position for the crack of length ℓ. Hence the crack set is given by $\mathscr{C}([\ell_0, \ell])$. The domain occupied by the cracked domain is denoted by $\Omega(\ell)$ and its boundary is decomposed into a portion over which displacements are prescribed, $\partial_d\Omega(\ell)$, and a portion over which boundary tractions are prescribed, $\partial_\tau\Omega(\ell)$. Further we let $\mathscr{C}([\ell_0, \ell]) \subseteq \partial_\tau\Omega(\ell)$.

The problem statement then reads: find the deformation $\boldsymbol{u}(\cdot, \ell) : \Omega(\ell) \to \mathbb{R}^2$, the load scaling factor $C : [\ell_0, \ell_{max}] \to \mathbb{R}$, and the crack set $\mathscr{C}([\ell_0, \ell_{max}])$ such that the following holds for all $\ell \in (\ell_0, \ell_{\max}]$:

$$\begin{aligned} -\nabla \cdot (\mathbb{C} : \nabla \boldsymbol{u}) &= \boldsymbol{b}(\ell), && \text{on } \Omega(\ell), \\ (\mathbb{C} : \nabla \boldsymbol{u})\boldsymbol{n} &= \boldsymbol{t}(\ell), && \text{on } \partial_\tau\Omega(\ell), \\ \boldsymbol{u} &= \boldsymbol{g}(\ell), && \text{on } \partial_d\Omega(\ell), \end{aligned}$$

$$
\begin{aligned}
K_I[\boldsymbol{u}] &= K_c,\\
K_{II}[\boldsymbol{u}] &= 0,\\
\mathscr{C}([\ell_0, \ell^-]) \subset \mathscr{C}&([\ell_0, \ell]), \qquad \forall \ell^- < \ell,
\end{aligned}
$$

where $\boldsymbol{b}(\ell) = C(\ell)\overline{\boldsymbol{b}}(\ell)$, $\boldsymbol{t}(\ell) = C(\ell)\overline{\boldsymbol{t}}(\ell)$, $\boldsymbol{g}(\ell) = C(\ell)\overline{\boldsymbol{g}}(\ell)$. Here $\overline{\boldsymbol{b}}(\cdot, \ell) : \Omega(\ell) \to \mathbb{R}^2$, $\overline{\boldsymbol{t}}(\cdot, \ell) : \partial_\tau \Omega(\ell) \to \mathbb{R}^2$, and $\overline{\boldsymbol{g}}(\cdot, \ell) : \partial_d \Omega(\ell) \to \mathbb{R}^2$ are arbitrary functions representing the "shape" of the body forces and boundary conditions. Effectively, for every crack length we know the "shape" of the applied body force ($\boldsymbol{b}$), boundary tractions ($\boldsymbol{t}$) and displacements ($\boldsymbol{g}$), and we must solve for the linearly scaling coefficient $C(\ell)$ such that the condition $K_I[\boldsymbol{u}] = K_c$ is always met, where $K_I[\boldsymbol{u}]$ and $K_{II}[\boldsymbol{u}]$ are the mode I and II stress intensity factors. The condition $K_{II}[\boldsymbol{u}] = 0$ dictates the direction of crack propagation following the Principle of Local Symmetry [19].

The "always propagating crack" problem circumvents some of the more delicate issues in crack propagation, such as crack arrest and catastrophic crack propagation, regularity of the crack path, and competition among multiple cracks. The algorithm introduced next is applicable to this simpler class of problems.

3.2 Crack Propagation Algorithm

There are three critical steps in the computation of the evolution of brittle crack paths: (1) the generation of a triangulation that conforms to the cracked domain, (2) the calculation of the elasticity fields, and (3) the evaluation of the stress intensity factors for curvilinear cracks. The steps are highlighted in Fig. 3.

We construct a triangulation that conforms to each cracked domains from a universal mesh, as described in Sect. 2. To ensure that the elasticity fields are sufficiently resolved, we draw on a class of finite element methods for domains with corners or cracks that retain optimal convergence rate for elements of any order in the face of singular solutions [10], in contrast to standard methods or methods with enrichments. Lastly, given that we count with higher order solutions of the elasticity fields, we employ a family of *interaction integrals* specifically designed to handle curvilinear cracks [9] which yield stress intensity factors that converge rapidly to the exact ones (namely, they converge with twice the rate of convergence of the derivatives of the solution of the elasticity fields; a motivation to the use higher-order finite element methods).

We approximate the crack set $\mathscr{C}([\ell_0, \ell])$ by a cubic spline interpolant Γ_ℓ^h of a finite set of crack tips $\mathscr{A}_\ell = \{x_n\}_{n=0}^{(\ell-\ell_0)/\Delta\ell}$, where ℓ_0 is the initial crack length and $\Delta\ell > 0$ is a crack discretization parameter. For a discrete crack Γ_ℓ^h, ℓ indicates the chord length (the length along the polygonal line formed by points in $\mathscr{A}_\ell$, plus the initial crack length ℓ_0) instead of its length. At any value of $\ell \geq \ell_0$, the algorithm proceeds through the following three steps:

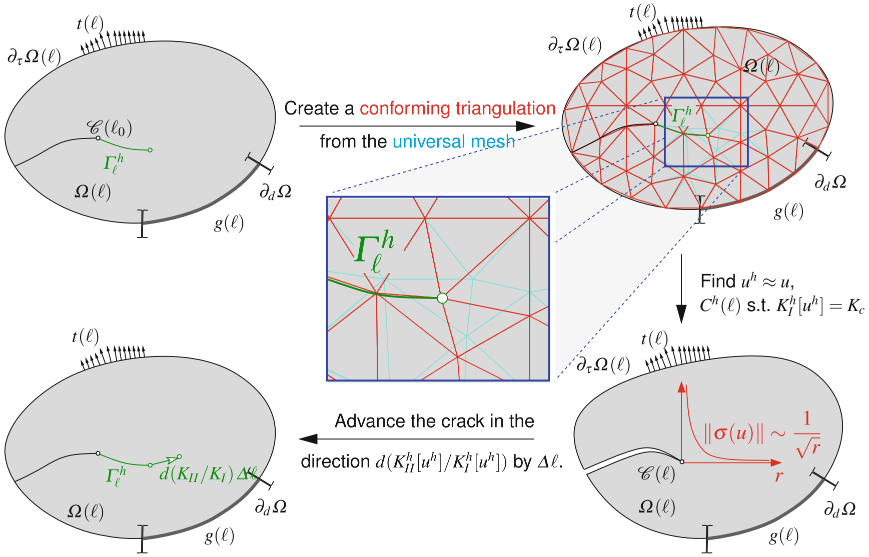


Fig. 3 The critical steps in the crack advancement algorithm

1. Generate a conforming triangulation to the crack Γ_ℓ^h.
2. Find $\boldsymbol{u}^h(\ell) \approx \boldsymbol{u}(\ell)$ and $C^h(\ell) \approx C(\ell)$.
3. Advance the crack in the direction $\boldsymbol{d}(K_{II}^h[\boldsymbol{u}^h]/K_I^h[\boldsymbol{u}^h])$ by $\Delta\ell$ (namely $\mathscr{A}_{\ell+\Delta\ell} = \mathscr{A}_\ell \cup \{x_{(\ell-\ell_0)/\Delta\ell} + \boldsymbol{d}(K_{II}^h[\boldsymbol{u}^h]/K_I^h[\boldsymbol{u}^h])\,\Delta\ell\}$).

Here the direction $d : \mathbb{R} \to S^1$ is chosen such that an infinitesimally short kink at the chosen angle satisfies $K_{II} = 0$ up to first order in the kink angle itself. Evaluating d in this way sidesteps the computationally intensive alternative of explicitly solving for the direction d, but it likely restricts the order of convergence of the algorithm.

3.3 Examples

We next showcase the application of the algorithm for the simulation of brittle fracture to two examples. The first example is a crack propagating along a circular arc, which we compare against an exact solution, and the second is a crack propagating in a perforated plate undergoing three-point bending, which we compare against experimental results. We also show (preliminary) results on the application of a slight modification of this algorithm to the computation of crack paths in a rapidly quenched plate [8].

3.3.1 A Crack Propagating Along a Circular Arc

The displacement and stress fields of an infinite medium that contains a crack shaped as a circular arc subjected to far-field stresses and traction-free faces was computed in [25]. The corresponding stress intensity factors can be found in [11]. We use this solution to construct a loading history that, when is applied as Dirichlet boundary conditions to a square-shaped domain, as illustrated in Fig. 4a, causes the crack to

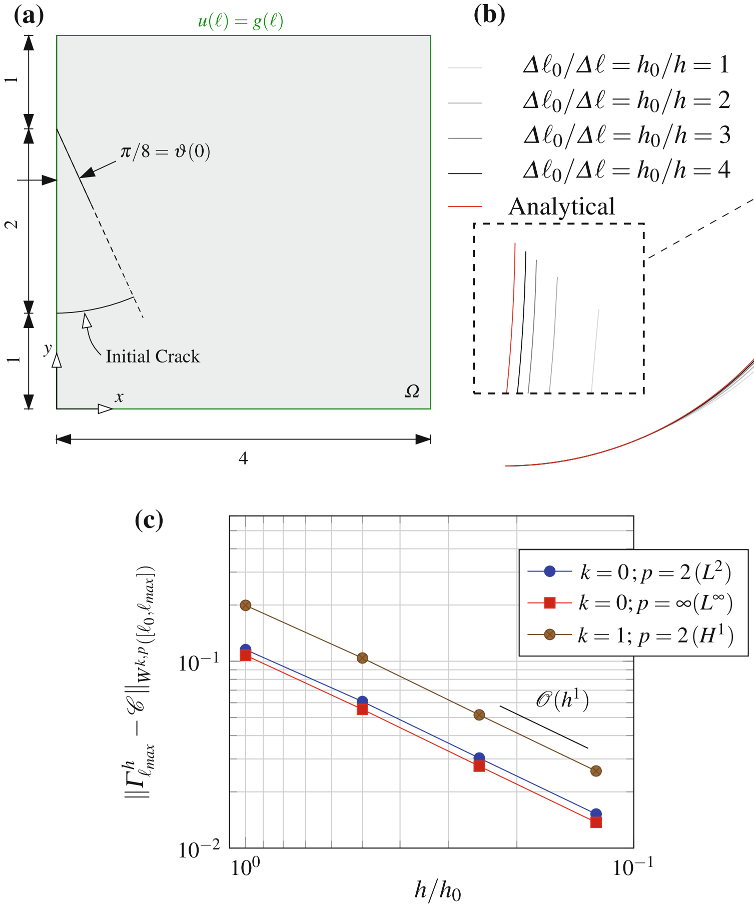


Fig. 4 The problem of a crack propagating along a circular arc. By imposing appropriate boundary conditions $\boldsymbol{g}(\ell)$, with the knowledge of the analytical solution as a function of the angle subdued by the crack, and hence ℓ, we can guide the crack to propagate along a circular arc. This problem provides a benchmark to establish convergence of the computed crack paths. Four level of refinements were used to perform the convergence of the computed crack path and convergence rates were observed to be of the order $\mathscr{O}(h^1)$. **a** Modeled problem, **b** convergence of the solution, **c** convergence of the crack path

propagate along a circular arc. For details on the construction of such a loading history we refer the interested reader to [35].

Figure 4a shows a square-shaped domain Ω with a pre-existing crack of radius $R = 2$ and angular span $\vartheta_0 = \pi/8$.The analyses were carried out with four progressively refined universal meshes. The coarsest background mesh as well as the conformed mesh are shown in Fig. 5, and their refinements were constructed by recursively subdividing each triangle of the background mesh into four similar ones. The ratio $\Delta\ell/h \approx 2$ was kept constant over all simulations, where as usual h denotes the maximum diameter of an element in a triangulation. As shown in Fig. 4b, the crack path converges to a circular arc, and the convergence curves for the $L^p([\ell_0, \ell_{max}]),\ p = 2, \infty$ and $H^1([\ell_0, \ell_{max}])$ norms are shown in Fig. 4c. Notably, convergence of the tangents to the crack path is also obtained.

This simple, nonetheless illustrative example, suggests that the algorithm is indeed convergent, and hence that the computed paths are largely independent of the chosen mesh.

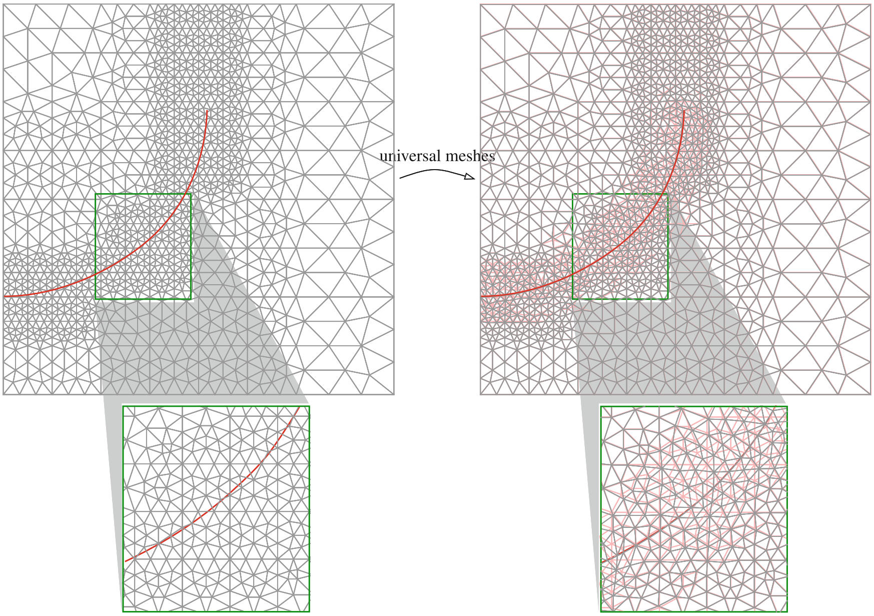
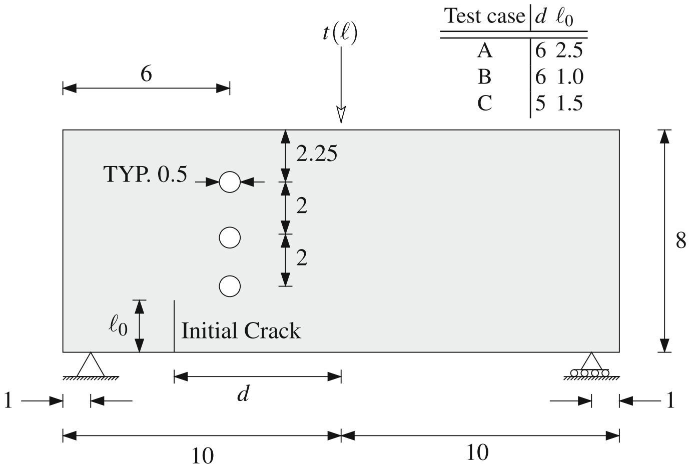


Fig. 5 In reference to the problem of a crack propagating along a circular arc, as presented in Sect. 3.3.1, we showcase the universal mesh used for the entire crack propagation simulation (*left*). Namely, the single background mesh (*left*) was used to generate a conforming triangulation to the family of cracks $\{\Gamma_\ell^h\}_\ell$. On the *right* we showcase the conformed triangulation at the end of the simulation with the background mesh in *red*

3.3.2 Perforated Plate

We next present the problem of a perforated plate undergoing three-point-bending. The problem setup is illustrated in Fig. 6. We performed the simulations for three configurations of the initial crack position (d) and length (ℓ_0). The values are tabulated in Fig. 6. In Fig. 7 we illustrate a universal mesh employed for one of the three simulations. It is worthwhile to note the adaptive nature of the background triangulation; in fact universal meshes can be easily adopted in conjunction with adaptive refinement. For each of the three simulations we generated a slightly different universal mesh to comply with the varying location of the initial crack.

Experimental results for this test setup are available in [5, 20]. The experiments were performed on polymethyl methacrylate (PMMA) plates. A comparison of the computed crack paths with digitized points from [5, 20] show a good agreement with experimental results. Further, relative convergence studies were performed on the computed crack paths, and the results, that can be found in [35], show a similar behavior to the one observed in Fig. 4.



Test case	d	ℓ_0
A	6	2.5
B	6	1.0
C	5	1.5

Fig. 6 Geometry for a plate with holes

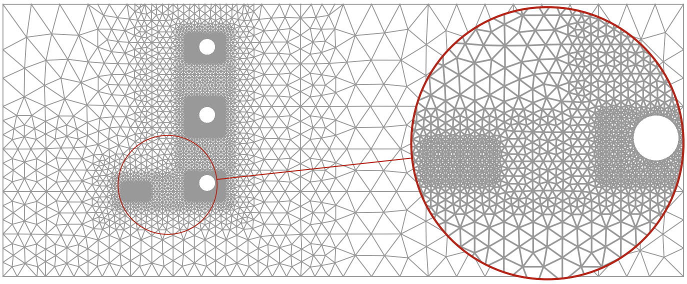


Fig. 7 Universal mesh for a plate with holes

3.3.3 Crack Path Instabilities in a Quenched Plate

Lastly we concisely present the problem of a thermally driven fracture in a quenched plate. The problem consists of a plate of finite width cracked along its center line, with low toughness (K_c), that, after being heated to temperature θ^+, is immersed in an ice bath at temperature θ^- with a constant velocity (v). Refer to Fig. 9 to supplement the above description (Fig. 8).

Depending on the material parameters, the presence of small deviations from the idealized descriptions above, and the configuration of the experiment, the crack path is expected to develop oscillations. Figure 10 showcases the results of experiments performed by [38]. In Fig. 11 we showcase some snapshots of the computed crack path for one set of inputs. These were computed through a modification to time-dependent problems of the algorithm introduced here. Details will appear in [8].

Although not shown here, the crack paths are converged up to a small tolerance. In our experience, this problem benefitted immensely from the high-order computation of the stress intensity factors; our previous attempts with low-order methods required excessively refined meshes to begin displaying some form of mesh-independent results. We hope to use this platform to better understand the underlying physics.

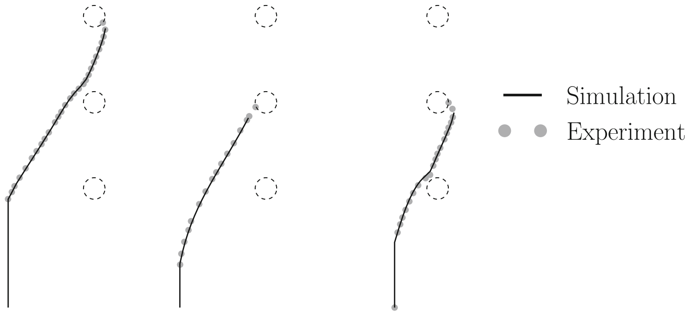


Fig. 8 Comparison between experimental results and computed crack paths for the problem of three-point-bending of a perforated plate. Experimental results were digitized from [5, 20]

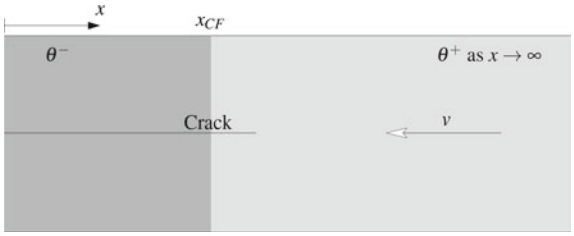


Fig. 9 Quenched plate problem setup

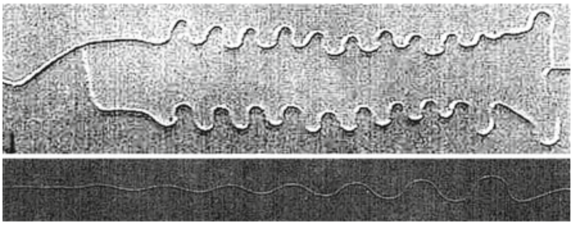


Fig. 10 Representative results of experiments of wavy crack patterns in rapidly quenched plates [38]. In both cases shown above, crack propagation along a straight crack is unstable. These cases correspond to different immersion speeds

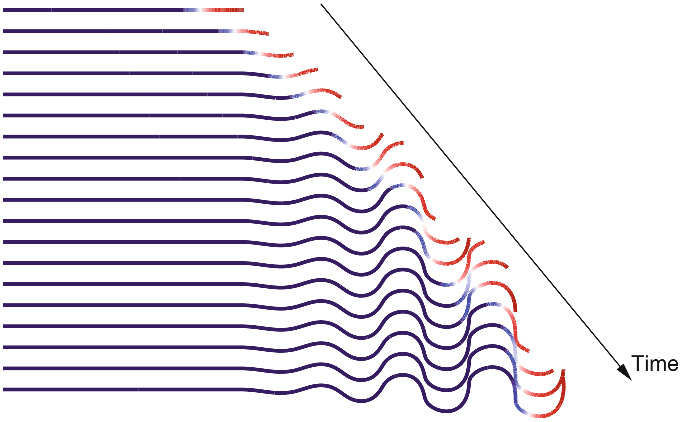


Fig. 11 Computed evolution of a thermally driven crack in a quenched plate. The contours show the temperature profile along the crack with *dark blue* representing θ^- and *dark red* representing θ^+

4 Beyond Brittle Fracture: Moving Boundary Problems

In addition to crack propagation, a variety of problems in science and engineering involve partial differential equations posed on domains that change with time. Such problems, collectively referred to as *moving-boundary problems*, appear in studies of fluid-structure interaction, phase-transitions, free-surface flows, aeroelasticity, and biolocomotion, to name a few. In this section, we demonstrate the applicability of universal meshes to this broader class of problems.

4.1 Examples: Flow Past Moving Obstacles

In the setting of fluid-structure interaction, universal meshes provide a conforming discretization of the fluid domain at all times, allowing finite element spaces of any desired order of accuracy to be used to spatially discretize the Navier-Stokes equations. This conforming discretization can be made to deform smoothly over time intervals that are short in comparison to the mesh spacing, thereby allowing standard numerical integrators to be used to solve the resulting system of ordinary differential equations. A projector (such as the nodal interpolation operator) is then used to transfer information between finite element spaces each time nodal positions change discontinuously. Details of this procedure are given in [13, 16], and rigorous theoretical bounds for a wide class of linear problems guarantee the high-order nature of the resulting numerical scheme when high-order finite elements are adopted [14, 15].

As an example, we consider in Fig. 12 the solution of incompressible, viscous flow past a rotating propeller at Reynolds number $Re = 290$. We solved the problem using a universal mesh having adaptive refinement in a neighborhood of the propeller, together with Taylor-Hood finite elements. Figure 12 shows contours of the vorticity at two instants in time. The robust nature of the method is patent in this example, as traditional deforming-mesh methods could easily encounter difficulties with mesh entanglement upon rotation of the propeller.

As a second example, we consider in Fig. 12 the solution of incompressible, viscous flow past a pair of NACA0015 airfoils that change their pitch sinusoidally in time. We solved the problem using a universal mesh together with Taylor-Hood finite elements. For simplicity, the tips of the airfoils were blunted so that the algorithm described in Sect. 2 (which applies to smooth geometries) could be applied in its most basic form. Figure 12 shows contours of the vorticity at two instants in time corresponding to the maximum and minimum pitch ($17°$ and $-17°$, respectively) of the airfoils.

Finally, we consider the solution of incompressible, viscous flow past an oscillating disk with unit diameter at Reynolds number $Re = 185$. We solved the problem using a universal mesh having adaptive refinement near the disk (see Fig. 13a), together with Taylor-Hood finite elements [16]. The disk's motion was prescribed using a sinusoidally varying vertical displacement with amplitude 0.2 and frequency equal to 0.8 times the natural shedding frequency of a fixed disk of the same diameter. Figure 13b shows a snapshot of the contours of the vorticity. Figure 13c shows the observed convergence of the drag and lift coefficient time series under refinement of the mesh, which were computed via direct integration over the boundary of the disk.

5 Outlook

Clearly for a universal mesh to be useful in engineering practice, it needs to be able to handle evolving geometries in three-dimensions. We show next some incipient results in this direction.

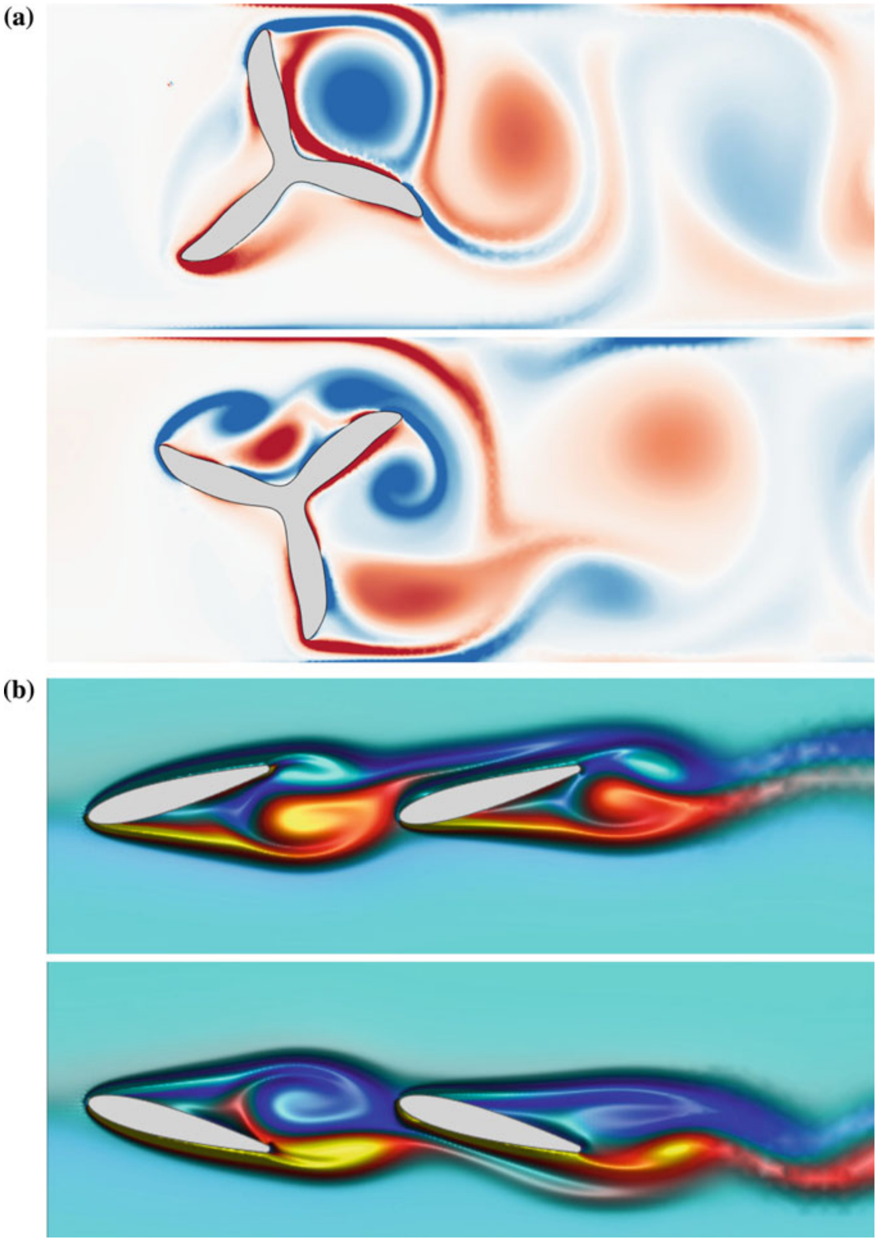


Fig. 12 Vorticity contours for two representative examples of flow past a moving obstacle. The simulations consist of incompressible viscous flow, computed using a universal mesh together with Taylor-Hood finite elements. **a** Flow past a rotating propeller, **b** flow past pitching airfoils

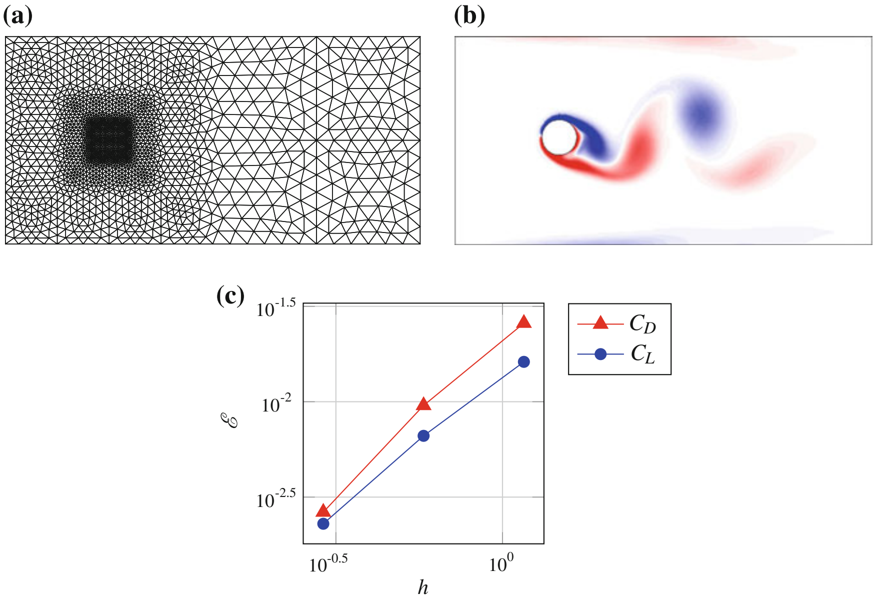


Fig. 13 Numerical simulation of incompressible, viscous flow past an oscillating disk using a universal mesh. In **a**, the background mesh adopted during the simulation is shown. In **b**, a snapshot of vorticity contours are shown. In **c**, the convergence of the drag and lift coefficient time series under mesh refinement is shown. The reported error $\mathscr{E}$ is the square root of the integrated squared error $(C_i(t) - \bar{C}_i(t))^2$, $i = L, D$, over the time interval [0, 1], relative to a reference solution $\bar{C}_i(t)$ obtained from a fine mesh with $h = 0.145$. Nearly quadratic convergence is observed. See [16] for details

5.1 *Universal Meshes for Smooth Three-Dimensional Domains*

The construction of a universal mesh in three dimensions follows the steps described in Sect. 2. Namely, given a smooth closed surface $\Gamma \subset \mathscr{D} \subset \mathbb{R}^3$ immersed in a mesh of tetrahedra $\mathscr{T}_h$, we first identify a set of faces Γ_h in $\mathscr{T}_h$ that lie near Γ. These faces are then mapped onto Γ via the closest point projection, and nearby nodes are adjusted via a relaxation step that ensures the quality of the resulting mesh.

In analogy with the algorithm presented in Sect. 2, Γ_h is chosen as the union of *positive faces* of positively cut tetrahedra in $\mathscr{T}_h$. A tetrahedron in $\mathscr{T}_h$ is called *positively cut* if it has three nodes on the non-negative side of Γ and one on the negative side. A face is then called a *positive face* if it belongs to a positively cut tetrahedron and all three of its vertices lie on the non-negative side of Γ. The closest point projection defines a one-to-one mapping between a positive face and its image on Γ provided that the mesh size is small compared to the local radius of curvature, and more importantly, provided some dihedral angles in the mesh are acute [21].

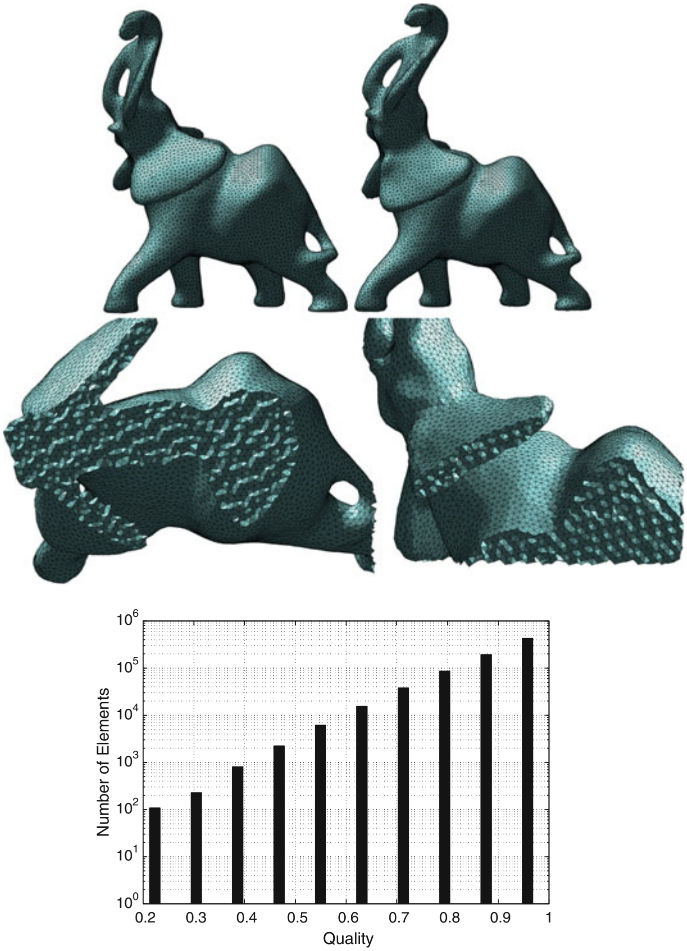


Fig. 14 The *top row* shows the tetrahedral meshes conforming to the geometry of an elephant in two different postures. They were obtained from the same background universal mesh, discarding exterior elements. In the *middle row* we show two cuts of the mesh displayed in the *top right*. In the *bottom row* we showcase the distribution of the quality of the tetrahedra on a logarithmic scale

Some examples that illustrate the use of a universal mesh in three dimensions are given in Figs. 14 and 15. In Fig. 14, two meshes of tetrahedra conforming to an elephant undergoing changes in its posture were obtained from a single universal mesh. In Fig. 15, the same procedure was used to construct conforming meshes of tetrahedra of a human upper airway.

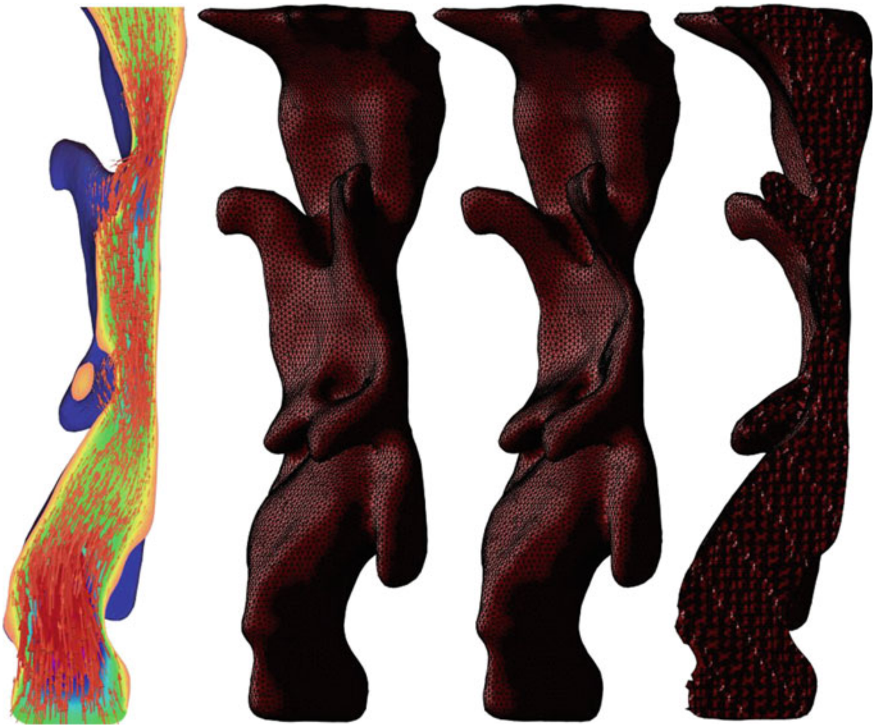


Fig. 15 The two figures in the *middle* show conforming meshes for two different configurations of a human upper airway. The figures on the *far left* and the *far right* show a cut through each tetrahedral mesh. We used a CT scan of a patient as an input for this example. The figure on the *far left* shows the contours of the velocity field computed by solving the Navier-Stokes equations inside. Coupling this tool to a solid mechanics analysis code for the upper airway would be useful to study the collapse of the upper airway, quite often the area of interest for patients diagnosed with obstructive sleep apnea

5.2 Universal Meshes for Evolving Curves on Surfaces

With an eye towards evolving crack fronts in three dimensions, we show next some early results on how a universal mesh can conform to a smooth curve drawn over a smooth surface, triangulating the interior of the curve over the surface, and conforming the tetrahedra to the surface and the mesh as well, see Fig. 16. To do so, a planar parametrization of the surface was constructed, and a smooth approximation of the given curve immersed in it. A conforming surface triangulation to the curve was then achieved by using a modification of the algorithm in two dimensions, not described here, and mapping the resulting planar triangulation back to the surface.

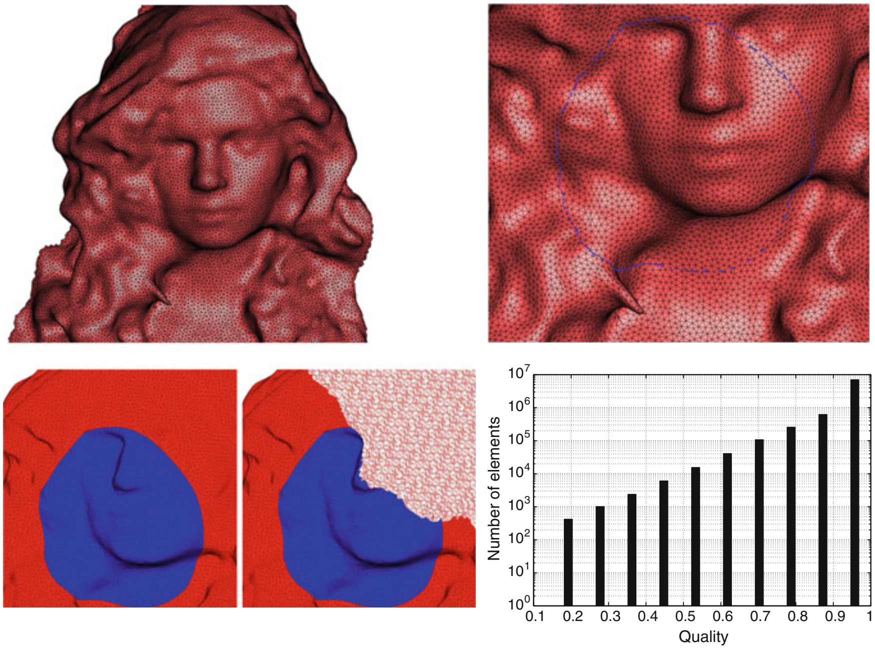


Fig. 16 A conforming mesh of the given surface is generated from the background universal mesh, as shown in the *top left* and *top right* figures. We then construct a planar parametrization from the corresponding surface triangulation, and conform the mapped mesh to the surface in the parametric planar domain. Mapping then back to the real space we obtain a surface triangulation that conforms to the curve on the surface, and after a relaxation step of the nodes near the surface to ensure good quality of the tetrahedra, we obtain the resulting mesh (*bottom row*). The distribution of the qualities of tetrahedra in this mesh is shown at the *bottom right*

Acknowledgments This work was supported by the Office of Technology Licensing Stanford Graduate Fellowship to Maurizio M. Chiaramonte, the National Science Foundation Graduate Research Fellowship to Evan S. Gawlik, and the Franklin P. Johnson Jr. Stanford Graduate Fellowship to Hardik Kabaria. Adrian J. Lew acknowledges the support of National Science Foundation; contract/grant number CMMI-1301396.

References

1. Angelillo, M., Babilio, E., & Fortunato, A. (2012). Numerical solutions for crack growth based on the variational theory of fracture. *Computational Mechanics*, *50*(3), 285–301.
2. Armando Duarte, C., & Tinsley Oden, J. (1996). An h-p adaptive method using clouds. *Computer Methods in Applied Mechanics and Engineering*, *139*(1–4), 237–262.
3. Azócar, D., Elgueta, M., & Rivara, M. C. (2010). Automatic LEFM crack propagation method based on local Lepp-Delaunay mesh refinement. *Advances in Engineering Software*, *41*, 111–119.

4. Belytschko, T., & Black, T. (1999). Elastic crack growth in finite elements with minimal remeshing. *International Journal for Numerical Methods in Engineering*, *45*(5), 601–620.
5. Bittencourt, T. N., Wawrzynek, P. A., Ingraffea, A. R., & Sousa, J. L. (1996). Quasi-automatic simulation of crack propagation for 2D LEFM problems. *Engineering Fracture Mechanics*, *55*, 321–334.
6. Bourdin, B., Francfort, G. A., & Marigo, J.-J. (2000). Numerical experiments in revisited brittle fracture. *Journal of the Mechanics and Physics of Solids*, *48*(4), 797–826.
7. Camacho, G. T., & Ortiz, M. (1996). Computational modelling of impact damage in brittle materials. *International Journal of Solids and Structures*, *33*(20–22), 2899–2938.
8. Chiaramonte, M. M., Keer, L. M., & Lew, A. J. (2015). Crack path instabilities in thermoelasticity.
9. Chiaramonte, M. M., Shen, Y., Keer, L. M., & Lew, A. J. (2015). Computing stress intensity factors for curvilinear cracks. *International Journal For Numerical Methods in Engineering*.
10. Chiaramonte, M. M., Shen, Y., & Lew, A. J. (2015). The h-version of the method of auxiliary mapping for higher order solutions of crack and re-entrant corner problem.
11. Cotterell, B., & Rice, J. R. (1965). Slightly curved or kinked cracks. *International Journal of Fracture Mechanics*, *16*, 155–168.
12. Fraternali, F. (2007). Free discontinuity finite element models in two-dimensions for in-plane crack problems. *Theoretical and Applied Fracture Mechanics*, *47*(3), 274–282.
13. Gawlik, E. S., & Lew, A. J. (2014). High-order finite element methods for moving boundary problems with prescribed boundary evolution. *Computer Methods in Applied Mechanics and Engineering*, *278*, 314–346.
14. Gawlik, E. S., & Lew, A. J. (2015). Supercloseness of orthogonal projections onto nearby finite element spaces. *Mathematical Modelling and Numerical Analysis*, *49*, 559–576.
15. Gawlik, E. S. & Lew, A. J. (2015). Unified analysis of finite element methods for problems with moving boundaries. *SIAM Journal on Numerical Analysis*.
16. Gawlik, E. S., Kabaria, H., & Lew, A. J. (2015). High-order methods for low reynolds number flows around moving obstacles based on universal meshes. *International Journal for Numerical Methods in Engineering*.
17. Germanovich, L., Murdoch, L. C., & Robinowitz, M. (2014). Abyssal sequestration of nuclear waste and other types of hazardous waste.
18. Giacomini, A., & Ponsiglione, M. (2006). Discontinuous finite element approximation of quasistatic crack growth in nonlinear elasticity. *Mathematical Models and Methods in Applied Sciences*, *16*(01), 77–118.
19. Gol'dstein, R. V., & Salganik, R. L. (1974). Brittle fracture of solids with arbitrary cracks. *International Journal of Fracture*, *10*(4), 507–523.
20. Ingraffea, A. R., & Grigoriu, M. (1990). Probabilistic fracture mechanics: A validation of predictive capability. Technical report, Cornell University.
21. Kabaria, H., & Lew, A. J. (In preparation). Universal meshes for smooth surfaces with no boundary in three dimensions. *International Journal for Numerical Methods in Engineering*.
22. Melenk, J. M., & Babuška, I. (1996). The partition of unity finite element method: Basic theory and applications. *Computer Methods in Applied Mechanics and Engineering*, *139*(1–4), 289–314.
23. Miehe, C., & Gürses, E. (2007). A robust algorithm for configurational-force-driven brittle crack propagation with r-adaptive mesh alignment. *International Journal for Numerical Methods in Engineering*, *72*(2), 127–155.
24. Moes, N., Dolbow, J., & Belytschko, T. (1999). A finite element method for crack growth without remeshing. *International Journal for Numerical Methods in Engineering*, *46*, 131–150.
25. Muskhelishvili, N. I. (Ed.). (1977). *Some Basic Problems of the Mathematical Theory of Elasticity: Fundamental Equations, Plane Theory of Elasticity, Torsion, and Bending (translated from Russian)* (2nd ed.). Leyden, The Netherlands: Noordhoff International Publishing.
26. Negri, M. (2005). A discontinuous finite element approach for the approximation of free discontinuity problems. *Advances in Mathematical Sciences and Applications*, *15*, 283–306.

27. Ortiz, M. (1996). Computational micromechanics. *Computational Mechanics*, *18*(5), 321–338.
28. Pandolfi, A., & Ortiz, M. (1998). Solid modeling aspects of three-dimensional fragmentation. *Engineering with Computers*, *14*(4), 287–308.
29. Pandolfi, A., & Ortiz, M. (2012). An eigenerosion approach to brittle fracture. *International Journal for Numerical Methods in Engineering*, *92*(8), 694–714.
30. Pandolfi, A., Krysl, P., & Ortiz, M. (1999). Finite element simulation of ring expansion and fragmentation: The capturing of length and time scales through cohesive models of fracture. *International Journal of Fracture*, *95*(1–4), 279–297.
31. Pandolfi, A., Li, B., & Ortiz, M. (2012). Modeling fracture by material-point erosion. *International Journal of Fracture*, *184*(1–2), 3–16.
32. Phongthanapanich, S., & Dechaumphai, P. (2004). Adaptive Delaunay triangulation with object-oriented programming for crack propagation analysis. *Finite Elements in Analysis and Design*, *40*, 1753–1771.
33. Rangarajan, R. & Lew, A. J. (2013). Analysis of a method to parameterize planar curves immersed in triangulations. *SIAM Journal on Numerical Analysis*, *51*(3), 1392–1420.
34. Rangarajan, R., & Lew, A. J. (2014). Universal meshes: A method for triangulating planar curved domains immersed in nonconforming triangulations. *International Journal for Numerical Methods in Engineering*, *98*(4), 236–264.
35. Rangarajan, R., Chiaramonte, M. M., Hunsweck, M. J., Shen, Y., & Lew, A. J. (2014). Simulating curvilinear crack propagation in two dimensions with universal meshes. *International Journal for Numerical Methods in Engineering*.
36. Ruiz, G., Pandolfi, A., & Ortiz, M. (2001). Three dimensional cohesive modeling of dynamic mixed-mode fracture. *International Journal for Numerical Methods in Engineering*, *52*(12), 97–120.
37. Schmidt, B., Fraternali, F., & Ortiz, M. (2009). Eigenfracture: An eigendeformation approach to variational fracture. *Multiscale Modeling & Simulation*, *7*(3), 1237–1266.
38. Yuse, A., & Sano, M. (1997). Instabilities of quasi-static crack patterns in quenched glass plates. *Physica D: Nonlinear Phenomena*, *108*(4), 365–378.
39. Zielonka, M. G., Ortiz, M., & Marsden, J. E. (2008). Variational r-adaption in elastodynamics. *International Journal for Numerical Methods in Engineering*, *74*(7), 1162–1197.

Free Energy, Free Entropy, and a Gradient Structure for Thermoplasticity

Alexander Mielke

Abstract In the modeling of solids the free energy, the energy, and the entropy play a central role. We show that the free entropy, which is defined as the negative of the free energy divided by the temperature, is similarly important. The derivatives of the free energy are suitable thermodynamical driving forces for reversible (i.e. Hamiltonian) parts of the dynamics, while for the dissipative parts the derivatives of the free entropy are the correct driving forces. This difference does not matter for isothermal cases nor for local materials, but it is relevant in the non-isothermal case if the densities also depend on gradients, as is the case in gradient thermoplasticity. Using the total entropy as a driving functional, we develop gradient structures for quasistatic thermoplasticity, which again features the role of the free entropy. The big advantage of the gradient structure is the possibility of deriving time-incremental minimization procedures, where the entropy-production potential minus the total entropy is minimized with respect to the internal variables and the temperature. We also highlight that the usage of an auxiliary temperature as an integrating factor in [30] serves exactly the purpose to transform the reversible driving forces, obtained from the free energy, into the needed irreversible driving forces, which should have been derived from the free entropy. This reconfirms the fact that only the usage of the free entropy as driving functional for dissipative processes allows us to derive a proper variational formulation.

1 Introduction

The mathematical theory of plasticity has its origin in the 1970 s based on the work Moreau [22], Johnson [12], and Gröger [10], which treated the small-strain case with quadrtic energies and fixed elastic domains. They developed a rich theory based on convex analysis and monotone operators which allowed for significant generalizations, but still staying in the small-strain regime, see e.g. [1]. Finite-strain elastoplas-

A. Mielke (✉)
WIAS, Mohrenstraße 39, 10117 Berlin, Germany
e-mail: alexander.mielke@wias-berlin.de

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_7

ticity also plays a fundamental role in engineering applications, and many algorithms were derived starting in the 1980s, see e.g. [14, 29]. A major breakthrough was the discovery in [24, 25] that incremental problems in finite-strain elastoplasticity can be formulated as minimization problems jointly for the elastic and the plastic updates. This means that elastoplasticity can be formulated in terms of a generalized gradient system with a dissipation potential $\mathscr{R}$ and a free energy $\mathscr{F}$ such that it reads

$$\mathrm{D}_u\mathscr{F}(t,u,z)=0,\quad 0\in\partial_{\dot z}\mathscr{R}(u,z;\dot z)+\mathrm{D}_z\mathscr{F}(t,u,z),$$

where u is the displacement, and z contains all internal (dissipative) variables. This theory even led to the first mathematical existence results for the rate-independent case, see [9, 13, 20].

However, the whole theory is restricted to the isothermal case, and it remains a challenging problem to find a corresponding mechanical and mathematical theory for thermoplasticity. There are major differences in the mechanical modeling between the isothermal and the non-isothermal case. In the former case there is one free energy, and time-incremental minimization procedures can be formulated by minimizing the sum of the free energy plus the dissipation in the time step. In the non-isothermal case one has to take care of the mechanical forces still given by the free energy, but instead of dissipation one now has to model entropy production. A time-incremental minimization procedure should involve the entropy production minus the total entropy. A first step in this direction was done in [30], and here we connect our work [16, 17] with the latter.

The major observation is that one has to formulate thermoplasticity in a suitable thermodynamically consistent way, in order to recast it in variational form. For this, we start from the GENERIC framework (General Equations for Non-Equilibrium Reversible Irreversible Coupling). For this we use the total energy $\mathscr{E}$ and the total entropy $\mathscr{S}$ as functionals with densities E and S, respectively, depending on the displacement u, the vector z of internal variables, and a thermodynamical variable r. Using the entropy-production potential $\mathscr{P}$ and its Legendre dual $\mathscr{P}^*$, we find the form (cf. [16])

$$\begin{aligned}
\rho\ddot u &= -\big(\mathrm{D}_u\mathscr{E}(q)-\Theta{*}\mathrm{D}_u\mathscr{S}(q)\big) &&= -\mathrm{D}_u\overline{\mathscr{F}}\big(u,z,\Theta(u,z,r)\big),\\
\dot z &= \partial_{\xi_z}\mathscr{P}^*_z\Big(q,\mathrm{D}_z\mathscr{S}(q)-\frac{1}{\Theta}{*}\mathrm{D}_z\mathscr{E}(q)\Big) &&= \partial_{\xi_z}\mathscr{P}^*_z\big(q,\mathrm{D}_z\overline{\mathscr{H}}(u,z,\Theta(u,z,r))\big),\\
\dot r &= -\frac{\Delta_u S(q)[\dot u]}{\partial_r S(q)}-\frac{\Delta_z E(q)[\dot z]}{\partial_r E(q)}-\frac{1}{\partial_r E}\,\mathrm{div}\Big(\kappa(q)\nabla\frac{\partial_r S}{\partial_r E}\Big),
\end{aligned}$$

where $\overline{\mathscr{F}}$ and $\overline{\mathscr{H}}$ are the total free energy and total free entropy expresses in terms of the temperature $\theta=\Theta(u,z,r)$, where r is an arbitrary scalar thermodynamic variable, such that Gibbs' relation $\theta=\Theta(u,z,r)=\frac{\mathrm{D}_r\mathscr{E}(u,z,r)}{\mathrm{D}_r\mathscr{S}(u,z,r)}$ holds.

The above form clearly shows the role of the free energy as the driving functional for the reversible elastodynamics, while the free entropy

$$\overline{\mathscr{H}}(u,z,\theta)=\int_\Omega \overline{H}(u,\nabla u,z,\nabla z,\theta)\,\mathrm{d}x \quad \text{with}\, \overline{H}(W,\theta)=-\frac{\overline{F}(W,\theta)}{\theta}$$

is the driving functional for the dissipative variables z (like the plastic tensor or the hardening variables). Locally the free entropy is simply given as 'minus the free energy divided by the temperature', but for functional derivatives, which involve integration by parts, new terms appear and the naive relation $\mathrm{D}_z\overline{\mathscr{H}}(u,z,\theta)=-\frac{1}{\theta}\mathrm{D}_z\overline{\mathscr{F}}(u,z,\theta)$ may be wrong. With $\overline{\mathscr{F}}(u,z,\theta)=\int_\Omega \overline{F}(u,\nabla u,z,\nabla z,\theta)\,\mathrm{d}x$ we have

$$\mathrm{D}_z\overline{\mathscr{H}}(u,z,\theta)=-\frac{1}{\theta}\mathrm{D}_z\overline{\mathscr{F}}(u,z,\theta)+\partial_{\nabla z}\overline{F}\big(u,\nabla u,z,\nabla z,\theta\big)\nabla\Big(\frac{1}{\theta}\Big),$$

where the last term vanishes only in two important cases: (i) in the isothermal case where $\nabla\theta\equiv 0$ and (ii) in the case "local case" where $\overline{F}$ does not depend on ∇z. In these two cases, it is correct to use the derivative of the free energy and put the factor θ into the dual entropy-production potential (thus turning it into a dual dissipation potential). However, in all other cases, one has to distinguish the free energy and the free entropy as driving functionals. Moreover, the inverse temperature $1/\theta$ is the driving force for heat transfer:

driving force for revers. dynamics :

$$\mathrm{D}_u\overline{\mathscr{F}}(u,z,\theta)=\mathrm{D}_uE(u,z,r)-\Theta(u,z,r){*}\mathrm{D}_uS(u,z,r),$$

driving force for dissip. dynamics :

$$\mathrm{D}_z\overline{\mathscr{H}}(u,z,\theta)=\mathrm{D}_zS(u,z,r)-\tfrac{1}{\Theta(u,z,r)}{*}\mathrm{D}_zE(u,z,r),$$

driving force for energy transport :

$$1/\theta=1/\Theta(u,z,r)=\tfrac{\partial_rS(u,z,r)}{\partial_rE(u,z,r)}.$$

An important fact is that the terms on the right-hand side are independent of the choice of the thermodynamical variable r, see Theorem 3, which gives a great flexibility in the mathematical approaches.

Turning to the quasistatic case, we drop the interia term $\rho\ddot{u}$ and rewrite the remaining system in the form

$$\begin{aligned}&0=\mathrm{D}_u\mathscr{E}(q)-A(q)\mathrm{D}_r\mathscr{E}(q),\quad \dot{z}=\partial_{\eta_z}\mathscr{P}_Z^*\Big(q;\mathrm{D}_z\mathscr{S}(q){-}B(q)\mathrm{D}_r\mathscr{S}(q)\Big),\\ &\dot{r}=-A(q)^*\dot{u}-B(q)^*\dot{z}-C(q)^*\operatorname{div}\big(\kappa(q)\nabla(C(q)\mathrm{D}_rS(q))\big)\\ &\text{with}\, A(q)\xi_r=\big(\tfrac{\xi_r}{\mathrm{D}_r\mathscr{S}(q)}\big)*\mathrm{D}_u\mathscr{S}(q),\ \ B(q)\xi_r=\big(\tfrac{\xi_r}{\mathrm{D}_r\mathscr{E}(q)}\big)*\mathrm{D}_z\mathscr{E}(q),\\ &\text{and}\quad C(q)\xi_r=\frac{\xi_r}{\mathrm{D}_r\mathscr{E}(q)}.\end{aligned}\tag{1}$$

Assuming that the first relation, which is static, can be solved in the form $u = \mathrm{U}(z, r)$, Theorem 4 shows that this system can be formulated as a gradient system as follows

$$\begin{pmatrix} \dot{z} \\ \dot{r} \end{pmatrix} = \partial_\xi \mathfrak{P}^*\big(z, r; \mathrm{D}\mathfrak{S}(z, r)\big) \ \text{ with } \mathfrak{S}(z, r) = \mathscr{S}(\mathrm{U}(z, r), z, r),$$

where $\mathfrak{P}^*$ is a suitably reduced dual entropy-production potential, and the reduced energy $\mathfrak{E}(z, r) = \mathscr{E}(\mathrm{U}(z, r), z, r)$ is conserved.

In general, this gradient structure is highly nonlocal, where $\mathfrak{P}^*$ involves the derivatives $\mathrm{D}_z\mathrm{U}(z, r)$ and $\mathrm{D}_r\mathrm{U}(z, r)$, and thus less useful. However, in the case $A(u, z, r) \equiv 0$, which occurs for a choice of r such that $\mathrm{D}_u\mathscr{S}(u, z, r) \equiv 0$, one obtains a system that allows for local approaches. Thus, the freedom of choosing r as freely as possible is essential. For that case, we propose the time-incremental minimization procedure

$$\begin{aligned} \begin{pmatrix} z^{k+1} \\ r^{k+1} \end{pmatrix} &\in \underset{(z,r)}{\mathrm{Arg\,min}} \Big\{ \tau\, \mathscr{P}\Big(q^k; \frac{1}{\tau}\begin{pmatrix} z-z^k \\ r-r^k \end{pmatrix}\Big) - \mathscr{S}(u^k, z, r) \Big\}, \\ u^{k+1} &\in \underset{u}{\mathrm{Arg\,min}}\, \mathscr{E}(u, z^{k+1}, r^{k+1}), \end{aligned} \tag{2}$$

where $\mathscr{P}$ is the primal entropy-production potential obtained from the dual potential $\mathscr{P}^*(u, z, r; \xi_z, \xi_r) = \mathscr{P}^*_Z(q; \xi_z - B(q)\xi_r) + \frac{1}{2}\int_\Omega \nabla(C(q)\xi_r)\cdot\kappa(q)\nabla(C(q)\xi_r)\,\mathrm{d}x$ via Legendre-Fenchel transform.

We discuss the abstract theory along specific thermomechanical examples. The simplest is a spring-damper system, see Examples 5 and 7. Section 4.1 discusses the Penrose-Fife model and shows how in [21] the gradient structure is exploited to do a rigorous homogenization, where the effective entropy functional is obtained by averaging of the free energy density. Section 4.2 treats a time-dependent thermoplastic model where the gradient structure in terms of the entropy involves a time-dependent entropy-production because of the elimination $u(t) = \mathbf{U}(z(t), \ell(t))$, where ℓ is the mechanical loading. Finally, a plastic model with thermal expansion is considered in Sect. 4.3.

For all these models we need a specific and problem-dependent choice of the thermodynamic variable r, which highlights the importance of a clear modeling in terms of the free energy and free entropy giving the driving forces $\mathrm{D}_u E(u, z, r) - \Theta(u, z, r) * \mathrm{D}_u S(u, z, r)$, $\mathrm{D}_z S(u, z, r) - \frac{1}{\Theta(u,z,r)} * \mathrm{D}_z E(u, z, r)$, and $1/\theta = \frac{\partial_r S(u,z,r)}{\partial_r E(u,z,r)}$.

2 Heat Equation as a Starting Example

As a first example we treat the heat equation with energy density e and heat flux $\mathbf{q}$:

$$\dot{e} + \operatorname{div} \mathbf{q} = 0 \ \text{ in } \Omega, \qquad \mathbf{q}\cdot n = 0 \text{ on } \partial\Omega.$$

Subsequently, we will drop all boundary conditions (like $\mathbf{q}\cdot n = 0$) and assume that we have no-flux boundary conditions for all quantities, such that the system is thermodynamically closed. We describe the energy density by an arbitrary scalar field r, which may be the energy density e itself, the absolute temperature θ, the coldness $1/\theta$, or the entropy density s. This means that we have constitutive functions

$$\theta = \Theta(r), \quad e = E(r), \quad s = S(r).$$

Of course, by Gibbs' relation $\theta \mathrm{d}s = \mathrm{d}e$ we have the compatibility $\Theta(r) = E'(r)/S'(r)$. Often this last relation is seen as the definition of the temperature. Note that already here the inverse of the absolute temperature plays the role of an integrating factor such that $1/\theta\ \mathrm{d}e$ is the total differential $\mathrm{d}s$, cf. [5].

In the classical form of the heat equation, the heat flux $\mathbf{q}$ is a linear function of the temperature gradient, which is called Fourier's law. In terms of r we arrive at

$$E'(r)\dot{r} - \operatorname{div}\big(k(r)\nabla\Theta(r)\big) = 0, \tag{3}$$

where $k \in \mathbb{R}^{d\times d}$ is the symmetric and positive definite heat conductivity matrix. However, for a proper coupling to other mechanical effects, we want to have a gradient flow formulation in terms of the total entropy $\mathscr{S}$ as a driving functional, while the total energy $\mathscr{E}$ should be conserved

$$\mathscr{S}(r) = \int_\Omega s(r(x))\,\mathrm{d}x \quad \text{and} \quad \mathscr{E}(r) = \int_\Omega E(r(x))\,\mathrm{d}x.$$

Hence, an *entropic gradient structure* must have the form

$$\dot{r} = \mathbb{K}(r)\mathrm{D}\mathscr{S}(r), \tag{4}$$

where $\mathbb{K}$ is a selfadjoint positive definite operator that maps the field $\xi_r = \mathrm{D}\mathscr{S}(r)$ to the rate $\dot{r}$, where ξ_r is the thermodynamical driving force associated with r. The operator $\mathbb{K}$ will be called *Onsager operator*, since Onsager showed that such linear operators should be symmetric. Indeed, in [23] the symmetry $\mathbb{K} = \mathbb{K}^*$ is called "reciprocal relation".

The positive semidefiniteness $\langle \xi_r, \mathbb{K}(r)\xi_r\rangle \geq 0$ guarantees that the second law of thermodynamics is satisfied. Note that energy conservation needs the relation $\mathbb{K}(r)\mathrm{D}\mathscr{E}(r) \equiv 0$. Using the variational derivative $\mathrm{D}\mathscr{E}(r) \equiv E'(r)$ we see that the only Onsager operators which are compatible with the usual heat equation (3) and the gradient structure (4) have the form

$$\mathbb{K}(r)\xi_r = -\frac{1}{E'(r)}\operatorname{div}\Big(\kappa(r)\nabla\big(\frac{\xi_r}{E'(r)}\big)\Big),$$

where $\kappa(r) \in \mathbb{R}^{d\times d}_{\mathrm{spd}}$ still can be chosen suitably.

As a result we see that the heat equation takes the general structure

$$\dot r = -\frac{1}{E'(r)}\,\mathrm{div}\Big(\kappa(r)\nabla\Big[\frac{S'(r)}{E'(r)}\Big]\Big) = \mathbb{K}(r)\mathrm{D}\mathscr{S}(r),$$

since $\mathrm{D}\mathscr{S}(r) \equiv S'(r)$. In this general form we see that $S'(r)/E'(r) = 1/\Theta(r)$ is the function under the spatial gradient, i.e. the heat flux has the form

$$\mathbf{q} = \kappa(r)\nabla\big(1/\theta\big) = -k\nabla\theta \quad \text{with } \kappa(r) = k(r)\Theta(r)^2.$$

Thus, we see that κ has to be chosen as $k(r)\Theta(r)^2$. More importantly, we see that the inverse temperature $1/\theta$ is the *driving force for energy flow*, independently of the choice of the scalar thermodynamical variable r.

To connect our theory to the work in [30] we introduce the *dual entropy-production potential (EPP)*, also called kinetic potential there, namely

$$\mathscr{P}^*(r,\xi) := \frac{1}{2}\langle\xi,\mathbb{K}(r)\xi\rangle = \frac{1}{2}\int_\Omega \nabla\Big(\frac{\xi}{E'(r)}\Big)\cdot\kappa(r)\nabla\Big(\frac{\xi}{E'(r)}\Big)\,\mathrm{d}x.$$

By Legendre transform we can define the (primal) entropy-production potential via

$$\mathscr{P}(r,\dot r) = \sup_\xi\big(\langle\xi,\dot r\rangle - \mathscr{P}^*(r,\xi)\big) \;=\; \sup_\xi\int_\Omega \xi\dot r - \frac{1}{2}\nabla\tfrac{\xi}{E'}\cdot\kappa\nabla\tfrac{\xi}{E'}\,\mathrm{d}x,$$

which is nonlocal in $\dot r$, since the maximizer ξ is obtained by solving the elliptic partial differential equation $-\,\mathrm{div}\big(\kappa\nabla(\xi/E')\big) = \dot r E'$.

The gradient flow $\dot r = \mathbb{K}(r)\mathrm{D}\mathscr{S}(r)$ can be rewritten in the four fully equivalent forms:

$$\begin{aligned} &\text{(i) } \dot r = \partial_\xi\mathscr{P}^*(r,\mathrm{D}\mathscr{S}(r)), &&\text{(ii) } \dot r = \mathrm{Arg\,min}_v\,\mathscr{P}(r,v) - \langle\mathrm{D}\mathscr{S}(r),v\rangle,\\ &\text{(iii) } \partial_{\dot r}\mathscr{P}(r,\dot r) = \mathrm{D}\mathscr{S}(r), &&\text{(iv) } \mathrm{D}\mathscr{S}(r) \in \mathrm{Arg\,min}_\xi\,\mathscr{P}^*(r,\xi) - \langle\xi,\dot r\rangle. \end{aligned}$$

Here the equivalence of (i) and (iii) is the Fenchel equivalence for the Legendre transformation, while (ii) and (iv) are simply equivalent to (i) and (iii), respectively, using the convexity of the EPPs $\mathscr{P}$ and $\mathscr{P}^*$. To calculate the rate $\dot r$ from the nonlocal minimum principle (ii), the following local inf-sup formulation was introduced in [30]:

$$(\widehat{\xi},\dot r) \;\text{ solves }\; \inf_v\Big(\sup_\xi\Big[\langle\xi,v\rangle - \mathscr{P}^*(r,\xi) - \langle\mathrm{D}\mathscr{S}(r),v\rangle\Big]\Big).$$

3 Non-isothermal Dissipative Material Models

We now consider general elastic materials with internal parameters describing dissipative effects such as plasticity, phase transformation, damage, magnetization, or polarization, see e.g. [20]. We follow the approach presented in [16] but do not emphasize the very useful framework GENERIC explicitly. Nevertheless we will see remainders of the reversible (i.e. Hamiltonian) dynamics in the quasistatic elastic force balance and of the irreversible dynamics in the flow rules for the internal variable z and the heat equation.

We consider a body in the reference configuration Ω, which is a bounded domain with Lipschitz boundary. The displacement is denoted by $u : \Omega \subset \mathbb{R}^d \to \mathbb{R}^d$, and $\mathbf{e}(u) = \frac{1}{2}(\nabla u + \nabla u^{\mathsf{T}})$ is the linearized strain tensor. All the internal variables (also called dissipative variables) are included in the variable $z : \Omega \to \mathbb{R}^m$, which may include plastic strains, phase indicators, or damage variables. By a general scalar field $r : \Omega \to \mathbb{R}$ we describe the thermodynamical properties, e.g. r can be either the temperature θ, the internal energy density e, or the entropy density s.

We consider a closed system, which means that we have no-flux boundary conditions. The total energy and total entropy are given by

$$\mathscr{E}_{\text{kin-pot}}(u, \dot{u}, z, r) = \mathscr{E}_{\text{kin}}(\dot{u}) + \mathscr{E}(u, z, r) \ \text{ and } \ \mathscr{S}(u, z, r) = \int_\Omega S(\nabla u, z, \nabla z, r)\,\mathrm{d}x$$

$$\text{where } \mathscr{E}_{\text{kin}}(\dot{u}) := \int_\Omega \frac{\rho}{2}|\dot{u}|^2\,\mathrm{d}x \text{ and } \mathscr{E}(u, \dot{u}, z, r) = \int_\Omega E(u, \nabla u, z, \nabla z, r)\,\mathrm{d}x,$$

where the consitutive laws E and S are related by Gibbs' relation $\theta = \Theta(q) := \frac{\partial_r E(q)}{\partial_r S(q)}$.

3.1 Free Energy and Free Entropy as Driving Functionals

Before we discuss the equations of motions for such material models, we introduce the free energy f and the free entropy s and discuss their role in continuum mechanics:

free energy $f = e - \theta s$ (Gibbs 1873, Helmholtz 1882),
free entropy $h = -f/\theta = s - e/\theta$ (Massieu 1869).

As is common for the free energy, we also consider the free entropy only as a function of $r = \theta$ and use the densities (with $W = (u, \nabla u, z, \nabla z)$)

$$f = \overline{F}(W, \theta) = \overline{E}(W, \theta) - \theta\,\overline{S}(W, \theta),$$

$$h = \overline{H}(W, \theta) = -\frac{\overline{F}(W, \theta)}{\theta} = \overline{S}(W, \theta) - \frac{\overline{E}(W, \theta)}{\theta}$$

as fields on Ω and define total free energy $\overline{\mathscr{F}}$ and the total free entropy $\overline{\mathscr{H}}$ via

$$\overline{\mathscr{F}}(u,z,\theta)=\int_\Omega \overline{F}(u,\nabla u,z,\nabla z,\theta)\,\mathrm{d}x \ \text{ and } \ \overline{\mathscr{H}}(u,z,\theta)=\int_\Omega \overline{H}(u,\nabla u,z,\nabla z,\theta)\,\mathrm{d}x.$$

The major point we want to address here is that $\overline{\mathscr{F}}$ and $\overline{\mathscr{H}}$ can serve as driving functionals, since their partial derivatives with respect to any of the variables $W=(W_1,\dots,W_k)=(u,\nabla u,z,\nabla z)$ are independent of the particular choice of the thermodynamic quantity r. The physical requirement for a driving force is that it takes the same physical value, independent of the choice of the thermodynamic quantity. The main observation is the following lemma which relies on the Gibbs relation.

Lemma 1 *Consider smooth functions $E:(W,r)\mapsto E(W,r)$ and $S:(W,r)\mapsto S(W,r)$ such that $\Theta(W,r):=\partial_r E(W,r)/\partial_r S(W,r)>0$. Consider any transformation $r=R(W,\rho)$ with $\partial_r R(W,r)\neq 0$ and define*

$$\widetilde{E}(W,\rho)=E(W,R(W,\rho)) \ \text{ and } \ \widetilde{S}(W,\rho)=S(W,R(W,\rho)).$$

Then, we have the identities

$$\begin{aligned}&\mathrm{D}_W E(W,r)-\Theta(W,r)\mathrm{D}_W S(W,r)=\mathrm{D}_W\widetilde{E}(W,\rho)-\widetilde{\Theta}(W,\rho)\mathrm{D}_W\widetilde{S}(W,\rho)\\ &\text{and } \ \widetilde{\Theta}(W,\rho)=\frac{\partial_\rho\widetilde{E}(W,\rho)}{\partial_\rho\widetilde{S}(W,\rho)}=\Theta(W,R(W,\rho)) \quad \text{if } r=R(W,\rho).\end{aligned}$$

In particular, for $R(W,\rho)=\Theta(W,\theta)=\theta$ and $\overline{F}(W,\theta)=\overline{E}(W,\theta)-\theta\overline{S}(W,\theta)$ we obtain

$$\mathrm{D}_W\overline{F}(W,\theta)=\mathrm{D}_W E(W,\rho)-\theta\mathrm{D}_W S(W,\rho), \quad \text{if } \theta=\Theta(W,\rho)=\frac{\partial_\rho E(W,\rho)}{\partial_\rho S(W,\rho)}.$$

Proof For the first result, we first establish the Gibbs relation using the chain rule:

$$\frac{\partial_\rho\widetilde{E}(W,\rho)}{\partial_\rho\widetilde{S}(W,\rho)}=\frac{\partial_r E(W,R(W,r))\partial_\rho R(W,\rho)}{\partial_r S(W,R(W,r))\partial_\rho R(W,\rho)}=\frac{\partial_r E(W,R(W,r))}{\partial_r S(W,R(W,r))}=\Theta(W,R(W,\rho)).$$

For the driving forces for W we again use the chain rule to obtain

$$\begin{aligned}\mathrm{D}_W\widetilde{E}(W,\rho)&=\mathrm{D}_W E(W,R(W,\rho))+\partial_r E(W,R(W,\rho))\mathrm{D}_W R(W,\rho),\\ \mathrm{D}_W\widetilde{S}(W,\rho)&=\mathrm{D}_W S(W,R(W,\rho))+\partial_r S(W,R(W,\rho))\mathrm{D}_W R(W,\rho).\end{aligned}$$

Thus, taking the linear combination $\mathrm{D}_W\widetilde{E}-\widetilde{\Theta}\mathrm{D}_W\widetilde{S}$ and using Gibbs' relation for $\widetilde{\Theta}$ we see that all terms involving $\mathrm{D}_W R$ cancel and the result follows. Finally choosing $R(W,\rho)=\Theta(W,\rho)=:\theta$ and setting $\overline{F}(W,\theta)=E(W,\theta)-\theta S(W,\theta)$ we obtain the desired result since for $\overline{\Theta}(W,\theta)=\theta$ we have $\mathrm{D}_W\overline{\Theta}(W,\theta)\equiv 0$. □

To highlight the previous result we consider the following simple case.

Example 2 We consider $\overline{E}(z,\theta) = \frac{2}{3}a(z)\theta^{3/2}$ and $\overline{S}(z,\theta) = a(z)\theta^{1/2}$, which gives the free energy $\overline{F}(z,\theta) = -\frac{1}{3}a(z)\theta^{3/2}$ and the free entropy $\overline{H}(z,\theta) = \frac{1}{3}a(z)\theta^{1/2}$.

Now consider r such that $\theta = R(z,r) = b(z)^2r^2$ giving $E(z,r) = \frac{2}{3}ab^3r^3$ and $S(z,r) = abr$. We can easily check the identity $\theta = b^2r^2 = \Theta(z,r) = \partial_r E/\partial_r S$ and find

$$\partial_z E(z,r) - \Theta(z,r)\partial_z S(z,r) = \tfrac{1}{3}a'(z)b(z)^3r^3 = \partial_z\overline{F}(z,b(z)^2r^2),$$

i.e. the driving forces coincide as desired. However, for $b'(z) \neq 0$ we have

$$\partial_z E(z,r) \neq \partial_z\overline{E}(z,b(z)^2r^2),\ \ \partial_z S(z,r) \neq \partial_z\overline{S}(z,b(z)^2r^2),\ \ \partial_z F(z,r) \neq \partial_z\overline{F}(z,b(z)^2r^2)$$

where $F(z,r) = E(z,r) - \Theta(z,r)S(z,r)$ is the free energy expressed in r. □

As a consequence of the previous theorem we see that the only mechanically relevant driving forces must be the derivative of the free energy $\mathrm{D}_W\overline{F}(W,\theta) = \mathrm{D}_W E(w,r) - \Theta(W,r)\mathrm{D}_W S(W,r)$ or the temperature $\theta = \partial_r E(W,r)/\partial_r S(W,r)$ or any W-independent combination of these two. In fact, we will see that the following three combinations are the most common:

driving force for reversible dynamics:

$$\mathrm{D}_W\overline{F}(W,\theta) = \mathrm{D}_W E(W,r) - \Theta(W,r)\mathrm{D}_W S(W,r),$$

driving force for dissipative dynamics:

$$\mathrm{D}_W\overline{H}(W,\theta) = \mathrm{D}_W S(W,r) - \frac{\mathrm{D}_W E(W,r)}{\Theta(W,r)},$$

driving force for energy transport: $\displaystyle \frac{1}{\theta} = \frac{1}{\Theta(W,r)} = \frac{\partial_r S(W,r)}{\partial_r E(W,r)}.$

However, there is still a major issue when considering fields over a body Ω and considering the total free energy $\overline{\mathscr{F}}$ and the total free entropy $\overline{\mathscr{H}}$. If we consider variations of these functionals the variational derivatives involve integrations by part, namely

$$\mathrm{D}_z\overline{\mathscr{H}}(u,z,\theta) = \partial_z\overline{H}(u,\nabla u,z,\nabla z,\theta) - \operatorname{div}\big(\partial_{\nabla z}\overline{H}(u,\nabla u,z,\nabla z,\theta)\big).$$

Now using the relation $\overline{H} = -\overline{F}/\theta$ we see that the differentials of $\overline{\mathscr{F}}$ and $\overline{\mathscr{H}}$ are not simply related by multiplying with temperature, since we have

$$\mathrm{D}_z\overline{\mathscr{H}}(u,z,\theta) = -\frac{1}{\theta}\mathrm{D}_z\overline{\mathscr{F}}(u,z,\theta) + \partial_{\nabla z}\overline{F}(u,\nabla u,z,\nabla z,\theta)\nabla\Big(\frac{1}{\theta}\Big).$$

The last term, which destroys the naive relation $\mathrm{D}_z\overline{\mathscr{H}}(u,z,\theta) = -\frac{1}{\theta}\mathrm{D}_z\overline{\mathscr{F}}(u,z,\theta)$, vanishes in two important cases: (i) in the isothermal case where $\nabla\theta \equiv 0$ and (ii) in the case “local case” where $\overline{F}$ does not depend on ∇z. *In all other cases, we have to be careful and distinguish the free energy and the free entropy as driving functionals.*

In many situations it is helpful to use other thermodynamical fields r instead of θ, in particular the internal-energy density $e = \overline{E}(W, \theta)$ or the entropy density $s = \overline{S}(W, \theta)$ are often relevant. For these situations it is better to use the total energy $\mathscr{E}$ and the total entropy $\mathscr{S}$ as function of (u, z, r). Hence, we need to adapt the nice cancellation properties derived in Lemma 1 by introducing a multiplication “$*$” for scalar fields α and variational derivatives $\delta_z G$ as follows:

$$\Big(\alpha * \delta_z G(z, \nabla z)\Big)(x) := \alpha(x)\partial_z G(z(x), \nabla z(x)) - \mathrm{div}\big(\alpha \partial_{\nabla z} G(z, \nabla z)\big)(x).$$

We also write $\alpha * \mathrm{D}_z \mathscr{G}(u, z, r)$ for $\alpha * \delta_z G(u, \nabla u, z, \nabla z, r)$ and obtain the important identities (5a, 5b) below. We should consider “$\alpha * \delta_z$” or “$\alpha * \mathrm{D}_z$” as one operator acting on functions G or functionals $\mathscr{G}$, respectively; see [20, Sect. 5.3] for a fully abstract definition.

Theorem 3 *Using the above definitions we have*

$$\mathrm{D}_u \overline{\mathscr{F}}(u, z, \Theta(u, z, r)) = \mathrm{D}_u \mathscr{E}(u, z, r) - \Theta(u, z, r) * \mathrm{D}_u \mathscr{S}(u, z, r), \tag{5a}$$

$$\mathrm{D}_z \overline{\mathscr{H}}(u, z, \Theta(u, z, r)) = \mathrm{D}_z \mathscr{S}(u, z, r) - \frac{1}{\Theta(u, z, r)} * \mathrm{D}_z \mathscr{E}(u, z, r). \tag{5b}$$

Proof The right-hand side in the first line can be written in full detail as

$$\mathrm{RHS} := \partial_u E(W, r) - \Theta(W, r)\partial_u S(W, r) - \mathrm{div}\big(\partial_{\nabla u} E(W, r) - \Theta(W, r)\partial_{\nabla z} S(W, r)\big),$$

where $W = (u, \nabla u, z, \nabla z)$. Using Lemma 1 we can apply the relation for ∂_u and $\partial_{\nabla u}$ independently and find

$$\begin{aligned} \mathrm{RHS} &= \partial_u \overline{F}(W, \Theta(W, r)) - \mathrm{div}\big(\partial_{\nabla u} \overline{F}(W, \Theta(W, r))\big) \\ &= \delta_u \overline{F}(W, \theta))|_{\theta=\Theta(W,r)} = \mathrm{D}_u \overline{\mathscr{F}}(u, z, \Theta(u, z, r)). \end{aligned}$$

This proves (5a), and the relation (5b) follows analogously. □

The importance of the formulas in Theorem 3 is that we are able to choose an arbitrary thermodynamics field r of describing the heat effects in our material model. This will allows us to find new mathematical formulations that cannot be accessed by using the temperature θ, the energy density e, or the entropy density s, only.

We remark that in many papers and textbooks only the free energy is used as driving functionals and that $\mathrm{D}_z \overline{\mathscr{F}}(u, z, r)$ is used as the driving force for the dissipative variable. This is correct for the cases of isothermal models or if $\overline{F}$ is local, i.e. $\partial_{\nabla z} \overline{F} \equiv 0$. In these two cases one has the relation $\mathrm{D}_z \overline{\mathscr{H}} = -\frac{1}{\theta} \mathrm{D}_z \overline{\mathscr{F}}$, and the factor $-1/\theta$ can be compensated in the dissipation potential, see [30] for the relevance of the “integrating factor θ”.

However, in other cases the usage of $\mathrm{D}_z\overline{\mathscr{F}}$ leads to equations that are thermodynamically inconsistent for the local balance laws, while the total energy conservation and total entropy production may still be valid, see the discussion in [16, Remark 4.1].

3.2 The Balance Equations for Dissipative Material Models

Acoording to [11, 16] the GENERIC framework suggests to write the coupling of elastodynamics for u, dissipative dynamics for z, and energy transport for r in the form

$$\begin{aligned}
\rho\ddot{u} &= -\big(\mathrm{D}_u\mathscr{E}(q) - \varTheta{*}\mathrm{D}_u\mathscr{S}(q)\big) && = -\mathrm{D}_u\overline{\mathscr{F}}(u, z, \varTheta(u, z, r)),\\
\dot{z} &= \partial_{\xi_z}\mathscr{P}_z^*\Big(q, \mathrm{D}_z\mathscr{S}(q) - \frac{1}{\varTheta}{*}\mathrm{D}_z\mathscr{E}(q)\Big) && = \partial_{\xi_z}\mathscr{P}_z^*\big(q, \mathrm{D}_z\overline{\mathscr{H}}(u, z, \varTheta(u, z, r))\big),\\
\dot{r} &= -\frac{\varDelta_u S(q)[\dot{u}]}{\partial_r S(q)} - \frac{\varDelta_z E(q)[\dot{z}]}{\partial_r E(q)} - \frac{1}{\partial_r E}\,\mathrm{div}\Big(\kappa(q)\nabla\Big[\frac{\partial_r S}{\partial_r E}\Big]\Big),
\end{aligned}$$

where the directional derivatives $\varDelta_u S(q)[\dot{u}]$ and $\varDelta_z E(q)[\dot{z}]$ are defined via

$$\varDelta_w G(w)[v] := \partial_w G(w, \nabla w)\cdot v + \partial_{\nabla w} G(w, \nabla w){:}\nabla v.$$

Here the first equation described elastodynamics and contains the Hamiltonian part. In particular, we see that the reversible (i.e. Hamiltonian) part of the dynamics is driven by the derivative $\mathrm{D}_u\overline{\mathscr{F}}(u, z, \varTheta(u, z, r))$ of the free energy $\overline{\mathscr{F}}$. In contrast, the dissipative effects described by the internal variable z and the thermodynamical field r are driven by $\mathrm{D}_z\overline{\mathscr{H}}(u, z, \varTheta(u, z, r))$ and $1/\varTheta = \partial_r S/\partial_r E$, respectively. In particular, we can define a joint dual entropy-production potential (EPP) $\mathscr{P}^*$ via

$$\begin{aligned}
&\mathscr{P}^*(u, z, r; \xi_u, \xi_z, \xi_r) = \mathscr{P}_0^*(u, z, r; \mathbf{M}(u, z, r)(\xi_z, \xi_r)^\mathsf{T}) \text{ with}\\
&\mathscr{P}_0^*(u, z, r; \eta_z, \eta_r) = \mathscr{P}_z^*(u, z, r; \eta_z) + \int_\varOmega \frac{1}{2}\nabla\eta_r\cdot\kappa\nabla\eta_r\,\mathrm{d}x \;\text{ and}\\
&\mathbf{M}(u, z, r) = \begin{pmatrix} I & \frac{-\square}{\partial_r E(\dots)}{*}\mathrm{D}_z\mathscr{E}(u, z, r)\\ 0 & \frac{\square}{\partial_r E(\dots)}\end{pmatrix},
\end{aligned}$$

where $\square$ indicates the slot, where the corresponding argument (here ξ_r) has to be inserted.

We now discuss the two first terms on the right-hand side of the energy balance for r, namely $\varDelta_u S(u, z, r)[\dot{u}]/\partial_r S$ and $\varDelta_z E(u, z, r)[\dot{z}]/\partial_r E$. The first term can be seen as the latent-heat production term that is dual to the term $\frac{\partial_r E}{\partial_r S} * \mathrm{D}_u\mathscr{S}$ in the linear momentum balance and thus belongs to the reversible (=Hamiltonian) part of dynamics. In particular it disappears completely if we choose $r = s$, which means that it does not change the entropy. We refer to [16] for more details. In contrast, the second term $\Delta_z E(u,z,r)[\dot z]/\partial_r E$ is an entropy-production term that is dual to the term $\frac{\partial_r S}{\partial_r E} * \mathrm{D}_z\mathscr{E}$ appearing in $\mathscr{P}^*_z$. We can now rewrite the system for (z,r) in the form

$$\begin{pmatrix}\dot z\\ \dot r\end{pmatrix} = \begin{pmatrix}0\\ -\frac{\Delta_u S[\dot u]}{\partial_r S}\end{pmatrix} + \mathbf{M}(u,z,r)^*\partial_{\boldsymbol{\xi}}\mathscr{P}^*_0\Big(u,z,r;\mathbf{M}(u,z,r)\begin{pmatrix}\mathrm{D}_z\mathscr{S}(u,z,r)\\ \mathrm{D}_r\mathscr{S}(u,z,r)\end{pmatrix}\Big) \quad (6)$$

Before restricting to the quasistatic case with $\rho = 0$ we look at the total energy balance and the total entropy production using the given abstract form. First we observe that along solutions $q(t) = (u(t), z(t), r(t))$ we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\Big(\int_\Omega \frac{\rho}{2}|\dot u|^2\,\mathrm{d}x + \mathscr{E}(q(t))\Big)\\
&= \int_\Omega \rho\ddot u\cdot\dot u\,\mathrm{d}x + \langle \mathrm{D}_u\mathscr{E}(q),\dot u\rangle + \langle \mathrm{D}_z\mathscr{E}(q),\dot z\rangle + \langle \mathrm{D}_r\mathscr{E}(q),\dot r\rangle\\
&\overset{(1)}{=} -\langle \Theta * \mathrm{D}_u\mathscr{S},\dot u\rangle + \Big\langle \binom{\mathrm{D}_z\mathscr{E}}{\partial_r E}, \binom{0}{-\frac{\Delta_u S[\dot u]}{\partial_r S}} + \mathbf{M}(q)^*\,\partial_{\boldsymbol{\eta}}\mathscr{P}^*_0\Big[q,\mathbf{M}(q)\binom{\mathrm{D}_z\mathscr{S}}{\partial_r S}\Big]\Big\rangle\\
&\overset{(2)}{=} 0 + 0 + \Big\langle \mathbf{M}(q)\binom{\mathrm{D}_z\mathscr{E}}{\mathrm{D}_r\mathscr{E}}, \partial_{\boldsymbol{\eta}}\mathscr{P}^*_0\Big[q,\mathbf{M}\binom{\mathrm{D}_z\mathscr{S}}{\partial_r S}\Big]\Big\rangle \overset{(3)}{=} 0,
\end{aligned}$$

where we used the momentum balance and (6) in (1). Equality (2) follows from Gibbs' relation $\Theta = \partial_r E/\partial_r S$ and the definition of "$*$", whereas (3) uses the special form of $\mathbf{M}$ giving $\mathbf{M}(q)(\mathrm{D}_z\mathscr{E},\mathrm{D}_r\mathscr{E})^\mathsf{T} = (0,1)^\mathsf{T}$ and the energy conservation $\langle(0,1)^\top,\partial_\eta\mathscr{P}^*(q,\xi)\rangle = 0$, which follows from the translationally symmetry $\mathscr{P}^*_0(q,\boldsymbol{\eta}{+}\lambda(0,1)^\mathsf{T}) = \mathscr{P}^*_0(q,\boldsymbol{\eta})$ for all $\lambda\in\mathbb{R}$.

Similarly, the total entropy production can be calculated as follows:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{S}(q(t)) &= \langle \mathrm{D}_u\mathscr{S}(q),\dot u\rangle + \langle \mathrm{D}_z\mathscr{S}(q),\dot z\rangle + \langle \mathrm{D}_r\mathscr{S}(q),\dot r\rangle\\
&\overset{(\mathrm{i})}{=} \langle \mathrm{D}_u\mathscr{S}(q),\dot u\rangle + \Big\langle \binom{\mathrm{D}_z\mathscr{S}}{\partial_r S}, \binom{0}{-\frac{\Delta_u S[\dot u]}{\partial_r S}} + \mathbf{M}(q)^*\partial_{\boldsymbol{\eta}}\mathscr{P}^*_0\Big(q,\mathbf{M}(q)\binom{\mathrm{D}_z\mathscr{S}}{\partial_r S}\Big)\Big\rangle\\
&= 0 + 0 + \Big\langle \mathbf{M}(q)\binom{\mathrm{D}_z\mathscr{S}}{\mathrm{D}_r\mathscr{S}}, \partial_{\boldsymbol{\eta}}\mathscr{P}^*_0\Big(q,\mathbf{M}\binom{\mathrm{D}_z\mathscr{S}}{\partial_r S}\Big)\Big\rangle \overset{(\mathrm{ii})}{\geq} 0,
\end{aligned}$$

where we used (6) for (i) and the fact that $\mathscr{P}^*_0$ is a dual dissipation potential in (ii), i.e. $\mathscr{P}^*_0(\boldsymbol{\eta}) \geq \mathscr{P}^*_0(0) = 0$ and convexity of $\mathscr{P}^*_0$ imply $\langle\boldsymbol{\eta},\partial\mathscr{P}^*_0(\boldsymbol{\eta})\rangle \geq 0$.

3.3 A Gradient Structure for the Quasistatic Case

Subsequently we choose the quasistatic approximation and neglect the kinetic energy, i.e. we set the density $\rho = 0$. It was shown already in [17] that, after elimination of the displacement u, the remaining equation for (z,r) is a gradient system if one uses the specific choice $r = s$ (the density of the entropy). Here we follow [16] and show that the result holds for any choice of r, which is extremely helpful, since traditionally one prefers $r = \theta$ (the temperature) and more recently also the choice $r = e$ (the density of the internal energy), but general r gives more flexibility, see e.g. Sects. 4.2 and 4.3. To simplify the formulas we restrict our subsequent discussion to the simpler case

$$\mathscr{P}_0^*(q;\eta_z,\eta_r) = \mathscr{P}_Z^*(q;\eta_z) + \int_\Omega \frac{1}{2}\nabla\eta_r\cdot\kappa(q)\nabla\eta_r\,\mathrm{d}x.$$

The quasistatic thermomechanical system for $q = (u, z, r)$ takes the form

$$0 = \mathrm{D}_u\mathscr{E}(q) - \Theta(q) * \mathrm{D}_u\mathscr{S}(q), \tag{7a}$$

$$\dot z = \partial_{\eta_z}\mathscr{P}_Z^*\Big(q;\mathrm{D}_z\mathscr{S}(q) - \frac{1}{\Theta(q)} * \mathrm{D}_z\mathscr{E}(q)\Big), \tag{7b}$$

$$\dot r = -\frac{\Delta_u S(q)[\dot u]}{\partial_r S(q)} - \frac{\Delta_z E(q)[\dot z]}{\partial_r E(q)} - \frac{1}{\partial_r E}\operatorname{div}\Big(\kappa(q)\nabla\frac{\partial_r S}{\partial_r E}\Big), \tag{7c}$$

still displaying the driving forces in terms of free energy and free entropy. However, the *special GENERIC structure* discussed in [16, Sect. 2.4] guides us to write the system in the form

$$0 = \mathrm{D}_u\mathscr{E}(q) - A(q)\mathrm{D}_r\mathscr{E}(q), \tag{8a}$$

$$\dot z = \partial_{\eta_z}\mathscr{P}_Z^*\Big(q;\mathrm{D}_z\mathscr{S}(q) - B(q)\mathrm{D}_r\mathscr{S}(q)\Big), \tag{8b}$$

$$\dot r = -A(q)^*\dot u - B(q)^*\dot z - C(q)^*\operatorname{div}\big(\kappa(q)\nabla(C(q)\mathrm{D}_r S(q))\big), \tag{8c}$$

where the operators $A(q)$, $B(q)$, and $C(q)$ are defined via

$$A\xi_r = \big(\tfrac{\xi_r}{\mathrm{D}_r\mathscr{S}(q)}\big) * \mathrm{D}_u\mathscr{S}(q),\quad B\xi_r = \big(\tfrac{\xi_r}{\mathrm{D}_r\mathscr{E}(q)}\big) * \mathrm{D}_z\mathscr{E}(q),\quad C\xi_r = \tfrac{\xi_r}{\mathrm{D}_r\mathscr{E}(q)}. \tag{8d}$$

By definition we have the following identities

$$\begin{array}{lll} A\mathrm{D}_r\mathscr{E} = \Theta * \mathrm{D}_u\mathscr{S}, & B\mathrm{D}_r\mathscr{E} = \mathrm{D}_z\mathscr{E}, & C\mathrm{D}_r\mathscr{E} = 1, \\ A\mathrm{D}_r\mathscr{S} = \mathrm{D}_u\mathscr{S}, & B\mathrm{D}_r\mathscr{S} = \big(\tfrac{1}{\Theta}\big) * \mathrm{D}_z\mathscr{E}, & C\mathrm{D}_r\mathscr{S} = 1/\Theta. \end{array} \tag{9}$$

For the following we assume that (7a) or (8a) can be solved uniquely in the form $u = \mathrm{U}(z, r)$. As a shorthand, we also write $q = \mathrm{Q}(z, r) = (\mathrm{U}(z, r), z, r)$.

Theorem 4 *Assume that the mapping $\xi_r \mapsto \xi_r + \mathrm{D}_r U(z, r)^* A(\mathrm{Q}(z, r))\xi_r$ is invertible and denote the inverse by $\mathbf{J}(z, r)$. Defining the functionals*

$$\mathfrak{S}(z, r) = \mathscr{S}(\mathrm{U}(z, r), z, r),\quad \mathfrak{E}(z, r) = \mathscr{E}(\mathrm{U}(z, r), z, r),\quad \textit{and}$$

$$\mathfrak{P}^*(z, r;\xi) = \mathscr{P}_Z^*\big(\mathrm{Q}(z, r);\xi_z - \mathbf{B}(z, r)\xi_r\big) + \int_\Omega \tfrac{1}{2}\nabla\big(\mathbf{C}(z, r)\xi_r\big)\cdot\kappa(\mathrm{Q})\nabla\big(\mathbf{C}(z, r)\xi_r\big)\,\mathrm{d}x,$$

$$\mathbf{B}(z, r) = \big(B(\mathrm{Q}(z, r)) + \mathrm{D}_z\mathrm{U}(z, r)^* A(\mathrm{Q}(z, r))\big)\mathbf{J}(z, r)\ \textit{and}\ \mathbf{C}(z, r) = C(\mathrm{Q}(z, r))\mathbf{J}(z, r),$$

we obtain the following gradient structure:

$$\text{(7a)}-\text{(7b)} \quad \Longleftrightarrow \quad \Big[\, u = \mathrm{U}(z,r) \ \ \textit{and} \ \ \frac{\mathrm{d}}{\mathrm{d}t}\binom{z}{r} = \partial_{\boldsymbol{\xi}}\mathfrak{P}^*\big(z,r;\mathrm{D}\mathfrak{S}(z,r)\big) \,\Big],$$

and we have energy conservation via $\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathfrak{P}^*(z,r;\boldsymbol{\xi}{+}\lambda\mathrm{D}\mathfrak{E}(z,r)) = 0.$

Proof The last relation follows from the definition of $\mathfrak{P}^*$ and the identities

$$\begin{aligned}
&\mathbf{J}\mathrm{D}_r\mathfrak{E} = \mathrm{D}_r\mathscr{E}(\mathrm{Q}), \quad \mathbf{C}\mathrm{D}_r\mathfrak{E}(z,r) = C(\mathrm{Q})\mathrm{D}_r\mathscr{E}(\mathrm{Q}) \equiv 1, \ \text{and} \\
&\mathrm{D}_z\mathfrak{E}{-}\mathbf{B}\mathrm{D}_r\mathfrak{E} = \mathrm{D}_z\mathrm{U}^*\mathrm{D}_u\mathscr{E} + \mathrm{D}_z\mathscr{E} - (B{+}\mathrm{D}_z\mathrm{U}^*A)\mathrm{D}_r\mathscr{E}) \\
&\qquad\qquad\quad = \mathrm{D}_z\mathrm{U}^*(\mathrm{D}_u\mathscr{E}{-}\Theta{*}\mathrm{D}_u\mathscr{S}) + \mathrm{D}_z\mathscr{E} - B\mathrm{D}_r\mathscr{E} = 0,
\end{aligned}$$

where we used (7a) and $(9)_2$, respectively.

To see the equivalence between the evolution equations, first note

$$\dot z = \partial_{\xi_z}\mathfrak{P}^*(z,r;\mathrm{D}\mathfrak{S}(z,r)) = \partial_{\eta_z}\mathscr{P}^*_Z(\mathrm{Q};\mathrm{D}_z\mathfrak{S} - \mathbf{B}\mathrm{D}_r\mathfrak{S}).$$

Proceeding as for $\mathrm{D}_z\mathfrak{E}{-}\mathbf{B}\mathrm{D}_r\mathfrak{E}$ we obtain $\mathrm{D}_z\mathfrak{S}{-}\mathbf{B}\mathrm{D}_r\mathfrak{S} = \mathrm{D}_z\mathscr{S} - B\mathrm{D}_r\mathscr{S}|_{q=\mathrm{Q}(z,r)}$, which is the physically correct driving force, namely the derivative of the free entropy. Thus, the equation for z is identical to (8b).

For the r-equation we first observe $\mathbf{C}\mathrm{D}_r\mathfrak{S} = C\mathrm{D}_r\mathscr{S}(\mathrm{Q}) = 1/\Theta(\mathrm{Q})$, which is the correct driving force for heat conduction. Thus, the gradient-flow equation for r reads

$$\dot r = -\mathbf{B}^*\partial_{\xi_z}\mathfrak{P}^*(..) - \mathbf{C}^*\,\mathrm{div}\,\big(\kappa\nabla(1/\Theta)\big) = \mathbf{J}^*\Big(-(A^*\mathrm{D}_z\mathrm{U}{+}B^*)\dot z - C^*\,\mathrm{div}\,\big(\kappa\nabla(1/\Theta)\big)\Big).$$

Now we use that by definition $\mathbf{J}^*$ is the inverse of $I + A^*\mathrm{D}_r\mathrm{U}$. Thus, we can rewrite the last equation in the form

$$\dot r = -A(\mathrm{Q})^*(\mathrm{U}(z,r)\dot z{+}\mathrm{D}_r\mathrm{U}(z,r)\dot r) - B(\mathrm{Q})^*\dot z - C(\mathrm{Q})^*\,\mathrm{div}\,\big(\kappa(\mathrm{Q})\nabla(1/\Theta(\mathrm{Q}))\big),$$

which is the desired Eq. (8c), because of $\dot u = \mathrm{D}_z\mathrm{U}(z,r)\dot z + \mathrm{D}_r\mathrm{U}(z,r)\dot r$. □

Before going into more details, we present a simple finite-dimensional example, where the reduction and the induced gradient structure can be calculated explicitly.

Example 5 We explain the derivation of the gradient structure by considering a simple mass-spring-damper system, where we assume energy conservation, i.e. the damping mechanics heats up the device, which additionally contains some thermal expansion, see Fig. 1. To keep the model as simple as possible we choose the free energy

$$F(u,z,\theta) = \frac{1}{2}u^2 + \alpha u\theta + \frac{1}{2}(u{-}z)^2 - \frac{4c}{3}\theta^{3/2},$$

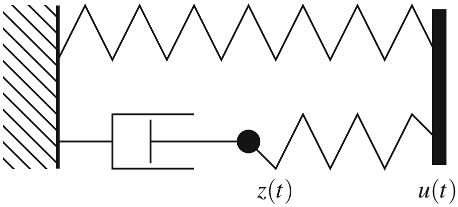


Fig. 1 A system with two springs and one damper. The upper spring undergoes thermal expansion. The mass at $u(t)$ is neglected in our quasistatic setting

where α is the thermal expansion coefficient. The classical force balances are

$$0 = \partial_u F(u, z, \theta) = u + \alpha\theta + u - z \quad \text{and} \quad 0 = \mu\dot z + z - u.$$

The evolution of θ will be determined by energy conservation.

For this we will transform the system into the above structure. First observe that

$$E(u, z, \theta) = \frac{1}{2}u^2 + \frac{1}{2}(u{-}z)^2 + \frac{2c}{3}\theta^{3/2} \quad \text{and} \quad S(u, z, u) = 2c\theta^{1/2} - \alpha u.$$

The driving force for the damper is the derivative of the free entropy with respect to z, which is

$$\partial_z H(u, z, \theta) = \partial_z S(u, z, \theta) - \theta\partial_z E(u, z, \theta) = (u{-}z)/\theta.$$

This is consistent with the choice of the EPP which differs from the dissipation potential $\mathscr{R}(\dot z) = \frac{\mu}{2}\dot z^2$ by a factor of temperature, namely $\mathscr{P}(\theta; \dot z) = \frac{\mu}{2\theta}\dot z^2$ and $\mathscr{P}^*(\theta; \xi) = \frac{\theta}{2\mu}\xi^2$. Together, the equations take the form (8), namely

$$\begin{aligned}
&0 = \mathrm{D}_u F(u, z, \theta) = 2u{-}z{+}\alpha\theta,\\
&\dot z = \frac{\theta}{\mu}\partial_z H(u, z, \theta) = \frac{1}{\mu}(u{-}z),\\
&\dot\theta = -A(u, z, \theta)^*\dot u - B(u, z, \theta)^*\dot z, \ \text{where}\ A(q) = -\frac{\alpha\theta^{1/2}}{c} \ \text{and}\ B(q) = \frac{z{-}u}{c\theta^{1/2}}.
\end{aligned}$$

There is no heat conduction term, since the temperature is the same in the whole system.

For the reduction we immediately find $u = \mathrm{U}(z, \theta) = \frac{1}{2}(z{-}\alpha\theta)$. Inserting this relation into the above system we see that the reduced ODE is given by

$$\dot z = -\frac{1}{2\mu}(z{+}\alpha\theta) \ \text{and} \ \Big(1{+}\frac{\alpha^2}{2c}\theta^{1/2}\Big)\dot\theta = \frac{-z}{2c\theta^{1/2}}\dot z, \tag{10}$$

where the last equation is equivalent to $\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(z(t), \theta(t)) = 0$ after multiplication with $c\theta^{1/2}$.

On the other hand, Theorem 4 provides a gradient structure via

$$\mathfrak{E}(z,\theta)=\frac{1}{4}z^2+\frac{\alpha^2}{4}\theta^2+\frac{2c}{3}\theta^{3/2},\quad \mathfrak{S}(z,\theta)=-\frac{\alpha}{2}z+\frac{\alpha^2}{2}\theta+2c\theta^{1/2},$$
$$\mathfrak{P}^*(\theta;\xi_z,\xi_\theta)=\frac{\theta}{2\mu}\big(\xi_z-\mathbf{B}(z,\theta)\xi_\theta\big)^2,\quad \mathbf{B}(z,r)=\frac{z}{2c\theta^{1/2}+\alpha^2\theta}.$$

It is easily checked that the gradient flow

$$\dot z=\partial_{\xi_z}\mathscr{P}^*\big(\theta;\ \mathrm{D}_z\mathfrak{S}(z,r)-\mathbf{B}(z,\theta)\mathrm{D}_\theta\mathfrak{S}(z,\theta)\big)\ \text{ and }\ \dot r=-\mathbf{B}(z,\theta)\dot z$$

is the same as (10), while the individual driving forces $\mathrm{D}_z\mathfrak{S}(z,\theta)=-\alpha/2$ and $\mathrm{D}_\theta\mathfrak{S}(z,\theta)=c/\theta^{1/2}+\alpha^2/2$ are different from what one might naively expect. □

The above abstract result is a beautiful and mathematically clean reduction of the quasistatically coupled system of elastostatics and dissipative material behavior to a perfect gradient system driven by the physical entropy $\mathfrak{S}$. However, in practice it is of limited use because of the involved nonlocal functionals. In particular, $\mathrm{U}(z,r)$ depends nonlocally on (z,r), since it is obtained by solving an elliptic boundary value problem. As a consequence, the operators $\mathbf{J}$, $\mathbf{B}$, and $\mathbf{C}$ are nonlocal as well.

Fortunately, there are cases, where the nonlocality disappears or is reduced to a minimum. The most important case occurs if the entropy functional is independent of u:

$$\alpha*\mathrm{D}_u\mathscr{S}(u,z,r)=0\text{ for all }\alpha\quad\Longrightarrow\quad A(q)\equiv 0.$$

As a consequence we obtain $\mathbf{B}(z,r)=B(\mathrm{Q}(z,r))$, $\mathbf{C}(z,r)=C(\mathrm{Q}(z,r))$, and $\mathbf{J}=\mathrm{id}$. Moreover, the elastostatic equation reduces to $\mathrm{D}_u\mathscr{E}(u,z,r)=0$. Here we see the advantage of using general thermodynamical variables r, since the form of $A(u,z,r)$ strongly depends on r: we have $A(u,z,r)\equiv 0$ only for specific choices, see Example 7 and Sects. 4.2 and 4.3.

3.4 A Time-Incremental Minimization Procedure

If we are able to find a formulation with $\alpha*\mathrm{D}_u\mathscr{S}\equiv 0$, we can take advantage of the gradient structure derived in Theorem 4, even without eliminating u explicitly. Hence, we start with system (8), but now under the simplifying assumption $A(q)=0$.

We first construct the (primal) entropy-production potential $\mathscr{P}(u,z,r;\dot z,\dot r)$. For this we introduce the dual potential for the heat transfer, which is quadratic, namely

$$\mathscr{P}^*_{\text{heat}}(q;\xi_r):=\frac{1}{2}\int_\Omega\nabla\big(C(q)\xi_r\big)\cdot\kappa(q)\nabla\big(C(q)\xi_r\big)\,\mathrm{d}x$$

and denote by $\mathscr{P}_{\text{heat}}$ its Legendre transform, i.e. $\mathscr{P}_{\text{heat}}(q;\dot r) = \sup_{\xi_r}\big(\int_\Omega \xi_r \dot r\,\mathrm{d}x - \mathscr{P}^*_{\text{heat}}(q;\xi_r)\big)$. Since $\mathscr{P}^*_{\text{heat}}(q;\mathrm{D}_r\mathscr{S})$ corresponds to an H^1 norm of $C(q)\mathrm{D}_r\mathscr{S} = 1/\Theta$, the quadratic form $\mathscr{P}_{\text{heat}}(q;\dot r)$ corresponds to an H^{-1} norm of $\partial_r E(q)\dot r$.

Recall that the full dual EPP has the form $\mathscr{P}^*(q;\xi_z,\xi_r) = \mathscr{P}^*_Z(q;\xi_z - B(q)\xi_r) + \mathscr{P}^*_{\text{heat}}(\xi_r)$; hence the associated primal EPP reads

$$\mathscr{P}(q;\dot z,\dot r) = \mathscr{P}_Z(q;\dot z) + \mathscr{P}_{\text{heat}}(q;\dot r + B(q)^*\dot z).$$

Using the Fenchel equivalence $\xi \in \partial\mathscr{P}(v) \Leftrightarrow v \in \mathscr{P}^*(\xi)$, we find that the system (8) with $A(q) \equiv 0$ for $q = (u,z,r)$ can be rewritten as follows:

$$\mathrm{D}_u\mathscr{E}(q) = 0, \qquad \begin{pmatrix}\mathrm{D}_z\mathscr{S}(q)\\ \mathrm{D}_r\mathscr{S}(q)\end{pmatrix} \in \partial_{(\dot z,\dot r)}\mathscr{P}(q;\dot z,\dot r) = \begin{pmatrix} I & B(q)\\ 0 & I\end{pmatrix}\begin{pmatrix}\partial\mathscr{P}_Z(q;\dot z)\\ \partial\mathscr{P}_{\text{heat}}(q;\dot r + B^*\dot z)\end{pmatrix}.$$

We see that both relations are variational in the sense that derivatives of functionals determine the solutions.

In particular, we can discretize the system in time such that we obtain time-incremental minimization principles that are useful for proving existence of solutions or for numerical simulation of concrete models.

Time-incremental minimization procedure for the case $A \equiv 0$: *Consider a time step* $\tau > 0$ *and assume that the initial condition* $q^0 = (u^0, z^0, r^0)$ *is given such that* $\mathrm{D}_u\mathscr{E}(q^0) = 0$. *We define* q^k *for* $k \in \mathbb{N}$ *incrementally as follows:*

$$\text{(TIMP)}\quad \begin{cases} \begin{pmatrix} z^{k+1}\\ r^{k+1}\end{pmatrix} & \in \operatorname*{Arg\,min}_{(z,r)}\Big\{\,\tau\,\mathscr{P}\Big(q^k; \dfrac{1}{\tau}\begin{pmatrix} z - z^k\\ r - r^k\end{pmatrix}\Big) - \mathscr{S}(u^k, z, r)\,\Big\},\\ \qquad u^{k+1} & \in \operatorname*{Arg\,min}_{u}\mathscr{E}(u, z^{k+1}, r^{k+1}). \end{cases} \tag{11}$$

Note that we do not enforce energy conservation, which could be done as well. However, it is better to use the errors in the energy conservation as quality control for the numerical accuracy, see the example below.

Remark 6 It may be tempting to write a similar time-incremental minimization procedure also in the case $A(q) \neq 0$. However, we see that the term $A(q)^*\dot u$ needs to be approximated. One way would be to use the consistent tangents $\mathrm{D}_z\mathrm{U}$ and $\mathrm{D}_r\mathrm{U}$ and to replace the term by $A(q^k)\big(\mathrm{D}_z\mathrm{U}(z^k, r^k)\dot z + \mathrm{D}_r\mathrm{U}(z^k, r^k)\dot r\big)$ before discretizing the derivatives by time increments. However, the numerical calculation of the tangents $\mathrm{D}_z\mathrm{U}$ and $\mathrm{D}_r\mathrm{U}$ seems to be very inefficient. Moreover, it is not clear whether the update $u^{k+1} = u^k + \tau\big(\mathrm{D}_z\mathrm{U}(z^k, r^k)(z^{k+1} - z^k) + \mathrm{D}_r\mathrm{U}(z^k, r^k)(r^{k+1} - r^k)\big)$ is consistent enough with the elastostatic equation $\mathrm{D}_u\mathscr{E}(q) - A(q)\mathrm{D}_r\mathscr{E}(q) = 0$.

To highlight the usefulness of the algorithm we return to the spring-damper model treated in Example 5. We will take advantage of using a suitable thermodynamic variable r, namely $r = s$.

Example 7 (Continuation of Example 5*)* The model is originally formulated in (u, z, θ) but $\partial_u S(u, z, \theta) = -\alpha$ does not vanish, so the model cannot be treated with these variables. Thus, we will use the entropy density s as the thermodynamical variable r:

$$s = R(z, \theta) := 2c\theta^{1/2} - \alpha u \quad \rightsquigarrow \quad \theta = \Theta(u, s) = \Big(\frac{s{+}\alpha u}{2c}\Big)^2.$$

Hence, we find the following relations

$$\widetilde{E}(u, z, s) = \frac{1}{2}u^2 + \frac{1}{2}(u{-}z)^2 + \widetilde{e}(s{+}\alpha u) \text{ with } \widetilde{e}(y) = \begin{cases} y^3/(12c^2) & \text{for } y \geq 0, \\ \infty & \text{for } y < 0, \end{cases}$$
$$\widetilde{S}(u, z, s) = s, \qquad \widetilde{A}(u, z, s) = 0, \qquad \widetilde{B}(u, z, s) = 4c^2\frac{z{-}u}{(s{+}\alpha u)^2} = \frac{z{-}u}{\Theta(u, s)}.$$

The full coupled system takes the form

$$0 = \mathrm{D}_u\widetilde{E}(u, z, s) = 2u - z + \frac{\alpha}{4c^2}(s{+}\alpha u)^2, \tag{12a}$$
$$\dot{z} = \frac{\Theta(u, s)}{\mu}\big(\partial_z\widetilde{S} - \widetilde{B}\partial_s\widetilde{S}\big) = \frac{\Theta}{\mu}\big(0 - \widetilde{B}(u, z, s)\,1\big) = \frac{u{-}z}{\mu}, \tag{12b}$$
$$\dot{s} = -\widetilde{A}(u, z, s)\,\dot{u} - \widetilde{B}(u, z, s)\dot{z} = 0 + \frac{\Theta}{\mu}\big(\widetilde{B}(u, z, s)\big)^2. \tag{12c}$$

Since the dual EPP $\mathscr{P}^*$ has the form $\mathscr{P}^*(\theta; \xi_z, \xi_s) = \frac{\Theta}{2\mu}(\xi_z{-}\widetilde{B}\xi_s)^2$ the primal EPP reads

$$\mathscr{P}(u, z, s; \dot{z}, \dot{s}) = \begin{cases} \dfrac{\mu}{2\Theta(u, s)}|\dot{z}|^2 & \text{if } \dot{s}{+}\widetilde{B}(u, z, s)\dot{z} = 0, \\ \infty & \text{else.} \end{cases}$$

Using the explicit constraint $\dot{s} = \widetilde{B}(u, z, s)\dot{z}$, the time-incremental minimization procedure of (11) takes the explicit form:

$$\text{(1) Find } (z^{k+1}, s^{k+1}) \text{ as minimizer of } \frac{\mu}{2\Theta(u^k, s^k)}\,\frac{1}{\tau}\big(z - z^k\big)^2 - s$$
$$\text{subject to } s - \widetilde{B}(u^k, z^k, s^k)z = s^k - \widetilde{B}(u^k, z^k, s^k)z^k;$$
$$\text{(2) find } u^{k+1} \text{ as minimizer of } E(u, z^{k+1}, s^{k+1}).$$

Here the first minimization problem is quadratic, and the explicit solution can be determined. In the second minimization problem the functional is cubic in u, so the unique minimizer $u^{k+1} = \mathrm{U}(z^{k+1}, s^{k+1})$ can be obtained by solving (12a). Thus, we find the incremental update formulas

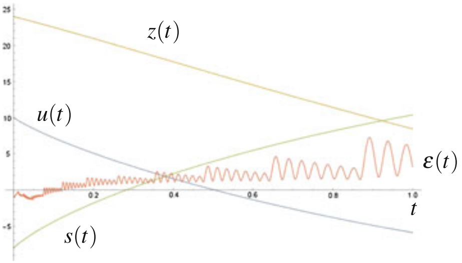
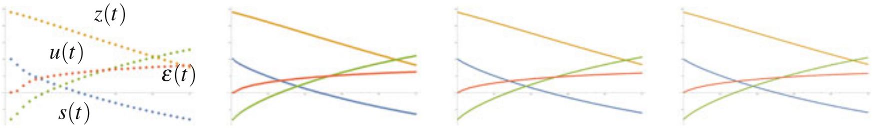


Fig. 2 Numerical solution $q(t) = (u(t), z(t), s(t))$ of the ODE (12) for parameters $\mu = \alpha = 1$ and $c = 1/2$. Here $\varepsilon(t) = 10^6(\widetilde{E}(q(t)) - \widetilde{E}(q(0)))$



$\tau = 1/30$ $\tau = 1/100$ $\tau = 1/300$ $\tau = 1/1000$

Fig. 3 Numerical simulation via (13) for time steps $\tau = 1/30,\ 1/100,\ 1/300,\ 1/1000$. Energy conservation is checked via $\varepsilon_\tau(k\tau) = (\widetilde{E}(q^k) - \widetilde{E}(q^0))/(\tau \widetilde{E}(q^0))$, hence $\varepsilon_\tau(1) \approx 5\tau$

$$z^{k+1} = z^k + \tau \frac{\Theta(q^k)}{\mu} \widetilde{B}(q^k), \quad s^{k+1} = s^k + \tau \frac{\Theta(q^k)}{\mu} \widetilde{B}(q^k)^2, \quad u^{k+1} = \mathrm{U}(z^{k+1}, s^{k+1}). \tag{13}$$

Inserting the explicit form of $\widetilde{B}$ we find the relation $z^{k+1} = z^k + \frac{\tau}{\mu}(u^k - z^k)$, which is an explicit discretization of (12b). Nevertheless, by construction of our algorithm we know that it is entropy increasing. Figure 2 shows the numerical solution of the ODE (12), while Fig. 3 shows numerical approximations obtain via the TIMP (11), which yields (13). We observe that the scheme is only of first order in the time step. However, we expect that it is stable even when treating fully nonlinear thermomechanical systems. □

4 Gradient Structures for Thermomechanical Systems

In this section we give three examples of temperature dependent models that can be rewritten in terms of entropic gradient flows.

4.1 Homogenization of the Penrose-Fife System

This model is without any elastic deformation, so there is no need to eliminate the variable u and the condition $\mathrm{D}_u \mathscr{S} \equiv 0$ is trivially satisfied.

The Penrose-Fife model was introduced in [27] to resolve a long-lasting debate concerning thermodynamically correct couplings between phase transitions models and the heat equation, see [19, 28] and [16, Remark 4.1] for details. Typically the free energy $F(z, \nabla z, \theta) = -c\theta \log\theta + \psi_0(z) + \theta\psi_1(z) + \theta\frac{\alpha}{2}|\nabla z|^2$ is used which leads to

$$E(z,\theta) = c\theta + \psi_0(z) \quad \text{and} \quad S(z, \nabla z, \theta) = c\log\theta + c - \psi_1(z) - \frac{\nu}{2}|\nabla z|^2. \tag{14}$$

The equations in the variables z and θ take the form

$$\text{(PF)} \quad \begin{cases} \dot z = m\big(\delta_z S(z,\theta) - \frac{1}{\theta}\partial_z E\big) \;=\; m\Big(\nu\Delta z - \psi_1'(z) - \frac{1}{\theta}\psi_0'(z)\Big) \\ c\dot\theta = -\psi_0'(z)\dot z + \operatorname{div}\big(k(z,\theta)\nabla\theta\big). \end{cases} \tag{15}$$

Almost all mathematical work is restricted to the case $E(z,\theta) = c\theta + \lambda z$, which is physically only relevant in a small temperature range. In particular, the logarithmic entropy $\sigma(\theta) = c\log\theta$ is only good for gases, while for solids one should have $\sigma(0) = 0$, e.g. $\sigma(\theta) = c\theta^\alpha$ for $\alpha \in]0, 1[$ is more appropriate.

In [21] we consider the internal energy e as thermodynamic variable $r = e$, namely $\mathscr{E}(z,e) = \int_\Omega e(x)\,\mathrm{d}x$ and $\mathscr{S}(z,e) = \int_\Omega \widehat{S}(z, \nabla z, e)\,\mathrm{d}x$. Indeed, the above case (14) can be rewritten in terms of e via $s = \widehat{S}(z, \nabla z, e) = c\log(e - \psi_0(z)) - c\log c - \psi_1(z) - \frac{\alpha}{2}|\nabla z|^2$, but much more general functions $\widehat{S}$ are possible.

The Penrose-Fife system (15) can be formulated as gradient system via the EPP

$$\mathscr{P}^*(z, e; \xi_z, \xi_e) = \frac{1}{2}\langle\boldsymbol{\xi}, \mathbb{K}(z,e)\boldsymbol{\xi}\rangle = \frac{1}{2}\int_\Omega m(x, z, e))\xi_z(x)^2 + \nabla\xi_e(x)\cdot\kappa(z,e)\nabla\xi_e(e)\,\mathrm{d}x.$$

In particular, one has the explicit form

$$\begin{aligned} \text{(PF)} \iff \begin{pmatrix}\dot z\\ \dot e\end{pmatrix} &= \mathbb{K}(z,e)\mathrm{D}\mathscr{S}(z,e) = \partial_{\boldsymbol{\xi}}\mathscr{P}^*\big(z, e; \mathrm{D}\mathscr{S}(z,e)\big) \\ &= \begin{pmatrix} m & 0 \\ 0 & -\operatorname{div}(\kappa\nabla\square)\end{pmatrix}\begin{pmatrix}\delta_z S\\ \partial_e S\end{pmatrix}. \end{aligned}$$

There is one special case where the gradient system can be rewritten as an *evolutionary variational inequality (EVI)*, cf. [2, 18]. For this we have to assume that $\mathbb{K}$ (or equivalently $\mathscr{P}^*$) does not depend on the state (z, e). Moreover, one needs to assume that $(z, e) \mapsto -\mathscr{S}(z, e)$ is λ-convex, i.e. for some $\lambda \in \mathbb{R}$ the function $(z, e) \mapsto -\mathscr{S}(z, e) - \lambda\mathscr{P}(z, e)$ is convex, where $\mathscr{P}$ is the primal EPP. Under these assumptions, a curve $q = (z, e) : [0, T] \mapsto \mathbf{X}_{\mathrm{PF}} := \mathrm{H}^1(\Omega) \times \mathrm{L}^1(\Omega)$ is a solution of (PF), if and only if

$$\begin{aligned} \text{(EVI)} \quad \mathrm{e}^{\lambda(t-s)}\mathscr{P}(q(t) - w) + \mathscr{P}(q(s) - w) &\le \textstyle\int_0^{t-s}\mathrm{e}^{\lambda r}\,\mathrm{d}r\,\big(\mathscr{S}(q(t)) - \mathscr{S}(w)\big) \\ &\text{for all } 0 \le s < t \text{ and all } w = (\widetilde{z}, \widetilde{e}) \in \mathbf{X}_{\mathrm{PF}}. \end{aligned}$$

This variational formulation of the Penrose-Fife model is ideal for coarse graining. Assuming that the entropy density $\widehat{S}$, the mobility m, and the heat conduction tensor κ depend periodically on a microscopical variable in the form

$$\widehat{S}_\varepsilon = \mathbb{S}(\tfrac{1}{\varepsilon}x, z, e) - \frac{1}{2}\nabla z{\cdot}\mathbb{A}(\tfrac{1}{\varepsilon}x)\nabla z, \quad m_\varepsilon(x) = \mathbb{M}(\tfrac{1}{\varepsilon}x), \quad \kappa_\varepsilon(x) = \mathbb{H}(\tfrac{1}{\varepsilon}x),$$

one can pass to the homogenization limit $\varepsilon \to 0$ using the abstract methods for evolutionary Γ-convergence described in [18]. In [21] it is shown that solutions $(z^\varepsilon, e^\varepsilon)$ for the gradient system $(\mathbf{X}_{\rm PF}, \mathscr{S}_\varepsilon, \mathscr{P}_\varepsilon)$ converge to the unique solution (z^0, e^0) of the limiting gradient system $(\mathbf{X}_{\rm PF}, \mathscr{S}_0, \mathscr{P}_0)$, if this is true for the initial conditions.

The effective entropy functional $\mathscr{S}_0$ and the effective EPP $\mathscr{P}_0$ are

$$\mathscr{S}_0(z, e) = \int_\Omega S_{\rm eff}(z, e) - \tfrac{1}{2}\nabla z{\cdot}A_{\rm hom}\nabla z \,\mathrm{d}x \;\text{ and}$$
$$\mathscr{P}_0^*(\xi_z, \xi_e) = \tfrac{1}{2}\int_\Omega m_{\rm harm}\xi_z^2 + \nabla\xi_e{\cdot}H_{\rm hom}\nabla\xi_e \,\mathrm{d}x.$$

Here $A_{\rm hom}$ and $H_{\rm hom}$ are the classical homogenized effective tensors obtained from the periodic functions $\mathbb{A}$ and $\mathbb{H}$, respectively. Moreover, $m_{\rm harm}$ is the harmonic mean of $\mathbb{M}$. More interesting is the homogenization of the nonlinear function $\mathbb{S}$ to obtain $S_{\rm eff}$. Here one takes advantage of the concavity of the mapping $e \mapsto \mathbb{S}(y, z, e)$. Doing a partial Legendre transform of $-\mathbb{S}$ with dual variable τ, one obtains the free energy evaluated at $\theta = -1/\tau$. After simply averaging $\mathbb{F}(y, z, -1/\tau)$ over the periodicity cell, one can reverse the Legendre transform and obtains $S_{\rm eff}(z, e)$. We refer to [21] for more details.

4.2 A Time-Dependent Thermoplasticity Model

We consider a special case of a linearized non-isothermal elastoplastic material, where the coupling between the strain $\mathbf{e}(u)$ and the temperature is only indirect via the plastic tensor z, cf. [3]. In contrast to the theory so far, we also allow for a time-dependent loading $\ell(t)$. Hence, we consider the functionals

$$\mathscr{E}(t, u, z, \theta) = \int_\Omega \frac{1}{2}|\mathbf{e}(u){-}z|_{\mathbb{C}}^2 + \Phi(z, \theta)\,\mathrm{d}x - \langle \ell(t), u\rangle \;\text{ and }\; \mathscr{S}(u, z, \theta) = \int_\Omega S(z, \theta)\,\mathrm{d}x,$$

where $|\mathbf{e}|_{\mathbb{C}}^2 = \mathbf{e}{:}\mathbb{C}\mathbf{e}$ and Gibbs' relation $\partial_\theta\Phi = \theta\partial_\theta S$. Setting $E(e, z, \theta) = \frac{1}{2}|\mathbf{e}{-}z|_{\mathbb{C}}^2 + \Phi(z, \theta)$, we will explicitly use the decouplings $\partial_{\mathbf{e}}\partial_\theta E = 0 = \partial_{\mathbf{e}} S$.

We note that $\mathrm{D}_u\mathscr{S} \equiv 0$ implies that the elastic equilibrium takes the form

$$0 = \mathrm{D}_u\mathscr{E}(t, u, z, \theta) - \theta * \mathrm{D}_u\mathscr{S}(u, z, \theta) = \mathrm{D}_u\mathscr{E}(t, u, z, \theta) = -\operatorname{div}\big(\mathbb{C}(\mathbf{e}(u){-}z)\big) - \ell(t).$$

In particular, we are able to solve this equation for u as a nonlocal function of z and the loading $\ell(t)$, namely $u(t) = \mathbf{U}(z(t), \ell(t))$, where $\mathbf{U} : \mathrm{L}^2(\Omega; \mathbb{R}^{d\times d}_{0,\mathrm{sym}}) \times (\mathrm{H}^1_\mathrm{D}(\Omega; \mathbb{R}^d))^* \to \mathrm{H}^1_\mathrm{D}(\Omega; \mathbb{R}^d)$ is a bounded linear operator.

Respecting the energy conservation, we can take the dual EPP $\mathscr{P}^*$ in the form

$$\mathscr{P}^*(q; \xi_u, \xi_z, \xi_\theta) = \mathscr{P}^*_0\Big(q; \xi_z - \frac{\xi_\theta}{\partial_\theta \Phi(z, \theta)} * \mathrm{D}_z \mathscr{E}(t, q), \frac{\xi_\theta}{\partial_\theta \Phi(z, \theta)}\Big),$$

which clearly satisfies $\mathscr{P}^*(q; \boldsymbol{\xi}{+}\lambda \mathrm{D}\mathscr{E}(t, q)) = \mathscr{P}^*(q, \boldsymbol{\xi})$.

Defining the reduced energy $\widehat{\mathscr{E}}(t, z, \theta) := \mathscr{E}(t, \mathbf{U}(z, \ell(t)), \theta)$ we find the relations

$$\widehat{\mathscr{E}}(t, z, \theta) = \int_\Omega \frac{1}{2} z{:}\mathbf{A}z + \Phi(z, \theta)\,\mathrm{d}x - \langle z, a(t)\rangle,$$

where $\mathbf{A}$ is a bounded, symmetric, non-negative linear operator from $\mathrm{L}^2(\Omega; \mathbb{R}^{d\times d}_{0,\mathrm{sym}})$ into itself and $a(t) = \mathbf{K}\ell(t)$ for a suitable bounded linear operator $\mathbf{K}$. Thus, Theorem 4 yields the gradient-flow equation

$$\begin{aligned}\begin{pmatrix}\dot{z}\\ \dot{\theta}\end{pmatrix} &= \partial_{\boldsymbol{\xi}} \mathscr{P}^*(z, \theta; \mathrm{D}\mathscr{S}(z, \theta))\\ &= \begin{pmatrix} I & 0\\ \frac{-1}{\partial_\theta \Phi}\mathrm{D}_z\widehat{\mathscr{E}} & \frac{1}{\partial_\theta \Phi}\end{pmatrix} \partial_{\boldsymbol{\xi}} \mathscr{P}^*_0\left(z, \theta; \begin{pmatrix} I & \frac{-1}{\partial_\theta \Phi} * \mathrm{D}_z\widehat{\mathscr{E}}\\ 0 & \frac{1}{\partial_\theta \Phi}\end{pmatrix}\begin{pmatrix}\mathrm{D}_z\mathscr{S}\\ \mathrm{D}_\theta\mathscr{S}\end{pmatrix}\right).\end{aligned}$$

We emphasize here that the transformation inside $\partial_{\boldsymbol{\xi}}\mathscr{P}^*$ via $\mathrm{D}_z\widehat{\mathscr{E}}{=}\mathbf{A}z + \partial_z\Phi(z, \theta) - a(t)$ is time-dependent and nonlocal because of $\mathbf{A}$.

To simplify the gradient structure we reformulate it using the internal energy

$$e(x) := \frac{1}{2} z(x) : (\mathbf{A}z)(x) + \Phi(z(x), \theta(x)),$$

where we note that the nonlocal operator $\mathbf{A}$ has to be taken with care. This relation can be inverted to express the temperature as function of z and e as follows. Denote by $\theta = \widetilde{\Theta}(z, \widetilde{e})$ the unique solution of $\widetilde{e} = \Phi(z, \theta)$ and define $\widetilde{S}(z, \widetilde{e}) = S(z, \widetilde{\Theta}(z, \widetilde{e}))$, which gives $\partial_{\widetilde{e}}\widetilde{S}(z, \widetilde{e}) = 1/\widetilde{\Theta}(z, \widetilde{e})$ by Gibbs' relation $\partial_\theta\Phi = \theta\partial_\theta S$. Then, with $\widetilde{e} = e - \frac{1}{2} z : \mathbf{A}z$, the total energy, total entropy, and the dual EPP read

$$\begin{aligned}&\widetilde{\mathscr{E}}(t, z, e) = \int_\Omega e - a(t){:}z\,\mathrm{d}x, \qquad \widetilde{\mathscr{S}}(z, e) = \int_\Omega \widetilde{S}(z, e{-}\tfrac{1}{2}z{:}\mathbf{A}z)\,\mathrm{d}x, \quad \text{and}\\ &\widetilde{\mathscr{P}}^*(t, z, e; \xi_z, \xi_e) := \widetilde{\mathscr{P}}^*_0(z, e; \xi_z{+}\xi_e a(t), \xi_e) = \widetilde{\mathscr{P}}^*_0\left(z, e; \begin{pmatrix} I & a(t)\\ 0 & 1\end{pmatrix}\begin{pmatrix}\xi_z\\ \xi_e\end{pmatrix}\right).\end{aligned}$$

The energy balance $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{E}(t, z(t), e(t)) = \partial_t\mathscr{E}(t, z(t), e(t))$ along solutions still follows from the relation $\widetilde{\mathscr{P}}^*_0(z, e; \boldsymbol{\xi}{+}(0, \lambda)^\mathsf{T}) = \widetilde{\mathscr{P}}^*_0(z, e; \boldsymbol{\xi})$ for all constants $\lambda \in \mathbb{R}$.

The primal EPP $\widetilde{\mathscr{P}}$ takes a similar time-dependent form

$$\widetilde{\mathscr{P}}(t,z,e;\dot z,\dot e) = \widetilde{\mathscr{P}}_0(z,e;\dot z,\dot e{-}a(t){:}\dot z) = \widetilde{\mathscr{P}}_0\big(z,e;\mathbf{N}(t)\big(\begin{smallmatrix}\dot z\\ \dot e\end{smallmatrix}\big)\big),\ \ \mathbf{N}(t) := \begin{pmatrix} I & 0\\ -a(t){:}\square & 1\end{pmatrix}$$

where $\mathscr{P}_0(z,e;v,w)=\infty$ if $\int_\Omega w\,\mathrm{d}x \neq 0$, which enforces energy conservation.

In total, the generalized gradient flow for this simple thermoplastic model can be written in the following two equivalent forms

$$\begin{pmatrix}\dot z\\ \dot e\end{pmatrix} = \mathbf{N}(t)^{-1}\partial_{\boldsymbol{\xi}}\widetilde{\mathscr{P}}_0^*\left(z,e;\begin{pmatrix}\mathrm{D}_z\widetilde{\mathscr{S}}{+}\mathrm{D}_e\widetilde{\mathscr{S}}\,a(t)\\ \mathrm{D}_e\widetilde{\mathscr{S}}(z,e)\end{pmatrix}\right) \quad\Longleftrightarrow$$
$$\begin{pmatrix}0\\ 0\end{pmatrix} \in \mathbf{N}(t)^*\partial_{\mathbf{v}}\widetilde{\mathscr{P}}_0\left(z,e;\mathbf{N}(t)\begin{pmatrix}\dot z\\ \dot e\end{pmatrix}\right) - \begin{pmatrix}\mathrm{D}_z\widetilde{\mathscr{S}}(z,e)\\ \mathrm{D}_e\widetilde{\mathscr{S}}(z,e)\end{pmatrix}.$$

We consider a specially simple case of thermo-viscoplastic gradient plasticity by choosing

$$\widetilde{\mathscr{P}}_0(z,e;v,w) = \sigma_{\text{yield}}\|v\|_{\mathrm{L}^1} + \frac{\mu}{2}\|v\|^2_{\mathrm{L}^2} + \frac{\kappa}{2}\|w\|^2_{\mathrm{H}^{-1}} \ \text{ and}$$
$$\widetilde{\mathscr{S}}(z,e) = \int_\Omega \widetilde{S}(z,e{-}\frac{1}{2}z{:}\mathbf{A}z) - \frac{\nu}{2}|\nabla z|^2\,\mathrm{d}x,$$

where $\|w\|^2_{\mathrm{H}^{-1}} = \|\nabla\phi\|^2_{\mathrm{L}^2}$ if $\Delta\phi = w$ in Ω and $\nabla\phi\cdot n = 0$ on $\partial\Omega$. This leads to the generalized gradient flow equation

$$\begin{aligned}&0 = \sigma_{\text{yield}}\,\mathrm{Sign}(\dot z) + \mu\dot z + \partial_z\widehat{S}(z,e{-}\tfrac{1}{2}z{:}\mathbf{A}z) - \tfrac{1}{2}\big(\Xi\,(\mathbf{A}z) + \mathbf{A}(\Xi z)\big) + \nu\Delta z,\\ &0 = \dot e - a(t){:}\dot z - \kappa\Delta\Xi, \qquad \text{where } \Xi = \partial_{\tilde e}\widetilde{S}\big(z,e{-}\tfrac{1}{2}z{:}\mathbf{A}z\big).\end{aligned} \tag{16}$$

Here $\Xi = \partial_{\tilde e}\widetilde{S}$ denotes the inverse temperature $1/\theta$.

In particular, the second formulation gives rise to a simple time-incremental minimization procedure, which is well-known in isothermal elastoplasticity (cf. [6–8, 13, 15, 24, 25], but is new for the non-isothermal case:

$$(\text{TIMP})_* \begin{pmatrix}z^{k+1}\\ e^{k+1}\end{pmatrix} \in \operatorname*{Arg\,min}_{(z,e)}\ (t^{k+1}{-}t^k)\widetilde{\mathscr{P}}_0\left(z^k,e^k;\tfrac{1}{t^{k+1}-t^k}\mathbf{N}(t^k)\begin{pmatrix}z{-}z^k\\ e{-}e^k\end{pmatrix}\right) - \widetilde{\mathscr{S}}(z,e).$$

We emphasize that $(\text{TIMP})_*$ is not equivalent to the one proposed in (11), since here we eliminated u beforehand by using the nonlocal operator $\mathbf{A}$. So, $(\text{TIMP})_*$ should be preferable if $\mathbf{A}$ is available. Again, we observe that the concavity of $\widetilde{\mathscr{S}}$ implies that the minimum problem is convex. In the case of viscoplasticity, $\widetilde{\mathscr{P}}_0$ is even strictly convex, so there is a unique minimizer in each time step. Thus, it should be possible to show existence of solutions for the thermo-viscoplastic system in (16). Unfortunately, the methods developed in [21] and based on the (EVI) are not applicable because of the nonquadratic behavior of $\widetilde{\mathscr{P}}_0$ due to $\sigma_{\text{yield}} > 0$.

4.3 A Thermoplastic Model with Thermal Expansion

Finally, we consider a classical plasticity model (see e.g. [4]) where thermal expansion leads to a stronger coupling of elastostatics and heat conduction. As usual we again start with a free energy containing a thermal expansion tensor $\mathbb{E} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$ in the form

$$F(\nabla u, z, \nabla z, \theta) = \frac{1}{2}|\mathbf{e}(u)-z|^2_{\mathbb{C}} + \psi_1(\theta)\mathbb{E}{:}\mathbf{e}(u) + H(z) + \frac{\sigma\theta}{2}|\nabla z|^2 - \frac{c}{\alpha(1+\alpha)}\theta^{1+\alpha},$$

with $c > 0$ and $\alpha \in \,]0,1[$. We obtain the energy and entropy functionals

$$\mathscr{E}(u, z, \theta) = \int_\Omega \frac{1}{2}|\mathbf{e}(u)-z|^2_{\mathbb{C}} + \widetilde{\psi}_1(\theta)\mathbb{E}{:}\mathbf{e}(u) + H(z) + \frac{c\theta^{1+\alpha}}{1+\alpha}\,\mathrm{d}x,$$
$$\mathscr{S}(u, z, \theta) = \int_\Omega \frac{c}{\alpha}\theta^\alpha - \psi_1'(\theta)\mathbb{E}{:}\mathbf{e}(u) - \frac{\nu}{2}|\nabla z|^2\,\mathrm{d}x,$$

where $\widetilde{\psi}_1(\theta) = \psi_1(\theta) - \theta\psi_1'(\theta)$. Clearly, $\alpha * \mathrm{D}_u\mathscr{S}(u, z, \theta) = -\operatorname{div}(\alpha\psi_1'(\theta)\mathbb{E})$ is non-zero, so the reduction to a local gradient system is not possible, unless we replace the temperature θ by a more convenient thermodynamically variable r. A possible choice is

$$r = R(\mathbf{e}(u), \theta) := \frac{c}{\alpha}\theta^\alpha - \psi_1'(\theta)\mathbb{E}{:}\mathbf{e}(u).$$

Since we also need the inverse transformation $\theta = \Theta(\mathbf{e}(u), r)$, we assume $\psi_1(\theta) = \theta$ for notational simplicity. Then $\widetilde{\psi}_1 \equiv 0$ and $\theta = \Theta(\mathbf{e}, r) = \big(\alpha(r{+}\mathbb{E}{:}\mathbf{e})/c\big)^{1/\alpha}$, and the functionals take the form

$$\mathscr{E}(u, z, r) = \int_\Omega \frac{1}{2}|\mathbf{e}(u)-z|^2_{\mathbb{C}} + H(z) + \frac{\alpha c_\alpha}{1+\alpha}\big(r{+}\mathbb{E}{:}\mathbf{e}(u)\big)^{1+1/\alpha}\,\mathrm{d}x,$$
$$\mathscr{S}(u, z, r) = \int_\Omega r - \frac{\nu}{2}|\nabla z|^2\,\mathrm{d}x, \text{ where } c_\alpha = (\alpha/c)^{1/\alpha}.$$

This choice now guarantees that $A(u, z, r) \equiv 0$ and the reduction to a local gradient system for (z, r) can be done as described at the end of Sect. 3.3. In particular the solution $u = \mathrm{U}(z, r)$ can be obtained as the unique minimizer of the convex functional $u \mapsto \mathscr{E}(u, z, r)$. The corresponding Euler-Lagrange equation reads

$$-\operatorname{div}\big(\mathbb{C}(\mathbf{e}(u)-z) + c_\alpha(r{+}\mathbb{E}{:}\mathbf{e}(u))^{1/\alpha}\mathbb{E}\big) = 0.$$

Acknowledgments The author is grateful to Michael Ortiz for many stimulating discussions on the interaction of mathematics and mechanics. Moreover, he thanks Ulisse Stefanelli for the fruitful collaboration concerning variational methods for evolutionary systems. The resarch was partially supported by the European Research Council via the grant ERC-2010-AdG no.267802 *Analysis of Multiscale Systems Driven by Functionals*.

References

1. Alber, H. -D. (1998). *Materials with memory* (Vol. 1682), Lecture Notes in Mathematics. Berlin: Springer.
2. Ambrosio, L., Gigli, N., & Savaré, G. (2005). *Gradient flows in metric spaces and in the space of probability measures.*, Lectures in Mathematics Basel: ETH Zürich. Birkhäuser Verlag.
3. Bartels, S., & Roubíček, T. (2008). Thermoviscoplasticity at small strains. *ZAMM—Journal of Applied Mathematics and Mechanics*, *88*, 735–754.
4. Bartels, S., & Roubíček, T. (2011). Thermo-visco-elasticity with rate-independent plasticity in isotropic materials undergoing thermal expansion. *Mathematical Modelling and Numerical Analysis (M2AN)*, *45*, 477–504.
5. Carathéodory, C. (1909). Untersuchungen über die Grundlagen der Thermodynamik. *Mathematische Annalen*, *67*, 355–386.
6. Carstensen, C., Hackl, K., & Mielke, A. (2002). Non-convex potentials and microstructures in finite-strain plasticity. *Proceedings of the Royal Society of London Series A*, *458*(2018), 299–317.
7. Dal Maso, G., DeSimone, A., & Mora, M. G. (2006). Quasistatic evolution problems for linearly elastic-perfectly plastic materials. *Archive for Rational Mechanics and Analysis*, *180*(2), 237–291.
8. Dal Maso, G., DeSimone, A., & Solombrino, F. (2011). Quasistatic evolution for cam-clay plasticity: a weak formulation via viscoplastic regularization and time parametrization. *Calculus of Variations and Partial Differential Equations*, *40*(2), 125–181.
9. Gürses, E., Mainik, A., Miehe, C., & Mielke, A. (2006). Analytical and numerical methods for finite-strain elastoplasticity. In R. Helmig, A. Mielke, & B. Wohlmuth (Eds.), *Multifield problems in solid and fluid mechanics* (pp. 443–481). Berlin: Springer.
10. Gröger, K. (1978). Zur Theorie des quasi-statischen Verhaltens von elastisch-plastischen Körpern. *ZAMM—Journal of Applied Mathematics and Mechanics*, *58*(2), 81–88.
11. Hütter, M., & Svendsen, B. (2012). Thermodynamic model formulation for viscoplastic solids as general equations for non-equilibrium reversible-irreversible coupling. *Continuum Mechanics and Thermodynamics*, *24*, 211–227.
12. Johnson, C. (1976). Existence theorems for plasticity problems. *Journal de Mathematiques Pures et Appliques (9)*, *55*(4), 431–444.
13. Mainik, A., & Mielke, A. (2009). Global existence for rate-independent gradient plasticity at finite strain. *Journal of Nonlinear Science*, *19*(3), 221–248.
14. Miehe, C., & Stein, E. (1992). A canonical model of multiplicative elasto–plasticity. Formulation and aspects of numerical implementation. *European Journal of Mechanics endash; A/Solids*, *11*, 25–43.
15. Mielke, A. (2003). Energetic formulation of multiplicative elasto-plasticity using dissipation distances. *Continuum Mechanics and Thermodynamics*, *15*, 351–382.
16. Mielke, A. (2011). Formulation of thermoelastic dissipative material behavior using GENERIC. *Continuum Mechanics and Thermodynamics*, *23*(3), 233–256.
17. Mielke, A. (2011). On thermodynamically consistent models and gradient structures for thermoplasticity. *GAMM—Mitteilungen*, *34*(1), 51–58.
18. Mielke, A. (2016). On evolutionary Γ-convergence for gradient systems. In A. Muntean, J. Rademacher, & A. Zagaris (Eds.), *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity, Lecture Notes in Applied Math. Mechanics, 3*, 187–249. Springer.
19. Mielke, A. (2013). Thermomechanical modeling of energy-reaction-diffusion systems, including bulk-interface interactions. *Discrete and Continuous Dynamical Systems—Series S*, *6*(2), 479–499.
20. Mielke, A., & Roubíček, T. (2015). Rate-independent systems: theory and application. *Applied Mathematical Sciences*, *193*. Springer.
21. Mielke, A., & Stefanelli, U. (2015). Homogenizing the penrose-fife system via its gradient structure. *In preparation*.

22. Moreau, J.-J. (1974). 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)* (pp. 171–322). Rome: Edizioni Cremonese.
23. Onsager, L. (1931). Reciprocal relations in irreversible processes, I+II. *Physical Review*, *37*, 405–426. (part II, 38:2265–2279).
24. Ortiz, M., & Repetto, E. (1999). Nonconvex energy minimization and dislocation structures in ductile single crystals. *Journal of the Mechanics and Physics of Solids*, *47*(2), 397–462.
25. Ortiz, M., & Stainier, L. (1999). The variational formulation of viscoplastic constitutive updates. *Computer Methods in Applied Mechanics and Engineering*, *171*(3–4), 419–444.
26. Ortiz, M., Repetto, E., & Stainier, L. (2000). A theory of subgrain dislocation structures. *Journal of the Mechanics and Physics of Solids*, *48*, 2077–2114.
27. Penrose, O., & Fife, P. C. (1990). Thermodynamically consistent models of phase-field type for the kinetics of phase transitions. *Physica D*, *43*(1), 44–62.
28. Penrose, O., & Fife, P. C. (1993). On the relation between the standard phase-field model and a "thermodynamically consistent" phase-field model. *Physica D*, *69*(1–2), 107–113.
29. Simo, J., & Ortiz, M. (1985). A unified approach to finite deformation elastoplastic analysis based on the use of hyperelastic constitutive relations. *Computer Methods in Applied Mechanics and Engineering*, *49*, 221–245.
30. Yang, Q., Stainier, L., & Ortiz, M. (2006). A variational formulation of the coupled thermomechanical boundary-value problem for general dissipative solids. *Journal of the Mechanics and Physics of Solids*, *54*, 401–424.

Comparison of Isotropic Elasto-Plastic Models for the Plastic Metric Tensor $C_p = F_p^T F_p$

Patrizio Neff and Ionel-Dumitrel Ghiba

Abstract We discuss in detail existing isotropic elasto-plastic models based on 6-dimensional flow rules for the positive definite plastic metric tensor $C_p = F_p^T F_p$ and highlight their properties and interconnections. We show that seemingly different models are equivalent in the isotropic case.

1 Introduction

Since the early days of the introduction of the multiplicative decomposition into computational elasto-plasticity, the need was felt to reduce the level of complexity and to discard the concept of a plastic rotation in the completely isotropic setting. This means to consider a flow rule not for the **plastic distortion** F_p (9-dimensional) [5, 6, 11, 30, 34, 36, 43, 44], but to consider directly a flow rule for the **plastic metric tensor** $C_p = F_p^T F_p \in \mathrm{PSym}(3)$ (6-dimensional) [1, 10, 38, 39, 45], which is then automatically invariant under left-multiplication of F_p with a plastic rotation. The plastic distortion is in general incompatible $F_p \neq \nabla\psi_p$, as is the plastic metric $C_p \neq \nabla\psi_p^T\nabla\psi_p$. A formulation in the plastic metric C_p is particular attractive because it circumvents problems associated with the intermediate configuration introduced by the multiplicative decomposition, which is trivially non-unique since

P. Neff · I.-D. Ghiba (✉)
Lehrstuhl für Nichtlineare Analysis und Modellierung, Fakultät für Mathematik,
Universität Duisburg-Essen, Thea-Leymann Str. 9, 45127 Essen, Germany
e-mail: dumitrel.ghiba@uni-due.de; dumitrel.ghiba@uaic.ro

P. Neff
e-mail: patrizio.neff@uni-due.de

I.-D. Ghiba
Department of Mathematics, Alexandru Ioan Cuza University of Iaşi,
Blvd. Carol I, No. 11, 700506 Iaşi, Romania

I.-D. Ghiba
Octav Mayer Institute of Mathematics of the Romanian Academy, Iaşi Branch,
700505 Iaşi, Romania

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_8

$$F = F_e \cdot F_p = F_e \cdot Q^T \cdot Q \cdot F^p = F_e^* \cdot F_p^*, \quad Q \in \mathrm{SO}(3).$$

Several proposals with the aim of removing the non-uniqueness of the intermediate configuration have been given in the literature. Our comparative study is related to the following models: Simo's model [41] (Reese and Wriggers [36], Miehe [21]); Miehe's model [22]; Lion's model [17] (Helm [12]), (Dettmer-Reese [6]); Simo and Hughes' model [42]; Helm's model [12] (Vladimirov, Pietryga and Reese [45], Shutov and Kreißig [39], Reese and Christ [35], Brepols, Vladimirov and Reese [1], Shutov and Ihlemann [38]); Grandi and Stefanelli's model [10] (Frigeri and Stefanelli [7]). All these models are given with respect to different configurations, either the reference configuration, the intermediate configuration or the current configuration. In order to be able to compare them, it is necessary to transform all to the same configuration for that purpose. In our case we choose the reference configuration. Moreover, any explicit dependence on F_p instead of C_p in the model formulation must be able to be subsumed into a dependence on C_p alone in the isotropic case. A major body of our work consists in showing this for the models under consideration.

The paper is structured as follows. After a paragraph giving some definitions which generalize the concepts from small strain-additive plasticity to finite strain plasticity we established some auxiliary results. Then we discuss existing 6-dimensional flow rules from the literature. The main properties of the investigated isotropic plasticity models are summarized in Figs. 1 and 2. Finally, in the appendix, we obtain explicit formulas for some of the isotropic plasticity models.

1.1 Consistent Isotropic Finite Plasticity Model for the Plastic Metric Tensor C_p

In this paper, we use the standard Euclidean scalar product on $\mathbb{R}^{3\times3}$ given by $\langle X, Y\rangle := \mathrm{tr}(XY^T)$, and thus the Frobenius tensor norm is $\|X\|^2 = \langle X, X\rangle$. The identity tensor on $\mathbb{R}^{3\times3}$ will be denoted by $\mathbb{1}$, so that $\mathrm{tr}(X) = \langle X, \mathbb{1}\rangle$. We let $\mathrm{Sym}(3)$ and $\mathrm{PSym}(3)$ denote the symmetric and positive definite symmetric tensors respectively. We adopt the usual abbreviations of Lie-group theory. Here and in the following the superscript ^{T} is used to denote transposition, $\mathrm{sym}\, X = \frac{1}{2}(X + X^T)$ denotes the symmetric part of the matrix $X \in \mathbb{R}^{3\times3}$, while $\mathrm{dev}_3\, X = X - \frac{1}{3}\,\mathrm{tr}(X)\cdot\mathbb{1}$ represents the deviatoric (trace free) part of the matrix X.

The classical concept of **associated perfect plasticity** is uniquely defined in the case of small strain-additive plasticity. In this case the total symmetric strain is decomposed additively into elastic and plastic parts $\varepsilon = \varepsilon_e + \varepsilon_p$ and the rate-independent evolution law for the symmetric plastic strain ε_p is given in subdifferential format

$$\frac{\mathrm{d}}{\mathrm{dt}}[\varepsilon_p] \in \partial\chi(\Sigma_{\mathrm{lin}}), \qquad \mathrm{tr}(\varepsilon_p) = 0,$$

where $\partial \chi$ is the subdifferential of the indicator function χ of the convex elastic domain

$$\mathscr{E}_{\mathrm{e}}(\Sigma_{\mathrm{lin}}, \frac{2}{3}\boldsymbol{\sigma}_{\mathrm{y}}^2) = \left\{\Sigma_{\mathrm{lin}} \in \mathrm{Sym}(3) \Big| \; \|\operatorname{dev}_3 \Sigma_{\mathrm{lin}}\|^2 \le \frac{2}{3}\boldsymbol{\sigma}_{\mathrm{y}}^2\right\} \subset \mathrm{Sym}(3)$$

and $\Sigma_{\mathrm{lin}} := -D_{\varepsilon_p}[W_{\mathrm{lin}}(\varepsilon - \varepsilon_p)]$ is the thermodynamic driving stress of the plastic process. Here, Σ_{lin} is clearly symmetric.

In such a way, the principle of **maximum dissipation** (equivalent to the convexity of the elastic domain and normality of the flow direction) is satisfied. The structure of associated flow rules in geometrically nonlinear theories is by far not as trivial as in the geometrically linear models. However, in this work we use:

Definition 1 (*geometrically nonlinear associated plastic flow*) We call a plastic flow rule for some plastic variable P (whether symmetric or not) associated, whenever the flow rule can be written as

$$\frac{\mathrm{d}}{\mathrm{dt}}[P]\,P^{-1} \in \partial\chi(\Sigma) \qquad \text{or} \qquad \sqrt{P}\frac{\mathrm{d}}{\mathrm{dt}}[P^{-1}]\,\sqrt{P} \in f = \partial\chi(\Sigma),$$

where Σ is some symmetric or non-symmetric stress tensor. Here, $\frac{\mathrm{d}}{\mathrm{dt}}[P^{-1}]\,P$ is the correct format for the time derivative (it will lead to an exponential update, see the implicit method based on the exponential mapping considered in [40]). Moreover, we require that χ is the indicator function of some **convex** domain in the Σ-stress space.

After liniarization (small strain-additive approximation) this condition is equivalent to classical associated plasticity. Further, let us also remark that a metric is by definition symmetric and positive definite, i.e. $C_p \in \mathrm{PSym}(3)$.

Definition 2 (*consistent isotropic finite plasticity model for plastic metric tensor C_p*) We say that an associated plastic flow rule, in the sense of Definition 1, for the plastic metric tensor C_p is consistent, whenever:

(i) it is thermodynamically correct, i.e. the reduced dissipation inequality is satisfied;
(ii) plastic incompressibility: the constraint $\det C_p(t) = 1$ for all $t \ge 0$ follows from the flow rule;
(iii) $C_p(t) \in \mathrm{PSym}(3)$ for all $t > 0$ if $C_p(0) \in \mathrm{PSym}(3)$.

As we will see from the next Lemma 4, our requirement (iii) follows if $C_p(t) \in \mathrm{Sym}(3)$ for all $t \ge 0$, $C_p(0) \in \mathrm{PSym}(3)$ and if (ii) is satisfied.
We finish our setup of preliminaries with the following definitions:

Definition 3 (*reduced dissipation inequality-thermodynamic consistency*) For a given energy W, we say that the reduced dissipation inequality along the plastic evolution is satisfied if and only if

$$\frac{\mathrm{d}}{\mathrm{dt}}[W(F\,F_p^{-1}(t)] = \frac{\mathrm{d}}{\mathrm{dt}}[\widetilde{W}(C\,C_p^{-1}(t)] = \frac{\mathrm{d}}{\mathrm{dt}}[\Psi(C, C_p(t)] \leq 0$$

for all constant in time F (viz. $C = F^T F$), depending in which format the elastic energy is given.

Definition 4 (*Loss of ellipticity in the elastic domain*) We say that the elasto-plastic formulation preserves ellipticity in the elastic domain whenever the purely elastic response in elastic unloading of the material remains rank-one convex for arbitrary large given plastic pre-distortion, see [9, 29].

1.2 Auxiliary Results

We consider the multiplicative decomposition of the deformation gradient [13–16, 28, 31] and we define, accordingly, the elastic and plastic strain tensors

$$\begin{aligned} C_e &:= F_e^T F_e \in \mathrm{PSym}(3), \qquad B_e := F_e\, F_e^T \in \mathrm{PSym}(3),\\ C_p &:= F_p^T F_p \in \mathrm{PSym}(3). \end{aligned}$$

Let us also define the stress tensors

$$\begin{aligned} \Sigma :&= 2\,C\,D_C[\widehat{W}(C)] = 2\,D_{\log C}[\overline{W}(\log C)] = D_{\log U}[\check{W}(\log U)]\\ &= U\,D_U[W(U)] = F^T D_F[W(F)]\,,\\ \tau :&= 2\,D_B[\widehat{W}(B)]\,B = 2\,D_{\log B}[\overline{W}(\log B)] = D_{\log V}[\check{W}(\log V)]\\ &= V\,D_V[W(V)] = 2\,F\,D_C[\widehat{W}(C)]\,F^T. \end{aligned}$$

The tensor $\Sigma = C \cdot S_2(C)$, where $S_2 = 2\,D_C[W(C)]$ is the second Piola-Kirchhoff stress tensor, is sometimes called the **Mandel stress tensor** and it holds $\mathrm{dev}_3\,\Sigma_e = \mathrm{dev}_3\,\Sigma_{\mathrm{E}}$, where Σ_{E} is the elastic **Eshelby tensor**

$$\Sigma_{\mathrm{E}} := F_e^T D_{F_e}[W(F_e)] - W(F_e)\cdot \mathbb{1} = D_{\log C_e}[\overline{W}(\log C_e)] - \overline{W}(\log C_e)\cdot \mathbb{1},$$

driving the plastic evolution (see e.g. [2–4, 20, 28]), while τ is the **Kirchhoff stress tensor** and Σ_e is defined in Remark 1.

Remark 1 We also need to consider the following elasto-plastic stress tensors:

$$\begin{aligned} \Sigma_e :&= 2\,C_e\,D_{C_e}[\widehat{W}(C_e)] = 2\,D_{\log C_e}[\overline{W}(\log C_e)] = D_{\log U_e}[\check{W}(\log U_e)]\\ &= U_e\,D_{U_e}[W(U_e)] = F_e^T D_{F_e}[W(F_e)]\,,\\ \tau_e :&= 2\,D_{B_e}[\widehat{W}(B_e)]\,B_e = 2\,D_{\log B_e}[\overline{W}(\log B_e)] = D_{\log V_e}[\check{W}(\log V_e)]\\ &= V_e\,D_{V_e}[W(V_e)] = 2\,F_e\,D_{C_e}[\widehat{W}(C_e)]\,F_e{}^T. \end{aligned}$$

The following relation holds true:

$$\Sigma = F^T \tau\, F^{-T}, \qquad \Sigma_e = F_e^T \tau_e\, F_e^{-T}. \tag{1}$$

Note that (1) is not at variance with symmetry of Σ and Σ_e in case of isotropy.

Using the fact that for given $F_e \in \mathrm{GL}^+(3)$ it holds $\|F_e^T S F_e^{-T}\|^2 \geq \frac{1}{2}\|S\|^2$ for all $S \in \mathrm{Sym}(3)$, the constant being independent of F_e [30], we obtain the estimate

$$\|\operatorname{dev}_3 \Sigma_e\| = \|F_e^T(\operatorname{dev}_3 \tau_e)F_e^{-T}\| \geq \frac{1}{\sqrt{2}}\|\operatorname{dev}_3 \tau_e\|,$$

which is valid for general anisotropic materials. Indeed, since

$$\begin{aligned}\operatorname{dev}_3 \Sigma_e = \operatorname{dev}_3(F_e^T \tau_e F_e^{-T}) &= F_e^T \tau_e F_e^{-T} - \frac{1}{3}\mathrm{tr}(F_e^T \tau_e F_e^{-T}) \cdot \mathbb{1} \\ &= F_e^T(\tau_e - \frac{1}{3}\mathrm{tr}(\tau_e)) \cdot \mathbb{1})F_e^{-T},\end{aligned}$$

we have

$$\operatorname{dev}_3 \Sigma_e = F_e^T(\operatorname{dev}_3 \tau_e)F_e^{-T}, \qquad \operatorname{dev}_3 \tau_e = F_e^{-T}(\operatorname{dev}_3 \Sigma_e)F_e^T, \qquad \mathrm{tr}(\Sigma_e) = \mathrm{tr}(\tau_e).$$

However, $\|\operatorname{dev}_3 \Sigma_e\| \neq \|\operatorname{dev}_3 \tau_e\|$ for general anisotropic materials. Let us remark that for elastically isotropic materials we have from the representation formula for isotropic tensor functions

$$\begin{aligned} D_{C_e}[\widehat{W}(C_e)] &= \alpha_1\, \mathbb{1} + \alpha_2\, C_e + \alpha_3\, C_e^2 \in \mathrm{Sym}(3), \qquad (2)\\ \Sigma_e = 2\, C_e \cdot D_{C_e}[\widehat{W}(C_e)] &= 2\, C_e\, (\alpha_1\, \mathbb{1} + \alpha_2\, C_e + \alpha_3\, C_e^2) \in \mathrm{Sym}(3),\end{aligned}$$

where

$$\alpha_1 = \frac{2}{I_3^{1/2}(C_e)}\left[I_2(C_e)\,\frac{\partial W}{\partial I_2(C_e)} + I_3(C_e)\frac{\partial W}{\partial I_3(C_e)}\right], \quad \alpha_2 = \frac{2}{I_3^{1/2}(C_e)}\frac{\partial W}{\partial I_1(C_e)},$$
$$\alpha_3 = -2\, I_3^{1/2}(C_e)\frac{\partial W}{\partial I_2(C_e)}$$

are scalar functions of the invariants of C_e, which are functions of $C\, C_p^{-1}$, see Lemma 2. This leads us to

Lemma 1 *For the isotropic case* $\|\operatorname{dev}_3 \Sigma_e\| = \|\operatorname{dev}_3 \tau_e\|$.

Proof For the isotropic case we have $\tau_e\, B_e = B_e\, \tau_e$, which implies

$$\begin{aligned}\|\operatorname{dev}_3 \Sigma_e\|^2 &= \langle F_e^T\, (\operatorname{dev}_3 \tau_e)\, F_e^{-T}, F_e^T\, (\operatorname{dev}_3 \tau_e)\, F_e^{-T}\rangle = \langle B_e\, (\operatorname{dev}_3 \tau_e), (\operatorname{dev}_3 \tau_e)\, B_e^{-1}\rangle \\ &= \|\operatorname{dev}_3 \tau_e\|^2.\end{aligned}$$

We also consider the following tensor

$$\widetilde{\Sigma} := 2\,C\,D_C[\widetilde{W}(C\,C_p^{-1})] = 2\,C\,D[\widetilde{W}(C\,C_p^{-1})]\,C_p^{-1} \notin \mathrm{Sym}(3), \tag{3}$$

which is not symmetric, in general. For instance, for the simplest Neo-Hooke energy $W(F_e) = \mathrm{tr}(C_e) = \mathrm{tr}(C\,C_p^{-1})$ we have $D\widetilde{W}(C\,C_p^{-1}){=}\mathbb{1}$ and $\widetilde{\Sigma}{=}2\,C\,C_p^{-1} \notin \mathrm{Sym}(3)$.

Lemma 2 *Any isotropic and objective free energy W defined in terms of F_e can be expressed as*

$$W(F_e) = \widetilde{W}(C\,C_p^{-1}) = \widetilde{W}(F^T F(F_p^T F_p)^{-1}). \tag{4}$$

Proof It is clear that any objective elastic energy $W(F_e)$ which is isotropic w.r.t. F_e, can be expressed in terms of the invariants of C_e, i.e.

$$\begin{aligned}
&W(F_e) = \Psi(I_1(C_e), I_2(C_e), I_3(C_e)),\\
&I_1(C_e) = \mathrm{tr}(C_e) = \mathrm{tr}(B_e), \quad I_2(C_e) = \mathrm{tr}(\mathrm{Cof}\,C_e) = \mathrm{tr}(\mathrm{Cof}\,B_e),\\
&I_3(C_e) = \det C_e = \det B_e.
\end{aligned}$$

Now every invariant can be rewritten as follows

$$\begin{aligned}
I_1(C_e) &= \langle C_e, \mathbb{1}\rangle = \langle F_e^T F_e, \mathbb{1}\rangle = \langle F_p^{-T} F^T\,(F\,F_p^{-1}), \mathbb{1}\rangle = \langle C, C_p^{-1}\rangle = \mathrm{tr}(C\,C_p^{-1})\\
&= I_1(C\,C_p^{-1}), \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5)\\
I_2(C_e) &= \langle \mathrm{Cof}\,C_e, \mathbb{1}\rangle = \det C_e\,\langle C_e^{-T}, \mathbb{1}\rangle = \det(F_p^{-T} C\,F_p^{-1})\,\langle [F_p^{-T} C\,F_p^{-1}]^{-T}, \mathbb{1}\rangle\\
&= \det C \det C_p^{-1}\,\langle C^{-T}, F_p^T\,F_p\rangle = \det(C\,C_p^{-1})\,\langle C^{-T} C_p^T, \mathbb{1}\rangle = \mathrm{tr}(\mathrm{Cof}(C\,C_p^{-1}))\\
&= I_2(C\,C_p^{-1}),\\
I_3(C_e) &= \det\,C_e = \det(F_p^{-T} C\,F_p^{-1}) = \det\,C\,\det\,C_p^{-1} = I_3(C\,C_p^{-1}).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
W(F_e) &= \Psi(I_1(C_e), I_2(C_e), I_3(C_e))\\
&= \Psi(I_1(C\,C_p^{-1}), I_2(C\,C_p^{-1}), I_3(C\,C_p^{-1})) = \widetilde{W}(C\,C_p^{-1}),
\end{aligned}$$

and the proof is complete. □

Remark 2 Since the principal invariants $I_k, k = 1, 2, 3$ are the coefficients of the characteristic polynomial and $I_1(C_e) = I_1(C\,C_p^{-1})$, $I_2(C_e) = I_2(C\,C_p^{-1})$, $I_3(C_e) = I_3(C\,C_p^{-1})$, the eigenvalues of C_e and $C\,C_p^{-1}$ coincide. Clearly, $C_e \in \mathrm{PSym}(3)$, however $C\,C_p^{-1} \notin \mathrm{Sym}(3)$ in general, unless C and C_p^{-1} commute.

Lemma 3 *The introduced stress tensors Σ_e, $\widetilde{\Sigma}$, τ_e are related as follows*

$$\Sigma_e = F_p^{-T}\,\widetilde{\Sigma}\,F_p^T, \qquad \widetilde{\Sigma} = F^T\,\tau_e\,F^{-T}.$$

Proof For arbitrary increment $H \in \mathbb{R}^{3\times 3}$, we compute

$$\begin{aligned}\langle D_F[W(F_e)], H\rangle = \langle D_F[W(F\,F_p^{-1})], H\rangle &= \langle D_{F_e}[W(F_e)], H\,F_p^{-1}\rangle\\ &= \langle D_{F_e}[W(F_e)]\,F_p^{-T}, H\rangle.\end{aligned}$$

On the other hand, we deduce

$$\begin{aligned}\langle D_F[\widetilde{W}(C\,C_p^{-1})], H\rangle &= \langle D_F[\widetilde{W}(F^T F\,C_p^{-1})], H\rangle\\ &= \langle D[\widetilde{W}(C\,C_p^{-1})], F^T H C_p^{-1} + H^T F\,C_p^{-1}\rangle\\ &= 2\,\langle F\,\mathrm{sym}[D[\widetilde{W}(C\,C_p^{-1})]C_p^{-1}], H\rangle,\end{aligned}$$

for all $H \in \mathbb{R}^{3\times 3}$. In view of Lemma 2 we have $W(F_e) = \widetilde{W}(C\,C_p^{-1})$. Therefore, we obtain

$$2\,F\,\mathrm{sym}[D[\widetilde{W}(C\,C_p^{-1})]C_p^{-1}] = D_{F_e}[W(F_e)]\,F_p^{-T},$$

and further

$$F_e^T\,D_{F_e}[W(F_e)]\,F_p^{-T} = 2\,F_e^T F\,\mathrm{sym}[D[\widetilde{W}(C\,C_p^{-1})]C_p^{-1}] = 2\,F_p^{-T} C\,\mathrm{sym}[D[\widetilde{W}(C\,C_p^{-1})]C_p^{-1}].$$

The above relation implies

$$\Sigma_e = F_e^T\,D_{F_e}[W(F_e)] = 2\,F_p^{-T} C\,D_C[\widetilde{W}(C\,C_p^{-1})]\,F_p^T = F_p^{-T}\,\widetilde{\Sigma}\,F_p^T.$$

Therefore, using Remark 1 the proof is complete. □

Next, we introduce a helpful lemma.

Lemma 4 *If* $t \mapsto C_p(t) \in \mathbb{R}^{3\times 3}$ *is continuous and satisfies:*

$$\left.\begin{array}{r}\det C_p(t) = 1 \quad \textit{for all} \quad t > 0,\\ C_p(0) \in \mathrm{PSym}(3),\\ C_p(t) \in \mathrm{Sym}(3) \quad \textit{for all} \quad t > 0\end{array}\right\} \quad \Rightarrow \quad C_p(t) \in \mathrm{PSym}(3) \quad \textit{for all} \quad t > 0.$$

Proof Using Cardano's formula and due to the symmetry of C_p, the continuity of the map $t \mapsto C_p(t)$ implies the continuity of mappings $t \mapsto \lambda_i(t), i = 1, 2, 3$, where $\lambda_i(t) \in \mathbb{R}$ are the eigenvalues of $C_p(t)$. Since $\lambda_i(0) > 0$ and $\lambda_1(t)\lambda_2(t)\lambda_3(t) = 1$ for all $t > 0$, it follows that $\lambda_i(t) > 0$ for all $t > 0$ and the proof is complete. □

We can slightly weaken the assumption in the previous lemma: $\det C_p(t) > 0$ for all $t > 0$ is sufficient.

2 The Simo-Miehe 1992 Spatial Model

In the remainder of this paper we discuss different proposal from the literature for plasticity models in C_p. Simo [41] (see also Reese and Wriggers [36] and Miehe [21, p. 72, Proposition 5.25]) considered the spatial flow rule in the form

$$-\frac{1}{2}\,\mathscr{L}_v(B_e) = \lambda_{\mathrm{p}}^{+}\,\partial_{\tau_e}\Phi(\tau_e)\cdot B_e, \tag{6}$$

where the Lie-derivative $\mathscr{L}_v(B_e)$ is given by $\mathscr{L}_v(B_e) := F\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\,F^T \in \mathrm{Sym}(3)$, the tensor $\tau_e = 2\,\partial_{B_e}W(B_e)\cdot B_e$ is the symmetric Kirchhoff stress tensor, the yield function $\Phi(\tau_e) = \|\operatorname{dev}_3\tau_e\| - \sqrt{\frac{2}{3}}\sigma_{\mathbf{y}}$ and the plastic multiplier λ_{p}^{+} satisfies the Karush-Kuhn-Tucker (KKT)-optimality constraints

$$\lambda_{\mathrm{p}}^{+} \geq 0, \qquad \Phi(\tau_e) \leq 0, \qquad \lambda_{\mathrm{p}}^{+}\,\Phi(\tau_e) = 0. \tag{7}$$

The flow rule (6) is equivalent with

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] &= -2\,\lambda_{\mathrm{p}}^{+}\,F^{-1}[\partial_{\tau_e}\Phi(\tau_e)\cdot B_e]\,F^{-T}\\ &= -2\,\lambda_{\mathrm{p}}^{+}\,F^{-1}\left[\frac{\operatorname{dev}_3\tau_e}{\|\operatorname{dev}_3\tau_e\|}\cdot B_e\right]F^{-T},\end{aligned} \tag{8}$$

which, in view of the properties (7) of λ_{p}^{+}, can be written with a subdifferential

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] \in -2\,F^{-1}\left[\partial_{\tau_e}\chi\,(\operatorname{dev}_3\tau_e)\cdot B_e\right]\,F^{-T}, \tag{9}$$

where χ is the indicator function of the elastic domain

$$\mathscr{E}_{\mathrm{e}}(\tau_{\mathrm{e}}, \frac{2}{3}\,\boldsymbol{\sigma}_{\mathbf{y}}^2) = \left\{\tau_e \in \mathrm{Sym}(3)\,\middle|\,\|\operatorname{dev}_3\tau_e\|^2 \leq \frac{2}{3}\,\boldsymbol{\sigma}_{\mathbf{y}}^2\right\} = \{\tau_e \in \mathrm{Sym}(3)\,|\,\Phi(\tau_e) \leq 0\}.$$

The subdifferential $\partial\chi\,(\operatorname{dev}_3\tau_e)$ of the indicator function χ is the normal cone

$$\mathscr{N}(\mathscr{E}_{\mathrm{e}}(\tau_e, \frac{2}{3}\,\boldsymbol{\sigma}_{\mathbf{y}}^2); \operatorname{dev}_3\tau_e) = \begin{cases} 0, & \tau_e \in \operatorname{int}(\mathscr{E}_{\mathrm{e}}(\tau_e, \frac{2}{3}\boldsymbol{\sigma}_{\mathbf{y}}^2))\\ \{\lambda_{\mathrm{p}}^{+}\,\frac{\operatorname{dev}_3\tau_e}{\|\operatorname{dev}_3\tau_e\|}\,|\,\lambda_{\mathrm{p}}^{+} \in \mathbb{R}_+\}, & \tau_e \notin \operatorname{int}(\mathscr{E}_{\mathrm{e}}(\tau_e, \frac{2}{3}\boldsymbol{\sigma}_{\mathbf{y}}^2)).\end{cases}$$

We deduce (see the model Eq. (5.25) from [21]) an equivalent definition for $\mathscr{L}_v(B_e)$ given by

$$-\frac{1}{2}\,\mathscr{L}_v(B_e) = \lambda_{\mathrm{p}}^{+}\,\frac{\operatorname{dev}_3\tau_e}{\|\operatorname{dev}_3\tau_e\|}\cdot B_e.$$

Since $C_p = F^T\, B_e^{-1}\, F$ we have $\mathscr{L}_v(B_e){=}F\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\,F^T{=}F\,\left(\frac{\mathrm{d}}{\mathrm{dt}}[F^{-1}B_eF^{-T}]\right)\,F^T$. On the other hand, from (8) it follows that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\,C_p &= -2\,\lambda_{\mathrm{p}}^+\,F^{-1}\left[\frac{\mathrm{dev}_3\,\tau_e}{\|\,\mathrm{dev}_3\,\tau_e\|}\cdot B_e\right]F^{-T}\,F^T\,B_e^{-1}\,F\\ &\in -2\,F^{-1}\,\partial_{\tau_e}\chi\,(\mathrm{dev}_3\,\tau_e)\,F.\end{aligned}\tag{10}$$

Since

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[\det C_p^{-1}] &= \langle \mathrm{Cof}\,C_p^{-1},\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle = \det C_p^{-1}\langle C_p,\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle\\ &= \det C_P^{-1}\langle \mathbb{1},\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]C_p\rangle,\end{aligned}\tag{11}$$

from the flow rule (8) together with $\det C_p(0) = 1$ and $\mathrm{tr}(F^{-1}\,\mathrm{dev}_3\,\tau_e\,F) = 0$ it follows at once that $\det C_p(t) = 1$ for all $t \geq 0$.

The next step is to prove that the flow rule (6) implies $\frac{\mathrm{d}}{\mathrm{dt}}[W(F_e)] \leq 0$ at fixed F, i.e. the reduced dissipation inequality is satisfied. We compute for fixed in time F

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[W(FF_p^{-1})] &= \langle D_{F_e}W(F_e),\,F\frac{\mathrm{d}}{\mathrm{dt}}[F_p^{-1}]\rangle = \langle D_{F_e}W(F_e),\,FF_p^{-1}F_p\frac{\mathrm{d}}{\mathrm{dt}}[F_p^{-1}]\rangle\\ &= \langle F_e^T D_{F_e}W(F_e),\,F_p\frac{\mathrm{d}}{\mathrm{dt}}[F_p^{-1}]\rangle = \langle \Sigma_e,\,F_p\frac{\mathrm{d}}{\mathrm{dt}}[F_p^{-1}]\rangle\\ &= -\langle \Sigma_e,\,\underbrace{\mathrm{sym}(\frac{\mathrm{d}}{\mathrm{dt}}[F_p]F_p^{-1})}_{D_p}\rangle,\end{aligned}\tag{12}$$

since $\Sigma_e \in \mathrm{Sym}(3)$. We also have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[C_p] &= \frac{\mathrm{d}}{\mathrm{dt}}[F_p^T F_p] = \frac{\mathrm{d}}{\mathrm{dt}}[F_p^T]F_p + F_p^T\frac{\mathrm{d}}{\mathrm{dt}}[F_p]\\ &= F_p^T\left(F_p^{-T}\frac{\mathrm{d}}{\mathrm{dt}}[F_p^T]\right)F_p + F_p^T\left(\frac{\mathrm{d}}{\mathrm{dt}}[F_p]F_p^{-1}\right)F_p = 2\,F_p^T D_p F_p,\end{aligned}$$

where $D_p := \mathrm{sym}\left(\frac{\mathrm{d}}{\mathrm{dt}}[F_p]F_p^{-1}\right)$. Hence, we easily deduce the representation $D_p = \frac{1}{2}\,F_p^{-T}\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p]F_p^{-1}$. Therefore, with (12) we obtain

$$\frac{\mathrm{d}}{\mathrm{dt}}[W(FF_p^{-1})] = -\langle \Sigma_e,\,\frac{1}{2}\,F_p^{-T}\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p]F_p^{-1}\rangle.\tag{13}$$

Moreover, since $\Sigma_e = F_e^T \tau_e\, F_e^{-T}$, we deduce

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[W(FF_p^{-1})] &= -\frac{1}{2}\langle F_e^T \tau_e\, F_e^{-T},\ F_p^{-T}\frac{\mathrm{d}}{\mathrm{dt}}[C_p]F_p^{-1}\rangle = \frac{1}{2}\langle F_e^T \tau_e\, F_e^{-T},\ F_p\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]F_p^T\rangle \\ &= \frac{1}{2}\langle\, F_p^T\, F_e^T \tau_e\, F_e^{-T} F_p, \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle = \frac{1}{2}\langle\, F^T \tau_e\, B_e^{-1}\, F, \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle \\ &= \frac{1}{2}\langle \tau_e, F\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\, F^T\, B_e^{-1}\rangle. \end{aligned} \tag{14}$$

The flow rule (8) implies

$$\frac{\mathrm{d}}{\mathrm{dt}}[W(FF_p^{-1})] = -\lambda_\mathrm{p}^+\,\langle \tau_e, \frac{\mathrm{dev}_3\,\tau_e}{\|\,\mathrm{dev}_3\,\tau_e\|}\rangle = -\lambda_\mathrm{p}^+\,\|\,\mathrm{dev}_3\,\tau_e\| \le 0. \tag{15}$$

In view of the definition of $\Sigma_e = F_e^T\,\tau_e\,F_e^{-T}$ we have $F^{-1}[\,\tau_e\,B_e]F^{-T} = F_p^{-1}[\Sigma_e]\,F_p^{-T}$. For the isotropic case it holds $\tau_e\,B_e = B_e\,\tau_e$. Hence,

$$\begin{aligned} F^{-1}[\tau_e\,B_e]F^{-T} &= F^{-1}[\,B_e\tau_e]F^{-T} = F_p^{-1}F_e^{-1}[\,F_e\,F_e^T\tau_e]F_e^{-T}\,F_p^{-T} \\ &= F_p^{-1}[\,F_e^T\tau_e F_e^{-T}]F_p^{-T} = F_p^{-1}[\Sigma_e]F_p^{-T}. \end{aligned}$$

We also observe $F^{-1}[\,\mathrm{tr}(\tau_e)\,B_e]F^{-T} = F_p^{-1}[\mathrm{tr}(\Sigma_e)]F_p^{-T}$. Thus, we obtain

$$F^{-1}[\mathrm{dev}_3\,\tau_e\,B_e]F^{-T} = F_p^{-1}[\mathrm{dev}_3\,\Sigma_e]F_p^{-T}.$$

Together with Remark 1 this implies that

$$F^{-1}\left[\frac{\mathrm{dev}_3\,\tau_e}{\|\,\mathrm{dev}_3\,\tau_e\|}\,B_e\right]F^{-T} = F_p^{-1}\left[\frac{\mathrm{dev}_3\,\Sigma_e}{\|\,\mathrm{dev}_3\,\Sigma_e\|}\right]F_p^{-T}. \tag{16}$$

Therefore, in the isotropic case, the flow rule (9) has a subdifferential structure:

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] \in -2\,F_p^{-1}\,[\partial_{\Sigma_e}\chi\,(\mathrm{dev}_3\,\Sigma_e)]\,F_p^{-T}, \tag{17}$$

where χ is the indicator function of the elastic domain

$$\mathscr{E}_\mathrm{e}(\Sigma_\mathrm{e}, \frac{2}{3}\boldsymbol{\sigma}_\mathrm{y}^2) = \left\{\Sigma_e \in \mathrm{Sym}(3)\middle|\ \|\,\mathrm{dev}_3\,\Sigma_e\|^2 \le \frac{2}{3}\,\boldsymbol{\sigma}_\mathrm{y}^2\right\}.$$

In view of the above equivalent representations of the flow rule, we may summarize the properties of the Simo-Miehe 1992 model:

(i) from (8) it follows, in the isotropic case (in which τ_e and B_e commute), that $C_p(t) \in \mathrm{Sym}(3)$;

(ii) plastic incompressibility: from (10) and (11) it follows that $\det C_p(t) = 1$, since the right hand side is trace-free;

(iii) for the isotropic case, the right hand-side of (8) is a function of C_p^{-1} and C alone, since $B_e = F\,C_p^{-1}\,F^T$ and $F^{-1}B_e F = F^{-1}F\,C_p^{-1}\,F^T F = C_p^{-1}\,C$;

(iv) from (i) and (ii) together and using Lemma 4 it follows that $C_p(t) \in \mathrm{PSym}(3)$;

(v) it is thermodynamically correct;

(vi) the right hand side in the representation (8) is not the subdifferential of the indicator function of some convex domain in some stress space. However, this model is an associated plasticity model in the isotropic case, see Propositions 1 and 3.

3 The Miehe 1995 Referential Model

Shutov [37] interpreted that Miehe in [22] considered the flow rule[1]

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\,C_p = -\lambda_{\mathrm{p}}^{+}\,D_{\widetilde{\Sigma}}\Phi(\widetilde{\Sigma}), \tag{18}$$

where $\widetilde{\Sigma} = 2\,C\,D_C[\widetilde{W}(C\,C_p^{-1})]$ and

$$\Phi(\widetilde{\Sigma}) = \|\,\mathrm{dev}_3\,\tau_e\| - \sqrt{\frac{2}{3}}\,\sigma_{\mathbf{y}} = \sqrt{\mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2)} - \sqrt{\frac{2}{3}}\,\sigma_{\mathbf{y}}.$$

In this model, it is important to note that it is not the Frobenius norm of $\mathrm{dev}_3\,\widetilde{\Sigma}$ which is used in the yield function Φ. Instead, in the denominator $\mathscr{F} := \sqrt{\mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2)}$ is considered, see Eq. (52) from [39]. Since $\mathrm{dev}_3\,\widetilde{\Sigma} \notin \mathrm{Sym}(3)$, it follows that $\mathscr{F} := \sqrt{\mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2)} \neq \|\,\mathrm{dev}_3\,\widetilde{\Sigma}\|$. Indeed, we have

$$\begin{aligned}\sqrt{\mathrm{tr}[(\mathrm{dev}_3\,\widetilde{\Sigma})^2]} = \|\,\mathrm{dev}_3(\widetilde{\Sigma})\| \quad &\Leftrightarrow \quad \langle \mathrm{dev}_3\,\widetilde{\Sigma}, (\mathrm{dev}_3\,\widetilde{\Sigma})^T\rangle = \langle \mathrm{dev}_3\,\widetilde{\Sigma}, \mathrm{dev}_3\,\widetilde{\Sigma}\rangle \\ &\Leftrightarrow \quad \langle \mathrm{dev}_3\,\widetilde{\Sigma}, \mathrm{skew}(\mathrm{dev}_3\,\widetilde{\Sigma})\rangle = 0 \quad \Leftrightarrow \quad \mathrm{dev}_3\,\widetilde{\Sigma} \in \mathrm{Sym}(3) \\ &\Leftrightarrow \quad \widetilde{\Sigma} \in \mathrm{Sym}(3).\end{aligned}$$

For the simplest Neo-Hooke elastic energy considered in Appendix A.2, $W(F_e) = \mathrm{tr}(C_e) = \widetilde{W}(C\,C_p^{-1}) = \frac{1}{2}\,\mathrm{tr}(C\,C_p^{-1})$, we have $\widetilde{\Sigma} = C\,C_p^{-1}$, which is not symmetric. Hence $\sqrt{\mathrm{tr}[(\mathrm{dev}_3\,\widetilde{\Sigma})^2]} \neq \|\,\mathrm{dev}_3\,\widetilde{\Sigma}\|$. Let us again remark that $\widetilde{\Sigma}$ is not necessarily symmetric for general C_p. However, using Lemma 3, we deduce

[1] Miehe [22] only defines the elastic domain $\mathscr{E}_{\mathrm{e}}(\widetilde{\Sigma}, \frac{2}{3}\,\sigma_{\mathbf{y}}^2) := \left\{\widetilde{\Sigma} \in \mathbb{R}^{3\times 3} \,\middle|\, \mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2) \leq \frac{2}{3}\,\sigma_{\mathbf{y}}^2\right\}$ in terms of τ_e, i.e. $\mathscr{E}_{\mathrm{e}}(\tau_{\mathrm{e}}, \frac{2}{3}\,\boldsymbol{\sigma}_{\mathbf{y}}^2) = \left\{\tau \in \mathrm{Sym}(3) \,\middle|\, \|\,\mathrm{dev}_3\,\tau\|^2 \leq \frac{2}{3}\,\boldsymbol{\sigma}_{\mathbf{y}}^2\right\}$. He uses the same notation for the referential quantities. Therefore, we have two interpretations at hand $\Phi(\widetilde{\Sigma}) = \|\,\mathrm{dev}_3\,\tau_e\| - \frac{2}{3}\,\sigma_{\mathbf{y}}^2 = \sqrt{\mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2)} - \frac{2}{3}\,\sigma_{\mathbf{y}}^2$. On the other hand, in the isotropic case, we have also $\Phi(\widetilde{\Sigma}) = \|\,\mathrm{dev}_3\,\Sigma_e\| - \frac{2}{3}\,\sigma_{\mathbf{y}}^2$.

$$\widetilde{\Sigma}\, C_p = F_p^T\, \Sigma_e\, F_p^{-T}\, C_p = F_p^T\, \Sigma_e\, F_p \in \mathrm{Sym}(3) \quad \Rightarrow \quad \mathrm{dev}_3\, \widetilde{\Sigma} \cdot C_p \in \mathrm{Sym}(3), \tag{19}$$
$$C_p^{-1}\, \widetilde{\Sigma} = C_p^{-1}\, F_p^T\, \Sigma_e\, F_p^{-T} = F_p^{-1} \Sigma_e\, F_p^{-T} \in \mathrm{Sym}(3) \quad \Rightarrow \quad C_p^{-1}\, \mathrm{dev}_3\, \widetilde{\Sigma} \in \mathrm{Sym}(3).$$

In the following, we discuss first the sign of the quantity[2] $\mathscr{F}^2 := \mathrm{tr}((\mathrm{dev}_3\, \widetilde{\Sigma})^2)$. First, we deduce

$$\begin{aligned} \mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2] &= \langle (\mathrm{dev}_3\, \widetilde{\Sigma})\, (\mathrm{dev}_3\, \widetilde{\Sigma}), \mathbb{1}\rangle = \langle \widetilde{\Sigma}\, (\mathrm{dev}_3\, \widetilde{\Sigma}), \mathbb{1}\rangle = \langle C_p^{-1}\, \widetilde{\Sigma}\, (\mathrm{dev}_3\, \widetilde{\Sigma})\, C_p, \mathbb{1}\rangle \\ &= \langle C_p^{-1}\, \widetilde{\Sigma}\, (\mathrm{dev}_3\, \widetilde{\Sigma} \cdot C_p)^T, \mathbb{1}\rangle = \langle C_p^{-1}\, \widetilde{\Sigma}\, C_p\, (\mathrm{dev}_3\, \widetilde{\Sigma})^T, \mathbb{1}\rangle \\ &= \langle C_p^{-1}\, \widetilde{\Sigma}\, C_p, \mathrm{dev}_3\, \widetilde{\Sigma}\rangle. \end{aligned} \tag{20}$$

We further see that

$$\begin{aligned} \langle C_p^{-1}\, \widetilde{\Sigma}\, C_p, \mathrm{dev}_3\, \widetilde{\Sigma}\rangle &= \langle U_p^{-1} U_p^{-1}\, \widetilde{\Sigma}\, U_p\, U_p, \mathrm{dev}_3\, \widetilde{\Sigma}\rangle = \langle U_p^{-1}\, \widetilde{\Sigma}\, U_p, U_p^{-1}\, \mathrm{dev}_3\, \widetilde{\Sigma}\, U_p\rangle \\ &= \langle U_p^{-1}\, \widetilde{\Sigma}\, U_p, U_p^{-1} \widetilde{\Sigma}\, U_p - \frac{1}{3}\mathrm{tr}(\widetilde{\Sigma}) \cdot \mathbb{1}\rangle \\ &= \langle U_p^{-1}\, \widetilde{\Sigma}\, U_p, U_p^{-1} \widetilde{\Sigma}\, U_p - \frac{1}{3}\mathrm{tr}(U_p^{-1}\, \widetilde{\Sigma}\, U_p) \cdot \mathbb{1}\rangle \\ &= \langle U_p^{-1}\, \widetilde{\Sigma}\, U_p, \mathrm{dev}_3(U_p^{-1} \widetilde{\Sigma}\, U_p)\rangle = \langle \mathrm{dev}_3(U_p^{-1}\, \widetilde{\Sigma}\, U_p), \mathrm{dev}_3(U_p^{-1} \widetilde{\Sigma}\, U_p)\rangle \\ &= \|\, \mathrm{dev}_3(U_p^{-1}\, \widetilde{\Sigma}\, U_p)\|^2 \geq 0, \end{aligned} \tag{21}$$

where $U_p^2 = C_p$. Thus $\mathscr{F}^2$ is positive and $\mathscr{F}$ is well defined.

Since $D_{\widetilde{\Sigma}}\Phi(\widetilde{\Sigma}) = \dfrac{1}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2]}}\, (\mathrm{dev}_3\, \widetilde{\Sigma})^T$ the flow rule (18) becomes

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\, C_p &= -\frac{\lambda_{\mathrm{p}}^+}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2]}}\, (\mathrm{dev}_3\, \widetilde{\Sigma})^T \\ \Leftrightarrow \qquad C_p\, \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] &= -\frac{\lambda_{\mathrm{p}}^+}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2]}}\, \mathrm{dev}_3\, \widetilde{\Sigma}. \end{aligned} \tag{22}$$

[2] If we are not looking for the sign of $\mathrm{tr}((\mathrm{dev}_3\, \widetilde{\Sigma})^2)$ for all $\mathrm{dev}_3\, \widetilde{\Sigma} \in \mathbb{R}^{3\times 3}$, then considering two particular values of $\mathrm{dev}_3\, \widetilde{\Sigma}$, e.g.

$$\mathrm{dev}_3\, \widetilde{\Sigma} = \begin{pmatrix} -\frac{1}{2} & 1 & 2 \\ -2 & -\frac{1}{2} & 3 \\ -1 & -3 & -\frac{1}{2} \end{pmatrix} \qquad \text{and} \qquad \mathrm{dev}_3\, \widetilde{\Sigma} = \begin{pmatrix} -\frac{1}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & -\frac{1}{3} \end{pmatrix},$$

we obtain $\mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2] = -2$ and $\mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2] = \frac{2}{3}$, respectively. Hence, $\mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2]$ is not positive for all $\widetilde{\Sigma} \in \mathbb{R}^{3\times 3}$.

Further, in view of (19), we obtain

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = -\frac{\lambda_\mathrm{p}^+}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2]}}\, C_p^{-1}\ \mathrm{dev}_3\, \widetilde{\Sigma}$$

$$\Leftrightarrow \qquad \frac{\mathrm{d}}{\mathrm{dt}}[C_p] = \frac{\lambda_\mathrm{p}^+}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2]}}(\mathrm{dev}_3\, \widetilde{\Sigma})\, C_p \in \mathrm{Sym}(3). \qquad (23)$$

Using Lemma 4 we obtain that $C_p \in \mathrm{PSym}(3)$.

We remark that the flow rule considered by Miehe [22] (in this interpretation) coincides with the flow rule (47) considered by Helm [12], see Proposition 2.

Remark 3 Although the flow rule considered in this interpretation of the Miehe 1995 model [22] has a subdifferential structure, the yield-function Φ is not convex. Hence, the flow rule is not a convex flow rule. In order to see the non-convexity of $\Phi(\widetilde{\Sigma})$ we observe first by looking at sublevel-sets that

$$\Phi(\widetilde{\Sigma}) = \sqrt{\mathrm{tr}((\mathrm{dev}_3\, \widetilde{\Sigma})^2)} - \frac{2}{3}\,\sigma_\mathrm{y}^2 \text{ is convex} \quad \Leftrightarrow \quad \widetilde{\Phi}(\widetilde{\Sigma}) = \mathrm{tr}[(\mathrm{dev}_3\, \widetilde{\Sigma})^2] \text{ is convex.}$$

The second derivative for the simpler function $\widetilde{\Phi}(\widetilde{\Sigma})$ is

$$D^2_{\widetilde{\Sigma}}\widetilde{\Phi}(\widetilde{\Sigma}).(H, H) = \langle (\mathrm{dev}_3\, H)^T, \mathrm{dev}_3\, H\rangle = \mathrm{tr}[(\mathrm{dev}_3\, H)^2], \qquad \forall\ \widetilde{\Sigma},\ H \in \mathbb{R}^{3\times 3}.$$

We know that $\mathrm{tr}[(\mathrm{dev}_3\, H)^2]$ is not positive for all $H \in \mathbb{R}^{3\times 3}$, since for the previous considered matrix H, such that

$$\mathrm{dev}_3\, H = \begin{pmatrix} -\frac{1}{2} & 1 & 2 \\ -2 & -\frac{1}{2} & 3 \\ -1 & -3 & -\frac{1}{2} \end{pmatrix},$$

we obtain $\mathrm{tr}[(\mathrm{dev}_3\, H)^2] = -2$. Therefore $\widetilde{\Phi}(\widetilde{\Sigma})$ is not convex, and thus $\Phi(\widetilde{\Sigma})$ cannot be convex.

4 The Lion 1997 Multiplicative Elasto-Plasticity Formulation in Terms of the Plastic Metric $C_p = F_p^T F_p$

This derivation was given by Lion [17, Eq.(47.2)] in the general form (see also [12, Eq.(6.33)]) and by Dettmer-Reese [6] in the isotropic case. Following [6] we consider a perfect plasticity model for the plastic metric C_p based on the flow rule

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] \in -F_p^{-1}\, \partial\chi\, (\mathrm{dev}_3\, \Sigma_e)\, F_p^{-T} \in \mathrm{Sym}(3) \quad \text{for} \quad \Sigma_e \in \mathrm{Sym}(3). \qquad (24)$$

Again, in this model it is not clear from the outset, that it is a formulation in C_p alone. The goal of such a 6-dimensional formulation is to avoid any explicit computation of the plastic distortion F_p. However, the right hand side of the above proposed flow rule is, in fact, a multivalued function in C and C_p^{-1} alone. Hence, we can express the flow rule (24) entirely in the form[3]

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\, C_p \in f(C, C_p^{-1}). \tag{25}$$

In order to show this remarkable property (satisfied only for isotropic response), and to determine the explicit form of the function $f(C, C_p^{-1})$, in view of (2) we remark that

$$\begin{aligned}
F_p^{-1} \frac{\operatorname{dev}_3 \Sigma_e}{\|\operatorname{dev}_3 \Sigma_e\|} F_p^{-T} &= \frac{1}{\|\operatorname{dev}_3 \Sigma_e\|}\left[F_p^{-1} \Sigma_e F_p^{-T} - \frac{1}{3}\operatorname{tr}(\Sigma_e)\, C_p^{-1}\right],\\
\operatorname{tr}(\Sigma_e) &= 2\,\langle \mathbb{1}, C_e\,(\alpha_1\,\mathbb{1} + \alpha_2\, C_e + \alpha_3\, C_e^2),\\
\|\Sigma_e\|^2 &= 4\langle C_e\,(\alpha_1\,\mathbb{1} + \alpha_2\, C_e + \alpha_3\, C_e^2), C_e\,(\alpha_1\,\mathbb{1} + \alpha_2\, C_e + \alpha_3\, C_e^2)\rangle,\\
\|\operatorname{dev}_3 \Sigma_e\| &= \sqrt{\|\Sigma_e\|^2 - \frac{1}{3}[\operatorname{tr}(\Sigma_e)]^2}.
\end{aligned}$$

It is clear that $F_p^{-1} \Sigma_e F_p^{-T} = 2\, F_p^{-1}\, C_e\,(\alpha_1\,\mathbb{1} + \alpha_2\, C_e + \alpha_3\, C_e^2)\, F_p^{-T} \in \operatorname{Sym}(3)$, and

$$\begin{aligned}
C_e &= F_e^T F_e = F_p^{-T} F^T F\, F_p^{-1} = F_p^{-T} C\, F_p^{-1},\\
\Sigma_e &= 2\, F_p^{-T}(\alpha_1\, C + \alpha_2\, C\, C_p^{-1} C + \alpha_3\, C\, C_p^{-1} C\, C_p^{-1} C)\, F_p^{-1},\\
\operatorname{tr}(\Sigma_e) &= 2\operatorname{tr}(\alpha_1\, C\, C_p^{-1} + \alpha_2\, C\, C_p^{-1} C\, C_p^{-1} + \alpha_3\, C\, C_p^{-1} C\, C_p^{-1} C\, C_p^{-1}),\\
\|\Sigma_e\|^2 &= 4\langle C_p^{-1} \widehat{f}, \widehat{f} C_p^{-1}\rangle.
\end{aligned}$$

where

$$\widehat{f} := \alpha_1\, C + \alpha_2\, C\, C_p^{-1} C + \alpha_3\, C\, C_p^{-1} C\, C_p^{-1} C.$$

Hence, we deduce

$$\begin{aligned}
F_p^{-1} \Sigma_e F_p^{-T} &= 2\, C_p^{-1}\, \widehat{f}(C, C_p^{-1})\, C_p^{-1} \in \operatorname{Sym}(3), \\
\operatorname{tr}(\Sigma_e) &= 2\operatorname{tr}(\widehat{f}(C, C_p^{-1})\, C_p^{-1}), \qquad \|\Sigma_e\|^2 = 4\langle C_p^{-1}\, \widehat{f}(C, C_p^{-1}), \widehat{f}(C, C_p^{-1})\, C_p^{-1}\rangle,\\
\|\operatorname{dev}_3 \Sigma_e\| &= 2\sqrt{\operatorname{tr}[(\widehat{f}(C, C_p^{-1})\, C_p^{-1})^2] - \frac{1}{3}[\operatorname{tr}(\widehat{f}(C, C_p^{-1}) C_p^{-1})]^2},
\end{aligned} \tag{26}$$

[3] Note carefully, that $f(C, C_p^{-1})$ is not necessarily symmetric. Moreover $C_p^{-1}\, \widehat{f}(C, C_p^{-1}) \notin \operatorname{Sym}(3)$ in general.

where

$$\widehat{f}(C, C_p^{-1}) := \alpha_1\, C + \alpha_2\, C\, C_p^{-1} C\; + \alpha_3\, C\, C_p^{-1} C\, C_p^{-1} C \in \operatorname{Sym}(3), \tag{27}$$

and $\alpha_i = \alpha_i(I_1(C_e), I_2(C_e), I_3(C_e))$, according to (2). Therefore, the multivalued function $f(C, C_p^{-1})$ is given by

$$f(C, C_p^{-1}) = \left\{ \frac{-\lambda_{\rm p}^+ \ \operatorname{dev}_3(\, C_p^{-1}\, \widehat{f}(C, C_p^{-1}))}{\sqrt{\operatorname{tr}[(\widehat{f}(C, C_p^{-1}) C_p^{-1})^2] - \frac{1}{3}[\operatorname{tr}(\widehat{f}(C, C_p^{-1}) C_p^{-1})]^2}} \;\Big|\; \lambda_{\rm p}^+ \in \mathbb{R}_+ \right\}. \tag{28}$$

In Appendix A.1 we give the specific expression for the functions $f(C, C_p^{-1})$ and $\widehat{f}(C, C_p^{-1})$ in case of the Neo-Hooke energy.

On the other hand, in view of Eq. (24) we also have

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p]\, C_p^{-1} = -C_p \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] \in C_p\, F_p^{-1}\, \partial\chi\, (\operatorname{dev}_3 \varSigma_e)\, F_p^{-T} = F_p^T \partial\chi\, (\operatorname{dev}_3 \varSigma_e) F_p^{-T}.$$

Hence, it follows that

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p] \in F_p^T\, \partial\chi\, (\operatorname{dev}_3 \varSigma_e)\, F_p^{-T} C_p = F_p^T \partial\chi\, (\operatorname{dev}_3 \varSigma_e) F_p \;\in\; \operatorname{Sym}(3), \tag{29}$$

which establishes symmetry of C_p whenever $C_p(0) \in \operatorname{Sym}(3)$.

Another important question is whether the solution C_p of the flow rule (24) is such that $\det C_p(t) = 1$, for all $t \geq 0$. Let C_p be the solution of the flow rule (24). Then, we have

$$\begin{aligned} C_p \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] &= -\lambda_{\rm p}^+\, C_p\, F_p^{-1} \frac{\operatorname{dev}_3 \varSigma_e}{\|\operatorname{dev}_3 \varSigma_e\|}\, F_p^{-T} = -\lambda_{\rm p}^+\, F_p^T F_p F_p^{-1} \frac{\operatorname{dev}_3 \varSigma_e}{\|\operatorname{dev}_3 \varSigma_e\|}\, F_p^{-T} \\ &= -\frac{\lambda_{\rm p}^+}{2} \frac{\operatorname{dev}_3(F_p^T \varSigma_e\, F_p^{-T})}{\|\operatorname{dev}_3 \varSigma_e\|}, \end{aligned} \tag{30}$$

which implies on the one hand

$$\langle \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\, C_p, \mathbb{1} \rangle = \langle \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}], C_p \rangle = 0.$$

On the other hand, the flow rule (24) together with $\det\, C_p(0) = 1$ leads to $\det\, C_p(t) = 1$, for all $t \geq 0$.

Let us remark that, in view of (25) and (28) we have for the flow rule (24)

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = \frac{-\lambda_{\mathrm{p}}^{+}}{\sqrt{\mathrm{tr}[(\widehat{f}(C, C_p^{-1})C_p^{-1})^2] - \frac{1}{3}[\mathrm{tr}[\widehat{f}(C, C_p^{-1})C_p^{-1}]]^2}}\ \underbrace{\mathrm{dev}_3[\underbrace{C_p^{-1}\widehat{f}(C, C_p^{-1})}_{\notin \mathrm{Sym}(3)}] \cdot C_p^{-1}}_{\in \mathrm{Sym}(3)},$$

which is in concordance with the requirement $C_p \in \mathrm{Sym}(3)$, as can be seen from (29). Note that the above formula cannot be read as

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = -\lambda_{\mathrm{p}}^{+} \frac{\mathrm{dev}_3 \varSigma}{\|\mathrm{dev}_3 \varSigma\|} \cdot C_p^{-1},$$

for some $\varSigma$, since

$$[\mathrm{tr}(\widehat{f}(C, C_p^{-1})C_p^{-1})^2] - \frac{1}{3}[\mathrm{tr}(\widehat{f}(C, C_p^{-1})C_p^{-1})]^2 \neq \|\,\mathrm{dev}_3(\widehat{f}(C, C_p^{-1})C_p^{-1})\|^2.$$

To see this, assume to the contrary that equality holds. Then we deduce

$$\begin{aligned}&[\mathrm{tr}(\widehat{f}(C, C_p^{-1})C_p^{-1})^2] - \frac{1}{3}[\mathrm{tr}(\widehat{f}(C, C_p^{-1})C_p^{-1})]^2 = \|\,\mathrm{dev}_3(\widehat{f}(C, C_p^{-1})C_p^{-1})\|^2\\ &\qquad \Leftrightarrow \quad \langle \widehat{f}(C, C_p^{-1})C_p^{-1}, (\widehat{f}(C, C_p^{-1})C_p^{-1})^T\rangle = \pm\langle \widehat{f}(C, C_p^{-1})C_p^{-1}, \widehat{f}(C, C_p^{-1})C_p^{-1}\rangle.\end{aligned} \tag{31}$$

Since $\widehat{f}(C, C_p^{-1}) \in \mathrm{Sym}(3)$, we obtain

$$\begin{aligned}\mathrm{tr}[(\widehat{f}(C, C_p^{-1})C_p^{-1})^2] &= \langle C_p^{-1}\,\widehat{f}(C, C_p^{-1})C_p^{-1}\,(\widehat{f}(C, C_p^{-1}), \mathbb{1}\rangle\\ &= \langle C_p^{-1}\,\widehat{f}(C, C_p^{-1}), \widehat{f}(C, C_p^{-1})C_p^{-1}\rangle.\end{aligned} \tag{32}$$

Using that $C_p \in \mathrm{PSym}(3)$, we further deduce that

$$\begin{aligned}\langle C_p^{-1}\,\widehat{f}(C, C_p^{-1}), \widehat{f}(C, C_p^{-1})C_p^{-1}\rangle &= \langle U_p^{-1}\,\widehat{f}(C, C_p^{-1})U_p^{-1}, U_p^{-1}\widehat{f}(C, C_p^{-1})U_p^{-1}\rangle\\ &= \|U_p^{-1}\,\widehat{f}(C, C_p^{-1})U_p^{-1}\|^2,\end{aligned} \tag{33}$$

where $U_p^2 = C_p$. Therefore, from (31) we deduce

$$\begin{aligned}&\langle \widehat{f}(C, C_p^{-1})C_p^{-1}, (\widehat{f}(C, C_p^{-1})C_p^{-1})^T\rangle = \langle \widehat{f}(C, C_p^{-1})C_p^{-1}, \widehat{f}(C, C_p^{-1})C_p^{-1}\rangle\\ &\quad \Leftrightarrow \quad \widehat{f}(C, C_p^{-1})C_p^{-1} \in \mathrm{Sym}(3),\end{aligned} \tag{34}$$

which is not true, in general. However, it is an associated plasticity model in the sense of Definition 1, see Proposition 3. We also remark that

$$C_p \, \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = \frac{-\lambda_{\mathrm{p}}^+ \, \mathrm{dev}_3[\widehat{f}(C, C_p^{-1}) \, C_p^{-1}]}{\sqrt{\mathrm{tr}[(\widehat{f}(C, C_p^{-1}) C_p^{-1})^2] - \frac{1}{3}[\mathrm{tr}[\widehat{f}(C, C_p^{-1}) C_p^{-1}]]^2}} \,. \tag{35}$$

In conclusion, using Lemma 4, we have

Remark 4 Any continuous solution $C_p \in \mathrm{Sym}(3)$ of the flow rule (24) belongs in fact to $\mathrm{PSym}(3)$.

As for the thermodynamical consistency, we remark that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}}[\widetilde{W}(C\,C_p^{-1})] &= \langle D[\widetilde{W}(C\,C_p^{-1})], C \, \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle = \langle C \, D_C[\widetilde{W}(C\,C_p^{-1})]\,C_p, \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle \\ &= \frac{1}{2}\langle \widetilde{\Sigma}\,C_p, \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle = \frac{1}{2}\langle C_p^{-1}\,\widetilde{\Sigma}\,C_p, C_p \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle \\ &= -\frac{1}{2}\langle C_p^{-1}\,\widetilde{\Sigma}\,C_p, \frac{\mathrm{d}}{\mathrm{dt}}[C_p]\,C_p^{-1}\rangle = -\frac{1}{4}\,\frac{\lambda_{\mathrm{p}}^+}{\|\,\mathrm{dev}_3\,\widetilde{\Sigma}\|}\langle C_p^{-1}\,\widetilde{\Sigma}\,C_p, \mathrm{dev}_3\,\widetilde{\Sigma}\rangle, \end{aligned} \tag{36}$$

which, using the formula $F_p^T\,\Sigma_e\,F_p^{-T} = \widetilde{\Sigma}$, leads to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}}[\widetilde{W}(C\,C_p^{-1})] &= -\frac{1}{4}\,\frac{\lambda_{\mathrm{p}}^+}{\|\,\mathrm{dev}_3(F_p^T\,\Sigma_e\,F_p^{-T})\|}\langle C_p^{-1}\,F_p^T\,\Sigma_e\,F_p^{-T}\,C_p, \mathrm{dev}_3(F_p^T\,\Sigma_e\,F_p^{-T})\rangle \\ &= -\frac{1}{4}\,\frac{\lambda_{\mathrm{p}}^+}{\|\,\mathrm{dev}_3(F_p^T\,\Sigma_e\,F_p^{-T})\|}\langle \Sigma_e, \Sigma_e - \frac{1}{3}\mathrm{tr}(\Sigma_e)\cdot \mathbb{1}\rangle \\ &= -\frac{1}{4}\,\frac{\lambda_{\mathrm{p}}^+}{\|\,\mathrm{dev}_3(F_p^T\,\Sigma_e\,F_p^{-T})\|}\|\,\mathrm{dev}_3\,\Sigma_e\|^2 \le 0. \end{aligned} \tag{37}$$

Note that this proof of thermodynamical consistency may be criticized because it involves the variable F_p, which should not appear at all. However, we may also use (20) and (21) to obtain

$$\frac{\mathrm{d}}{\mathrm{dt}}[\widetilde{W}(C\,C_p^{-1})] = -\frac{1}{4}\,\frac{\lambda_{\mathrm{p}}^+}{\|\,\mathrm{dev}_3\,\widetilde{\Sigma}\|}\,\mathrm{tr}[(\mathrm{dev}_3\,\widetilde{\Sigma})^2] \tag{38}$$

$$= -\frac{1}{4}\,\frac{\lambda_{\mathrm{p}}^+}{\|\,\mathrm{dev}_3\,\widetilde{\Sigma}\|}\,\|\,\mathrm{dev}_3(U_p^{-1}\,\widetilde{\Sigma}\,U_p)\|^2 \le 0, \tag{39}$$

We may summarize the properties of the Lion 1997 model:

(i) from (24) it follows that $C_p(t) \in \mathrm{Sym}(3)$;
(ii) plastic incompressibility: from (24) together with $\det C_p(0) = 1$ it follows that $\det C_p(t) = 1$;
(iii) for the isotropic case, the right hand-side of (24) is a function of C_p^{-1} and C alone;

(iv) from (i) and (ii) together and using Lemma 4 it follows that $C_p(t) \in \mathrm{PSym}(3)$;
(v) it is thermodynamically correct;
(vi) it is an associated plasticity model in the sense of Definition 1, see Proposition 3.

Remark 5 (Simo-Miehe 1992 model vs. Lion 1997 model) In the anisotropic case, the flow rule proposed by Simo and Miehe [41] (and later by Reese and Wriggers [36] and Miehe [21]) is not completely equivalent with the flow rule proposed by Lion (see also [1, 6]), since $\| \mathrm{dev}_3\, \tau_e\| \neq \| \mathrm{dev}_3\, \Sigma_e\|$ does not hold true in general. However, the difference is nearly absorbed by the positive plastic multipliers. The models may differ due to different yield conditions, but the flow rules are similar, having the same performance with respect to the thermodynamic consistency. Both models are consistent according to our Definition 2, but we may not switch between them, since different elastic domains are considered, namely $\mathscr{E}_{\Sigma_e}$ and $\mathscr{E}_{\tau_e}$, respectively. This is in fact the main difference between these two models. Having different elastic domains we have different boundary points, since a point of the boundary of $\mathscr{E}_{\tau_e}$ is not necessarily on the boundary of $\mathscr{E}_{\tau_e}$. Hence, in these two flow rules we have a different behaviour corresponding to the indicator function of different domains. The material may reach the boundary of the elastic domain $\mathscr{E}_{\tau_e}$, while it is strictly inside the elastic domain $\mathscr{E}_{\Sigma_e}$, for the same local response.

However, we have the following result:

Proposition 1 *In the isotropic case the flow rule proposed by Simo and Miehe [41] is equivalent with the flow rule proposed by Lion [17].*

Proof We compare the flow rules (17) and (24) and the proof is complete. □

5 The Simo and Hughes 1998 Plasticity Formulation in Terms of a Plastic Metric

The book [42] has been edited years after the untimely death of J.C. Simo. In this book also a finite strain plasticity model is proposed. However, this model has a subtle fundamental deficiency which we aim to describe in the interest of the reader. The flow rule considered in [42, p. 310] is

$$\frac{\mathrm{d}}{\mathrm{dt}}[\overline{C}_p^{\,-1}] = -\frac{2}{3}\lambda_\mathrm{p}^+ \,\mathrm{tr}(B_e)\, F^{-1} \frac{\mathrm{dev}_n\, \tau_e}{\| \mathrm{dev}_n\, \tau_e\|} F^{-T}, \qquad \overline{C}_p = \frac{C_p}{\det C_p^{1/3}}\,, \tag{40}$$

where $B_e = F_e F_e^T$, $\tau_e = 2\, F_e\, D_{C_e}[W(C_e)]\, F_e^T = 2\, B_e\, D_{B_e}[W(B_e)]$ is the elastic Kirchhoff stress tensor and $\lambda_\mathrm{p}^+ \geq 0$ is the consistency parameter. If the plastic flow is isochoric then $\det F_p = \det C_p = 1$. However, we must always have $\det \overline{C}_p = 1 = \det \overline{C}_p^{\,-1}$ by definition of $\overline{C}_p$. Since $F_e = F\, F_p^{-1}$, we have $\mathrm{tr}(B_e) = \langle F_p^{-T} F^T\, F\, F_p^{-1}\rangle = \langle \mathbb{1}, C\, C_p^{-1}\rangle = \mathrm{tr}(C\, C_p^{-1})$. Moreover, note that for elastically isotropic materials it holds

$$D_{C_e}[W(C_e)] = \alpha_1\, \mathbb{1} + \alpha_2\, C_e + \alpha_3\, C_e^2 \in \mathrm{Sym}(3), \tag{41}$$
$$\tau_e = 2\, F_e\, [\alpha_1\, \mathbb{1} + \alpha_2\, C_e + \alpha_3\, C_e^2]\, F_e^T,$$

where $\alpha_1, \alpha_2, \alpha_3$ are scalar functions of the invariants of C_e which are functions of $C\, C_p^{-1}$, see Lemma 2. Since $C_e = F_p^{-T} C\, F_p^{-1}$, we obtain

$$\begin{aligned}\tau_e &= 2\, F\, F_p^{-1}\, [\alpha_1\, \mathbb{1} + \alpha_2\, F_p^{-T} C\, F_p^{-1} + \alpha_3\, F_p^{-T} C\, F_p^{-1} F_p^{-T} C\, F_p^{-1}] F_p^{-T}\, F^T \\ &= 2\, F\, f_1(C, C_p^{-1})\, F^T,\end{aligned} \tag{42}$$

with $f_1(C, C_p^{-1}) :=\alpha_1\, C_p^{-1}+\alpha_2\, C_p^{-1} C\, C_p^{-1} + \alpha_3\, C_p^{-1} C\, C_p^{-1}\, C\, C_p^{-1} \in \mathrm{Sym}(3)$.Thus, for elastically isotropic materials we deduce

$$\begin{aligned}F^{-1}\, [\mathrm{dev}_n\, \tau_e]\, F^{-T} &=2\, f_1(C, C_p^{-1}) - \frac{2}{3}\mathrm{tr}(F\, f_1(C, C_p^{-1})\, F^T)\, C^{-1} \\ &=2\, f_1(C, C_p^{-1}) - \frac{2}{3}\langle f_1(C, C_p^{-1}), C\rangle\, C^{-1} \in \mathrm{Sym}(3), \\ \|\, \mathrm{dev}_3\, \tau_e \| &=2\sqrt{\langle f_1(C, C_p^{-1}) \cdot C, C \cdot\, f_1(C, C_p^{-1})\rangle^2 - \frac{1}{9}\langle f_1(C, C_p^{-1}), C\rangle^2}.\end{aligned} \tag{43}$$

Hence, $F^{-1}\, \frac{\mathrm{dev}_n\, \tau_e}{\|\, \mathrm{dev}_n\, \tau_e \|}\, F^{-T} \in \mathrm{Sym}(3)$ is a function of C, C_p^{-1}. Therefore the flow rule (40) can entirely be expressed in terms of C and C_p^{-1} alone.

Remark 6 It is not true, in general, that the right hand side of the flow rule (40) is in concordance with $\det \overline{C}_p(t) = 1$, assuming that $\det \overline{C}_p(0) = 1$.

Proof From (40) we obtain by right multiplication with $\overline{C}_p$

$$\frac{\mathrm{d}}{\mathrm{dt}}[\overline{C}_p^{\,-1}]\, \overline{C}_p = -\frac{2}{3}\, (\det C_p)^{-1/3}\, \lambda_{\mathrm{p}}^+\, \mathrm{tr}(C\, C_p^{-1})\, F_p^{-1} F_e^{-1} \frac{\mathrm{dev}_n\, \tau_e}{\|\, \mathrm{dev}_n\, \tau_e \|} F_e^{-T}\, F_p. \tag{44}$$

On the other hand, we have[4]

$$\frac{\mathrm{d}}{\mathrm{dt}}[\det \overline{C}_p^{\,-1}] = \langle \mathrm{Cof}\ \overline{C}_p^{\,-1}, \frac{\mathrm{d}}{\mathrm{dt}}[\overline{C}_p^{\,-1}]\rangle = \det \overline{C}_P^{\,-1} \langle \overline{C}_p, \frac{\mathrm{d}}{\mathrm{dt}}[\overline{C}_p^{\,-1}]\rangle = \det \overline{C}_P^{\,-1} \langle \mathbb{1}, \frac{\mathrm{d}}{\mathrm{dt}}[\overline{C}_p^{\,-1}]\overline{C}_p\rangle.$$

Hence, we deduce

$$\frac{\mathrm{d}}{\mathrm{dt}}[\det \overline{C}_p^{\,-1}] = -\frac{2}{3}\, (\det C_p)^{-1/3}\, \lambda_{\mathrm{p}}^+\, \mathrm{tr}(C\, C_p^{-1}) \langle\, F_e^{-1} \frac{\mathrm{dev}_n\, \tau_e}{\|\, \mathrm{dev}_n\, \tau_e \|} F_e^{-T}, \mathbb{1}\rangle. \tag{45}$$

Since $F_e^{-1} \frac{\mathrm{dev}_n\, \tau_e}{\|\, \mathrm{dev}_n\, \tau_e \|} F_e^{-T}$ is not necessarily a trace free matrix, we can not conclude that $\det \overline{C}_p^{\,-1}(t) = \mathit{const.}$ for all $t > 0$. For instance, for elastically isotropic materials (see (43)) we have

[4] Let us remark that $\frac{\mathrm{d}}{\mathrm{dt}}[\det \overline{C}_p^{\,-1}] = \det \overline{C}_p^{\,-1} \langle \mathbb{1}, \frac{\mathrm{d}}{\mathrm{dt}}[\overline{C}_p^{\,-1}]\overline{C}_p\rangle$ shows that $\det \overline{C}_p^{\,-1} > 0$ by direct integration of the ordinary differential equation.

$$\begin{aligned}
\operatorname{dev}_3 \tau_e =&2\langle f_1(C, C_p^{-1}), C_p\rangle - \frac{2}{3}\langle f_1(C, C_p^{-1}), C\rangle\langle C^{-1}, C_p\rangle\\
=&2\,\alpha_1\left[\langle C_p^{-1}, C_p\rangle - \frac{1}{3}\langle C_p^{-1}, C\rangle\langle C^{-1}, C_p\rangle\right]\\
&+2\,\alpha_2\left[\langle C_p^{-1} C\, C_p^{-1}, C_p\rangle - \frac{1}{3}\langle C_p^{-1} C\, C_p^{-1}, C\rangle\langle C^{-1}, C_p\rangle\right]\\
&+2\,\alpha_3\left[\langle C_p^{-1} C\, C_p^{-1}\, C\, C_p^{-1}, C_p\rangle - \frac{1}{3}\langle C_p^{-1} C\, C_p^{-1}\, C\, C_p^{-1}, C\rangle\langle C^{-1}, C_p\rangle\right],
\end{aligned} \tag{46}$$

which shows that $F_e^{-1}\frac{\operatorname{dev}_n \tau_e}{\|\operatorname{dev}_n \tau_e\|}F_e^{-T}$ is not necessarily a trace free matrix, see Appendix A.4. □

Summarizing the properties of the flow rule (40) we have:

(i) it is thermodynamically correct;
(ii) the right hand side is a function of C and $\overline{C}_p^{-1}$ only;
(iii) from this flow rule it follows $\overline{C}_p(t) \in \operatorname{Sym}(3)$ and $\det \overline{C}_p(t) > 0$. Hence, it follows that $\overline{C}_p(t) \in \operatorname{PSym}(3)$;
(iv) plastic incompressibility: however, it does **not follow** from the flow rule that $\boldsymbol{\det \overline{C}_p(t) = 1}$ (which must hold by the very definition of $\overline{C}_p$, since the right hand side is not trace-free, in general;
(v) it is **not an associated plasticity model** in the sense of Definition 1.

6 The Helm 2001 Model

In this section we consider the model proposed by Helm [12], Vladimirov, Pietryga and Reese [45, Eq. 25] (see also [35] and [39, Eq. 55] and the model considered by Brepols, Vladimirov and Reese [1, page16], Shutov and Ihlemann [38, Eq. 80]). We prove later that this model is similar to the model considered by Miehe [22] in 1995, provided certain interpretations are included. Vladimirov, Pietryga and Reese [45, Eq. 25] considered the following flow rule

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p] = \lambda_\mathrm{p}^+ \frac{\operatorname{dev}_3 \widetilde{\Sigma}}{\sqrt{\operatorname{tr}((\operatorname{dev}_3 \widetilde{\Sigma})^2)}} \cdot C_p\,, \tag{47}$$

where $\widetilde{\Sigma} = 2\,C\,D_C[\widetilde{W}(C\,C_p^{-1})]$ is not necessarily symmetric for general $C_p \in \operatorname{PSym}(3)$, while $(\operatorname{dev}_3 \widetilde{\Sigma}) \cdot C_p \in \operatorname{Sym}(3)$, see Sect. 3. Therefore, we have $\frac{\mathrm{d}}{\mathrm{dt}}[C_p] \in \operatorname{Sym}(3)$. The flow rule (47) implies

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[\widetilde{W}(C\,C_p^{-1})] &= \langle D[\widetilde{W}(C\,C_p^{-1})],\, C\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle = \langle C\,D_C[\widetilde{W}(C\,C_p^{-1})]\,C_p,\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle \\ &= \frac{1}{2}\langle\widetilde{\Sigma}\,C_p,\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle = \frac{1}{2}\langle C_p^{-1}\,\widetilde{\Sigma}\,C_p,\,C_p\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle \qquad (48)\\ &= -\frac{1}{2}\frac{\lambda_{\mathrm{p}}^{+}}{\sqrt{\mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2)}}\langle C_p^{-1}\,\widetilde{\Sigma}\,C_p,\,\mathrm{dev}_3\,\widetilde{\Sigma}\rangle.\end{aligned}$$

Thus, using (20) and (21) we deduce

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[\widetilde{W}(C\,C_p^{-1})] &= -\frac{1}{2}\frac{\lambda_{\mathrm{p}}^{+}}{\sqrt{\mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2)}}\,\mathrm{tr}[(\mathrm{dev}_3\,\widetilde{\Sigma})^2]\\ &= -\frac{\lambda_{\mathrm{p}}^{+}}{2}\|\,\mathrm{dev}_3(U_p^{-1}\,\widetilde{\Sigma}\,U_p)\| \le 0, \qquad (49)\end{aligned}$$

which shows thermodynamical consistency.

Summarizing, the Helm 2001 (Reese 2008 and Shutov-Ihlemann 2014) model has the following properties:

(i) from (49) it follows that it is thermodynamically correct;
(ii) plastic incompressibility: from (47) and (11) it follows that $\det C_p(t) = 1$;
(iii) for the isotropic case, the right hand-side of the flow rule (47) is a function of C_p^{-1} and C alone;
(iv) from $\frac{\mathrm{d}}{\mathrm{dt}}[C_p] \in \mathrm{Sym}(3)$. it follows, in the isotropic case, that $C_p(t) \in \mathrm{Sym}(3)$;
(v) from (ii) and (iii) together and using Lemma 4 it follows that $C_p(t) \in \mathrm{PSym}(3)$;
(vi) it has formally subdifferential structure, see Proposition 4. However, the elastic domain $\mathscr{E}_{\mathrm{e}}(\widetilde{\Sigma},\,\frac{2}{3}\,\sigma_{\mathbf{y}}^2)$ is not convex w.r.t $\widetilde{\Sigma}$, see Remark 3.

Moreover, we see that the following result holds:

Proposition 2 *The flow rule considered by Helm [12] coincides with the flow rule* (47)*, i.e. with the interpretation of Miehe's proposal [22] presented in Sect. 3.*

Proof The proof follows from (47) and combined with (23). □

7 The Grandi-Stefanelli 2015 Model

In this section we present a model based on one representation used by Grandi and Stefanelli [10] and previously used by Frigeri and Stefanelli [7, p. 7]. We start by computing

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}\widetilde{W}(C\,C_p^{-1}) &= \langle D\widetilde{W}(C\,C_p^{-1}), C\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle = \langle \mathrm{sym}[C\,D\widetilde{W}(C\,C_p^{-1})], \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\rangle \\ &= \langle \underbrace{\sqrt{C_p}^{\,-1}\,\mathrm{sym}[C\,D\widetilde{W}(C\,C_p^{-1})]\,\sqrt{C_p}^{\,-1}}_{:=\frac{1}{2}\,\overset{\circ}{\Sigma}\,\in\,\mathrm{Sym}(3)}, \sqrt{C_p}\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\,\sqrt{C_p}\rangle.\end{aligned} \tag{50}$$

It is now easy to see that, if we choose

$$\sqrt{C_p}\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\,\sqrt{C_p} \in -\partial_{\overset{\circ}{\Sigma}}\chi\,(\mathrm{dev}_3\,\overset{\circ}{\Sigma}), \tag{51}$$

where $\chi\,(\mathrm{dev}_3\,\overset{\circ}{\Sigma})$ is the indicator function of the convex elastic domain

$$\overset{\circ}{\mathscr{E}}_e(\overset{\circ}{\Sigma}, \frac{1}{3}\,\sigma_{\mathbf{y}}^2) := \left\{ \overset{\circ}{\Sigma} \in \mathrm{Sym}(3)\;|\; \|\,\mathrm{dev}_3\,\overset{\circ}{\Sigma}\|^2 \leq \frac{1}{3}\,\sigma_{\mathbf{y}}^2 \right\},$$

then $C_p \in \mathrm{Sym}(3)$ and the reduced dissipation inequality $\frac{\mathrm{d}}{\mathrm{dt}}\widetilde{W}(C\,C_p^{-1}) \leq 0$ is satisfied. Thus, the model is thermodynamically correct. We also remark that the flow rule (51) implies

$$\mathrm{tr}(\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\,C_p) = \langle \frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\sqrt{C_p}\,\sqrt{C_p}, \mathbb{1}\rangle = \langle \sqrt{C_p}\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\sqrt{C_p}, \mathbb{1}\rangle = 0.$$

Hence, we obtain $\det C_p(t) = 1$ and further $C_p(t) \in \mathrm{PSym}(3)$.

Using Lemma 3, we give next some new representations of the stress-tensor

$$\overset{\circ}{\Sigma} := 2\,\sqrt{C_p}^{\,-1}\,\mathrm{sym}[C\,D\widetilde{W}(C\,C_p^{-1})]\,\sqrt{C_p}^{\,-1} = 2\,\mathrm{sym}\left[\sqrt{C_p}^{\,-1}\,(C\,D\widetilde{W}(C\,C_p^{-1}))\,\sqrt{C_p}^{\,-1}\right]$$

in terms of the stress tensors $\Sigma_e, \widetilde{\Sigma}$ and τ_e, respectively. From (3) we obtain $C\,D\widetilde{W}(C\,C_p^{-1}) = \frac{1}{2}\,\widetilde{\Sigma}\,C_p$. We also use $F_p = R_p\,U_p = R_p\,\sqrt{C_p}$ and $F = F_eF_p$. Hence, we deduce

$$\begin{aligned}\overset{\circ}{\Sigma} &= \mathrm{sym}(\sqrt{C_p}^{\,-1}\,\widetilde{\Sigma}\,C_p\,\sqrt{C_p}^{\,-1}) = \mathrm{sym}(\sqrt{C_p}^{\,-1}\,\widetilde{\Sigma}\,\sqrt{C_p}), \\ \overset{\circ}{\Sigma} &= \mathrm{sym}(\sqrt{C_p}^{\,-1}\,F_p^T\,\Sigma_e\,F_p^{-T}\,\sqrt{C_p}) = \mathrm{sym}(\sqrt{C_p}^{\,-1}\,\sqrt{C_p}\,R_p^T\,\Sigma_e\,R_p\,\sqrt{C_p}^{\,-1}\,\sqrt{C_p}) \\ &= \mathrm{sym}(R_p^T\,\Sigma_e\,R_p), \\ \overset{\circ}{\Sigma} &= \mathrm{sym}(\sqrt{C_p}^{\,-1}\,F^T\,\tau_e\,F^{-T}\,\sqrt{C_p}) = \mathrm{sym}(\sqrt{C_p}^{\,-1}\,F_p^T F_e^T\,\tau_e\,F_e^{-T}F_p^{-T}\,\sqrt{C_p}) \\ &= \mathrm{sym}(R_p^T F_e^T\,\tau_e\,F_e^{-T}\,R_p).\end{aligned}$$

Note that $\varSigma_e$ is symmetric in case of elastic isotropy. Hence, for the isotropic case, we have

$$\overset{\circ}{\varSigma} = R_p^T\, \varSigma_e\, R_p, \qquad \overset{\circ}{\varSigma} = R_p^T\, F_e^T\, \tau_e\, F_e^{-T}\, R_p. \tag{52}$$

However, we have $\|\overset{\circ}{\varSigma}\|^2 = \|R_p^T\, \varSigma_e\, R_p\|^2 = \|\varSigma_e\|^2$, $\operatorname{tr}(\overset{\circ}{\varSigma}) = \operatorname{tr}(R_p^T\, \varSigma_e\, R_p) = \operatorname{tr}(\varSigma_e)$. Together, we obtain that

$$\|\operatorname{dev}_3 \overset{\circ}{\varSigma}\| = \|\operatorname{dev}_3 \varSigma_e\|.$$

In conclusion, for isotropic elastic materials we have the equivalence of the elastic domains

$$\overset{\circ}{\mathscr{E}}_e(\overset{\circ}{\varSigma}, \frac{1}{3}\boldsymbol{\sigma}_{\mathrm{y}}^2) = \mathscr{E}_{\mathrm{e}}(\varSigma_e, \frac{1}{3}\boldsymbol{\sigma}_{\mathrm{y}}^2). \tag{53}$$

Therefore, the flow rule (51) proposed by Grandi and Stefanelli [10] has the following properties:

(i) it is thermodynamically correct;
(ii) from this flow rule it follows $C_p(t) \in \operatorname{Sym}(3)$ and $\det C_p(t) = 1$. Hence, it follows that $C_p(t) \in \operatorname{PSym}(3)$;
(iii) the elastic domain $\overset{\circ}{\mathscr{E}}_e$ is convex w.r.t. $\overset{\circ}{\varSigma}$;
(iv) it is an associated plasticity model in the sense of Definition 1;
(v) it preserves ellipticity in elastic loading if the energy is elliptic throughout the domain $\overset{\circ}{\mathscr{E}}_e$ which makes it useful in association with the exponentiated Hencky energy W_{eH} [8, 9, 27, 29, 30, 32, 33].

We finish this section by comparing the Helm 2001 model and the Lion 1997 flow rule with the Grandi-Stefanelli 2015 model.

Proposition 3 *In the isotropic case, the Lion 1997 flow rule (i.e. the Dettmer-Reese 2004 model [6]) is equivalent with the Grandi-Stefanelli 2015 flow rule.*

Proof We recall that the flow rule of the Lion 1997 model is

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = -\lambda_{\mathrm{p}}^+\, F_p^{-1}\, \frac{\operatorname{dev}_3 \varSigma_e}{\|\operatorname{dev}_3 \varSigma_e\|}\, F_p^{-T}, \quad \lambda_{\mathrm{p}}^+ \in \mathbb{R}_+, \tag{54}$$

for $\varSigma_e \notin \operatorname{int}(\mathscr{E}_{\mathrm{e}}(\varSigma_e, \frac{1}{3}\boldsymbol{\sigma}_{\mathrm{y}}^2))$. Since $F_p = R_p\sqrt{C_p}$ and in the isotropic case $\varSigma_e = R_p\, \overset{\circ}{\varSigma}\, R_p^T$, using (53) we rewrite the Lion's flow rule in the form

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = -\lambda_{\mathrm{p}}^+\, \sqrt{C_p}^{-1}\, R_p^T\, \frac{\operatorname{dev}_3(R_p\, \overset{\circ}{\varSigma}\, R_p^T)}{\|\operatorname{dev}_3(R_p\, \overset{\circ}{\varSigma}\, R_p^T)\|}\, R_p\, \sqrt{C_p}^{-1}, \qquad \lambda_{\mathrm{p}}^+ \in \mathbb{R}_+,$$

for $R_p \, \mathring{\Sigma} \, R_p^T \notin \operatorname{int}(\mathring{\mathscr{E}}_e(\mathring{\Sigma}, \frac{1}{3}\boldsymbol{\sigma}_{\mathbf{y}}^2))$, which is equivalent with

$$\sqrt{C_p}\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\sqrt{C_p} = -\lambda_{\mathrm{p}}^{+} \, \frac{\operatorname{dev}_3 \mathring{\Sigma}}{\| \operatorname{dev}_3 \mathring{\Sigma} \|}, \qquad \lambda_{\mathrm{p}}^{+} \in \mathbb{R}_{+}, \tag{55}$$

for $R_p \, \mathring{\Sigma} \, R_p^T \notin \operatorname{int}(\mathring{\mathscr{E}}_e(\mathring{\Sigma}, \frac{1}{3}\boldsymbol{\sigma}_{\mathbf{y}}^2))$. Moreover, $R_p \, \mathring{\Sigma} \, R_p^T \in \operatorname{int}(\mathring{\mathscr{E}}_e(\mathring{\Sigma}, \frac{1}{3}\boldsymbol{\sigma}_{\mathbf{y}}^2)) \;\Leftrightarrow\; \mathring{\Sigma} \in \operatorname{int}(\mathring{\mathscr{E}}_e(\mathring{\Sigma}, \frac{1}{3}\boldsymbol{\sigma}_{\mathbf{y}}^2))$. Therefore, the flow rule (55) becomes

$$\sqrt{C_p}\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\sqrt{C_p} \in -\partial_{\mathring{\Sigma}} \chi(\operatorname{dev}_3 \mathring{\Sigma}), \tag{56}$$

which coincides with the Grandi-Stefanelli 2015 flow rule (51). □

Proposition 4 *In the isotropic case, the Helm 2001 flow rule is equivalent with the Grandi-Stefanelli 2015 flow rule, i.e. it is also equivalent with the Lion 1997 flow rule and the Dettmer-Reese 2004 model.*

Proof We have

$$\begin{aligned}\mathring{\Sigma} &= \operatorname{sym}(\sqrt{C_p}^{-1} \, \widetilde{\Sigma} \, \sqrt{C_p}) = \operatorname{sym}(\sqrt{C_p}^{-1} \, \widetilde{\Sigma} \, C_p \, C_p^{-1} \, \sqrt{C_p}) \\ &= \operatorname{sym}(\sqrt{C_p}^{-1} \, (\widetilde{\Sigma} \, C_p)\sqrt{C_p}^{-1}),\end{aligned}$$

and we recall that for isotropic materials

$$\widetilde{\Sigma} \, C_p = F_p^T \, \Sigma_e \, F_p^{-T} \, C_p = F_p^T \, \Sigma_e \, F_p \in \operatorname{Sym}(3)$$

holds. Hence, for isotropic materials

$$\mathring{\Sigma} = \sqrt{C_p}^{-1} \, (\widetilde{\Sigma} \, C_p)\sqrt{C_p}^{-1} = \sqrt{C_p}^{-1} \, \widetilde{\Sigma} \, \sqrt{C_p}, \qquad \widetilde{\Sigma} = \sqrt{C_p} \, \mathring{\Sigma} \, \sqrt{C_p}^{-1}.$$

Using the above identity, we may rewrite the Helm 2001-flow rule (47) in the form

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p] = \lambda_{\mathrm{p}}^{+} \, \frac{\operatorname{dev}_3(\sqrt{C_p} \, \mathring{\Sigma} \, \sqrt{C_p}^{-1})}{\sqrt{\operatorname{tr}((\operatorname{dev}_3(\sqrt{C_p} \, \mathring{\Sigma} \, \sqrt{C_p}^{-1}))^2)}} \cdot C_p \, . \tag{57}$$

We also have

$$\begin{aligned}
\operatorname{tr}(\sqrt{C_p}\,\overset{\circ}{\Sigma}\,\sqrt{C_p}^{-1}) &= \operatorname{tr}(\overset{\circ}{\Sigma}),\\
\operatorname{dev}_3(\sqrt{C_p}\,\overset{\circ}{\Sigma}\,\sqrt{C_p}^{-1}) &= \sqrt{C_p}\,(\operatorname{dev}_3\overset{\circ}{\Sigma})\,\sqrt{C_p}^{-1},\\
\operatorname{tr}([\operatorname{dev}_3(\sqrt{C_p}\,\overset{\circ}{\Sigma}\,\sqrt{C_p}^{-1})]^2) &= \operatorname{tr}([\sqrt{C_p}\,(\operatorname{dev}_3\overset{\circ}{\Sigma})\,\sqrt{C_p}^{-1}]^2) = \operatorname{tr}(\sqrt{C_p}\,(\operatorname{dev}_3\overset{\circ}{\Sigma})^2\,\sqrt{C_p}^{-1})\\
&= \operatorname{tr}((\operatorname{dev}_3\overset{\circ}{\Sigma})^2) = \langle(\operatorname{dev}_3\overset{\circ}{\Sigma})^2, \mathbb{1}\rangle\\
&= \langle\operatorname{dev}_3\overset{\circ}{\Sigma}, \operatorname{dev}_3\overset{\circ}{\Sigma}\rangle = \|\operatorname{dev}_3\overset{\circ}{\Sigma}\|^2.
\end{aligned}$$

Hence, Helm's flow rule (47) is equivalent with

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{dt}}[C_p] = \lambda_\mathrm{p}^+\,\sqrt{C_p}\,\frac{\operatorname{dev}_3\overset{\circ}{\Sigma}}{\|\operatorname{dev}_3\overset{\circ}{\Sigma}\|}\,\sqrt{C_p}\\
\Leftrightarrow\quad &\sqrt{C_p}\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\,\sqrt{C_p} = -\lambda_\mathrm{p}^+\,\frac{\operatorname{dev}_3\overset{\circ}{\Sigma}}{\|\operatorname{dev}_3\overset{\circ}{\Sigma}\|}\,, \qquad (58)
\end{aligned}$$

and the proof is complete. □

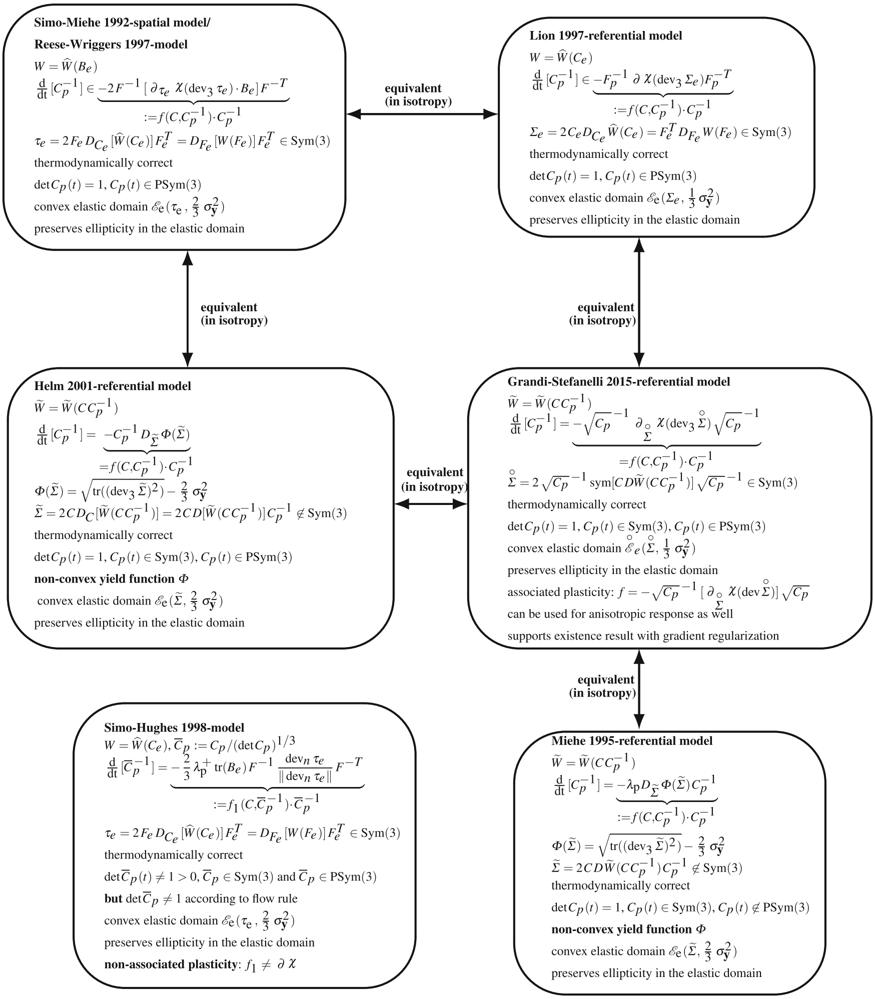
Remark 7 The equivalence is true for an isotropic formulation only. However, the Grandi-Stefanelli model will provide a consistent flow-rule for a plastic metric also in the anisotropic case.

An existence proof for the energetic formulation [7] of the model given by Grandi and Stefanelli [10] together with a full plastic strain regularization can be given along the lines of Mielke's energetic approach [18, 19, 24–26].

8 Summary

In isotropic elasto-plasticity it is common knowledge that a reduction to a *6-dimensional flow rule* for a *plastic metric* C_p is in principle possible. We have discussed several existing different models. Not all of them are free of inconsistencies. This testifies to the fact that setting up a consistent 6-dimensional flow-rule is not entirely trivial.

One problem which often occurs, is that the flow rule for C_p is written in terms of F_p, which however should not appear at all. One finding of our investigation is that, nevertheless, in the isotropic case, all consistent flow rules can be expressed in C_p alone and are equivalent. The Grandi-Stefanelli model [10] has the decisive



Simo-Miehe 1992-spatial model/

Reese-Wriggers 1997-model

$W = \widehat{W}(B_e)$

$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] \in \underbrace{-2F^{-1}[\,\partial_{\tau_e}\ \chi(\mathrm{dev}_3\,\tau_e)\cdot B_e]F^{-T}}_{:=f(C,C_p^{-1})\cdot C_p^{-1}}$

$\tau_e = 2F_e D_{C_e}[\widehat{W}(C_e)]F_e^T = D_{F_e}[W(F_e)]F_e^T \in \mathrm{Sym}(3)$

thermodynamically correct

$\det C_p(t) = 1,\ C_p(t) \in \mathrm{PSym}(3)$

convex elastic domain $\mathscr{E}_{\mathrm{e}}(\tau_{\mathrm{e}}\,,\,\frac{2}{3}\,\sigma_{\mathbf{y}}^2)$

preserves ellipticity in the elastic domain

Lion 1997-referential model

$W = \widehat{W}(C_e)$

$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] \in \underbrace{-F_p^{-1}\ \partial\,\chi(\mathrm{dev}_3\,\Sigma_e)F_p^{-T}}_{:=f(C,C_p^{-1})\cdot C_p^{-1}}$

$\Sigma_e = 2C_e D_{C_e}\widehat{W}(C_e) = F_e^T D_{F_e} W(F_e) \in \mathrm{Sym}(3)$

thermodynamically correct

$\det C_p(t) = 1,\ C_p(t) \in \mathrm{PSym}(3)$

convex elastic domain $\mathscr{E}_{\mathrm{e}}(\Sigma_e\,,\,\frac{1}{3}\,\sigma_{\mathbf{y}}^2)$

preserves ellipticity in the elastic domain

Helm 2001-referential model

$\widetilde{W} = \widetilde{W}(CC_p^{-1})$

$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = \underbrace{-C_p^{-1}D_{\widetilde{\Sigma}}\Phi(\widetilde{\Sigma})}_{=f(C,C_p^{-1})\cdot C_p^{-1}}$

$\Phi(\widetilde{\Sigma}) = \sqrt{\mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2)} - \frac{2}{3}\,\sigma_{\mathbf{y}}^2$

$\widetilde{\Sigma} = 2CD_C[\widetilde{W}(CC_p^{-1})] = 2CD[\widetilde{W}(CC_p^{-1})]C_p^{-1} \notin \mathrm{Sym}(3)$

thermodynamically correct

$\det C_p(t) = 1,\ C_p(t) \in \mathrm{Sym}(3),\ C_p(t) \in \mathrm{PSym}(3)$

non-convex yield function Φ

convex elastic domain $\mathscr{E}_{\mathrm{e}}(\widetilde{\Sigma},\,\frac{2}{3}\,\sigma_{\mathbf{y}}^2)$

preserves ellipticity in the elastic domain

Grandi-Stefanelli 2015-referential model

$\widetilde{W} = \widetilde{W}(CC_p^{-1})$

$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = \underbrace{-\sqrt{C_p}^{\,-1}\ \partial_{\mathring{\Sigma}}\ \chi(\mathrm{dev}_3\,\mathring{\Sigma})\sqrt{C_p}^{\,-1}}_{=f(C,C_p^{-1})\cdot C_p^{-1}}$

$\mathring{\Sigma} = 2\sqrt{C_p}^{\,-1}\,\mathrm{sym}[CD\widetilde{W}(CC_p^{-1})]\sqrt{C_p}^{\,-1} \in \mathrm{Sym}(3)$

thermodynamically correct

$\det C_p(t) = 1,\ C_p(t) \in \mathrm{Sym}(3),\ C_p(t) \in \mathrm{PSym}(3)$

convex elastic domain $\mathring{\mathscr{E}}_e(\mathring{\Sigma},\,\frac{1}{3}\,\sigma_{\mathbf{y}}^2)$

preserves ellipticity in the elastic domain

associated plasticity: $f = -\sqrt{C_p}^{\,-1}[\,\partial_{\mathring{\Sigma}}\ \chi(\mathrm{dev}\,\mathring{\Sigma})]\sqrt{C_p}$

can be used for anisotropic response as well

supports existence result with gradient regularization

Simo-Hughes 1998-model

$W = \widehat{W}(C_e),\ \overline{C}_p := C_p/(\det C_p)^{1/3}$

$\frac{\mathrm{d}}{\mathrm{dt}}[\overline{C}_p^{-1}] = \underbrace{-\frac{2}{3}\lambda_{\mathrm{p}}^{+}\,\mathrm{tr}(B_e)F^{-1}\,\frac{\mathrm{dev}_n\,\tau_e}{\|\mathrm{dev}_n\,\tau_e\|}\,F^{-T}}_{:=f_1(C,\overline{C}_p^{-1})\cdot\overline{C}_p^{-1}}$

$\tau_e = 2F_e D_{C_e}[\widehat{W}(C_e)]F_e^T = D_{F_e}[W(F_e)]F_e^T \in \mathrm{Sym}(3)$

thermodynamically correct

$\det\overline{C}_p(t) \neq 1 > 0,\ \overline{C}_p \in \mathrm{Sym}(3)$ and $\overline{C}_p \in \mathrm{PSym}(3)$

but $\det\overline{C}_p \neq 1$ according to flow rule

convex elastic domain $\mathscr{E}_{\mathrm{e}}(\tau_{\mathrm{e}}\,,\,\frac{2}{3}\,\sigma_{\mathbf{y}}^2)$

preserves ellipticity in the elastic domain

non-associated plasticity: $f_1 \neq \partial\,\chi$

Miehe 1995-referential model

$\widetilde{W} = \widetilde{W}(CC_p^{-1})$

$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = \underbrace{-\lambda_{\mathrm{p}}D_{\widetilde{\Sigma}}\Phi(\widetilde{\Sigma})C_p^{-1}}_{:=f(C,C_p^{-1})\cdot C_p^{-1}}$

$\Phi(\widetilde{\Sigma}) = \sqrt{\mathrm{tr}((\mathrm{dev}_3\,\widetilde{\Sigma})^2)} - \frac{2}{3}\,\sigma_{\mathbf{y}}^2$

$\widetilde{\Sigma} = 2CD\widetilde{W}(CC_p^{-1})C_p^{-1} \notin \mathrm{Sym}(3)$

thermodynamically correct

$\det C_p(t) = 1,\ C_p(t) \in \mathrm{Sym}(3),\ C_p(t) \notin \mathrm{PSym}(3)$

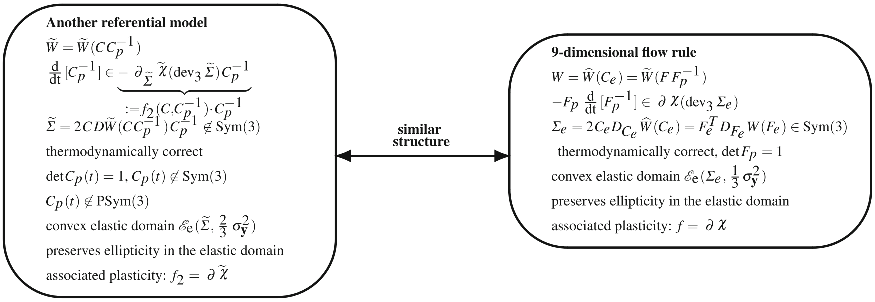
non-convex yield function Φ

convex elastic domain $\mathscr{E}_{\mathrm{e}}(\widetilde{\Sigma},\,\frac{2}{3}\,\sigma_{\mathbf{y}}^2)$

preserves ellipticity in the elastic domain

Fig. 1 Idealized, isotropic perfect plasticity models involving a 6-dimensional flow rule for C_p w.r.t. the reference configuration are considered. By definition, the trajectory for the plastic metric $C_p(t)$ should remain in PSym(3). λ_{p}^{+} is the plastic multiplier. We have recast all flow rules in the format $\frac{\mathrm{d}}{\mathrm{dt}}[P^{-1}]\,P \in -\partial\chi$ or $\sqrt{P}\,\frac{\mathrm{d}}{\mathrm{dt}}[P^{-1}]\sqrt{P} \in -\partial\chi$

advantage to be operable also in the anisotropic case. In Figs. 1 and 2 we summarize the investigated isotropic plasticity models and we indicate if the known conditions which make them consistent are satisfied.



Another referential model

$\widetilde{W} = \widetilde{W}(C C_p^{-1})$

$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] \in \underbrace{-\,\partial_{\widetilde{\Sigma}}\,\widetilde{\chi}(\mathrm{dev}_3\,\widetilde{\Sigma})\,C_p^{-1}}_{:=f_2(C,C_p^{-1})\cdot C_p^{-1}}$

$\widetilde{\Sigma} = 2 C D\widetilde{W}(C C_p^{-1}) C_p^{-1} \notin \mathrm{Sym}(3)$

thermodynamically correct

$\det C_p(t) = 1,\ C_p(t) \notin \mathrm{Sym}(3)$

$C_p(t) \notin \mathrm{PSym}(3)$

convex elastic domain $\mathscr{E}_{\mathrm{e}}(\widetilde{\Sigma}, \frac{2}{3}\,\sigma_{\mathbf{y}}^2)$

preserves ellipticity in the elastic domain

associated plasticity: $f_2 = \partial\,\widetilde{\chi}$

9-dimensional flow rule

$W = \widehat{W}(C_e) = \widetilde{W}(F F_p^{-1})$

$-F_p\,\frac{\mathrm{d}}{\mathrm{dt}}[F_p^{-1}] \in \partial\,\chi(\mathrm{dev}_3\,\Sigma_e)$

$\Sigma_e = 2 C_e D_{C_e}\widehat{W}(C_e) = F_e^T D_{F_e} W(F_e) \in \mathrm{Sym}(3)$

thermodynamically correct, $\det F_p = 1$

convex elastic domain $\mathscr{E}_{\mathrm{e}}(\Sigma_e, \frac{1}{3}\,\sigma_{\mathbf{y}}^2)$

preserves ellipticity in the elastic domain

associated plasticity: $f = \partial\,\chi$

Fig. 2 An inconsistent model and a 9-dimensional flow rule for F_p. They are associative, since both flow rules are in the format $\frac{\mathrm{d}}{\mathrm{dt}}[P]\,P^{-1} \in -\partial\chi$ or $\frac{\mathrm{d}}{\mathrm{dt}}[\varepsilon_p] \in \partial\chi$

Acknowledgments We would like to thank Prof. Stefanie Reese (RWTH Aachen), Prof. Jörn Ihlemann (TU Chemnitz), Dr. Alexey Shutov (TU Chemnitz) and Prof. Alexander Mielke (WIAS-Berlin) for in-depth discussion of flow rules in C_p. I.D. Ghiba acknowledges support from the Romanian National Authority for Scientific Research (CNCS-UEFISCDI), Project No. PN-II-ID-PCE-2011-3-0521.

Appendix

A.1 The Lion 1997 Model for the Neo-Hooke Elastic Energy

For a quick consistency check we exhibit the consistency of this model directly for a Neo-Hooke elastic energy and we give the concrete expression for the functions $f(C, C_p^{-1})$ and $\widehat{f}(C, C_p^{-1})$. To this end, we consider the energy

$$\widehat{W}_{\mathrm{NH}}(C_e) = \mu\,\mathrm{tr}\left(\frac{C_e}{\det C_e^{1/3}}\right) + h(\det C) \overset{\det C_p = 1}{=} \mu\,\mathrm{tr}\left(\frac{C_e}{\det C^{1/3}}\right) + h(\det C).$$

We deduce $\Sigma_e := 2\,C_e\,D_{C_e}[\widehat{W}(C_e)] = 2\,C_e\,\mu\,\frac{1}{\det C^{1/3}} \cdot \mathbb{1} = 2\,\mu\,\frac{1}{\det C^{1/3}} \cdot C_e$. Hence, the flow rule (24) can be written in the form

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] &= -\lambda_{\mathrm{p}}^{+}\,F_p^{-1}\,\frac{\mathrm{dev}\,C_e}{\|\,\mathrm{dev}\,C_e\|}\,F_p^{-T} \\ &= -\frac{\lambda_{\mathrm{p}}^{+}}{\|\,\mathrm{dev}\,C_e\|}\left(F_p^{-1} C_e\,F_p^{-T} - \frac{1}{3}\,\mathrm{tr}(C_e)\cdot F_p^{-1} F_p^{-T}\right) \\ &= -\frac{\lambda_{\mathrm{p}}^{+}}{\|\,\mathrm{dev}\,C_e\|}\left(C_p^{-1}\,C\,C_p^{-1} - \frac{1}{3}\,\mathrm{tr}(C_p^{-1}\,C)\cdot C_p^{-1}\right). \qquad (59)\end{aligned}$$

We also deduce

$$\begin{aligned}\|\operatorname{dev} C_e\|^2 &= \|C_e\|^2 - \frac{1}{3}\,[\operatorname{tr}(C_e)]^2 = \|F_p^{-T}\,C\,F_p^{-1}\|^2 - \frac{1}{3}\,[\operatorname{tr}(F_p^{-T}\,C\,F_p^{-1})]^2\\ &= \langle F_p^{-T}\,C\,F_p^{-1}, F_p^{-T}\,C\,F_p^{-1}\rangle - \frac{1}{3}\,\langle F_p^{-T}\,C\,F_p^{-1}), \mathbb{1}\rangle^2\\ &= \langle C_p^{-1}\,C, C\,C_p^{-1}\rangle - \frac{1}{3}\,[\operatorname{tr}(C_p^{-1}\,C)]^2 = [\operatorname{tr}(C_p^{-1}\,C)^2] - \frac{1}{3}\,[\operatorname{tr}(C_p^{-1}\,C)]^2.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] &= -\frac{\lambda_{\mathrm{p}}^+}{\sqrt{\operatorname{tr}[(C_p^{-1}\,C)^2] - \frac{1}{3}\,[\operatorname{tr}(C_p^{-1}\,C)]^2}}\,\left(C_p^{-1}\,C\, - \frac{1}{3}\operatorname{tr}(C_p^{-1}\,C)\cdot\mathbb{1}\right)\,C_p^{-1}\\ &= -\frac{\lambda_{\mathrm{p}}^+}{\sqrt{[\operatorname{tr}(C_p^{-1}\,C)^2] - \frac{1}{3}\operatorname{tr}[(C_p^{-1}\,C)]^2}}\,C_p^{-1}\,\left(C\,C_p^{-1} - \frac{1}{3}\operatorname{tr}(C\,C_p^{-1})\cdot\mathbb{1}\right).\end{aligned}\tag{60}$$

Comparing (35), (60), (25) and (28), we deduce

$$\begin{aligned}&\widehat{f}(C, C_p^{-1}) = C,\\ &f(C, C_p^{-1}) = \left\{\frac{-\lambda_{\mathrm{p}}^+}{\sqrt{\operatorname{tr}[(C\,C_p^{-1})^2] - \frac{1}{3}[\operatorname{tr}(C\,C_p^{-1})]^2}}\,\operatorname{dev}_3(\,C_p^{-1}\,C)\;\;\Big|\;\;\lambda_{\mathrm{p}}^+ \in \mathbb{R}_+\right\}.\end{aligned}$$

We clearly see that even for this simple energy, we have $\operatorname{tr}[(C\,C_p^{-1})^2] - \frac{1}{3}[\operatorname{tr}(C\,C_p^{-1})]^2 \neq \|\operatorname{dev}_3(C\,C_p^{-1})\|^2$, since if we assume the contrary we deduce

$$\begin{aligned}\operatorname{tr}[(C\,C_p^{-1})^2] - \frac{1}{3}[\operatorname{tr}(C\,C_p^{-1})]^2 &= \|C\,C_p^{-1}\|^2 - \frac{1}{3}[\operatorname{tr}(C\,C_p^{-1})]^2\\ \Leftrightarrow \langle C\,C_p^{-1}, (C\,C_p^{-1})^T\rangle &= \pm\langle C\,C_p^{-1}, C\,C_p^{-1}\rangle.\end{aligned}\tag{61}$$

On the other hand, we deduce

$$\begin{aligned}\operatorname{tr}[(C\,C_p^{-1})^2] &= \langle (C\,C_p^{-1})\,(C\,C_p^{-1}), \mathbb{1}\rangle = \langle C_p^{-1}\,C\,C_p^{-1}\,(C\,C_p^{-1})\,C_p, \mathbb{1}\rangle\\ &= \langle C_p^{-1}\,C\,C_p^{-1}\,C_p\,(C\,C_p^{-1})^T, \mathbb{1}\rangle = \langle C_p^{-1}\,C, C\,C_p^{-1}\rangle.\end{aligned}\tag{62}$$

Since from Remark 4 it follows that $C_p \in \operatorname{PSym}(3)$, we further deduce that

$$\begin{aligned}\langle C_p^{-1}\,C\,C_p^{-1}\,C_p, C\,C_p^{-1}\rangle &= \langle U_p^{-1}U_p^{-1}\,C\,C_p^{-1}\,U_p\,U_p, C\,C_p^{-1}\rangle\\ &= \langle U_p^{-1}\,C\,C_p^{-1}\,U_p, U_p^{-1}C\,C_p^{-1}\,U_p\rangle\\ &= \|U_p^{-1}\,C\,U_p^{-1}\|^2 \geq 0,\end{aligned}\tag{63}$$

where $U_p^2 = C_p$. Hence, $[\mathrm{tr}(C\,C_p^{-1})^2] \geq 0$ and from (61) we deduce

$$\begin{aligned}\langle C\,C_p^{-1}, (C\,C_p^{-1})^T\rangle = \langle C\,C_p^{-1}, C\,C_p^{-1}\rangle \quad &\Leftrightarrow \quad \langle C\,C_p^{-1}, \mathrm{skew}(C\,C_p^{-1})\rangle = 0\\ \Leftrightarrow \quad \mathrm{skew}(C\,C_p^{-1}) = 0 \quad &\Leftrightarrow \quad C\,C_p^{-1} \in \mathrm{Sym}(3). \end{aligned} \tag{64}$$

In conclusion, $\mathrm{tr}[(C\,C_p^{-1})^2] - \frac{1}{3}[\mathrm{tr}(C\,C_p^{-1})]^2 \neq \|\,\mathrm{dev}_3(C\,C_p^{-1})\|^2$ and the flow-rule does not have a subdifferential structure of the form $C_p\,\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] \in -\partial\chi\,(\mathrm{dev}\,\Sigma)$.

A.2 The Helm 2001 Model for the Neo-Hooke Energy

For the simplest Neo-Hooke elastic energy $W(F_e) = \mathrm{tr}(C_e) = \widetilde{W}(C\,C_p^{-1}) = \frac{1}{2}$ $\mathrm{tr}(C\,C_p^{-1})$, we have

$$D_C[\widetilde{W}(C\,C_p^{-1})] = \frac{1}{2}\,C_p^{-1} \qquad \Rightarrow \qquad \widetilde{\Sigma} = C\,C_p^{-1} \notin \mathrm{Sym}(3), \tag{65}$$

and the flow rule (47) implies

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[C_p] &= \lambda_\mathrm{p}^+\,\frac{\mathrm{dev}_3(C\,C_p^{-1})}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3(C\,C_p^{-1}))^2]}}\cdot C_p\\ &= \frac{\lambda_\mathrm{p}^+}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3(C\,C_p^{-1}))^2]}}\,[C - \frac{1}{3}\mathrm{tr}(C\,C_p^{-1})\cdot C_p] \in \mathrm{Sym}(3) \quad \Rightarrow \quad C_p \in \mathrm{Sym}(3),\end{aligned}$$

and also

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p]\,C_p^{-1} = \lambda_\mathrm{p}^+\,\frac{\mathrm{dev}_3(C\,C_p^{-1})}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3(C\,C_p^{-1}))^2]}} \quad \Rightarrow \quad \det C_p = 1. \tag{66}$$

The thermodynamical consistency may follow from (49). An alternative proof, directly for the Neo-Hooke case, results from (48) and (65), since we have at fixed in time C

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}[\widetilde{W}(C\,C_p^{-1})] &= -\frac{\lambda_\mathrm{p}^+}{4\sqrt{\mathrm{tr}[(\mathrm{dev}_3(C\,C_p^{-1}))^2]}}\,\langle C_p^{-1}\,\widetilde{\Sigma}\,C_p, \mathrm{dev}_3\,\widetilde{\Sigma}\rangle\\ &= -\frac{\lambda_\mathrm{p}^+}{\sqrt{\mathrm{tr}[(\mathrm{dev}_3(C\,C_p^{-1}))^2]}}\,\langle C_p^{-1}\,C, \mathrm{dev}_3(C\,C_p^{-1})\rangle\end{aligned}$$

$$
\begin{aligned}
&= -\frac{\lambda_{\rm p}^{+}}{\sqrt{\operatorname{tr}[(\operatorname{dev}_3(C\,C_p^{-1}))^2]}}\,\langle (C\,C_p^{-1})^T, \operatorname{dev}_3(C\,C_p^{-1})\rangle \qquad (67)\\
&= -\frac{\lambda_{\rm p}^{+}}{\sqrt{\operatorname{tr}[(\operatorname{dev}_3(C\,C_p^{-1}))^2]}}\,\langle C^{-1/2}\operatorname{dev}_3(C\,C_p^{-1})\,C^{1/2}\,C^{-1/2}\,\operatorname{dev}_3(C\,C_p^{-1})C^{1/2}, \mathbb{1}\rangle\\
&= -\frac{\lambda_{\rm p}^{+}}{\sqrt{\operatorname{tr}[(\operatorname{dev}_3(C\,C_p^{-1}))^2]}}\,\langle \operatorname{dev}_3(C^{1/2}\,C_p^{-1}\,C^{1/2})^T\ \operatorname{dev}_3(C^{1/2}\,C_p^{-1}\,C^{1/2}), \mathbb{1}\rangle\\
&= -\frac{\lambda_{\rm p}^{+}}{\sqrt{\operatorname{tr}[(\operatorname{dev}_3(C\,C_p^{-1}))^2]}}\,\|\operatorname{dev}_3(C^{1/2}\,C_p^{-1}\,C^{1/2})\|^2,
\end{aligned}
$$

which is negative.[5] Therefore, this model is thermodynamically correct as now shown also for the simple Neo-Hooke energy.

A.3 Another Referential Model

We recall that, in view of Lemma 2, any isotropic free energy W defined in terms of F_e can be expressed as $W(F_e) = \widetilde{W}(C\,C_p^{-1})$. In order to assume that the reduced dissipation inequality is satisfied, we compute

$$
\begin{aligned}
\frac{\rm d}{\rm dt}\widetilde{W}(C\,C_p^{-1}) &= \langle D\widetilde{W}(C\,C_p^{-1}), C\,\frac{\rm d}{\rm dt}[C_p^{-1}]\rangle\\
&= \langle C\,D\widetilde{W}(C\,C_p^{-1})\,C_p^{-1}, \frac{\rm d}{\rm dt}[C_p^{-1}]\,C_p\rangle = \langle \widetilde{\Sigma}, \frac{\rm d}{\rm dt}[C_p^{-1}]\,C_p\rangle.
\end{aligned}
$$

Here, $\widetilde{\Sigma} = 2\,C\,D_C[\widetilde{W}(C\,C_p^{-1})]$, as in the Reese 2008 and Shutov-Ihlemann 2014 model. It is tempting to assume the flow rule in the associated form (see e.g. the habilitation thesis of Miehe [21, p. 73, Satz 5.32] or [23] and also [22, Table 1])

$$
\frac{\rm d}{\rm dt}[C_p^{-1}]\,C_p \in -\,\partial_{\widetilde{\Sigma}}\chi\,(\operatorname{dev}_3\widetilde{\Sigma}), \qquad (68)
$$

where $\chi\,(\operatorname{dev}_3\widetilde{\Sigma})$ is the indicator function of the convex elastic domain

$$
\mathscr{E}_{\rm e}(\widetilde{\Sigma}, \frac{2}{3}\,\sigma_{\rm y}^2) := \left\{\widetilde{\Sigma} \in \mathbb{R}^{3\times 3}\mid \|\operatorname{dev}_3\widetilde{\Sigma}\|^2 \le \frac{2}{3}\,\sigma_{\rm y}^2\right\}.
$$

[5] Surprisingly, this follows even if C and C_p^{-1} do not commute in general. If C and C_p commute, then $X = C\,C_p^{-1} \in \operatorname{Sym}(3)$ and the quantity does have a sign, since then $\langle X^T, \operatorname{dev}_3 X\rangle = \|\operatorname{dev}_3 X\|^2 \ge 0$.

Note that this flow rule (68) is not the formulation which Miehe seemed to intend. We have discussed the correct interpretation in Sect. 3.

Regarding such a formulation we can summarize our observations:

(i) this flow rule is thermodynamically correct;
(ii) the right hand side is a function of C and C_p^{-1} only, i.e. $\widetilde{\Sigma} = \widetilde{\Sigma}(C, C_p^{-1})$;
(iii) plastic incompressibility: from this flow rule it follows that $\det C_p(t) = 1$, since the right hand side is trace-free;
(iv) however, the computed tensor $\boldsymbol{C_p(t)}$ **will not be symmetric** since $\widetilde{\Sigma}\, C_p^{-1} \notin \mathrm{Sym}(3)$ in general. For instance, for the simplest Neo-Hooke energy $W(F_e) = \mathrm{tr}(C_e) = \mathrm{tr}(C\, C_p^{-1})$ we have $\widetilde{\Sigma} = 2\, C\, C_p^{-1} \notin \mathrm{Sym}(3)$, $\widetilde{\Sigma}\, C_p^{-1} = 2\, C\, C_p^{-2} \notin \mathrm{Sym}(3)$, in general, and the flow rule becomes

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}] = -2\, \frac{\lambda_\mathrm{p}^+}{\|\, \mathrm{dev}(C\, C_p^{-1})\|}\, [C\, C_p^{-2} - \frac{1}{3}\, \mathrm{tr}(C\, C_p^{-1}) \cdot C_p^{-1}] \notin \mathrm{Sym}(3); \tag{69}$$

(v) it is an associated plasticity model in the sense of Definition 1.

In conclusion, this model is inconsistent with the requirement for a plastic metric, i.e. $C_p \in \mathrm{Psym}(3)$. Moreover, if we are looking to the flow rule in the associated form considered in the habilitation thesis of Miehe [21, p. 73, Satz 5.32] (see [23] and also [22, Table 1]), since the subdifferential $\partial_{\widetilde{\Sigma}} \chi\, (\mathrm{dev}_3\, \widetilde{\Sigma})$ of the indicator function χ is the normal cone

$$\mathscr{N}(\mathscr{E}_\mathrm{e}(\widetilde{\Sigma}, \frac{1}{3}\, \boldsymbol{\sigma}_\mathbf{y}^2); \mathrm{dev}_3\, \widetilde{\Sigma}) = \begin{cases} 0, & \widetilde{\Sigma} \in \mathrm{int}(\mathscr{E}_\mathrm{e}(\widetilde{\Sigma}, \frac{1}{3}\boldsymbol{\sigma}_\mathbf{y}^2)) \\ \{\lambda_\mathrm{p}^+ \, \frac{\mathrm{dev}_3\, \widetilde{\Sigma}}{\|\, \mathrm{dev}_3\, \widetilde{\Sigma}\|} \mid \lambda_\mathrm{p}^+ \in \mathbb{R}_+\}, & \widetilde{\Sigma} \notin \mathrm{int}(\mathscr{E}_\mathrm{e}(\widetilde{\Sigma}, \frac{1}{3}\boldsymbol{\sigma}_\mathbf{y}^2)). \end{cases}$$

the flow rule can be written in the form

$$\frac{\mathrm{d}}{\mathrm{dt}}[C_p^{-1}]\, C_p = -\lambda_\mathrm{p}^+\, \frac{\mathrm{dev}_3\, \widetilde{\Sigma}}{\|\, \mathrm{dev}_3\, \widetilde{\Sigma}\|}\,, \tag{70}$$

which is not equivalent with the flow rule (22) considered by Miehe in [22], since $\widetilde{\Sigma} \notin \mathrm{Sym}(3)$. Let us remark that we have the symmetries $\mathrm{dev}_3\, \widetilde{\Sigma} \cdot C_p \in \mathrm{Sym}(3)$, $C_p^{-1}\, \mathrm{dev}_3\, \widetilde{\Sigma} \in \mathrm{Sym}(3)$, but these do not assure that the flow rule (70) implies $C_p \in \mathrm{Sym}(3)$.

A.4 The Simo-Hughes 1998-Model for the Saint-Venant-Kirchhoff Energy and for the Neo-Hooke Energy

In order to see that the quantity $F_e^{-1}\, \frac{\mathrm{dev}_n\, \tau_e}{\|\, \mathrm{dev}_n\, \tau_e\|}\, F_e^{-T}$ which appears in the Simo-Hughes flow rule is not necessarily a trace free matrix, we consider two energies: the isotropic elastic Saint-Venant-Kirchhoff energy and the energy considered by Simo and Hughes [42, p. 307]. On the one hand, the well known isotropic elastic Saint-Venant-Kirchhoff energy is

$$W_{\text{SVK}} = \frac{\mu}{4}\,\|C_e - \mathbb{1}\|^2 + \frac{\lambda}{8}\,[\operatorname{tr}(C_e - \mathbb{1})]^2 = \frac{\mu}{4}\,\|B_e - \mathbb{1}\|^2 + \frac{\lambda}{8}\,[\operatorname{tr}(B_e - \mathbb{1})]^2,$$

and the corresponding Kirchhoff stress tensor is given by

$$\begin{aligned}\tau_e^{\text{SVK}}(U) = D_{B_e}[W^{\text{SVK}}(B_e)] &= \mu\,(F_e^{-T}\,C_e\,F_e^T - \mathbb{1}) + \frac{\lambda}{2}\operatorname{tr}(C_e - \mathbb{1})\cdot\mathbb{1}\\ &= \mu\,(F_e\,F_e^T - \mathbb{1}) + \frac{\lambda}{2}\operatorname{tr}(F_e\,F_e^T - \mathbb{1})\cdot\mathbb{1}.\end{aligned}$$

Hence, we deduce

$$\begin{aligned}F_e^{-1}\,[\operatorname{dev}_n \tau_e^{\text{SVK}}]\,F_e^{-T} = \mu\,F_e^{-1}\operatorname{dev}_n[\,F_e\,F_e^T\,]\,F_e^{-T} &= \mu\,F_e^{-1}\,[\,F_e\,F_e^T - \frac{1}{3}\operatorname{tr}(F_e\,F_e^T)\cdot\mathbb{1}]\,F_e^{-T}\\ &= \mu\,[\,\mathbb{1} - \frac{1}{3}\operatorname{tr}(F_e\,F_e^T)\cdot F_e^{-1}\,F_e^{-T}\,]\\ &= \mu\,[\,\mathbb{1} - \frac{1}{3}\operatorname{tr}(F_e\,F_e^T)\cdot F_e^{-1}\,F_e^{-T}\,],\end{aligned}$$

and further

$$\begin{aligned}\langle F_e^{-1}\,[\operatorname{dev}_n \tau_e^{\text{SVK}}]\,F_e^{-T}, \mathbb{1}\rangle &= \mu\,\langle\,\mathbb{1} - \frac{1}{3}\operatorname{tr}(F_e\,F_e^T)\cdot F_e^{-1}\,F_e^{-T}], \mathbb{1}\rangle\\ &= \mu\left[3 - \frac{1}{3}\operatorname{tr}(F_e\,F_e^T)\operatorname{tr}(F_e^{-1}\,F_e^{-T})\right]\\ &= \mu\left[3 - \frac{1}{3}\operatorname{tr}(C_e)\operatorname{tr}(C_e^{-1})\right]\\ &= \mu\left[3 - \frac{1}{3\,\det C_e}\operatorname{tr}(C_e)\operatorname{tr}(\operatorname{Cof} C_e)\right].\end{aligned}$$

We remark that $\langle F_e^{-1}\,[\operatorname{dev}_n \tau_e^{\text{SVK}}]\,F_e^{-T}, \mathbb{1}\rangle = 0$ if and only if $\operatorname{tr}(C_e)\operatorname{tr}(\operatorname{Cof} C_e) = 9\,\det C_e$, which does not hold true in general. Since C_e and $\operatorname{Cof} C_e$ are coaxial and symmetric, the problem can be reduced to the diagonal case, i.e. we may assume $C_e = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3), \lambda_i > 0$. Hence the condition $\operatorname{tr}(C_e)\operatorname{tr}(\operatorname{Cof} C_e) = 9\,\det C_e$, becomes

$$\begin{aligned}9\,\lambda_1\lambda_2\lambda_3 &= (\lambda_1 + \lambda_2 + \lambda_3)(\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_2)\\ \Leftrightarrow\quad 0 &= \lambda_1(\lambda_2 - \lambda_3)^2 + \lambda_2(\lambda_3 - \lambda_1)^2 + \lambda_3(\lambda_1 - \lambda_3)^2\end{aligned}$$

which is satisfied if and only if $\lambda_1 = \lambda_2 = \lambda_3$. Therefore, for the Saint-Venant-Kirchhoff energy, in this model, $\det \overline{C}_p = 1$ is only true for the conformal mapping $F_e = \lambda \cdot \mathrm{SO}(3) \in \mathbb{R}_+ \cdot \mathrm{SO}(3)$.

On the other hand, the energy considered by Simo and Hughes [42, p. 307] is

$$W_{\text{Simo}}(B_e) = \frac{\mu}{2}\langle \frac{B_e}{\det B_e^{1/3}} - \mathbb{1}, \mathbb{1}\rangle + \frac{\kappa}{4}\left[(\det B_e - 1) - \log(\det B_e)\right],$$

for which the Kirchhoff stress tensor is given by

$$\tau_e^{\text{Simo}} = \mu \operatorname{dev}_3\left(\frac{B_e}{\det B_e^{1/3}}\right) + \frac{\kappa}{2}\left(J_e - \frac{1}{J_e}\right)\cdot \mathbb{1}.$$

Hence, we deduce

$$\begin{aligned}
&\langle F_e^{-1}\,[\operatorname{dev}_n \tau_e^{\text{Simo}}]\,F_e^{-T}, \mathbb{1}\rangle = \mu\,\frac{1}{\det B_e^{1/3}}\,\langle F_e^{-1}[\operatorname{dev}_3 B_e]\,F_e^{-T}, \mathbb{1}\rangle \\
&= \mu\,\frac{1}{\det B_e^{1/3}}\,\langle \operatorname{dev}_3 B_e, F_e^{-T}\,F_e^{-1}\rangle \\
&= \mu\,\frac{1}{\det B_e^{1/3}}\,\langle \operatorname{dev}_3 B_e, B_e^{-1}\rangle = \mu\,\frac{1}{\det B_e^{1/3}}\left[\langle B_e, B_e^{-1}\rangle - \frac{1}{3}\operatorname{tr}(B_e)\operatorname{tr}(B_e^{-1})\right] \\
&= \mu\,\frac{1}{\det B_e^{1/3}}\left[3 - \frac{1}{3}\operatorname{tr}(B_e)\operatorname{tr}(B_e^{-1})\right] \\
&= \mu\,\frac{1}{\det B_e^{4/3}}\left[3\,\det B_e - \frac{1}{3}\operatorname{tr}(B_e)\operatorname{tr}(\operatorname{Cof} B_e)\right].
\end{aligned}$$

Therefore $\langle F_e^{-1}\,[\operatorname{dev}_n \tau_e^{\text{Simo}}]\,F_e^{-T}, \mathbb{1}\rangle = 0$ if and only if $9\,\det B_e = \operatorname{tr}(B_e)\operatorname{tr}(\operatorname{Cof} B_e)$. Similar as above, it follows that this holds true if and only if $F_e = \lambda \cdot \mathrm{SO}(3) \in \mathbb{R}_+ \cdot \mathrm{SO}(3)$.

References

1. Brepols, T., Vladimirov, I., & Reese, S. (2014). Numerical comparison of isotropic hypo-and hyperelastic-based plasticity models with application to industrial forming processes. *International Journal of Plasticity*, *63*, 18–48.
2. Cleja-Ţigoiu, S. (2003). Consequences of the dissipative restrictions in finite anisotropic elasto-plasticity. *International Journal of Plasticity*, *19*(11), 1917–1964.
3. Cleja-Ţigoiu, S., & Iancu, L. (2013). Orientational anisotropy and strength-differential effect in orthotropic elasto-plastic materials. *International Journal of Plasticity*, *47*, 80–110.
4. Cleja-Ţigoiu, S., & Maugin, G. A. (2000). Eshelby's stress tensors in finite elastoplasticity. *Acta Mechanica*, *139*(1–4), 231–249.
5. Cuitino, A., & Ortiz, M. (1992). A material-independent method for extending stress update algorithms from small-strain plasticity to finite plasticity with multiplicative kinematics. *Engineering Computations*, *9*(4), 437–451.
6. Dettmer, W., & Reese, S. (2004). On the theoretical and numerical modelling of Armstrong-Frederick kinematic hardening in the finite strain regime. *Computer Methods in Applied Mechanics and Engineering*, *193*(1), 87–116.

7. Frigeri, S., & Stefanelli, U. (2012). Existence and time-discretization for the finite-strain Souza-Auricchio constitutive model for shape-memory alloys. *Continuum Mechanics and Thermodynamics*, *24*(1), 63–77.
8. Ghiba, I. D., Neff, P., & Silhavy, M. (2015). The exponentiated Hencky-logarithmic strain energy. *International Journal of Non-Linear Mechanics*, *71*, 48–51.
9. Ghiba, I. D., Neff, P., & Martin, R. J. (2015). An ellipticity domain for the distortional Hencky-logarithmic strain energy. In *Proceedings of the Royal Society of London A, 471*(2184). doi:10.1098/rspa.2015.0510.
10. Grandi, D., & Stefanelli, U. (2015). Finite plasticity in $P^T P$. *Preprint*arXiv:1509.08681
11. Gupta, A., Steigmann, D. J., & Stölken, J. S. (2011). Aspects of the phenomenological theory of elastic-plastic deformation. *Journal of Elasticity*, *104*(1–2), 249–266.
12. Helm, D. (2001). *Formgedächtnislegierungen: experimentelle Untersuchung, phänomenologische Modellierung und numerische Simulation der thermomechanischen Materialeigenschaften*. Ph.D-Thesis: Universität Kassel.
13. Kröner, E. (1955). Der fundamentale Zusammenhang zwischen Versetzungsdichte und Spannungsfunktionen. *Zeitschrift fur Angewandte Mathematik und Physik*, *142*(4), 463–475.
14. Kröner, E. (1958). *Kontinuumstheorie der Versetzungen und Eigenspannungen*. Berlin: Springer.
15. Kröner, E. (1959). Allgemeine Kontinuumstheorie der Versetzungen und Eigenspannungen. *Archive for Rational Mechanics and Analysis*, *4*(1), 273–334.
16. Lee, E. H. (1969). Elastic-plastic deformation at finite strains. *Journal of Applied Mechanics*, *36*(1), 1–6.
17. Lion, A. (1997). A physically based method to represent the thermo-mechanical behaviour of elastomers. *Acta Mechanica*, *123*(1–4), 1–25.
18. Mainik, A., & Mielke, A. (2005). Existence results for energetic models for rate-independent systems. *Calculus of Variations and Partial Differential Equations*, *22*(1), 73–99.
19. Mainik, A., & Mielke, A. (2009). Global existence for rate-independent gradient plasticity at finite strain. *Journal of nonlinear science*, *19*(3), 221–248.
20. Maugin, G. (1994). Eshelby stress in elastoplasticity and ductile fracture. *International Journal of Plasticity*, *10*(4), 393–408.
21. Miehe, C. (1992). *Kanonische Modelle multiplikativer Elasto-Plastizität*. Thermodynamische Formulierung und numerische Implementation. Habilitationsschrift: Universität Hannover, Germany.
22. Miehe, C. (1995). A theory of large-strain isotropic thermoplasticity based on metric transformation tensors. *Archive of Applied Mechanics*, *66*, 45–64.
23. Miehe, C. (1998). A constitutive frame of elastoplasticity at large strains based on the notion of a plastic metric. *International Journal of Solids and Structures*, *35*(30), 3859–3897.
24. Mielke, A. (2003). Energetic formulation of multiplicative elasto-plasticity using dissipation distances. *Continuum Mechanics and Thermodynamics*, *15*(4), 351–382.
25. Mielke, A. (2004). Existence of minimizers in incremental elasto-plasticity with finite strains. *SIAM Journal on Mathematical Analysis*, *36*, 384–404.
26. Mielke, A., & Müller, S. (2006). Lower semicontinuity and existence of minimizers in incremental finite-strain elastoplasticity. *ZAMM - Journal of Applied Mathematics and Mechanics*, *86*, 233–250.
27. Montella, G., Govindjee, S., & Neff, P. (2015) The exponentiated Hencky strain energy in modelling tire derived material for moderately large deformations. *To appear in Journal of Engineering Materials and Technology-Transactions of the ASME*. Preprint arXiv:1509.06541.
28. Neff, P., Chełmiński, K., & Alber, H. D. (2009). Notes on strain gradient plasticity: finite strain covariant modelling and global existence in the infinitesimal rate-independent case. *Mathematical Models and Methods in Applied Sciences*, *19*, 307–346.
29. Neff, P., & Ghiba, I. D. (2014). Loss of ellipticity for non-coaxial plastic deformations in additive logarithmic finite strain plasticity. *International Journal of Non-Linear Mechanics, 81*, 122–128. Preprint arXiv:1410.2819.

30. Neff, P., & Ghiba, I. D. (2015). The exponentiated Hencky-logarithmic strain energy. Part III: Coupling with idealized isotropic finite strain plasticity. *Continuum Mechanics and Thermodynamics, 28*, 477–487. doi:10.1007/s00161-015-0449-y, the special issue in honour of D. J. Steigmann.
31. Neff, P., & Knees, D. (2008). Regularity up to the boundary for nonlinear elliptic systems arising in time-incremental infinitesimal elasto-plasticity. *SIAM Journal on Mathematical Analysis*, *40*(1), 21–43.
32. Neff, P., Ghiba, I. D., & Lankeit, J. (2015). The exponentiated Hencky-logarithmic strain energy. Part I. *Journal of Elasticity*, *121*, 143–234.
33. Neff, P., Ghiba, I. D., Lankeit, J., Martin, R., & Steigmann, D. J. (2015). The exponentiated Hencky-logarithmic strain energy. Part II: Coercivity, planar polyconvexity and existence of minimizers. *Zeitschrift fur Angewandte Mathematik und Physik*, *66*, 1671–1693.
34. Ortiz, M., & Simo, J. C. (1986). An analysis of a new class of integration algorithms for elastoplastic constitutive relations. *International Journal for Numerical Methods in Engineering*, *23*(3), 353–366.
35. Reese, S., & Christ, D. (2008). Finite deformation pseudo-elasticity of shape memory alloys-Constitutive modelling and finite element implementation. *International Journal of Plasticity*, *24*(3), 455–482.
36. Reese, S., & Wriggers, P. (1997). A material model for rubber-like polymers exhibiting plastic deformation: computational aspects and a comparison with experimental results. *Computer Methods in Applied Mechanics and Engineering*, *148*, 279–298.
37. Shutov, A.V. (2014). Personal comunication. 8/2014.
38. Shutov, A. V., & Ihlemann, J. (2014). Analysis of some basic approaches to finite strain elastoplasticity in view of reference change. *International Journal of Plasticity*, *63*, 183–197.
39. Shutov, A. V., & Kreißig, R. (2008). Finite strain viscoplasticity with nonlinear kinematic hardening: Phenomenological modeling and time integration. *Computer Methods in Applied Mechanics and Engineering*, *197*(21), 2015–2029.
40. Shutov, A. V., Landgraf, R., & Ihlemann, J. (2013). An explicit solution for implicit time stepping in multiplicative finite strain viscoelasticity. *Computer Methods in Applied Mechanics and Engineering*, *265*, 213–225.
41. Simo, J.C. (1993). Recent developments in the numerical analysis of plasticity. In E. Stein (ed.), *Progress in computational analysis of inelastic structures* (pp. 115–173). Springer.
42. Simo, J.C., & Hughes, J.R. (1998). *Computational Inelasticity.*, volume 7 of *Interdisciplinary Applied Mathematics*. Springer, Berlin.
43. Simo, J. C., & Ortiz, M. (1985). A unified approach to finite deformation elastoplastic analysis based on the use of hyperelastic constitutive equations. *Computer Methods in Applied Mechanics and Engineering*, *49*, 221–245.
44. Steigmann, D. J., & Gupta, A. (2011). Mechanically equivalent elastic-plastic deformations and the problem of plastic spin. *Theoretical and Applied Mechanics*, *38*(4), 397–417.
45. Vladimirov, I., Pietryga, M., & Reese, S. (2008). On the modelling of non-linear kinematic hardening at finite strains with application to springback-comparison of time integration algorithms. *International Journal for Numerical Methods in Engineering*, *75*(1), 1–28.

Quasi-Static Evolutions in Brittle Fracture Generated by Gradient Flows: Sharp Crack and Phase-Field Approaches

Matteo Negri

Abstract In this paper we will describe how gradient flows, in a suitable norm, are natural and helpful to generate quasi-static evolutions in brittle fracture. First, we will consider the case of a brittle crack running along a straight line according to Griffith's law. Then, we will see how the same approach leads to quasi-static evolutions in the phase field setting, taking into account the alternate minimization scheme. In the latter, the norm associated to the gradient flow is not "user supplied", however, the algorithm itself together with the separate quadratic structure of the energy defines a family of norms which, in the limit, characterizes the quasi-static evolution. Mathematically speaking, all of these evolutions are (parametrized) BV-evolutions.

1 Introduction

The idea of using monotone descent paths (among which the gradient flow) for quasi-static crack propagation goes back to the foundation of fracture mechanics: according to Griffith's principle [12] "the system can pass from the unbroken to the broken condition by a process involving a continuous decrease of potential energy". Choosing the gradient flow, in a suitable norm, as optimal and most common descent path, Griffith's criterion would be: "the system follows the gradient flow of the potential energy".

As a matter of fact, gradient flows usually refer to time dependent problems while Griffith's law does not make any reference to time, neither directly or indirectly (e.g. in terms of velocities). In our rate-independent setting, the gradient flows will provide in fact a *parametrization* of the path connecting "the unbroken" with "the broken condition". Such a parametrization will appear both in the construction of the solution, by time discretization, and in the quasi-static evolution itself, specifically in the instantaneous "catastrophic" propagations.

M. Negri (✉)
Department of Mathematics, University of Pavia, Via Ferrata 1, 27100 Pavia, Italy
e-mail: matteo.negri@unipv.it

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_9

Our construction of quasi-static propagations follows closely a well known scheme in computational mechanics: we employ a uniform time discretization, say $t_k = k\Delta t$, together with an incremental law based on Griffith's principle: at each discrete time t_k the crack advances along a decreasing path of the energy, if any, until it reaches a stationary point of the energy; decreasing paths will then be written as suitable gradient flows.

In particular, for a straight crack this scheme provides in the limit, as $\Delta t \to 0$, a quasi-static evolution which can be described rigorously in several equivalent ways: by means of Karush-Kuhn-Tucker (KKT) conditions (cf. Theorem 1 and [23]) by BV-solutions [18] (cf. Corollary 1 and [19]) or by parametrized BV-solutions [9] (cf. Corollary 2 and [22]). At this point it is important to remark that in the limit, as $\Delta t \to 0$, the quasi-static evolution can be discontinuous in time. This is a common feature of BV-solutions for rate-independent systems and, most important, it is not a pure mathematical artefact. In the rate independent setting discontinuities represent catastrophic propagations of the crack, which can happen in real life: a numerical example (cf. Sect. 2) shows a clear jump discontinuity in a standard ASTM compact tension test. Mathematically, jump discontinuities are characterized by unstable regimes of propagation where instantaneously the crack advances following a gradient flow, which is indeed "a process involving a continuous decrease of potential energy".

For the general situation in which the crack path is unknown, both mathematical and numerical models involving geometrical and topological features of the crack become sensibly harder. Facing these problems is challenging but in practice it is more convenient to employ regularized models [1, 5, 7, 13, 16, 17, 24, 26] which bypass the issues related to the morphology of the crack. One of the most successful choice is the phase field approach, which has been implemented in different ways and for several problems in fracture. Here we will focus our interest on the evolution obtained with a very efficient numerical method, known as *alternate minimization* [5]. In this scheme at each time t_k the evolution is obtained by a sequence of (quadratic) minimization problems, which produces a monotone decreasing path of the energy, in agreement with Griffith's criterion. Formally this scheme "defines" a discrete evolution law for the crack (represented by the phase-field variable). Our goal, in analogy with the straight crack problem, is to characterize the limit evolution obtained by letting $\Delta t \to 0$ and to show its main properties. First, we will see that in the limit we get a quasi-static BV-evolution, which in general does not coincide with the evolution obtained by global minimization problems [10]. Then, we recast the evolution by KKT conditions where it appears an energy release rate; in this respect, note that the alternate minimization algorithm does not employ explicitly any kind of energy release. Finally, we show that the irreversibility constraint is thermodynamically consistent and that the evolution of the displacement field follows a sort of visco-elastic flow in the jumps.

Mathematically, in order to characterize the limit it is fundamental to recast alternate minimization as a gradient flow, with respect to a suitable family of norms, induced by the separately quadratic structure of the phase-field energy; clearly this is a particular choice, which works extremely well, but other choices are also possible and worth studying, e.g. [22]. In this work, proofs and fine mathematical details are not included; the interested reader can make reference for instance to [14, 19, 22, 23].

2 Sharp Crack

2.1 Setting: Compact Tension

In order to avoid technical issues as much as possible, we will state our results only for a representative example, cf. Fig. 1. Denote by Ω the open set in Fig. 1 (obtained removing a couple of symmetric holes from a rectangle). Let $\partial\Omega = \partial_D\Omega \cup \partial_N\Omega$ where $\partial_N\Omega$ is the boundary of the rectangle while $\partial_D\Omega$ denotes (the union of) the boundaries of the circular holes. Assume that the initial crack K_0 is given by the line segment $(0, l_0] \times \{0\}$ for $l_0 > 0$. In our simple setting the crack will propagate horizontally, thus our family of admissible cracks will be given by the line segments of the form $K_l = (0, l] \times \{0\}$. Clearly such a family is simply parametrized by the scalar $l \in [l_0, L)$, which gives as well the position of the crack tip.

Consider on $\partial_D\Omega$ a proportional boundary condition of the form $u = \pm t\hat{e}$ where the sign ± is chosen as in Fig. 1.

We consider in-plane elasticity with linearised energy density

$$W(Du) = \tfrac{1}{2}\, Du : \mathbf{C}[Du] = \tfrac{1}{2}\, \boldsymbol{\varepsilon}(u) : \boldsymbol{\sigma}(u)$$

where $\boldsymbol{\varepsilon}(u) = (Du + Du^T)/2$ and $\mathbf{C}[Du] = \boldsymbol{\sigma}(u) = 2\mu\boldsymbol{\varepsilon}(u) + \lambda \mathrm{tr}(\boldsymbol{\varepsilon}(u))I$, for $\lambda, \mu > 0$ the Lamé coefficients. For $t \in [0, T]$ and $l \in [l_0, L)$ the space of admissible configurations is

$$\mathscr{U}_{t,l} = \{u \in H^1(\Omega\backslash K_l, \mathbf{R}^2) : u = \pm t\hat{e}\ \partial_D\Omega\}\,.$$

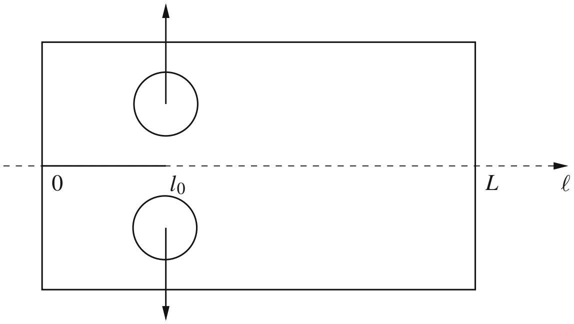


Fig. 1 An ASTM-compact tension geometry: the set $\Omega\backslash K_0$

Hence, for $u \in \mathscr{U}_{t,l}$ the elastic energy will be

$$E(u) = \int_{\Omega \setminus K_l} W(Du)\, dx\,.$$

Before proceeding it is convenient to introduce the *reduced elastic energy*: for $t \in [0, T]$ and $l \in [l_0, L)$ let

$$\mathscr{E}(t,l) = E(u_{t,l}),$$

where $u_{t,l} \in \operatorname{argmin}\{E(u) : u \in \mathscr{U}_{t,l}\}$. Note that in quasi-static evolutions it is not restrictive to employ $\mathscr{E}$ instead of E since it is assumed that the system is always in equilibrium and $u_{t,l}$ is indeed the only equilibrium point; in particular it solves the PDE

$$\begin{cases} \operatorname{div}(\boldsymbol{\sigma}(u_{t,l})) = 0 & \Omega \setminus K_l \\ u_{t,l} = \pm t\hat{e} & \partial_D \Omega \\ \boldsymbol{\sigma}(u_{t,l})\,\hat{n} = 0 & \partial_N \Omega \cup K_l^{\pm}. \end{cases} \tag{1}$$

Note that the Neumann homogeneous boundary condition holds on the boundary $\partial_N \Omega$ of the rectangle and on both the crack faces, above denoted by $K_l^{\pm}$.

Let us now turn to dissipation. Since we are interested in brittle fracture the energy dissipated by the crack will be provided by a potential $\mathscr{K} : [l_0, L) \to \mathbf{R}^+$ which is simply of the form $\mathscr{K}(l) = G_c(l - l_0)$, being $G_c > 0$ the material toughness.

In the sequel we will always work with the *reduced total energy* $\mathscr{F} : [0, T] \times [l_0, L) \to \mathbf{R}^+$ given by

$$\mathscr{F}(t,l) = \mathscr{E}(t,l) + \mathscr{K}(l).$$

Before proceeding it is fundamental to have at our disposal the partial derivatives of the energy $\mathscr{F}$.

Lemma 1 *The energy $\mathscr{F} : [0, T] \times [l_0, L) \to \mathbf{R}^+$ is differentiable with respect to both its variables with*

$$\partial_t \mathscr{F}(t,l) = \partial_t \mathscr{E}(t,l) = \int_{\partial_D \Omega} (\pm\hat{e}) \cdot \boldsymbol{\sigma}(u_{t,l})\,\hat{n}\, ds = \mathscr{P}^{ext}(t,l),$$
$$\partial_l \mathscr{F}(t,l) = -G(t,l) + G_c\,,$$

where $\mathscr{P}^{ext}$ is the power of the external forces while G denotes as usual the energy release rate. Moreover, $G(t,\cdot)$ is non-negative and locally Lipschitz continuous in $[l_0, L)$.

A proof can be adapted e.g. from [19] or [23]. In this setting, by irreversibility, equilibrium reads

$$\partial_l \mathscr{F}(t,l) = -G(t,l) + G_c \geq 0 \quad \Leftrightarrow \quad G(t,l) \leq G_c. \tag{2}$$

2.2 Discrete in Time Evolution

Now, we will define the discrete in time evolution by a sequence of incremental problems. Denote by ℓ the evolution in time. Given $\Delta t > 0$ let $t_k = k\Delta t$ for $k = 0, ..., [T/\Delta t]$ and let $\ell(t_0) = l_0$. Knowing $\ell(t_k)$ we define $\ell(t_{k+1})$ as

$$\ell(t_{k+1}) = \min\{l \geq \ell(t_k) : G(t_{k+1}, l) \leq G_c\}. \tag{3}$$

In other terms, we advance the crack up to the closest equilibrium point. This is in some sense a *return mapping* algorithm on the set $\{G(t_{k+1}, l) \leq G_c\}$ of equilibrium points at time t_{k+1}, which is usually at the core of many crack tracking algorithms.

Now, let us see how to recast the incremental problem as a gradient flow. By irreversibility the crack cannot heal, for this reason it is convenient to introduce the one sided "slope"

$$|\partial_l \mathscr{F}(t, l)|^-$$

where $|\cdot|^-$ denotes the negative part. Next, let us introduce an auxiliary parameter $s \in \mathbf{R}^+$ and an auxiliary function $l : \mathbf{R}^+ \to [l_0, L)$. We set $\ell(t_{k+1}) = \sup_s l(s)$ where l solves the gradient flow

$$\begin{cases} \dot{l}(s) = |\partial_l \mathscr{F}(t_{k+1}, l(s))|^- \\ l(0) = \ell(t_k). \end{cases}$$

In this simple setting the gradient flow boils down to an autonomous Cauchy problem for a non-linear, first order ODE. Intuitively, l grows when $\partial_l \mathscr{F}(t_{k+1}, l) < 0$, i.e. when $G(t_{k+1}, l) > G_c$ and thus when the crack is not in equilibrium. It is easy to see that there exists a unique solution and that the definition $\ell(t_{k+1}) = \sup_s l(s)$ coincides with (3) (for a proof, see [19]).

At this point we have defined $\ell(t_k)$ for $t_k = k\Delta t$. Now consider a sequence of time steps $\Delta t_n \searrow 0$ and denote by ℓ_n the corresponding discrete evolutions, defined in the discrete points $t_{n,k} = k\Delta t_n$. Denote again by $\ell_n : [0, T] \to [l_0, L)$ the piecewise affine interpolate of $\ell_n(t_{n,k})$. By Helly's Theorem it follows that (up to subsequences) ℓ_n converges pointwise to a limit evolution ℓ. Our goal is now the characterization of ℓ: we will provide two characterizations, the first in terms of KKT conditions, the second in terms of parametrized BV-evolutions. In order to better understand the meaning of these characterizations it is useful to show first an explicit example, which has been computed numerically.

2.3 Example

First, let us comment on Fig. 2. Remember that the "loading" is monotone increasing. The set of *critical points* of the energy, i.e. $\{(t, l) : G(t, l) = G_c\}$ is represented with a dotted curve. This curve splits the (t, l)-plane into two regions: on the left is the set

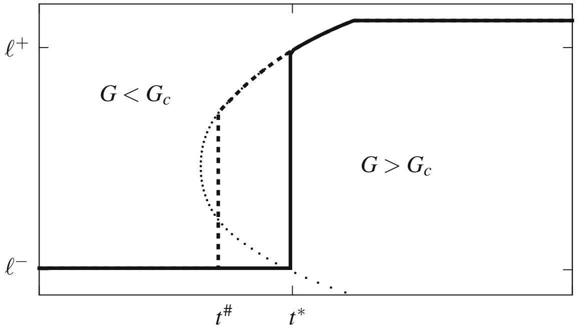


Fig. 2 Quasi-static evolutions: of energetic type (*dashed*) and of BV-type (*solid*)

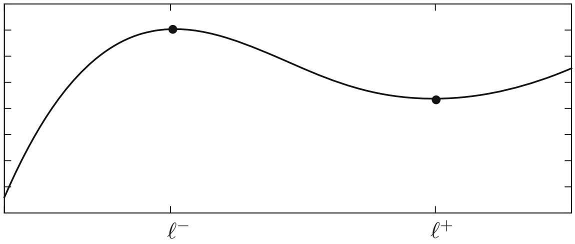


Fig. 3 A detail of the energy landscape $\mathscr{F}(t^*, l)$

$\{(t, l) : G(t, l) < G_c\}$ of *stable points* while on the right is the set $\{(t, l) : G(t, l) > G_c\}$ of *unstable points*. From the picture it is clear that up to time t^* the crack is not moving, since it is inside the *stable region*. At time t^* a "catastrophic evolution" occurs: in the instantaneous transition from ℓ^- to ℓ^+ the system crosses the unstable region, since $G(t^*, l) > G_c$ for every $l \in (\ell^-, \ell^+)$. The energy landscape at time t^* is reported in Fig. 3: it is clear that $\mathscr{F}(t^*, \cdot)$ is not convex and that ℓ makes a transition from ℓ^- to ℓ^+ following a descent path, in particular $\mathscr{F}(t^*, \ell^-) > \mathscr{F}(t^*, \ell^+)$.

Figure 2 shows (in dashed bold line) also the energetic evolution, obtained by global energy minimization; this evolution presents a discontinuity as well, however the qualitative behaviour is quite different: the system crosses first the *stable* and then the *unstable* region, in particular a propagation occurs even if $G(t^\#, \ell^-) < G_c$.

2.4 Characterization by Karush-Kuhn-Tucker Conditions

The next Theorem provides the first characterization of the limit evolution. For a proof, see [19].

Theorem 1 *The limit evolution ℓ, obtained letting $\Delta t_n \searrow 0$, is non-decreasing and belongs to $BV(0, T)$. Moreover*

$$G(t, \ell^-(t)) \leq G_c \quad \textit{for } t \in [0, T], \tag{4}$$

$$\big(G(t, \ell^-(t)) - G_c\big)\, d\ell(t) = 0 \quad \textit{in the sense of measures in } [0, T]. \tag{5}$$

Furthermore, for $t \in J(\ell)$ (the set of jumps) we have

$$G(t,l) \geq G_c \quad \textit{for every } l \in [\ell^-(t), \ell^+(t)]. \tag{6}$$

Let us make some comments on the above Theorem. Clearly (4) and (5) play the role of the "classical" KKT conditions, with some technical differences: the left limit ℓ^- is used instead of ℓ and the weak measure theoretic derivative $d\ell$ is used instead of the speed $\dot{\ell}$. Both these technical details are due to the fact that in general the evolution belongs to $BV(0,T)$ and thus it may have jump discontinuities. However the qualitative meaning of (4) and (5) is quite clear and consistent with standard KKT conditions. What is instead not common in the study of quasi-static evolutions is condition (6) which characterizes the behaviour in the jumps in terms of *unstable* branches of propagation (cf. also Fig. 2).

Finally, in terms of derivatives of the energy, the above Theorem reads as follows.

Corollary 1 *The limit ℓ, obtained letting $\Delta t_n \searrow 0$, is non-decreasing and belongs to $BV(0,T)$. Moreover*

$$\partial_l \mathscr{F}(t, \ell^-(t)) \geq 0 \quad \textit{for } t \in [0,T]\,, \tag{7}$$

$$\partial_l \mathscr{F}(t, \ell^-(t))\, d\ell(t) = 0 \quad \textit{in the sense of measures in } [0,T]. \tag{8}$$

Furthermore, for $t \in J(\ell)$ (the set of jumps) we have

$$\partial_l \mathscr{F}(t, \ell^-(t)) \leq 0 \quad \textit{for every } l \in [\ell^-(t), \ell^+(t)]. \tag{9}$$

At this point it is necessary to introduce the concept of BV-solution. Here we will give the "simplest" possible definition, for a general treatise see [18].

Corollary 2 *The limit ℓ is non-decreasing and of class $BV(0,T)$. Moreover*

$$|\partial_l \mathscr{F}(t, \ell^-(t))|^- = 0 \quad \textit{for } t \in [0,T] \tag{10}$$

and for every $t \in [0,T]$ the following energy identity holds:

$$\mathscr{F}(t, \ell^-(t)) = \mathscr{F}(0, l_0) + \int_0^t \partial_t \mathscr{F}(\tau, \ell(\tau))\, d\tau - \sum_{t \in J(\ell)} \operatorname{diss}(\mathscr{F}(t,\cdot)), \tag{11}$$

where

$$\operatorname{diss}(\mathscr{F}(t,\cdot)) = \int_{\ell^-(t)}^{\ell^+(t)} |\partial_l \mathscr{F}(t,l)|^-\, dl$$

denotes the "energy gap" in the discontinuity points.

An evolution ℓ which satisfies (10) and (11) is called a BV-solution.

It is important to note that equilibrium (10) and energy balance (11) provide a very concise and mathematically convenient way of characterizing the quasi-static evolution; they are indeed equivalent to the KKT conditions of Theorem 1. Without entering too much into the technical details (for a complete proof the reader can follow [20] or [22]) let us see how (7)–(9) follow from (10)–(11). First,

$$|\partial_l \mathscr{F}(t, \ell^-(t))|^- = 0 \;\Leftrightarrow\; \partial_l \mathscr{F}(t, \ell^-(t)) \geq 0\,,$$

gives (7). Next, by the chain rule in $BV(0,T)$

$$\begin{aligned}\mathscr{F}(t, \ell^-(t)) = \mathscr{F}(0, l_0) + \int_0^t \partial_t \mathscr{F}(\tau, \ell(\tau))\, d\tau \;+\; \int_0^t \partial_l \mathscr{F}(\tau, \ell^-(\tau))\, d_{ac}\ell(\tau) \;+ \\ + \sum_{t \in J(\ell)} [\![\mathscr{F}(t, \cdot)]\!],\end{aligned}$$

where $d_{ac}\ell$ denotes the (weak) measure theoretic derivative of ℓ in $[0, T] \setminus J(\ell)$. Comparing with (11) we get

$$\int_0^t \partial_l \mathscr{F}(\tau, \ell^-(\tau))\, d_{ac}\ell(\tau) + \sum_{t \in J(\ell)} [\![\mathscr{F}(t, \cdot)]\!] = -\sum_{t \in J(\ell)} \operatorname{diss}(\mathscr{F}(t, \cdot)).$$

Since the measure $d_{ac}\ell$ is supported in $[0, T] \setminus J(\ell)$ it follows that

$$\partial_l \mathscr{F}(t, \ell^-(t))\, d_{ac}\ell(t) = 0 \quad \text{in the sense of measures in } [0, T]$$

and that

$$[\![\mathscr{F}(t, \cdot)]\!] = \int_{\ell^-(t)}^{\ell^+(t)} \partial_l \mathscr{F}(t, l)\, dl = -\int_{\ell^-(t)}^{\ell^+(t)} |\partial_l \mathscr{F}(t, l)|^-\, dl, \quad \text{for every } t \in J(\ell).$$

The former leads to (8), thanks to the continuity of G, while the latter leads to

$$\partial_l \mathscr{F}(t, l) = -|\partial_l \mathscr{F}(t, l)|^- \quad \text{for every } l \in [\ell^-(t), \ell^+(t)],$$

which is in turn equivalent to

$$\partial_l \mathscr{F}(t, l) \leq 0 \quad \text{for every } l \in [\ell^-(t), \ell^+(t)],$$

that is (9).

2.5 Characterization as a Graph Parametrized BV-evolution

We have seen that the limit evolution ℓ can have jump discontinuities in time. For this reason it is convenient, both for theoretical and numerical purposes, to represent ℓ

by a (Lipschitz) parametrization of the form $s \mapsto (t(s), l(s))$ of the extended graph. Remember that the extended graph is just the set $\{(t, l) : \ell^-(t) \le l \le \ell^+(t)\}$ obtained "completing" the jumps with a vertical line segments (see Fig. 2). Therefore, using the map $s \mapsto (t(s), l(s))$ a jump of ℓ at time t^* will be characterized by $t(s) = t^*$ in $[s_1, s_2]$ together with $l(s_1) = \ell^-(t^*)$ and $l(s_2) = \ell^+(t^*)$. In essence, continuity points (in time) will correspond to points with $t'(s) > 0$ while discontinuity points (in time) will correspond to points with $t'(s) = 0$. The advantages of this representation, originally suggested in [9], will become more clear in Sect. 3 and in general are quite evident in the case of infinite dimensional systems [22].

Here for sake of simplicity we will skip any argument on the different ways which provide existence of a parametrized evolution. We will instead assume that $s \mapsto (t(s), l(s))$ is a parametrization of the extended graph of the solution ℓ, obtained as above by letting $\Delta t_n \searrow 0$. We will consider also $t' \ge 0$ (in order to avoid physically meaningless cases) and we will normalize the parametrization with $t'(s) + l'(s) = 1$ (for a.e. $s \in [0, S]$). The resulting parametrization satisfies the following properties.

Theorem 2 *The (normalized) parametrization $s \mapsto (t(s), l(s))$ satisfies $t'(s) \ge 0$, $l'(s) \ge 0$ and $t'(s) + l'(s) \le 1$. Moreover it satisfies the following conditions:*

$$|\partial_l \mathscr{F}(t(s), l(s))|^- = 0 \qquad \textit{for every } s \textit{ with } t'(s) > 0, \tag{12}$$

$$\begin{aligned} \mathscr{F}(t(s), l(s)) = \ & \mathscr{F}(0, l(0)) + \int_0^s \partial_t \mathscr{F}(t(r), l(r))\, t'(r)\, dr \, + \\ & - \int_0^s |\partial_\ell \mathscr{F}(t(r), l(r))|^- \, l'(r)\, dr \quad \textit{for every } s. \end{aligned} \tag{13}$$

A (normalized) parametrization $s \mapsto (t(s), \ell(s))$ which satisfies (10) *and* (11) *is called a parametrized BV-solution.*

It is not difficult to see that, upon choosing the right parametrization, (12)–(13) follows from (10)–(11) and vice versa (for a proof see [22]).

3 Alternate Minimization Scheme in the Phase-Field Approach

In this section we will deal with the phase-field approach for fracture, which in the last decade has been an effective and popular method for the simulation of crack propagation, see e.g. [1, 5, 7, 13, 16, 17] and many others. In particular we will consider evolutions defined by the alternate minimization scheme; we will see that in the limit they will provide indeed parametrized BV-evolution, with respect to a suitable family of norms.

3.1 Setting

For sake of simplicity let us consider the same geometry and the same boundary conditions of Sect. 2.1. In the phase-field framework it is however necessary to redefine the space of admissible displacements by

$$\mathscr{U}_t = \{u \in H^1(\Omega, \mathbf{R}^2) : u = \pm t\hat{e} \text{ on } \partial_D \Omega\}$$

and to introduce a set for the phase field variables

$$\mathscr{Z} = \{z \in H^1(\Omega) : 0 \le z \le 1\}.$$

By linearity we can always write $\mathscr{U}_t = t\mathscr{U}$ where

$$\mathscr{U} = \{u \in H^1(\Omega, \mathbf{R}^2) : u = \pm \hat{e} \text{ on } \partial_D \Omega\}.$$

In the sequel we will work indeed with the spaces $\mathscr{U}$ and $\mathscr{Z}$ which are independent of time. For $\varepsilon > 0$ and $\eta_\varepsilon > 0$, typically with $\eta_\varepsilon = o(\varepsilon)$, the phase field elastic and dissipated energy [2] will be respectively

$$\mathscr{E}_\varepsilon(t, u, z) = \tfrac{1}{2} \int_\Omega t^2 (z^2 + \eta_\varepsilon) W(Du)\, dx,$$
$$\mathscr{K}_\varepsilon(z) = \tfrac{1}{2} G_c \int_\Omega (z-1)^2/2\varepsilon + \varepsilon |\nabla z|^2\, dx.$$

The total energy $\mathscr{F}_\varepsilon : [0, T] \times \mathscr{U} \times \mathscr{Z} \to \mathbb{R}$ will be $\mathscr{F}_\varepsilon(t, u, z) = \mathscr{E}_\varepsilon(t, u, z) + \mathscr{K}_\varepsilon(z)$.

In the sequel it will be fundamental to have at hand the partial derivatives of the energy,

$$\partial_t \mathscr{F}_\varepsilon(t, u, z) = \int_\Omega t (z^2 + \eta_\varepsilon) W(Du)\, dx, \tag{14}$$

$$\partial_u \mathscr{F}_\varepsilon(t, u, z)[\phi] = \int_\Omega t^2 (z^2 + \eta_\varepsilon) \boldsymbol{\sigma}(u) : \boldsymbol{\varepsilon}(\phi)\, dx, \tag{15}$$

$$\partial_z \mathscr{F}_\varepsilon(t, u, z)\,[\xi] = \int_\Omega t^2 z \xi \; W(Du)\, dx + G_c \int_\Omega (z-1)\xi/2\varepsilon + \varepsilon \nabla z \cdot \nabla \xi\, dx. \tag{16}$$

For our purposes the spaces of admissible variations for $\mathscr{Z}$ and $\mathscr{U}$ will be provided respectively by

$$\Xi = \{\xi \in H^1(\Omega) : \xi \le 0\}, \qquad \Phi = \{\phi \in H^1(\Omega, \mathbf{R}^2) : \phi = 0 \text{ on } \partial_D \Omega\}.$$

To conclude this section, let us see how to define a notion of energy release, with respect to a variation ξ, and how to get the power of external forces. To this end, denoting $u(t,z) \in \operatorname{argmin}\{\mathscr{E}_\varepsilon(t,u,z) : u \in \mathscr{U}\}$ we will call "energy release functional" (with respect to a variation ξ)

$$\mathscr{G}_\varepsilon(t,z)[\xi] = -\lim_{h\to 0^+} \frac{\mathscr{E}_\varepsilon(t,u(t,z+h\xi),z+h\xi) - \mathscr{E}_\varepsilon(t,u(t,z),z)}{h}$$

In this way the displacement field changes "simultaneously" with the variation of the phase field variable. Actually, by minimality of $u(t,z)$, the derivative can be represented explicitly as (see e.g. [15])

$$\mathscr{G}_\varepsilon(t,z)[\xi] = -\partial_z \mathscr{E}_\varepsilon(t,u,z)\,[\xi] = -\int_\Omega t^2 z\xi\, W(Du)\, dx\,. \tag{17}$$

Now, let us introduce the set of normalized variations

$$\hat{\Xi}_z = \left\{\xi \in \Xi : \int_\Omega (z-1)\xi/4\varepsilon + \varepsilon\nabla z\cdot\nabla\xi\, dx \le 1\right\}.$$

Note that these variations normalize the "variation of crack length" since

$$d\mathscr{K}_\varepsilon(z)[\hat{\xi}] = G_c \int_\Omega (z-1)\hat{\xi}/4\varepsilon + \varepsilon\nabla z\cdot\nabla\hat{\xi}\, dx \le G_c\,.$$

Then, we can define the energy release as

$$G_\varepsilon(t,z) = \sup\{\mathscr{G}_\varepsilon(t,z)[\hat{\xi}] : \hat{\xi} \in \hat{\Xi}_z\}. \tag{18}$$

Finally, Green's formula allows to rewrite (14) as

$$\partial_t \mathscr{F}_\varepsilon(t,u(t,z),z) = \int_{\partial_D\Omega} (\pm\hat{e})\cdot\boldsymbol{\sigma}_z(tu(t,z))\,\hat{n}\, ds = \mathscr{P}^{ext}_\varepsilon(t,u(t,z),z), \tag{19}$$

where $\boldsymbol{\sigma}_z(w)$ denotes the phase field stress, that is

$$\boldsymbol{\sigma}_z(w) = (z^2+\eta_\varepsilon)\boldsymbol{\sigma}(w).$$

3.2 Discrete in Time Evolution by Alternate Minimization

As we did in Sect. 2.2, given $\Delta t > 0$, let $t_k = k\Delta t$ and set the initial conditions $u(t_0) = u_0$ and $z(t_0) = z_0$. Known $u(t_{k-1})$ and $z(t_{k-1})$ we will introduce a couple of auxiliary sequences, u^m and z^m, with $u^0 = u(t_{k-1})$ and $z^0 = z(t_{k-1})$ defined recursively by the following *alternate minimization scheme* [5]

$$\begin{cases} u^m \in \operatorname{argmin}\big\{\mathscr{F}_\varepsilon\big(t_k\,,\cdot\,,z^{m-1}\big) : u \in \mathscr{U}\big\}, \\ z^m \in \operatorname{argmin}\big\{\mathscr{F}_\varepsilon(t_k\,,u^m\,,\cdot) : z \in \mathscr{Z} \text{ with } z \le z^{m-1}\}, \end{cases} \tag{20}$$

where the constraint $z \le z^{m-1}$ models the irreversibility of the crack. Then we define the updates $u(t_k) = \lim_{m\to+\infty} u^m$ and $z(t_k) = \lim_{m\to+\infty} z^m$. More precisely, we have the following result (for a proof see [14]).

Proposition 1 *The sequence u^m converges to $u(t_k)$ strongly in $H^1(\Omega, \mathbf{R}^2)$ while z^m converges to $z(t_k)$ strongly in $H^1(\Omega)$. Further,*

$$\partial_u \mathscr{F}_\varepsilon(t_k, u(t_k), z(t_k))[\phi] = 0 \quad \textit{for every } \phi \in \Phi,$$
$$\partial_z \mathscr{F}_\varepsilon(t_k, u(t_k), z(t_k))[\xi] \ge 0 \quad \textit{for every } \xi \in \Xi.$$

In other terms, $u(t_k)$ is an equilibrium point for $\mathscr{F}_\varepsilon(t_k, \cdot, z(t_k))$ while $z(t_k)$ is an equilibrium point for $\mathscr{F}_\varepsilon(t_k, u(t_k), \cdot)$ (for the latter remember the irreversibility constraint).

In this way, given $\Delta t > 0$ a time discrete evolution is provided in terms of the equilibrium configurations $(u(t_k), z(t_k))$ for $t_k = k\Delta t$. In order to understand the limit evolution, obtained by letting $\Delta t \to 0$, we have first to recast the alternate minimization scheme as a gradient flow. This is the goal of the next section.

3.3 Minimization as a Gradient Flow

3.3.1 An Illustrative Example

In order to better understand the gradient flow structure behind (20) let us start with an example: the minimization of a quadratic functional in a finite dimensional setting. Let $F(x) = \frac{1}{2}x^T Ax + b^T x + c$ for $x \in \mathbb{R}^n$ and $A^T = A > 0$. Let $\|x\|_A = \sqrt{x^T Ax}$ be the "natural" norm induced by the symmetric, positive define matrix A, with associated scalar product $\langle\cdot,\cdot\rangle_A$.

Our problem is the following: given x_0 find the increment x_* in such a way that $x_0 + x_*$ is the minimizer of F. Since F is quadratic we can write

$$F(x_0 + x_*) = F(x_0) + \nabla F(x_0)x_* + \tfrac{1}{2}x_*^T Ax_*$$

and we can characterize x_* by stationarity of the energy, i.e.,

$$\nabla^T F(x_0) + Ax_* = 0 \quad \Leftrightarrow \quad x_* = -A^{-1}\nabla^T F(x_0) = -\nabla^T_A F(x_0).$$

In the last term of the previous row we have introduced the notation $\nabla_A^T F(x_0)$ which denotes the gradient of F (computed in x_0) with respect to the norm $\|\cdot\|_A$, that is the (unique) vector such that $dF(x_0)[x'] = \langle \nabla_A F(x_0), x'\rangle_A$ for every $x' \in \mathbb{R}^n$. For our purposes it is just important to remark that $\nabla_A F(x_0) = \nabla F(x_0) A^{-1}$ and that

$$-\widehat{\nabla}_A F(x_0) = -\nabla_A F(x_0)/\|\nabla_A F(x_0)\|_A = \operatorname{argmin}\big\{dF(x_0)[x'] : \|x'\|_A \le 1\big\}.$$

In other terms, the normalized gradient $\widehat{\nabla}_A F(x_0)$ provides the steepest descent direction with respect to the norm $\|\cdot\|_A$. Now, let $k = \|\nabla_A F(x_0)\|_A$ and define $x(s) = x_0 - \lambda(s)\, \widehat{\nabla}_A F(x_0)$ where $\lambda(s) = 1 - e^{-ks}$ for $s \in [0, +\infty)$. Then, by homogeneity, $x(s)$ solves the gradient flow

$$\begin{cases} x'(s) = -\nabla_A F(x(s)) & \text{for } s \in [0, +\infty), \\ x(0) = x_0. \end{cases} \tag{21}$$

Note that the right hand side is evaluated in $x(s)$, and not in x_0, and that $\lim_{s\to+\infty} x(s) = x_0 + x_*$ is exactly the minimizer. In other terms, the linear interpolation of the points x_0 and $x_0 + x_*$ with parametrization $\lambda(s)$ is the solution of the gradient flow (21).

It is now time to turn back to the phase field setting.

3.3.2 A Family of "Intrinsic Norms" for the Phase Field Energy

As we have seen above, in order to recast minimization as a gradient flow, as a first step it is necessary to single out an "intrinsic norm" which is nothing but the quadratic part of the energy. In our setting, for the phase field variable we will employ

$$\|z\|_{t,u}^2 = \int_\Omega z^2\big(G_c/2\varepsilon + t^2 W(Du)\big) + \varepsilon G_c |\nabla z|^2 \, dx,$$
$$\langle z, \xi\rangle_{t,u} = \int_\Omega z\xi\big(G_c/2\varepsilon + t^2 W(Du)\big) + \varepsilon \nabla z \cdot \nabla \xi \, dx,$$

while for the displacement field u we will employ

$$|u|_{t,z}^2 = \int_\Omega t^2 (z^2 + \eta_\varepsilon) W(Du)\, dx,$$
$$\langle u, \phi\rangle_{t,z} = \int_\Omega t^2 (z^2 + \eta_\varepsilon) \boldsymbol{\sigma}(u) : \boldsymbol{\varepsilon}(\phi)\, dx.$$

With the above definitions the quadratic structure of the energy looks very clear, indeed we can write the energy as

$$\mathscr{F}_\varepsilon(t,u,z) = \tfrac12 |u|^2_{t,z} + c_z \quad \text{for } c_z = \tfrac12 G_c \int_\Omega (z-1)^2/2\varepsilon + \varepsilon|\nabla z|^2\, dx,$$

$$\mathscr{F}_\varepsilon(t,u,z) = \tfrac12 \|z\|^2_{t,u} - b(z) + c_{t,u} \quad \text{for } \begin{cases} b(z) = G_c \displaystyle\int_\Omega z/2\varepsilon\, dx, \\ c_{t,u} = \dfrac12 \displaystyle\int_\Omega \eta_\varepsilon t^2 W(Du) + G_c/2\varepsilon\, dx. \end{cases}$$

Note that c_z is independent of u and vice versa $c_{t,u}$ is independent of z, while $b(\cdot)$ is linear. As a consequence, the partial derivatives as well take a particularly simple form, being

$$\partial_u \mathscr{F}_\varepsilon(t,u,z)[\phi] = \langle u, \phi\rangle_{t,z}, \qquad \partial_z \mathscr{F}_\varepsilon(t,u,z)[\xi] = \langle z, \xi\rangle_{t,u} - b(\xi).$$

We will see in Sect. 3.4.3 how to write more explicitly the gradient flows originating from alternate minimization.

Finally, it will be very convenient, if not necessary, to define a couple of slopes, with respect to the "norms" defined above, that is

$$|\partial_u \mathscr{F}_\varepsilon(t,u,z)|^-_{t,z} = |\min\{\partial_u \mathscr{F}_\varepsilon(t,u,z)[\phi] : \phi \in \Phi,\ |\phi|_{t,z} \le 1\}|^-,$$
$$|\partial_z \mathscr{F}_\varepsilon(t,u,z)|^-_{t,u} = |\min\{\partial_z \mathscr{F}_\varepsilon(t,u,z)[\xi] : \xi \in \Xi,\ \|\xi\|_{t,z} \le 1\}|^-,$$

where $|\cdot|^-$ is once again the negative part.

3.4 A Parametrized "BV-Evolution"

Consider a sequence $\Delta t_n \searrow 0$. For each Δt_n, let $u(t_{n,k})$ and $z(t_{n,k})$ (for $t_k = k\Delta t_n$) be given by the alternate minimization scheme (20). In order to define the limit as $\Delta t_n \searrow 0$, it is natural to introduce, for every time t_k, an arc length interpolation of the alternate minimizing path (u^m, z^m) which links $(u(t_{k-1}), z(t_{k-1})) = (u^0, z^0)$ with $(u(t_k), z(t_k)) = \lim_{m\to+\infty}(u^m, z^m)$. In the end, the interpolation provides a parametrization of the evolution, i.e. a map $(0, S_n) \ni s \mapsto (t_n(s), u_n(s), z_n(s))$ with $t_n(0) = 0$, $u_n(0) = u_0$, $v_n(0) = v_0$ and $t_n(S) = T$. In this respect, it is technically quite delicate to show that the lengths S_n of the interpolating curve are uniformly finite with respect to Δt_n (for the detail the reader should make reference to the forthcoming [14]). We can then apply an abstract results developed in [22] which yields the following Theorem.

Theorem 3 *Given $\Delta t_n \searrow 0$, let $s \mapsto (t_n(s), u_n(s), z_n(s))$ be the parameterizations of the discrete evolutions provided by the alternate minimization scheme* (20)*. Then, up to subsequences, there exists a limit normalized parametrization $s \mapsto (t(s), u(s), z(s))$ with $t'(s) \ge 0$, $z'(s) \le 0$ and $t'(s) + |u'(s)|_{t(s),z(s)} + \|z'(s)\|_{t(s),u(s)} \le 1$. Moreover, for every s with $t'(s) > 0$ the following equilibrium conditions holds*

$$|\partial_u \mathscr{F}_\varepsilon(t(s), u(s), z(s))|^-_{t(s),z(s)} = |\partial_z \mathscr{F}_\varepsilon(t(s), u(s), z(s))|^-_{t(s),z(s)} = 0. \tag{22}$$

Finally, for every s it holds the energy balance

$$\begin{aligned} \mathscr{F}(t(s), u(s), z(s)) = \mathscr{F}(0, u_0, z_0) + \int_0^s \partial_t \mathscr{F}(t(r), u(r), z(r))\, t'(r)\, dr + \\ - \int_0^s |\partial_u \mathscr{F}(t(r), u(r), z(r))|^-_{t(r),z(r)}\, |u'(r)|_{t(r),z(r)}\, dr + \\ - \int_0^s |\partial_v \mathscr{F}(t(r), u(r), z(r))|^-_{t(r),u(r)}\, \|z'(r)\|_{t(r),u(r)}\, dr. \end{aligned} \tag{23}$$

In the next section we will explain better the meaning of the previous Theorem, which is by itself quite technical. However, the analogy with Theorem 2 should be quite evident.

3.4.1 Continuity Points: Equilibrium

Let us start considering the points where $t'(s) > 0$, which correspond in the parametric setting to continuity points in time. Equation (22) gives equilibrium. Indeed, by definition of the slopes

$$|\partial_u \mathscr{F}_\varepsilon(t(s), u(s), z(s))|^-_{t(s),z(s)} = 0 \;\Leftrightarrow\; \partial_u \mathscr{E}_\varepsilon(t(s), u(s), z(s))[\phi] = 0 \quad \text{for } \phi \in \Phi.$$
$$|\partial_z \mathscr{F}_\varepsilon(t(s), u(s), z(s))|^-_{t(s),u(s)} = 0 \;\Leftrightarrow\; \partial_z \mathscr{F}_\varepsilon(t(s), u(s), z(s))[\xi] \geq 0 \quad \text{for } \xi \in \Xi.$$

Let us write more explicitly the equilibrium conditions. Introducing the phase field stress $\boldsymbol{\sigma}_z(u) = (z^2 + \eta_\varepsilon)\, \boldsymbol{\sigma}(u)$ we get

$$\partial_u \mathscr{E}_\varepsilon(t(s), u(s), z(s))[\phi] = \int_\Omega \boldsymbol{\sigma}_{z(s)}(u(s)) : \boldsymbol{\varepsilon}(\phi)\, dx = 0, \qquad \text{for} \phi \in \Phi,$$

and thus

$$\begin{cases} \operatorname{div}(\boldsymbol{\sigma}_{z(s)}(u(s))) = 0 & \Omega \\ u(s) = \pm \hat{e} & \partial_D \Omega \\ \boldsymbol{\sigma}_{z(s)}(u(s))\, \hat{n} = 0 & \partial_N \Omega, \end{cases} \tag{24}$$

which is the phase-field counterpart of (1).

Now, let us see discuss the physical meaning to the equilibrium condition with respect to the phase field variable z. By the definition (17) of energy release the equilibrium condition

$$\partial_z \mathscr{F}_\varepsilon(t(s), u(s), z(s))[\xi] \geq 0 \quad \text{for } \xi \in \Xi$$

reads

$$-\mathscr{G}_\varepsilon(t(s),z(s))[\xi]+\partial_z\mathscr{K}_\varepsilon(z(s))[\xi]\geq 0 \quad \text{for } \xi\in\Xi \tag{25}$$

where

$$\partial_z\mathscr{K}_\varepsilon(z(s))[\xi]=G_c\int_\Omega (z-1)\xi/4\varepsilon+\varepsilon\nabla z\cdot\nabla\xi\,dx\,.$$

Employing the normalized set $\hat{\Xi}_{z(s)}$ and taking the supremum with respect to $\hat{\xi}\in\hat{\Xi}_{z(s)}$ from (25) it follows

$$G_\varepsilon(t(s),z(s))\leq G_c, \tag{26}$$

which plays the role of (4).

To conclude this section, we remark that by the separate quadratic structure of the energy we have the following *separate minimality* property

$$u(s)\in\operatorname{argmin}\big\{\mathscr{E}_\varepsilon(t(s),u,z(s)):u\in\mathscr{U}\big\}$$

$$z(s)\in\operatorname{argmin}\big\{\mathscr{F}_\varepsilon(t(s),u(s),z):z\in\mathscr{Z}\,,z\leq z(s)\big\}.$$

Note that $z(s)$ is only a constrained minimizer and that in general it is not true that

$$(u(s),z(s))\in\operatorname{argmin}\big\{\mathscr{F}_\varepsilon(t(s),u,z):u\in\mathscr{U},\ z\in\mathscr{Z}\,,z\leq z(s)\big\}$$

as it would be in an energetic evolution.

3.4.2 Continuity Points: Thermodynamic Consistency

In this section we will discuss a couple of thermodynamic issues: the first is simply the energy balance (which will also lead to a KKT condition) while the second originates from the relationship between irreversibility constraint and dissipated energy. In both the cases we will assume that $t'(s)>0$ in a parametrization interval $[s_1,s_2]$ (so that the corresponding evolution is continuous in time). By (19) and (22) we can rewrite (23) as

$$\mathscr{F}_\varepsilon(t(s_2),u(s_2),z(s_2))=\mathscr{F}_\varepsilon(t(s_1),u(s_1),z(s_1))+\int_{s_1}^{s_2}\mathscr{P}^{ext}_\varepsilon(t(r),u(r),z(r))\,t'(r)\,dr,$$

which is the usual energy balance in parametrized integral form. Using the chain rule the above energy identity reads: for every $s\in(s_1,s_2)$

$$\begin{aligned}\mathscr{F}'_\varepsilon(t(s),u(s),z(s)) &= \partial_t\mathscr{F}_\varepsilon(t(s),u(s),z(s))\,t'(s)+\partial_u\mathscr{F}_\varepsilon(t(s),u(s),z(s))[u'(s)]+\\ &\quad+\partial_z\mathscr{F}_\varepsilon(t(s),u(s),z(s))[z'(s)]\\ &= \mathscr{P}^{ext}_\varepsilon(t(s),u(s),z(s))\,t'(s)\,.\end{aligned}$$

Again by (19) and (22) it follows that

$$\partial_z \mathscr{F}_\varepsilon(t(s), u(s), z(s))[z'(s)] = 0\,.$$

Here, note that equilibrium gives $\partial_z \mathscr{F}_\varepsilon(t(s), u(s), z(s))[\xi] \ge 0$ for every $\xi \in \Xi$. Using the notation of (25) the above identity in KKT fashion becomes

$$\big(\mathscr{G}_\varepsilon(t(s), z(s)) - \partial_z \mathscr{K}_\varepsilon(z(s))\big)[z'(s)] = 0\,, \tag{27}$$

which plays the role of (5).

Now, let us study the relationship between the irreversibility constraint and the dissipated energy $\mathscr{K}_\varepsilon(z)$. In general if $z_1 \le z_2$ it is not true that $\mathscr{K}_\varepsilon(z_1) \le \mathscr{K}_\varepsilon(z_2)$! This is simply due to the fact that

$$\mathscr{K}_\varepsilon(z) = \tfrac{1}{2} G_c \int_\Omega (z-1)^2/2\varepsilon + \varepsilon |\nabla z|^2 \, dx$$

includes a gradient term which is not monotone. For instance, consider $z_1 \le z_2$ with z_1 constant and with z_2 (highly) oscillating. Then

$$\int_\Omega (z_1 - 1)^2/4\varepsilon \, dx \ \ge \ \int_\Omega (z_2 - 1)^2/4\varepsilon \, dx$$

while

$$\int_\Omega \varepsilon |\nabla z_1|^2 \, dx \ < \ \int_\Omega \varepsilon |\nabla z_2|^2 \, dx.$$

If $z_1 \approx z_2$ but the energy of ∇z_2 is big enough it may be that $\mathscr{K}_\varepsilon(z_1) < \mathscr{K}_\varepsilon(z_2)$. Thus, the irreversibility constraint, given by the monotonicity of z, does not always match with the monotonicity of the dissipated energy. However, this is not what happens in the evolution, at least in those interval $[s_1, s_2]$ where $t'(s) > 0$. Indeed, $z'(s) \in \Xi$ and thus by (27)

$$\partial_z \mathscr{E}_\varepsilon(t(s), u(s), z(s))[z'(s)] + \partial_z \mathscr{K}_\varepsilon(z(s))[z'(s)] = 0.$$

Note that

$$\partial_z \mathscr{E}_\varepsilon(t(s), u(s), z(s))[z'(s)] = \int_\Omega t^2(s)\, z(s)\, z'(s)\, W(Du(s))\, dx \le 0,$$

because the only negative term is $z'(s)$. It follows that $\partial_z \mathscr{K}_\varepsilon(z(s))[z'(s)] \ge 0$ and thus the energy $s \mapsto \mathscr{K}_\varepsilon(z(s))$ is non-decreasing.

3.4.3 Discontinuity Points: Which Gradient Flow?

In this section we want to collect some properties of the evolution in the jumps, i.e., in the parametrization intervals $[s_1, s_2]$ where $t(s)$ is constant, and thus $t'(s) = 0$. It is fair to say that at the current stage the picture is not fully clear and detailed. In order to understand the main qualitative features it is not too restrictive to assume that for $s \in [s_1, s_2]$ it holds

$$|\partial_u \mathscr{F}_\varepsilon(t(s), u(s), z(s))|^-_{t(s),\, z(s)} = |\partial_v \mathscr{F}_\varepsilon(t(s), u(s), z(s))|^-_{t(s),\, u(s)} = 1\,.$$

Under these assumptions, (23) implies (by the chain rule and convexity arguments) that for a.e. $s \in [s_1, s_2]$ we have

$$u'(s) \in \operatorname{argmin} \{\partial_u \mathscr{F}_\varepsilon(t(s), u(s), z(s))[\phi] : \phi \in \Phi,\ |\phi|_{t(s),\, z(s)} = 1\},$$
$$z'(s) \in \operatorname{argmin} \{\partial_v \mathscr{F}_\varepsilon(t, u, v)[\xi] : \xi \in \Xi,\ \|\xi\|_{t(s),\, u(s)} = 1\}.$$

In other terms, u' and z' are the steepest descent direction for $\mathscr{F}_\varepsilon$ with respect to the "intrinsic norms". If the mathematical meaning is formally clear, the physical behaviour is understood only for the gradient flow of u. Indeed (cf. [14]) on the jumps the displacement field evolves like a "phase-field visco-elastic flow"

$$\begin{cases} \operatorname{div}(\boldsymbol{\sigma}_{z(s)}(u(s) + u'(s))) = 0 & \Omega \\ u(s) = \pm\hat{e} & \partial_D \Omega \\ \boldsymbol{\sigma}_{z(s)}(u(s) + u'(s))\,\hat{n} = 0 & \partial_N \Omega. \end{cases}$$

On the contrary, it seems not easy to provide a meaningful PDE for the evolution of the phase field variable z since Ξ is not a space, but just a convex set.

4 Open Problems

In a broader perspective, the most interesting, and probably most difficult, open problem is the convergence of the parametrized quasi-static BV-evolutions, say $s \mapsto (t_\varepsilon(s), u_\varepsilon(s), z_\varepsilon(s))$, as $\varepsilon \to 0$. On the base of Γ-convergence [2, 6] it is expected a sharp crack evolution, possibly in the space *SBD* [4] or *GSBD* [8]. However, Γ-convergence has been crafted to study the convergence of energies and global minimizers but it is not enough to provide convergence of equilibrium points and slopes which are the main ingredients for gradient flows and BV-evolutions. In general to have convergence of BV-evolutions it is necessary to have at least a sort of Γ-liminf inequality for the slopes [22], as it is for gradient flows [25]. However, it is not yet known any reasonable notion of slope in *SBD* spaces.

In this direction some partial results have been published: for instance [11] (in the one dimensional setting) [21, 27] (with geometrical restriction on the crack) and [3] (with a regularized energy).

References

1. Abdollahi, A., & Arias, I. (2012). Phase-field modeling of crack propagation in piezoelectric and ferroelectric materials with different electromechanical crack conditions. *Journal of the Mechanics and Physics of Solids*, *60*(12), 2100–2126.
2. Ambrosio, L., & Tortorelli, V. M. (1990). Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence. *Journal of the Mechanics and Physics of Solids*, *43*(8), 999–1036.
3. Babadjian, J. F. & Millot, V. (2014) Unilateral gradient flow of the Ambrosio-Tortorelli functional by minimizing movements. *Journals Annales de l'Institut Henri Poincar.é Annales: Analyse Non Lineaire*, *31*(4), 779–822.
4. Bellettini, G., Coscia, A., & Dal Maso, G. (1998). Compactness and lower semicontinuity properties in SBD(Ω). *Mathematische Zeitschrift*, *228*(2), 337–351.
5. Bourdin, B., Francfort, G. A., & Marigo, J.-J. (2000). Numerical experiments in revisited brittle fracture. *Journal of the Mechanics and Physics of Solids*, *48*(4), 797–826.
6. Chambolle, A. (2003). A density result in two-dimensional linearized elasticity and applications. *Archive for Rational Mechanics and Analysis*, *167*(3), 211–233.
7. Conti, S., Focardi, M. & Lurlano, F. Phase-field approximation of cohesive fracture energies.
8. Dal Maso, G. (2013). Generalised functions of bounded deformation. *Journal of the European Mathematical Society (JEMS)*, *15*(5):1943–1997.
9. Efendiev, M. A., & Mielke, A. (2006). On the rate-independent limit of systems with dry friction and small viscosity. *Archive for Rational Mechanics and Analysis*, *13*(1), 151–167.
10. Francfort, G. A., & Marigo, J.-J. (1998). Revisiting brittle fracture as an energy minimization problem. *Journal of the Mechanics and Physics of Solids*, *46*(8), 1319–1342.
11. Francfort, G. A., Le, N. Q., & Serfaty, S. (2009). Critical points of Ambrosio-Tortorelli converge to critical points of Mumford-Shah in the one-dimensional Dirichlet case. *Archive for Rational Mechanics and Analysis*, *15*(3), 576–598.
12. Griffith, A. A. (1920). The phenomena of rupture and flow in solids. *Journal of the Mechanics and Physics of Solids*, *18*, 163–198.
13. Hesch, C., & Weinberg, K. (2014). Thermodynamically consistent algorithms for a finite-deformation phase-field approach to fracture. *Journal of the Mechanics and Physics of Solids*, *99*(12), 906–924.
14. Knees, D. & Negri, M. Convergence of alternate minimization schemes for phase field fracture and damage.
15. Knees, D., Rossi, R., & Zanini, C. (2013). A vanishing viscosity approach to a rate-independent damage model. *Journal of the Mechanics and Physics of Solids*, *23*(4), 565–616.
16. Larsen, C. J., Ortner, C., & Süli, E. (2010). Existence of solutions to a regularized model of dynamic fracture. *Journal of the Mechanics and Physics of Solids*, *20*(7), 1021–1048.
17. Miehe, C., Welschinger, F., & Hofacker, M. (2010). Thermodynamically consistent phase-field models of fracture: Variational principles and multi-field FE implementations. *Journal of the Mechanics and Physics of Solids*, *83*(10), 1273–1311.
18. Mielke, A. (2011). Differential, energetic, and metric formulations for rate-independent processes. In nonlinear PDE's and applications, volume 2028 of lecture notes in mathematics (pp. 87–170). Heidelberg: Springer.
19. Negri, M. (2010). A comparative analysis on variational models for quasi-static brittle crack propagation. *Journal of the Mechanics and Physics of Solids*, *3*, 149–212.

20. Negri, M. (2010). From rate-dependent to rate-independent brittle crack propagation. *Journal of the Mechanics and Physics of Solids*, *98*(2), 159–178.
21. Negri, M. (2013). From phase-field to sharp cracks: Convergence of quasi-static evolutions in a special setting. *Journal of the Mechanics and Physics of Solids*, *26*, 219–224.
22. Negri, M. (2014). Quasi-static rate-independent evolutions: Characterization, existence, approximation and application to fracture mechanics. *Journal of the Mechanics and Physics of Solids*, *20*(4), 983–1008.
23. Negri, M., & Ortner, C. (2008). Quasi-static propagation of brittle fracture by Griffith's criterion. *Journal of the Mechanics and Physics of Solids*, *18*(11), 1895–1925.
24. Pandolfi, A., & Ortiz, M. (2012). An eigenerosion approach to brittle fracture. *Journal of the Mechanics and Physics of Solids*, *92*(8), 694–714.
25. Sandier, E., & Serfaty, S. (2004). Gamma-convergence of gradient flows with applications to Ginzburg-Landau. *Journal of the Mechanics and Physics of Solids*, *57*(12), 1627–1672.
26. Schmidt, B., Fraternali, F., & Ortiz, M. (2008). Eigenfracture: An eigendeformation approach to variational fracture. *Journal of the Mechanics and Physics of Solids*, *7*(3), 1237–1266.
27. Sicsic, P., & Marigo, J. -J. (2013). From gradient damage laws to Griffith's theory of crack propagation. *Journal of Elasticity*, *113*(1), 55–74.

Improving the Material-Point Method

Deborah Sulsky and Ming Gong

Abstract The material-point method (MPM) was introduced about 20 years ago and is a versatile method for solving problems in continuum mechanics. The flexibility of the method is achieved by combining two discretizations of the material. One is a Lagrangian description based on representing the continuum by a set of material points that are followed throughout the calculation. The second is a background grid that is used to solve the continuum equations efficiently. In its original form, some applications of the method appeared to be second order accurate while other tests showed poor or no convergence. This paper provides a framework for analyzing the errors in MPM. Moreover, the analysis suggests modifications to the algorithm to improve accuracy. The analysis also points to connections between MPM and other meshfree methods.

1 Introduction

The 1990s saw significant advances in the development of meshfree technologies for computational mechanics. Methods such as the Element Free Galerkin Method (EFG) [3, 4], Reproducing Kernel Particle Method (RKPM) [13, 14], h-p clouds [8], meshless local Petrov-Galerkin (MLPG) [2] and the Partition of Unity Method (PU) [15] rely on meshfree approximations of functions constructed from scattered data. Two mainly equivalent approaches to function reconstruction are used in this literature. The first is the moving least squares method (MLS) with polynomial basis functions and the second is reproducing kernel particle methods (RKPM). In MLS a functional is minimized to determine the particle shape functions whereas the guiding principle in RKPM is to determine shape functions that exactly reproduce polyno-

D. Sulsky (✉) · M. Gong
Department of Mathematics and Statistics, The University of New Mexico,
Albuquerque, NM 87131, USA
e-mail: sulsky@math.unm.edu

M. Gong
e-mail: gonmin77@unm.edu

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_10

mials up to a given order. These methods most often use a weak formulation of the governing equations resulting in an implementation that is similar in structure to finite element methods but with shape functions developed from the scattered data points. A continuing research topic is development of efficient and stable methods for the integration of the weak form equations [7]. More recently, the Optimal Transportation Meshfree (OTM) [12] scheme was developed based on max-ent interpolation [1]. The equations of motion are formulated from optimal transportation theory which discretizes the inertial action in space and time within a variational framework.

The material-point method (MPM) [20] was developed contemporaneously with the meshfree technologies but its formulation was inspired by the particle-in-cell method [6, 10]. Unlike general meshfree methods, MPM is restricted to solving problems in continuum mechanics. The method discretizes the continuum based on representing it by a set of Lagrangian material points that are followed throughout the calculation. Like meshfree methods, functions are reconstructed from these scattered data points. However, these functions are then evaluated to provide information for a grid where continuum equations of motion are solved efficiently.

Equations of motion are solved in an updated Lagrangian frame on the computational grid, using standard finite difference or finite element methods. Advection is modeled by moving the material points in the computed velocity field. Each numerical material point carries its material properties without error while it is advected. Since all the properties of the continuum are assigned to the numerical material points, the information carried by these points is enough to characterize the flow and the grid carries no permanent information. Accuracy and consistency of MPM have been examined with fragmentary results in the literature, e.g. [18, 19]. The goal of this paper is to establish a framework for studying the accuracy of MPM. The accuracy of each step in the algorithm is examined as well as how the steps interact, in order to make improvements.

For time-dependent problems, there are four steps in the algorithm: (i) choose a convenient computational grid; (ii) map information from the material points to the grid; (iii) solve the field equations on the grid; and (iv) update the material points based on the grid solution. Similar accuracy should be maintained by the function reconstruction in step (ii) and the grid-based solution method chosen to advance the solution in step (iii). Finally, a scheme of consistent accuracy must be employed to update the material points in step (iv) in order to maintain the overall accuracy of the method. Our analysis brings to light connections between MPM and other meshfree methods. Moreover, our analysis shows how to exploit the function reconstruction techniques used in meshfree methods to improve MPM.

2 Preliminaries

Let $\varOmega(0) \in \mathbb{R}^d$ denote the reference configuration of a continuum body in d dimensions, with material points labeled X. The set $\varOmega(0)$ is assumed open and bounded, with a smooth boundary $\varGamma_0 = \partial\varOmega(0)$. Assume the reference boundary $\varGamma_0$ is partitioned into disjoint subsets, $\Gamma_0 = \overline{\Gamma_u \bigcup \Gamma_t}$ and $\Gamma_u \bigcap \Gamma_t = \emptyset$, such that displacement is prescribed on Γ_u and traction is prescribed on Γ_t.

Let the spatial (current) configuration of the same body be $\Omega(t) \in \mathbb{R}^d$, with points labeled x. Assume there exists a smooth mapping, the motion of the body, $\varphi : \Omega(0) \times [0, \mathrm{T}] \to \mathbb{R}^d$, such that $\Omega(t) = \boldsymbol{\varphi}(\Omega(0), \mathrm{t})$ and $x = \boldsymbol{\varphi}(X, \mathrm{t})$, where $[0, \mathrm{T}] \subset \mathbb{R}$ is the time interval of interest. The current boundary is $\Gamma(t) = \partial\Omega(t)$ with $\Gamma(t) = \Gamma_t(t) \cup \Gamma_u(t)$ and $\Gamma_t(t) \cap \Gamma_u(t) = \emptyset$, where $\Gamma_u(t) = \varphi(\Gamma_u, t)$ and $\Gamma_t(t) = \varphi(\Gamma_t, t)$.

The deformation gradient of the motion is defined as

$$F(X, \mathrm{t}) = \mathrm{Grad}\ \varphi. \tag{1}$$

Given a reference density $\rho_0 : \Omega(0) \to \mathbb{R}^+$, the spatial density is $\rho = J^{-1}\rho_0$, where $J = \det\ F$.

MPM solves continuum mechanics problems for a body occupying $\Omega(t)$ at time t, of the form

$$\nabla \cdot \sigma + \rho b = \rho a \qquad \text{in } \Omega(t) \tag{2}$$

$$\sigma \cdot n = \bar{t} \qquad \text{on } \Gamma_t(t) \tag{3}$$

$$u = \overline{u} \qquad \text{on } \Gamma_u(t) \tag{4}$$

with initial conditions

$$v(x, 0) = v_0(x), \qquad x \in \Omega(0)$$

$$\sigma(x, 0) = \sigma_0(x), \qquad x \in \Omega(0).$$

In these equations, the mass density is $\rho(x, t)$, the acceleration is $a(x, t)$, the velocity is $v(x, t)$, the displacement is $u(x, t)$, the specific body force is $b(x, t)$ and the Cauchy stress is $\sigma(x, t)$. Boundary conditions can consist of applied traction, $\bar{t}$, on a portion of the boundary denoted by $\Gamma_t(t)$, with n being the unit outward normal to the boundary, and prescribed displacement, $\overline{u}$, on a portion denoted by $\Gamma_u(t)$. A constitutive equation for the stress is required to complete the specification. For this work, we will use a neo-Hookean constitutive model extended to the compressible range,

$$J\sigma = \lambda(\ln J)I + \mu\left(FF^T - I\right), \tag{5}$$

where μ and λ are the Lamé constants and I is the second order identity tensor.

The philosophy of MPM is to solve the equations of motion on a background grid, but to keep track of trajectories of a set of material points, where the material points represent the geometry of the body $\Omega(t)$ during deformation and carry the solution. The process is accomplished by mapping information between the grid and the material points. The details are given in the next section.

3 Main Steps of MPM

A body occupying $\Omega(0)$ is discretized into a finite set of N_p regions, $\Omega_p(0)$, with a material point at the centroid, with position $x_p(0)$, $p = 1, 2, \ldots, N_p$. Each of these points represents a volume of material, $V_p(0)$, the volume of $\Omega_p(0)$, with a mass, $m_p = \int_{\Omega_p(0)} \rho_0(X)\, dV$. The material points move in time and the associated volume moves with them. Mass conservation requires that m_p be constant in time. The aim is to determine the position, $x_p(t)$, the density, $\rho_p(t)$, the velocity, $v_p(t)$, and the stress, $\sigma_p(t)$, at time t associated with these points given the initial values, $\rho_p(0) = \rho_0(x_p(0))$, $v_p(0) = v_0(x_p(0))$ and $\sigma_p(0) = \sigma_0(x_p(0))$.

3.1 Function Reconstruction from Scattered Data

Mapping the information from the material points to the grid involves reconstructing a function from scattered data on the material points, and evaluating the reconstructed function on the background grid. That is, given data, u_p, $p = 1, 2, \ldots, N_p$ at material points, we wish to determine basis functions $\psi_p^{[r]}(x)$ so that the reconstructed function can be written in the form

$$u^R(x) = \sum_{p=1}^{N_p} \psi_p^{[r]}(x) u_p. \tag{6}$$

We base this function reconstruction on the idea of reproducing polynomials exactly up to a specified degree r, where r is a non-negative integer [11, 16]. The notation $\mathbb{P}_r = \mathbb{P}_r(\Omega)$ denotes the space of polynomials of degree less than or equal to r on Ω. The dimension of $\mathbb{P}_r$ is N_r,

$$N_r = \binom{r+d}{d} = \frac{(r+d)!}{r!d!}. \tag{7}$$

The polynomial reproducing conditions are

$$u(x) = \sum_{p=1}^{N_p} \psi_p^{[r]}(x) u(x_p) \quad \forall u \in \mathbb{P}_r. \tag{8}$$

In order to express these conditions concisely, introduce a multi-index, α. A multi-index is an ordered collection of non-negative integers, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$. The length of α is defined as $|\alpha| = \sum_{i=1}^{d} \alpha_i$. We also define, for any point $x \in \mathbb{R}^d$, the monomial, $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}$. A sequence of points $\xi_\alpha \in \mathbb{R}$, indexed by the multi-indices α with $|\alpha| \leq r$, is ordered with the terms in lexical order to form a vector $\xi \in \mathbb{R}^{N_r}$. Lexical order is $\alpha = (0, 0, \ldots, 0), (1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 1)$,

$(2, 0, 0, \ldots, 0)$, $(1, 1, 0, \ldots, 0)$, $\ldots$, $(0, 0, \ldots, 0, r)$. The basis functions in (6) are written in terms of a weight function $w_h(x)$ with compact support defined by h, and a local polynomial basis,

$$\psi_p^{[r]}(x) = w_h(x - x_p) \sum_{|\alpha| \le r} (x - x_p)^\alpha a_\alpha(x), \quad p = 1, 2, \ldots, N_p. \tag{9}$$

The reproducing conditions (8) introduce N_r conditions for the coefficients a_α,

$$\sum_{|\alpha| \le r} m_{\alpha+\beta}(x) a_\alpha(x) = \delta_{|\beta|,0}, \quad |\beta| \le r \tag{10}$$

where the moment functions are

$$m_\alpha(x) = \sum_{p=1}^{N_p} w_h(x - x_p)(x - x_p)^\alpha. \tag{11}$$

The above system of equations can be written in matrix form

$$M(x)a(x) = h(0) \tag{12}$$

by introducing the moment matrix $M(x)$,

$$M(x) = \sum_{p=1}^{N_p} w_h(x - x_p) h(x - x_p) h(x - x_p)^T, \tag{13}$$

where $h(z) = (z^\alpha)_{|\alpha| \le r} \in \mathbb{R}^{N_r}$. Then, the shape function is

$$\psi_p^{[r]}(x) = w_h(x - x_p) h^T(0) M^{-1}(x) h(x - x_p). \tag{14}$$

The moment matrix is symmetric and positive semi-definite and the sum of rank one matrices. A necessary condition for $M(x)$ to be nonsingular is that for any x there be at least N_r nonzero terms in the sum (13). The resulting shape functions $\psi_p^{[r]}(x)$ form a partition of unity and if $w_h \in C^k$ then $\psi_p^{[r]} \in C^k$, $p = 1, 2, \ldots, N_p$.

For this work, we restrict our attention to functions $u(x)$ in a Hilbert space and assume the boundary of the domain is smooth. Under appropriate hypotheses on the distribution of material points so that the moment matrix is nonsingular, we have the following error estimates for the function reconstruction [9]. For any $u \in H^{m+1}(\Omega)$ and $w \in C^k$

$$\|u - u^R\|_{H^l(\Omega)} \le c h^{\min(m+1,r+1)-l} |u|_{H^{\min(m+1,r+1)}(\Omega)} \quad \forall l \le \min(m+1, r+1, k). \tag{15}$$

If $m \geq r$ then the above reduces to

$$\|u - u^R\|_{H^l(\Omega)} \leq ch^{r+1-l}|u|_{H^{r+1}(\Omega)} \quad \forall l \leq \min(r+1, k); \tag{16}$$

in particular, for $l = 0$

$$\|u - u^R\|_{L_2(\Omega)} \leq ch^{r+1}|u|_{H^{r+1}(\Omega)}. \tag{17}$$

3.1.1 Standard MPM

The reconstructed function in the standard MPM algorithm is a particular case of Shepard function interpolation [17] which corresponds to $r = 0$. Shepard function interpolation is used to construct the velocity $v^R(x)$ from N_p discrete, scattered values v_p as

$$v^R(x) = \sum_{p=1}^{N_p} \psi_p^{[0]}(x) v_p, \tag{18}$$

where, from (13) and (14)

$$\psi_p^{[0]}(x) = \frac{w_h(x - x_p)}{\sum_{p=1}^{N_p} w_h(x - x_p)}, \tag{19}$$

for some weight function, w. The standard MPM uses a linear hat function weighted by the material-point mass for the weight function

$$w_h(x - x_p) = m_p s(x - x_p). \tag{20}$$

In one dimension, the linear hat function, written in terms of natural coordinates, is $s = \hat{s}(\xi)$,

$$\hat{s}(\xi) = \begin{cases} 1 - |\xi| & |\xi| \leq 1 \\ 0 & \text{otherwise} \end{cases}, \tag{21}$$

where $\xi = (x_p - x)/h$ is the natural coordinate with h being the support of s and the mesh spacing. In two space dimensions, the weight function is the tensor product of one-dimensional weight functions, $s(\xi, \eta) = \hat{s}(\xi) \otimes \hat{s}(\eta)$, with a similar construction in three dimensions.

3.1.2 Examples in One Dimension

Consider a function, $f(x) : \mathbb{R} \to \mathbb{R}$, defined on an interval $[a, b]$. Construct N_e elements of equal size $h = (b - a)/N_e$. The Ith element is the interval $[x_{I-1}, x_I]$ where the nodes are given by $x_I = a + Ih$, $I = 1, 2, \ldots, N_n = N_e + 1$. Within each

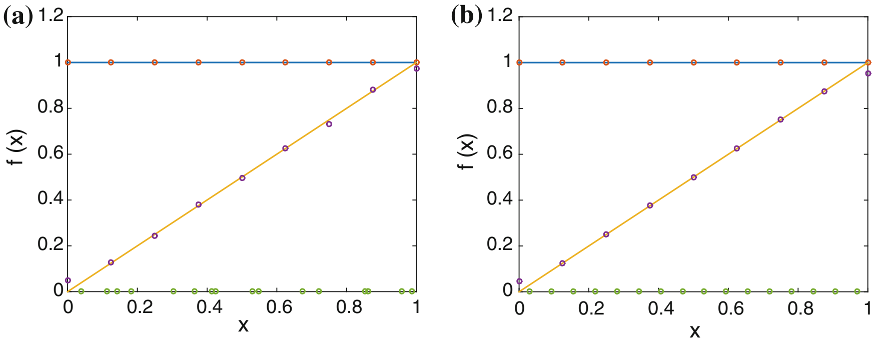


Fig. 1 Shepard interpolation using the functions $f(x) = 1$ and $f(x) = x$ (shown with *solid lines*) and **a** randomly-placed sampling points or **b** equally-spaced sampling points. The *circles* on the x-axis show the positions of the sampling points. The other *circles* in the plots show the reconstructed values $f^R(x_I)$. With equally-spaced points, the reconstructed function is exact for constant functions and exact at the interior nodes for linear functions

element, sample the function $f(x)$ at two randomly selected points. The sampling points, $x_p, p = 1, 2, \ldots, N_p = 2N_e$, play the role of material points. Set $f_p = f(x_p)$, $p = 1, 2, \ldots, N_p$. Now reconstruct the function using the material-point data to obtain $f^R(x)$. In MPM, we are interested in the value of the reconstructed function at nodes, $f^R(x_I)$.

By construction, Shepard interpolation reproduces constant functions. If we set $f(x) = 1$ and take $a = 0, b = 1$ and $N_e = 8$, Fig. 1a shows the exact function (solid line) and the reconstructed values at nodes, x_I, (circles). The circles along the x-axis indicate the position of the randomly selected material points. The figure shows that $f(x) = 1$ is reconstructed exactly. The figure also shows the reconstructed nodal values for $f(x) = x$ which are not exact. However, for equally-distributed sampling points, the Shepard interpolation reproduces linear functions except on the boundary of the domain. Figure 1b shows the reconstruction of $f(x) = 1$ and $f(x) = x$ when the material points are equally spaced with $N_p^e = 2$ material points per element located at $x_p = x_{I-1} + (2p^e - 1)h/N_p^e$, $p^e = 1, 2, \ldots, N_p^e$, $I = 1, 2, \ldots, N_e$ and $p = (I-1)N_p^e + p^e$. For linear functions, $f^R(x_I) = x_I$ exactly in the interior of the domain but there is an $O(h)$ error at the boundary points.

Figure 2 shows the same type of reconstruction except using the smooth function $f(x) = \sin \pi x$ on [0, 1], again with $N_e = 8$ and $N_p^e = 2$. For this example, the reconstruction is not exact whether or not the sampling points are equally spaced or randomly selected. However, the errors get smaller as h gets smaller, but the rate of convergence differs. We examine the L_2 convergence rate as given by the error estimate (17). For each h, we compute the L_2 error *excluding* the endpoints of the domain by approximating the integral

$$\int_{x_1}^{x_{N_e}} (f(x) - f^R(x))^{1/2}\, dx \tag{22}$$

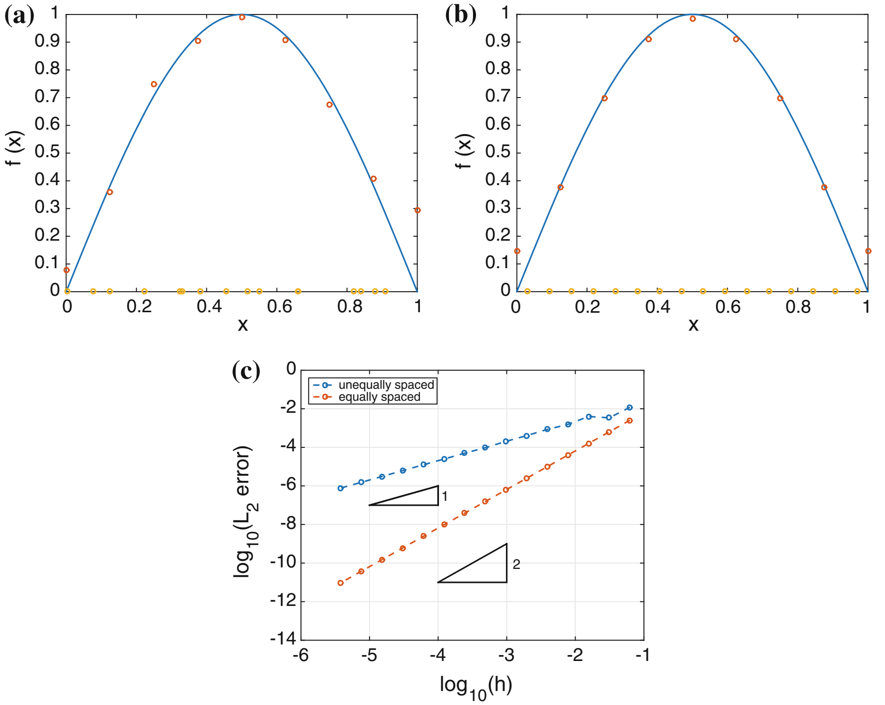


Fig. 2 Shepard interpolation using the function $f(x) = \sin \pi x$ (*solid line*) with **a** randomly placed sampling points and **b** equally spaced sampling points. The *circles* on the x-axis show the positions of the sampling points. The other *circles* in the plots show the reconstructed values $f^R(x_I)$. **c** Shows the convergence rate with h using random and equally spaced sampling points

using the trapezoidal rule based on the nodal values of the integrand. With randomly-selected sampling points, the convergence rate is $O(h)$ as expected from (17). With equally-spaced points the convergence rate is one order higher than expected, $O(h^2)$. Since there are $O(h)$ errors at the end points, if the end point values are included in the error estimate, the rate would be $O(h^{3/2})$ for one space dimension.

3.2 Solve the Momentum Equation on the Background Grid

In MPM, Eq. (2) is solved in weak form. In order to simplify the task, assume $\overline{u} = 0$ on $\Gamma_u(t)$ and $\Gamma_t(t) = \emptyset$. That is, displacement is prescribed over the entire boundary and the domain boundary is fixed in time. We also assume a nice boundary, such as a piecewise polygonal boundary with the background mesh conforming to the domain boundary. Under these conditions, the weak form of the governing equation is

$$\int_{\Omega(t)} \Big[\sigma \cdot \nabla \eta - \rho(b-a) \cdot \eta \Big] \, dv = 0 \tag{23}$$

for all admissible η. In particular, an admissible η is zero on the domain boundary.

Shape functions are used in the MPM to construct approximations for the finite element method on the background grid. The background grid is subdivided into elements, Ω^e, $e = 1, 2, \ldots, N_e$. The nodes of this mesh are $x_I(t)$, $I = 1, \ldots, N_n$. If each element Ω^e has m nodes then we can refer to the nodes belonging to an individual element with the notation $x_I^e(t)$, $I = 1, \ldots, m$. The nodal displacements are given by

$$u^h(x,t) = \sum_{I=1}^{N_n} u_I(t) N_I(x). \tag{24}$$

The velocity is represented by

$$v^h(x,t) = \frac{D}{Dt} u^h(x,t) = \sum_{I=1}^{N_n} \frac{d}{dt} u_I(t) N_I(x) = \sum_{I=1}^{N_n} v_I(t) N_I(x). \tag{25}$$

Likewise, the acceleration is

$$a^h(x,t) = \frac{D}{Dt} v^h(x,t) = \sum_{I=1}^{N_n} \frac{d}{dt} v_I(t) N_I(x) = \sum_{I=1}^{N_n} a_I(t) N_I(x). \tag{26}$$

Over a time step, the grid is Lagrangian. Thus, over the step, the shape function does not change with time. This fact is significant since a time derivative of the shape function is not required in the formulas for velocity and acceleration. Also, the natural coordinates for a material point in an element remain constant over the Lagrangian time step.

The approximations, (28)–(30), along with a representation for an admissible, smooth, virtual displacement η,

$$\boldsymbol{\eta}^h(x) = \sum_{I=1}^{N_n} \eta_I N_I(x), \tag{27}$$

are used to obtain semi-discrete equations of motion. Substitute the representations (24)–(27) into each term of equation (23) to obtain

$$\begin{aligned} \int_{\Omega(t)} \eta^h \cdot \rho(x,t) \frac{Dv^h}{Dt} dv &= \sum_{I=1}^{N_n} \eta_I \cdot \sum_{J=1}^{N_n} \int_{\Omega(t)} \rho(x,t) N_I(x) N_J(x) \frac{dv_J(t)}{dt} dv \\ &= \sum_{I=1}^{N_n} \eta_I \cdot \sum_{J=1}^{N_n} M_{IJ}(t) \frac{dv_J(t)}{dt}, \end{aligned} \tag{28}$$

$$-\int_{\Omega(t)} \nabla\eta^h : \sigma dv = -\sum_{I=1}^{N_n} \eta_I \cdot \int_{\Omega(t)} \nabla N_I(x) \cdot \sigma(x,t) dv \tag{29}$$

$$\int_{\Omega(t)} \eta^h \cdot \rho(x,t) b dv = \sum_{I=1}^{N_n} \eta_I \cdot \int_{\Omega(t)} \rho(x,t) b(x,t) N_I(x) dv. \tag{30}$$

In order to complete the spatial discretization a quadrature rule must be given to evaluate the integrals in Eqs. (28)–(30). In the standard MPM, the material points are used as quadrature points and the integrals become sums over material points. For example, to discretize the inertial term, Eq. (28), the components of the consistent mass matrix M_{IJ} are written as

$$\begin{aligned} M_{IJ}(t) &= \int_{\Omega(t)} \rho(x,t) N_I(x) N_J(x) dv \\ &\sim \sum_{p=1}^{N_p} \rho(x_p(t),t) N_I(x_p(t)) N_J(x_p(t)) V_p(t) \\ &= \sum_{p=1}^{N_p} m_p N_I(x_p) N_J(x_p), \end{aligned} \tag{31}$$

where the material-point mass is related to the density and volume through the equation $m_p = \rho(x_p(t),t)V_p(t)$.

The numerical simulations use the simpler lumped-mass matrix. This matrix is diagonal with the diagonal entries obtained by summing over the corresponding row of the consistent mass matrix and using the property $\sum_{J=1}^{N_n} N_J(x) = 1$,

$$M_I(t) = \int_{\Omega(t)} \rho(x,t) N_I(x) dv \sim \sum_{p=1}^{N_p} m_p N_I(x_p). \tag{32}$$

Equation (29) provides the nodal values of the internal forces

$$\begin{aligned} F_I^{\text{int}}(t) &= -\int_{\Omega(t)} \nabla N_I(x) \cdot \sigma(x,t) dv \\ &\sim -\sum_{p=1}^{N_p} \nabla N_I(x)|_{x=x_p} \sigma_p(t) V_p(t). \end{aligned} \tag{33}$$

In the above equation, the notation for the stress at the material point position, $\sigma_p(t) = \sigma(x_p(t),t)$ has been introduced.

Finally, the nodal values of the external forces arise from the body forces, Eq. (30),

$$\begin{aligned} F_I^{\text{ext}}(t) &= \int_{\Omega(t)} \rho(x,t) b(x,t) N_I(x) dv \\ &\sim \sum_{p=1}^{N_p} m_p b(x_p(t),t) N_I(x_p(t)). \end{aligned} \tag{34}$$

The weak form of the momentum balance equates (28) to the sum of the forces (29)–(30). Since the weak form must hold for arbitrary η_I, except at nodes on the boundary where the displacement is prescribed, we obtain the semi-discrete equation for the nodal acceleration at unconstrained nodes

$$\sum_{J=1}^{N_n} M_{IJ}(t) \frac{dv_J(t)}{dt} = F_I^{\text{int}}(t) + F_I^{\text{ext}}(t). \tag{35}$$

With the lumped mass matrix this equation becomes

$$M_I(t) \frac{dv_I(t)}{dt} = F_I^{\text{int}}(t) + F_I^{\text{ext}}(t). \tag{36}$$

The momentum equation is solved for the acceleration at unconstrained nodes, which is then integrated in time to obtain the corresponding velocity and displacement.

$$\begin{aligned} \frac{dv_I(t)}{dt} &= a^h(x_I,t) \\ \frac{du_I(t)}{dt} &= v^h(x_I,t). \end{aligned} \tag{37}$$

Nodes constrained by the displacement boundary conditions move according to those prescribed constraints.

3.3 *Update the Material-Point Information*

Once the equation of motion on the grid is solved, we need to update the material-point information. The material points move along with the flow within an element as it deforms in a Lagrangian manner over the time step. The material-point velocity and position are updated as

$$\begin{aligned} \frac{dv_p(t)}{dt} &= a^h(x_p,t) \\ \frac{dx_p(t)}{dt} &= v^h(x_p,t). \end{aligned} \tag{38}$$

The deformation gradient is also updated according to

$$\frac{dF_p(t)}{dt} = L_p(t)F_p(t), \quad L_p(t) = \nabla v^h(x,t)|_{x=x_p(t)}. \tag{39}$$

The density and volume are obtained from the algebraic update

$$V_p(t) = \det F_p(t)V_p(0), \quad \rho_p(t) = \rho_0(x_p(t))/V_p(t). \tag{40}$$

A time-integration scheme is needed to solve the semi-discrete equations numerically. The order of the time integration must match the order of the spatial discretization to maintain overall accuracy of the method.

3.4 Generate a New Grid

After the material-point information is updated, a new grid is generated for the next time step. Usually in MPM the grid points that have moved during the Lagrangian step are moved back to their previous locations. Note, however that one can generate any convenient grid instead.

4 Time Discretization

Although the theoretical framework above provides guidance on the construction of a method of any order, the focus in this work will be to obtain second order accuracy in space and time. A nominally second order, leapfrog scheme is used for the time discretization. In this scheme the velocity and displacement updates are staggered in time. A superscript is used to denote the time level in the discrete equations. Let $t^{n+\frac{1}{2}} = \frac{1}{2}(t^{n+1} + t^n)$, $\Delta t^n = t^{n+\frac{1}{2}} - t^{n-\frac{1}{2}}$, and $\Delta t^{n+\frac{1}{2}} = t^{n+1} - t^n$. The acceleration is obtained from Eq. (36) evaluated at time t^n

$$M_I^n a_I^n = F_I^{\text{int},n} + F_I^{\text{ext},n}, \tag{41}$$

where the mass comes from (32). The internal force comes from Eq. (33)

$$F_I^{\text{int},n} = -\sum_{p=1}^{N_p} \nabla N_I(x)|_{x=x_p^n} \sigma_p^n V_p^n, \tag{42}$$

and the external force comes from Eq. (34)

$$F_I^{\mathrm{ext},n} = \sum_{p=1}^{N_p} m_p b(x_p^n, t^n) N_I(x_p^n). \tag{43}$$

A centered difference formula for the acceleration is

$$a_I^n = \frac{v_I^{n+\frac{1}{2}} - v_I^{n-\frac{1}{2}}}{t^{n+\frac{1}{2}} - t^{n-\frac{1}{2}}} = \frac{1}{\Delta t^n}(v_I^{n+\frac{1}{2}} - v_I^{n-\frac{1}{2}}). \tag{44}$$

This formula can be converted into an update for the velocity

$$v_I^{n+\frac{1}{2}} = v_I^{n-\frac{1}{2}} + \Delta t^n a_I^n. \tag{45}$$

Similarly, the velocity can be obtained by differencing the displacement

$$v_I^{n+\frac{1}{2}} = \frac{u_I^{n+1} - u_I^n}{t^{n+1} - t^n} = \frac{1}{\Delta t^{n+\frac{1}{2}}}(u_I^{n+1} - u_I^n). \tag{46}$$

Likewise, this formula can be converted into an update for the displacement

$$u_I^{n+1} = u_I^n + \Delta t^{n+\frac{1}{2}} v_I^{n+\frac{1}{2}}. \tag{47}$$

The material points are transported in the same fields as the nodes. Since $a_I^n = a^h(x_I, t^n)$ and $v_I^{n+\frac{1}{2}} = v^h(x_I, t^{n+\frac{1}{2}})$, the material points are updated using

$$\begin{aligned}
v_p^{n+\frac{1}{2}} &= v_p^{n-\frac{1}{2}} + \Delta t^n a^h(x_p^n, t^n) = v_p^{n-\frac{1}{2}} + \Delta t^n \sum_{I=1}^{N_n} N_I(x_p^n) a_I^n, \\
x_p^{n+1} &= x_p^n + \Delta t^{n+\frac{1}{2}} v^h(x_p^{n+\frac{1}{2}}, t^{n+\frac{1}{2}}) = x_p^n + \Delta t^{n+\frac{1}{2}} \sum_{I=1}^{N_n} N_I(x_p^n) v_I^{n+\frac{1}{2}}, \\
u_p^{n+1} &= x_p^{n+1} - x_p^0,
\end{aligned} \tag{48}$$

Information from the grid is also used to update the material-point stress state. In this work, we will consider hyperelastic models that are specified through the deformation gradient. The deformation gradient is updated using the chain rule

$$F_p^{n+1} = \frac{\partial x^{n+1}}{\partial x^n} F_p^n, \tag{49}$$

and

$$\frac{\partial x^{n+1}}{\partial x^n} = I + \Delta t^{n+\frac{1}{2}} L_p^{n+\frac{1}{2}}, \tag{50}$$

where

$$L_p^{n+\frac{1}{2}} = \sum_{I=1}^{N_n} \nabla N_I(x_p^n) v_I^{n+\frac{1}{2}}. \tag{51}$$

Once the deformation gradient is advanced in time, we can compute the stress, σ_p^{n+1} at material points from the constitutive equation. Likewise the density is $\rho_p^{n+1} = \rho_0(x_p^0)/J_p^{n+1}$ with $J_p^{n+1} = \det(F_p^{n+1})$. Similarly, the volume corresponding to the material point is $V_p^{n+1} = J_p^{n+1} V_p(0)$.

5 Numerical Example

To keep the presentation simple while illustrating the ideas, a one-dimensional example will be used to demonstrate the convergence properties of MPM. The example is constructed using the method of manufactured solutions described in [18, 19] and outlined in the next section.

5.1 Manufactured Solution

The standard MPM and its improvements will be examined through the solution of a one-dimensional bar problem. We would like a problem with an exact solution so that we may accurately compute the order of convergence of the method. An example consisting of a longitudinally vibrating bar can be obtained by writing the problem in the reference configuration. Consider the time dependent problem for a 1D, elastic bar of length, L, with fixed ends, and constant cross-section. The equivalent equation to (2) for the bar problem in the reference configuration is

$$\frac{\partial P(X,t)}{\partial X} + \rho_0(X) B(X,t) = \rho_0(X) \frac{\partial V(X,t)}{\partial t}, \tag{52}$$

with boundary conditions

$$V(0,t) = V(L,t) = 0. \tag{53}$$

The initial conditions are

$$V(X,0) = v_0(X), \quad P(X,0) = P_0(X). \tag{54}$$

In these equations V is the velocity, ρ_0 is the original mass density, B is the body force per unit mass, and P is the 1st Piola-Kirchoff stress. The relationship between P and the Cauchy stress is

$$J\sigma = PF^T. \tag{55}$$

For the neo-Hookean model

$$P(X,t) = \left[\lambda(\ln J)I + \mu\left(FF^T - I\right)\right]F^{-T}. \tag{56}$$

For the manufactured solution, determine the body force so that the displacement, U, and velocity, V, are given by

$$U(X,t) = \frac{A}{C\pi}\sin\pi X\sin C\pi t, \quad V(X,t) = A\sin\pi X\cos C\pi t. \tag{57}$$

The corresponding initial conditions are

$$v_0(X) = A\sin\pi X, \quad P_0(X) = 0, \quad F(X,0) = 1, \tag{58}$$

and the boundary conditions $V(0,t) = V(L,t) = 0$ are satisfied.

The deformation gradient is a diagonal matrix with $\operatorname{diag}(F) = (F_{11}, 1, 1)$, where $F_{11} = 1 + (A/C)\cos\pi X\sin C\pi t$. Note that $J = \det(F) = F_{11}$ and

$$\begin{aligned}\frac{\partial P_{11}}{\partial X} &= \left[\frac{\lambda}{F_{11}^2}(1 - \ln F_{11}) + \frac{\mu}{F_{11}^2}(F_{11}^2 + 1)\right]\frac{\partial F_{11}}{\partial X}\\ &= \left[\frac{\lambda}{F_{11}^2}(1 - \ln F_{11}) + \frac{\mu}{F_{11}^2}(F_{11}^2 + 1)\right](-\pi^2 U).\end{aligned} \tag{59}$$

Similarly, the time derivative of the velocity is related to the displacement

$$\frac{\partial V}{\partial t} = -C^2\pi^2 U. \tag{60}$$

Therefore, to have a solution to equation (52), the body force must be

$$\rho_0 B = \pi^2 U\left[\frac{\lambda}{F_{11}^2}(1 - \ln F_{11}) + \frac{\mu}{F_{11}^2}(F_{11}^2 + 1) - \rho_0 C^2\right]. \tag{61}$$

This solution is valid for any choice of C^2. We will use $C^2 = E/\rho_0$, where E is the Young's modulus. With ν being Poison's ratio, we can set $\mu = E/(2(1+\nu))$ and $\lambda = E\nu/((1+\nu)(1-2\nu))$ in the usual way.

If we introduce dimensionless variables so that lengths are scaled by L, density by ρ_0, velocity by C, and stresses by E, we need only consider a bar of unit length with dimensionless density $\rho_0 = 1$, dimensionless modulus, $E = 1$ and therefore

$C = 1$. The only free parameter is the dimensionless amplitude of the initial velocity $\hat{A} = A/C$ which controls the amount of deformation in the bar.

The material-point solution is an approximation to the solution of the partial differential equations where $u_p^n \sim U(x_p^0, t^n)$, $v_p^{n+\frac{1}{2}} \sim V(x_p^0, t^{n+\frac{1}{2}})$, $\sigma_p^n \sim P(x_p^0, t^n)$, and $F_p^n \sim F(x_p^0, t^n)$. In one space dimension, the two-norm of the error can be computed using one-point quadrature to approximate the integral defining the norm. In order to measure the error in mesh-based quantities, the exact material-point quantities are projected to the grid. For example, the exact material-point velocity, $V(x_p^0, t^{n+\frac{1}{2}})$, is used to reconstruct a velocity field which is then evaluated at grid nodes and compared with the numerical solution, $v_I^{n+\frac{1}{2}}$.

6 Convergence Rate of Standard MPM

Standard MPM with leapfrog time differencing is given by the equations in Sect. 4 with function reconstruction based on $\psi^{[0]}$ as given in Sect. 3.1.1. To test convergence, the method is applied to solve the problem of Sect. 5.1. Since we wish to consider smooth solutions, the choice of the dimensionless parameter $\hat{A}$ should be less than one to avoid the formation of shocks. The last consideration in running the problem is stability. A necessary condition for the stability of the leapfrog scheme is that the Courant-Friedrichs-Levy (CFL) condition be satisfied. For this difference scheme, the CFL stability requirement is

$$\max \frac{c\Delta t}{h} \leq 1. \tag{62}$$

The maximum is over all x at time t and c is the maximum wave speed. In the one-dimensional case, the wave speed is $c = \sqrt{E/\rho} = C\sqrt{\rho_0/\rho} = C\sqrt{J}$. For the manufactured solution of Sect. 5.1, we can compute the maximum of $\sqrt{J}$ to be $\sqrt{1 + A/C}$. To insure stability for all cases, we choose $\Delta t = 0.7h$.

Figure 3a shows a plot of the logarithm of the L^2 norm of the error in various computed quantities versus the logarithm of the mesh size h for standard MPM. The error is the error in the computed solution to the problem from Sect. 5.1 solved using $E = 1$, $\nu = 0$, $\rho_0 = 1$, $L = 1$ and $\hat{A} = 0.001\pi$. The mesh sizes considered are $h = 2^{-(i+2)}$, $i = 1, 2, 3, 4$. There are initially two, equally-spaced, material points per element. The computed quantities considered in the figure are σ_p, ρ_p, v_p, and u_p—the stress, density, velocity and displacement stored at material points, and, v_I, the velocity stored on the grid. The simulations are run while the time satisfies $t \leq t_{\max}$ and $t_{\max}$ is set to a dimensionless time of 2. Thus, the simulations terminate at a time t_{end} such that $t_{\text{end}} \leq t_{\max}$ and $t_{\text{end}} + \Delta t > t_{\max}$. The error norms are computed at t_{end} for the density, stress and displacement, and at $t_{\text{end}} - \Delta t/2$ for the velocity. The observed convergence rate for the deformation gradient, density and stress is 1.8. The rate is the same for these quantities since the density and stress are algebraic

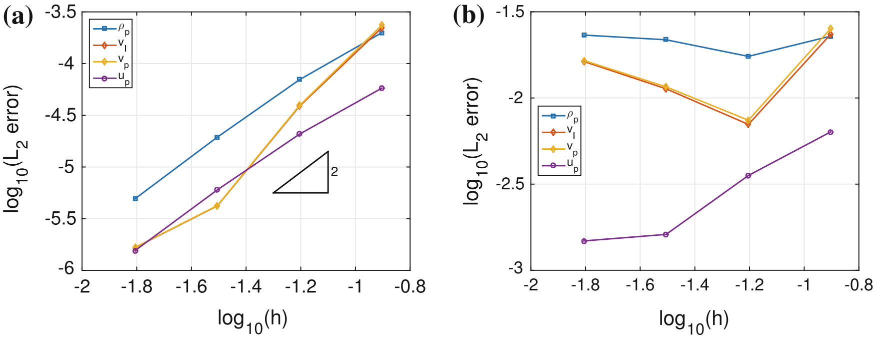


Fig. 3 Convergence of various computed quantities using the standard MPM for the problem in Sect. 5.1 with **a** $\hat{A} = 0.001\pi$ and **b** $\hat{A} = 0.1\pi$

functions of the deformation gradient. The figure shows data for the density only. The observed convergence rate for the displacement is 1.7. The convergence behavior for the velocity is a bit erratic but a least squares fit gives an overall convergence rate of 2.4.

The calculation is repeated in Fig. 3b except with a larger value of $\hat{A} = 0.1\pi$. The results shown in the plot indicate that MPM fails to converge. The difference in these outcomes is due to the change in the parameter $\hat{A}$. Note that the maximum displacement of a material point is $\hat{A}/\pi$ and the distance of a material point to the boundary of an element is initially $h/4$. Thus, when the mesh size is smaller than $h = 4\hat{A}/\pi$, material points cross element boundaries during the numerical solution procedure. In Fig. 3a, the mesh sizes used are such that no material points cross cell boundaries and in Fig. 3b all of the mesh sizes used result in material points crossing cell boundaries. These results emphasize the conditional consistency of the standard MPM and may explain seemingly contradictory reports of convergence and of non-convergence of the method in the literature.

7 Improving the Material-Point Method

In the standard MPM, there are two main steps that account for inaccuracies of the algorithm. The primary error corresponds to the quadrature rule used in standard MPM to approximate the nodal forces or nodal masses. The material points are used as the quadrature points. Since the material points move, they can be located arbitrarily in the computational element, degrading accuracy. Indeed, for large enough motion relative to the grid size, the method is inconsistent. The other source of error corresponds to the mapping from the material-point information to the grid. As we saw in Sect. 3.1, the standard algorithm uses Shepard interpolation which is first order for arbitrarily placed material points. The improvements considered in this section are intended to achieve consistent second order accuracy of the method.

7.1 Function Reconstruction

Since the function reconstruction outlined in Sect. 3.1 can, in principle, be used to reconstruct functions to any order of accuracy from scattered data, we use this property to obtain second order accuracy in MPM. The guiding principle is to regard the material points as sampling points, and use the function reconstruction to approximate quantities as needed. For example, use the material-point velocity to reconstruct the velocity field and evaluate the reconstructed velocity field at grid nodes to start a time step. Similarly, reconstruct a stress field from the material-point values and evaluate the reconstructed stress field at quadrature points when computing internal forces.

To obtain the nodal velocity, standard MPM uses a shape function determined from the general form given in Eq. (9) using $r = 0$ with the weight function w_h chosen to be the linear hat function given in Eq. (21), weighted by material-point mass. With $r = 0$, in general, only constant functions are reproduced exactly. To increase the order of accuracy, the improved method uses $\psi_p^{[1]}(x)$ determined from the same general form, but using a polynomial basis consisting of the constant and linear monomials. There is also the choice of weight function; and to construct $\psi_p^{[1]}(x)$ we use the linear hat function, Eq. (21), without the mass weighting. The reconstructed function is then continuous and the consistency error is $O(h^2)$.

Smoother reconstructed functions can be achieved using higher order splines for the weight function. For example, the cubic spline would give a C^2 reconstructed function. The cubic spline expressed in natural coordinates is

$$\hat{s}_3(\xi) = \begin{cases} \frac{2}{3} - \xi^2 + \frac{1}{2}|\xi|^3 & |\xi| \le 1 \\ \frac{4}{3} - 2|\xi| + \xi^2 - \frac{1}{6}|\xi|^3 & 1 \le |\xi| \le 2 \\ 0 & \text{otherwise} \end{cases} . \tag{63}$$

To distinguish a shape function constructed with the cubic spline as weight from $\psi_p^{[1]}(x)$ which uses the linear hat function for the weight function, introduce the notation $C_p^{[1]}(x)$ to denote the former. Note that both shape functions have the same $O(h^2)$ consistency error since they are constructed to reproduce constant and linear functions. In addition to being smoother, the wider support of the shape function $C_p^{[1]}(x)$ is also useful if a situation arises where there is only one material point in an element; since, in that case, the moment matrix for $\psi_p^{[1]}(x)$ would be singular.

The linear hat function or the cubic spline are convenient weight functions for reproducing nodal values since it is easy to determine which nodes are in the support of the shape function associated with a material point. To determine function values at the centers of elements (the quadrature points for one-point quadrature), it is numerically simpler to use a quadratic spline rather than the linear hat function or cubic spline since the quadratic spline does not require logic to determine which element centers are in the support of the weight function of a given material point. The shape function $Q_p^{[1]}(x)$ used to reconstruct values at the quadrature points is the same order as $\psi_p^{[1]}(x)$ or $C_p^{[1]}$ as indicated by the superscript [1], but uses a different weight function. The weight used to determine $Q_p^{[1]}(x)$ is

$$\hat{s}_2(\xi) = \begin{cases} \frac{3}{4} - \xi^2 & |\xi| \le \frac{1}{2} \\ \frac{1}{2}(\frac{3}{2} - |\xi|)^2 & \frac{1}{2} \le |\xi| \le \frac{3}{2} \\ 0 & \text{otherwise} \end{cases} . \tag{64}$$

The quadratic and cubic spline weight functions render the reconstruction nonlocal to the element and they require the use of a structured (logically rectangular) grid.

7.1.1 Examples in One Dimension

We again consider the function $f(x) = \sin \pi x$ on [0, 1]. Divide the domain into $N_e = 8$ elements and use $N_p^e = 2$ randomly chosen material points per element as sampling points. Figure 4a shows the reconstructed values $f^R(x_I)$ at nodes, using

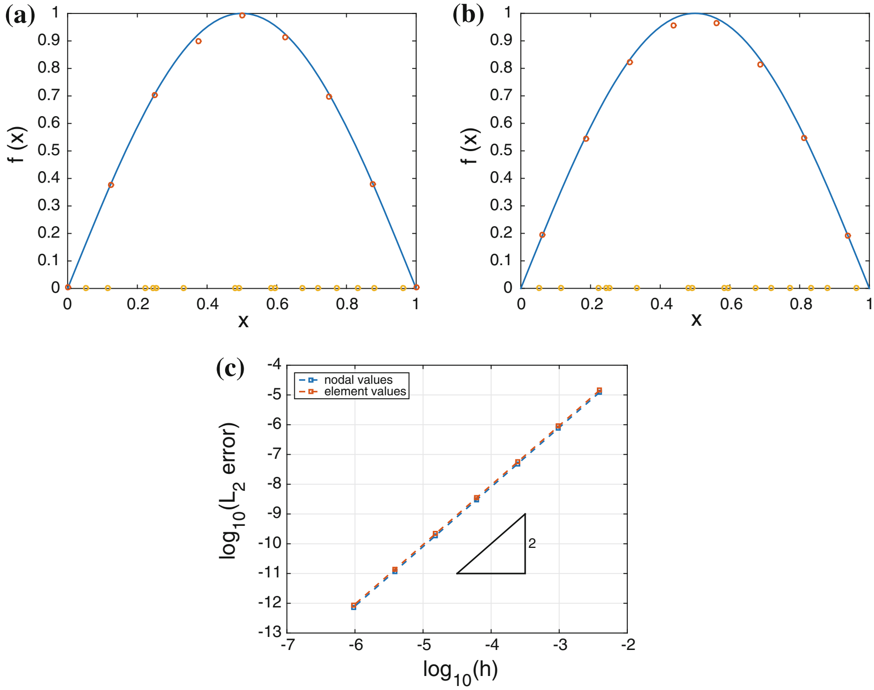


Fig. 4 Function reconstruction using the function$f(x) = \sin \pi x$ (*solid line*) with randomly chosen sampling points. The *circles* on the x-axis show the positions of the sampling points. The other *circles* in the plots show the reconstructed values **a** $f^R(x_I)$ using $\psi_p^{[1]}(x)$ and **b** $f^R(x_e)$ using $Q_p^{[1]}(x)$. **c** Shows the convergence rate with h for nodal and element values

$\psi^{[1]}(x)$ as the shape function. Similarly, Fig. 4b shows function values reconstructed at element centers, using $Q_p^{[1]}(x)$ as the shape function. The rate of convergence for both methods is shown in Fig. 4c. It is observed that the errors are nearly identical for both methods and both have $O(h^2)$ consistency errors.

7.2 Summary of the IMPM Algorithm

The main steps of IMPM algorithms are summarized as follows.

(1) Reconstruct velocity, stress and density from material-point data. The velocity is needed at grid nodes

$$v_I^{n-\frac{1}{2}} = v^R(x_I, t^{n-\frac{1}{2}}) = \sum_{p=1}^{N_p} \psi_p^{[1]}(x_I) v_p^{n-\frac{1}{2}}.$$

The density and the stress are needed at quadrature points for the elements in order to compute the mass matrix and the internal forces, respectively. For the one-dimensional problem, we use one-point quadrature with the quadrature point located at the center of the element. The reconstructed values are

$$\rho_e^n = \rho^R(x_e, t^n) = \sum_{p=1}^{N_p} Q_p^{[1]}(x_e)\rho_p^n,$$

$$\sigma_e^n = \sigma^R(x_e, t^n) = \sum_{p=1}^{N_p} Q_p^{[1]}(x_e)\sigma_p^n,$$

where x_e denotes the quadrature point.
Nodal forces and nodal masses are now assembled from the element contributions as in the standard finite element method. The internal and external forces are

$$f_I^{\text{int},n} = -\int_\Omega \sigma(x, t^n)\nabla N_I(x)\,dx \sim -\sum_{e=1}^{N_e} \nabla N_I(x_e)\sigma_e^n h, \tag{65}$$

$$f_I^{\text{ext},n} = \int_\Omega \rho(x, t^n) b(x, t^n) N_I(x)\,dx \sim \sum_{e=1}^{N_e} N_I(x_e)\rho_e^n b(x_e, t^n) h. \tag{66}$$

Similarly, the lumped nodal masses are

$$M_I^n = \int_\Omega \rho(x, t^n) N_I(x)\, dx \sim \sum_{e=1}^{N_e} \rho_e^n N_I(x_e) h. \tag{67}$$

(2) Solve the momentum equation on the finite element grid.

$$M_I^n a_I^n = f_I^{\mathrm{int},n} + f_I^{\mathrm{ext},n} \tag{68}$$

$$v_I^{n+\frac{1}{2}} = v_I^{n-\frac{1}{2}} + \Delta t^n a_I^n. \tag{69}$$

(3) Update the information on the material points. The update follows the same equations as the standard method given in Eq. (48) for the velocity, position and displacement. The density and deformation gradient are updated following Eq. (49).

(4) Generate a new grid.

7.3 Convergence Rate of Improved MPM

In this section we repeat the analysis of Sect. 6 using the improved method. We restrict attention to the larger deformation case where $\hat{A} = 0.1\pi$. Nodal values are obtained using $\psi_p^{[1]}$ and element values are obtained using $Q_p^{[1]}$. Figure 5 shows the logarithm of the L_2 error as a function of the logarithm of the mesh size for the material-point quantities, stress, velocity and displacement, and also for the grid stress, velocity and displacement. The mesh sizes used in this figure are $h = 2^{-(i+3)}$, $i = 1, 2, 3, 4$.

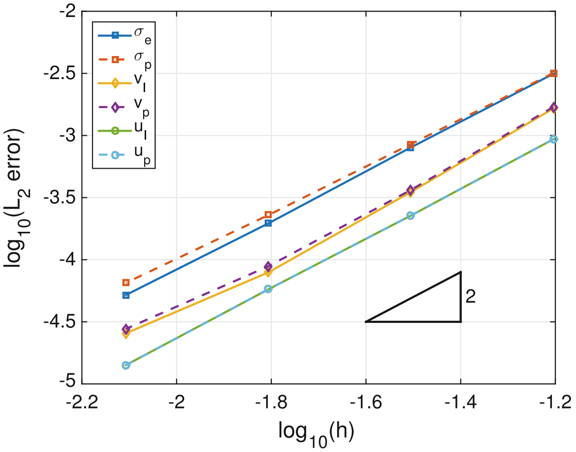


Fig. 5 Convergence of various computed quantities using the improved MPM for the problem in Sect. 5.1 with $\hat{A} = 0.1\pi$. The observed rates of convergence are σ_e (1.98), σ_p (1.87), v_I (2.02), v_p (1.98), u_i (2.01), u_p (2.01)

The observed accuracy is second order. This figure should be compared with Fig. 3b which shows failure of the standard method to converge for this large deformation case.

Similar behavior is obtained using $C_p^{[1]}$ for computing nodal values from the material point data.

8 Conclusions

This paper has presented a framework for improving the accuracy of MPM. The material points are viewed as providing data for a finite element method. In this finite element method, in essence, a new grid is used each step with the configuration at the n-th time step being the reference configuration for the step. Thus, it is necessary to initialize the values on the grid at the beginning of each step. The initialization is accomplished by evaluating a function at nodes that is reconstructed from the material-point data. This function reconstruction from scattered data appears in other contexts, notably other meshfree methods for continuum mechanics. We therefore draw on the literature to find methods that perform the reconstruction to any desired accuracy. A potential advantage to using MPM over other meshfree methods that rely on the same function reconstruction arises from the use of the finite element mesh to take gradients and perform quadrature. In MPM the shape functions for function reconstruction are not used as the basis functions for solving the equations of motion, avoiding costly differentiation of these functions. A potential advantage of MPM over using the finite element method on its own is in avoiding problematic mesh distortion when materials undergo large deformations.

To illustrate the ideas for improving MPM in one space dimension, two-noded displacement elements have been used with function reconstruction based on reproducing linear polynomials. B-splines have been used as the weight function in building the shape functions used for the function reconstruction. We have also used four-noded quadrilaterals in two space dimensions and will report the results elsewhere. Clearly, higher-order element formulations need to be paired with higher-order function reconstruction for overall higher-order accuracy of the method. Also, if higher-order accuracy is needed in time, then the leapfrog scheme considered in this paper also needs to be replaced with a method of appropriate order.

For small deformations relative to the grid size and equally-spaced, or nearly equally-spaced material points, it is possible to obtain apparent second-order accuracy with the standard MPM algorithm. We have demonstrated that second-order accuracy is possible with large deformations using the improved method. However, there is more work to be done to achieve second-order accuracy in general. The principal remaining issue is related to the fact that MPM represents the geometry of a body implicitly through the distribution of material points. This representation makes the treatment of general boundary conditions difficult, especially if one is interested in accuracy greater than first order. We also find apparent instabilities arise if we drive the simulations to very small mesh sizes. The source of the instability is unknown at this time. Two potential sources are ill-conditioning of the moment matrices or a ringing instability [5].

We have shown a path to obtaining methods with specified accuracy. It is of interest to provide a mathematical analysis of the methods to make precise the conditions for convergence at a given rate. This topic, as well as a stability analysis of the methods, remain topics of continued research.

Acknowledgments This work was partially supported by the National Science Foundation under grant ARC 1023667 to the University of New Mexico.

References

1. Arroyo, M., & Ortiz, M. (2006). Local maximum-entropy approximation schemes: A seamless bridge between finite elements and meshfree methods. *International Journal for Numerical Methods in Engineering*, *65*(13), 2167–2202.
2. Atluri, S. N., & Zhu, T. (1998). A new meshless local Petrov-Galerkin (MLPG) approach in computational mechanics. *Computational Mechanics*, *22*, 117–127.
3. Beltschko, T., Lu, Y. Y., & Gu, L. (1994). Element-free Galerkin methods. *International Journal of Numerical Methods in Engineering*, *37*, 229–256.
4. Belytschko, T., Krongauz, Y., Organ, D., Fleming, M., & Krysl, P. (1996). Meshless methods: An overview and recent developments. *Computer Methods in Applied Mechanics and Engineering*, *139*, 3–47.
5. Brackbill, J. U. (1988). The ringing instability in particle-in-cell calculations of low-speed flow. *Journal of Computational Physics*, *75*, 469–492.
6. Brackbill, J. U., & Ruppel, H. M. (1986). FLIP: A method for adaptively zoned, particle-in-cell calculations of fluid flows in two dimensions. *Journal of Computational Physics*, *65*, 314–343.
7. Duan, Q., Gao, X., Wang, B., Li, X., Zhang, H., Belytschko, T., et al. (2014). Consistent element-free Galerkin method. *International Journal for Numerical Methods in Engineering*, *99*, 79–101.
8. Duarte, A. C., & Oden, J. T. (1996). H-p clouds–an h-p meshless method. *Numerical Methods for Partial Differential Equations*, *12*(6), 673–705.
9. Han, W., & Meng, X. (2001). Error analysis of the reproducing kernel particle method. *Computer Methods in Applied Mechanics and Engineering*, *190*, 6157–6181.
10. Harlow, F. H. (1957). Hydrodynamic problems involving large fluid distortions. *Journal of the Association for Computing Machinery*, *4*, 137.
11. Li, S., & Liu, W. K. (2002). Meshfree and particle methods and their applications. *Applied Mechanics Reviews*, *55*(1), 1–34.
12. Li, B., Habbal, F., & Ortiz, M. (2010). Optimal transportation meshfree approximation schemes for fluid and plastic flows. *International Journal for Numerical Methods in Engineering*, *83*, 1541–1579.
13. Liu, W., Jun, S., Li, S., Adee, J., & Beltschko, T. (1995). Reproducing kernel particle methods for structural dynamics. *International Journal of Numerical Methods in Engineering*, *38*, 1655–1680.
14. Liu, W. K., Jun, S., & Zhang, Y. (1995). Reproducing kernel particle methods. *International Journal for Numerical Methods in Engineering*, *20*, 1081–1106.
15. Melenk, J. M., & Babuska, I. (1996). Partition of unity finite element method. *Computer Methods in Applied Mechanics and Engineering*, *139*, 314–389.
16. Nguyen, V. P., Rabczuk, T., Bordas, S., & Duflot, M. (2008). Meshless methods: A review and computer implementation aspects. *Mathematics and Computers in Simulation*, *79*(3), 793.

17. Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings—1968 ACM National Conference* (pp. 517–524).
18. Steffen, M., Kirby, R. M., & Berzins, M. (2008). Analysis and reduction of quadrature errors in the material point method (MPM). *International Journal for Numerical Methods in Engineering*, *76*(6), 922–948.
19. Steffen, M., Kirby, R. M., & Berzins, M. (2010). Decoupling and balancing of space and time errors in the material point method (MPM). *International Journal for Numerical Methods in Engineering*, *82*, 1207–1243.
20. Sulsky, D., Chen, Z., & Schreyer, H. L. (1995). Application of a particle-in-cell method to solid mechanics. *Computer Physics Communications*, *87*, 236–252.

Meshfree Methods Applied to Consolidation Problems in Saturated Soils

Pedro Navas, Susana López-Querol, Rena C. Yu and Bo Li

Abstract A meshfree numerical model, based on the principle of Local Maximum Entropy, with a B-Bar based algorithm to avoid instabilities, is applied to solve consolidation problems in saturated soils. This numerical scheme has been previously validated for purely elasticity problems without water (mono phase), as well as for steady seepage in elastic porous media. Hereinafter, the model is validated for well known consolidation theoretical problems, both static and dynamic, with known analytical solutions. For several examples, the solutions obtained with the new code are compared to PLAXIS (commercial software). Finally, after validated, solutions for dynamic radial consolidation and sinks, which have not been found in the literature, are presented as a novelty. This new numerical approach is demonstrated to be feasible for this kind of problems in porous media.

1 Introduction

The settlement of saturated soils under loading is caused by a gradual interchange between pore pressure and effective stress. Immediately after external loadings are applied to a saturated soil domain, all the external pressure transfers to water, and

P. Navas · R.C. Yu (✉)
School of Civil Engineering, University of Castilla La-Mancha, Avda. Camilo
Jose Cela s/n, 13071 Ciudad Real, Spain
e-mail: pedro.navas@uclm.es

R.C. Yu
e-mail: rena@uclm.es

S. López-Querol
Department of Civil, Environmental and Geomatic Engineering, University
College London, Gower Street, London WC1E 6BT, UK
e-mail: s.lopez-querol@ucl.ac.uk

B. Li
Department of Mechanical and Aerospace Engineering, Case Western
Reserve University, Cleveland, OH 44106, USA
e-mail: bo.li4@case.edu

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_11

some time is required for drainage to take place. When this drainage (i.e. dissipation of excess pore pressures) is complete, the solid phases totally takes the external pressure. This process is known as *soil consolidation* [1]. The implementation of the Biot's equations [2] is a well-known way to solve problems in porous media from a macro-scale point of view. The advantage of this method is the possibility of accounting for coupling between the fluid phase and the solid skeleton. The $u - p_w$ formulations, where u denotes the solid phase displacement, and p_w is the pore fluid pressure [3], have traditionally been employed for simulating coupled problems in saturated porous media since the final equations work with less degrees of freedom (three in 2D, four in 3D problems) compared with that of a complete formulation. The recent $u - w$ formulation, where w represents the relative fluid displacement with respect to the solid phase, which is usually referred as the displacement-based or complete formulation, has been employed in several numerical schemes (López-Querol et al. [4], and recently adopted by Cividini and Gioda [5]). Such a methodology is assumed in this work, first for its simplicity in imposing impervious boundary conditions compared to the $u - p_w$ approaches; second, as the free surface comes out naturally as the zero-pressure contour, no detection algorithm is necessary; third, it facilities the modelling of large and/or nonlinear deformations of the solid phase as well as the possible separation between the solid and fluid phases in the case of local failure (liquefaction or slope instability). Since meshfree numerical schemes have been known to perform particularly well in the regime of large deformations, we endeavour to apply such schemes to coupled problems in saturated porous media, using the $u - w$ formulation.

There are many different flavours of meshfree methods available. The present research has been carried out using the principle of maximum entropy [6], the shape functions developed by Arroyo and Ortiz [7], in particular, the OTM framework [8], for its numerous advantages in comparison with its alternatives. For example, the exact mass transport, the satisfaction of the continuity equation, exact linear and angular momentum conservation in order to solve different problems as spurious modes, tensile instabilities and unknown convergence or stability properties. Since the deformation and velocity fields are interpolated from nodal values using max-ent shape functions, the Kronecker-delta property at the boundary makes it possible for the direct imposition of essential boundary conditions. In the current work, an Eulerian framework is employed to solve the Biot's equations for porous media. In addition, the parameters pertinent to the local maximum entropy are obtained efficiently and independent of the support size through the Nelder-Mead algorithm [9].

Locking in near-incompressible materials is not unusual for the numerical methods based on either finite elements approaches [10–17] or meshfree approximation schemes [18]. In the case of flow through saturated media, in a displacement formulation, since both the undrained soil phase and the fluid phase are nearly incompressible, locking may also occur. The recent approach developed by Ortiz and Sukumar [18], avoids locking by averaging the volumetric part of the strain tensor with the value of the pressure. However, since the pressure term is part of the constitutive model employed, the constitutive model is necessarily modified. As an alternative, we implement a volumetric-strain average instead of a pressure average approach.

The proposed algorithm is independent of the constitutive model employed and it is generically applicable to solve other locking problems. The idea behind is inspired by that of Hughes [10] and the posterior developments of the B-Bar method [16]. The specific strategy is analogous to that of the diamond elements of Hauret, Kuhl and Ortiz [15] and it is the first B-Bar implementation for a pure displacement approach within the framework of meshfree methods. It is also straightforward to be extended for finite deformations and nonlinear applications. Extension of the formulation by the authors in [19] to axisymmetric framework is presented in the current work.

The rest of the paper is organised as follows. The mathematical framework, including the B-Bar based algorithm is presented next. Applications to various consolidation problems are illustrated in Sect. 3. Relevant conclusions are drawn in Sect. 4.

2 Mathematical Framework

In this section, we first summarise the governing equations for unconfined seepage problems, in particular, the Biot's equations, formulated in a $u - w$ framework, which have been successfully utilised by López-Querol et al. [20, 21], Cividini and Gioda [5]; next, the B-bar implementation in axisymmetric framework for elastic and porous media are given in detail.

2.1 The Biot's Equations: A $\boldsymbol{u} - \boldsymbol{w}$ Formulation

The Biot's equations [22] are based on formulating the mechanical behaviour of a solid-fluid mixture, the coupling between different phases, and the continuity of flux through a differential domain of saturated porous media. For clarity, bold symbols for notation of vectors and matrices, and regular letters to denote scalar variables, are used. Let ρ and ρ_f respectively represent the mixture and fluid phase densities; $\boldsymbol{b}$ and κ stand for the external acceleration vector and the permeability coefficient in [m^3· s/kg] (in civil engineering, however, the notion of hydraulic conductivity, $k = \kappa\rho_f g$, in [m/s], is often used instead), the three equations of Biot, which represent the mixture equilibrium, the fluid phase equilibrium and the continuity equation respectively, are expressed as follows:

$$\boldsymbol{S}^{\boldsymbol{T}}\boldsymbol{d\sigma} - \rho\boldsymbol{d\ddot{u}} - \rho_f\boldsymbol{d\ddot{w}} + \rho\boldsymbol{db} = 0, \tag{1}$$

$$-\nabla dp_w - \kappa^{-1}\boldsymbol{d\dot{w}} - \rho_f\boldsymbol{d\ddot{u}} - \frac{\rho_f}{n}\boldsymbol{d\ddot{w}} + \rho_f\boldsymbol{db} = 0, \tag{2}$$

$$\nabla \cdot \boldsymbol{d\dot{w}} + \boldsymbol{m} \cdot \boldsymbol{d\dot{\varepsilon}}^s + \frac{d\dot{p}_w}{Q} = 0, \tag{3}$$

where $\boldsymbol{S}$ is a differential operator, $\boldsymbol{u}$ is the displacement vector of the solid skeleton and $\boldsymbol{w}$ is the relative displacement of the fluid phase with respect to the solid one. By denoting $\boldsymbol{U}$ as the absolute displacement of the fluid phase, $\boldsymbol{w}$ is related with $\boldsymbol{U}$ through the soil porosity, n, as follows:

$$\boldsymbol{w} = n(\boldsymbol{U} - \boldsymbol{u}). \tag{4}$$

Additionally in Eq. (3), $\boldsymbol{m}$ represents the unit matrix expressed in Voigt form, whereas Q stands for the mixture compressibility, which is calculated as

$$Q = \left[K_s^{-1}(1-n) + nK_f^{-1}\right]^{-1}, \tag{5}$$

where K_s and K_f are the bulk modulus of the solid grains and the compressive modulus of the fluid phase.

A 2D approach is considered in the derivations presented hereinafter, therefore the differential operator, $\boldsymbol{S}$, and the unit matrix $\boldsymbol{m}$ are written as

$$\boldsymbol{S} = \begin{pmatrix} \frac{\partial}{\partial x} & 0 \\ 0 & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} \end{pmatrix}, \ \boldsymbol{m} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}. \tag{6}$$

Assuming tensile stresses (except pore pressure p_w, which is positive for compression) and strains as positive, the Terzaghi's effective stress [23] is defined as follows

$$\boldsymbol{\sigma} = \boldsymbol{\sigma'} - p_w\boldsymbol{m}, \tag{7}$$

where $\boldsymbol{\sigma'}$ and $\boldsymbol{\sigma}$ are the respective vectorial form in Voigt notation for the effective and total stress tensor. For the case of linear elasticity, the incremental relationship between stresses and strains is governed by:

$$\boldsymbol{d\sigma'} = \boldsymbol{D^e d\varepsilon^s}, \tag{8}$$

where $\boldsymbol{D^e}$ denotes the elastic tensor, which under plane strain conditions, it is given by:

$$\boldsymbol{D^e} = \frac{\lambda}{\nu}\begin{pmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & \frac{1-2\nu}{2} \end{pmatrix} \tag{9}$$

where ν is the Poisson's ratio, λ, the first constant of Lamé.

Rearranging the above equations, Eq. (1) can be re-written as

$$\boldsymbol{S^T D^e S du} - \nabla dp_w - \rho \boldsymbol{d\ddot{u}} - \rho_f \boldsymbol{d\ddot{w}} + \rho \boldsymbol{db} = 0. \tag{10}$$

In the $u - w$ approach, also known as the *complete* formulation (no additional assumption is required under plane strain conditions), each node has four degrees of freedom, $\boldsymbol{u}$ and $\boldsymbol{w}$ (two components each in 2D problems) and the scalar p_w, the pore pressure, is obtained afterwards. By comparison, in the traditional $u - p_w$ formulation, each node has only three degrees of freedom in 2D, but results in complications in imposing impervious boundary conditions.

Integrating Eq. (3) in time, and substituting dp_w in Eqs. (10) and (2), it yields:

$$\boldsymbol{S^T D^e S du} + Q\nabla\left(\nabla^{\boldsymbol{T}}\boldsymbol{du}\right) + Q\nabla\left(\nabla^{\boldsymbol{T}}\boldsymbol{dw}\right) - \rho \boldsymbol{d\ddot{u}} - \rho_f \boldsymbol{d\ddot{w}} + \rho \boldsymbol{db} = 0, \tag{11}$$

$$Q\nabla\left(\nabla^{\boldsymbol{T}}\boldsymbol{du}\right) + Q\nabla\left(\nabla^{\boldsymbol{T}}\boldsymbol{dw}\right) - k^{-1}\boldsymbol{d\dot{w}} - \rho_f \boldsymbol{d\ddot{u}} - \frac{\rho_f}{n}\boldsymbol{d\ddot{w}} + \rho_f \boldsymbol{db} = 0. \tag{12}$$

Note that an isotropic medium is assumed in the above equations. The final system of equations, once the elementary matrices have been assembled, can be expressed as:

$$\boldsymbol{Kdu} + \boldsymbol{Cd\dot{u}} + \boldsymbol{Md\ddot{u}} = \boldsymbol{df}, \tag{13}$$

where $\boldsymbol{K}$, $\boldsymbol{C}$ and $\boldsymbol{M}$ respectively denote stiffness, damping and mass matrices, $\boldsymbol{du}$ represents the vector of unknowns (containing both the solid phase and fluid displacements, $\boldsymbol{u}$ and $\boldsymbol{w}$), expressed incrementally, and $\boldsymbol{df}$ is the increment of the external forces vector, containing gravity acceleration, as well as boundary conditions for nodal forces.

2.2 B-Bar Formulation in Elastic Axisymmetric Problems

In axisymmetric problems, x direction is changed by r, y changes to z. Due to this fact, the shape function based on the principle of Local Maximum Entropy is similar to that of the 2D case. Consequently, the new displacement vector is calculated with the following equation:

$$\begin{bmatrix} u_r \\ u_z \end{bmatrix} = \begin{bmatrix} N_1 & 0 & N_2 & 0 & \cdots \\ 0 & N_1 & 0 & N_2 & \cdots \end{bmatrix} \begin{bmatrix} u_{r1}^h \\ u_{z1}^h \\ u_{r2}^h \\ u_{z2}^h \\ \vdots \end{bmatrix}, \tag{14}$$

where the superscript ^{h} denotes discrete nodal values. In an axisymmetric problem, a different ε matrix is obtained according to [24]:

$$\begin{bmatrix} \varepsilon_r \\ \varepsilon_z \\ \varepsilon_\theta \\ \gamma_{rz} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial r} & 0 \\ 0 & \frac{\partial}{\partial z} \\ \frac{1}{r} & 0 \\ \frac{\partial}{\partial z} & \frac{\partial}{\partial r} \end{bmatrix} \begin{bmatrix} u_r \\ u_z \end{bmatrix}. \tag{15}$$

Voigt notation is assumed in order to obtain the final B-bar matrix. The process to obtain B matrix in index notation is:

$$\varepsilon_l = S_{lj} u_j = S_{lj} N_{jk} u_k^h = B_{lk} u_k^h.$$

Thereby, the B matrix is the following one:

$$\mathbf{B} = \left[\begin{array}{cc|cc|c} \frac{\partial N_1}{\partial r} & 0 & \frac{\partial N_2}{\partial r} & 0 & \\ 0 & \frac{\partial N_1}{\partial z} & 0 & \frac{\partial N_2}{\partial z} & \cdots \\ \frac{N_1}{r} & 0 & \frac{N_2}{r} & 0 & \\ \frac{\partial N_1}{\partial z} & \frac{\partial N_1}{\partial r} & \frac{\partial N_2}{\partial z} & \frac{\partial N_2}{\partial r} & \end{array} \right]. \tag{16}$$

If σ is required, the following equation should be employed:

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_r \\ \sigma_z \\ \sigma_\theta \\ \tau_{rz} \end{bmatrix} = \boldsymbol{D\varepsilon}, \tag{17}$$

where

$$\boldsymbol{D} = \frac{\lambda}{\nu} \begin{bmatrix} 1-\nu & \nu & \nu & 0 \\ \nu & 1-\nu & \nu & 0 \\ \nu & \nu & 1-\nu & 0 \\ 0 & 0 & 0 & \frac{1-2\nu}{2} \end{bmatrix}. \tag{18}$$

Stiffness matrix is calculated by taking into account that the volume integral is extended around the whole ring of material as follows:

$$\mathbf{K^p} = 2\pi \int \mathbf{B^T D B}\, r\, dr\, dz, \tag{19}$$

where the superscript ^{p} represents the fact that the matrix is calculated for each material point within a patch. The final expression is written as:

$$\mathbf{K}^p = 2\pi \overline{\mathbf{B}}^{\mathbf{T}} \mathbf{D} \overline{\mathbf{B}}\, \bar{r} A, \tag{20}$$

where A is the associated area of the material point.

The external forces in Eq. (13) are calculated in the same way:

$$\mathbf{f} = \begin{bmatrix} 2\pi r f_r \\ 2\pi r f_z \end{bmatrix}, \tag{21}$$

where f_r and f_z respectively denote radial and vertical components of the external force.

If a B-Bar based algorithm is implemented in this problem, the starting point is similar to the plane strain one, which is based on the transformation of the $\boldsymbol{\varepsilon}$ tensor to the $\overline{\boldsymbol{\varepsilon}}$ tensor, a tensor where the volumetric part is obtained as an average of the neighbour integration points, as we can see in the following equation:

$$\overline{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} - \frac{1}{d} tr(\boldsymbol{\varepsilon})\mathbf{I} + \frac{1}{d}\overline{[tr(\boldsymbol{\varepsilon})]}^p \mathbf{I}, \tag{22}$$

where d is the dimension of the problem, in this case 3; and $\overline{[tr(\boldsymbol{\varepsilon})]}^p$ is the average trace of $\boldsymbol{\varepsilon}$ of the neighbour integration points in a chosen patch, calculated by:

$$\overline{[tr(\boldsymbol{\varepsilon})]}^p = \sum_{i=1}^{Nb} tr[\boldsymbol{\varepsilon}^{(i)}]\, w_i. \tag{23}$$

In addition, the trace could be obtained from the strain vector as:

$$\varepsilon_x + \varepsilon_y = \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_r \\ \varepsilon_z \\ \varepsilon_\theta \\ \gamma_{rz} \end{bmatrix} = \varepsilon_{kk} = m_k \varepsilon_k. \tag{24}$$

Rearranging the above equation using the B-matrix,

$$\varepsilon_{ll} = m_l \varepsilon_l = m_l B_{lk} u_k^h = T_k\; u_k^h, \tag{25}$$

where

$$T = \left[\, \tfrac{\partial N_1}{\partial r} + \tfrac{N_1}{r} \;\; \tfrac{\partial N_1}{\partial z} \,\middle|\, \tfrac{\partial N_2}{\partial r} + \tfrac{N_2}{r} \;\; \tfrac{\partial N_2}{\partial z} \,\middle|\, \cdots \right]. \tag{26}$$

Thus, the final lth-component of the tensor $\overline{\boldsymbol{\varepsilon}}$ (in Voigt form) for a single integration point i is calculated in the same way as in 2D problems:

$$\begin{aligned} \overline{\varepsilon}_l^{(i)} &= \left[B_{lk}^{(i)} - \frac{1}{d} m_l \left(T_k^{(i)} - \sum_{j=1}^{Nb} [T_k^{(j)} w^{(j)}] \right) \right] u_k^h \\ &= \overline{B}_{lk}^{(i)} u_k^h. \end{aligned}$$

2.3 B-Bar Implementation in $u - w$ Axisymmetric Problems

In order to apply the B-Bar method in a multiphase problem, we need to define a constitutive matrix first to relate stress with strain or displacements of the different phases. The proposed problem is the $u - w$ problem with soil and water phases. The effective stress tensor is calculated the same way as Eq. (7), if linear elasticity is assumed, the relationship between stresses and strains, expressed in its incremental form, is governed by Eq. (8), and p_w is obtained with the third Biot's equation:

$$\nabla \cdot d\dot{\boldsymbol{w}} + \boldsymbol{m}^T d\dot{\boldsymbol{\varepsilon}}^s + \frac{d\dot{p}_w}{Q} = 0$$

$$d\dot{p}_w = -Q \, [\nabla \cdot d\dot{\boldsymbol{w}} + \boldsymbol{m}^T d\dot{\boldsymbol{\varepsilon}}^s]$$

The final stress equation will be:

$$\boldsymbol{\sigma} = \boldsymbol{D^e} \; \boldsymbol{\varepsilon}^s + Q \; [tr(\boldsymbol{\varepsilon}^s) + tr(\boldsymbol{\varepsilon}^w)] \, \mathbf{I}.$$

If:

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon^s \\ \\ \varepsilon^w \end{bmatrix} = \begin{bmatrix} \varepsilon_r^s \\ \varepsilon_z^s \\ \varepsilon_\theta^s \\ \gamma_{rz}^s \\ \varepsilon_r^w \\ \varepsilon_z^w \\ \varepsilon_\theta^w \end{bmatrix} = \boldsymbol{S} \, \boldsymbol{u}$$

where $\boldsymbol{S}$ is the derivative matrix operator and $\boldsymbol{u}$ is a vector of displacements of both phases:

$$\begin{bmatrix} \varepsilon_r^s \\ \varepsilon_z^s \\ \varepsilon_\theta^s \\ \gamma_{rz}^s \\ \varepsilon_r^w \\ \varepsilon_z^w \\ \varepsilon_\theta^w \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial r} & 0 & 0 & 0 \\ 0 & \frac{\partial}{\partial z} & 0 & 0 \\ \frac{1}{r} & 0 & 0 & 0 \\ \frac{\partial}{\partial z} & \frac{\partial}{\partial r} & 0 & 0 \\ 0 & 0 & \frac{\partial}{\partial r} & 0 \\ 0 & 0 & 0 & \frac{\partial}{\partial z} \\ 0 & 0 & \frac{1}{r} & 0 \end{bmatrix} \begin{bmatrix} u_r \\ u_z \\ w_r \\ w_z \end{bmatrix}.$$

In addition, the summation of traces of strain could be done with a $\boldsymbol{m^*}$ operator:

$$tr(\boldsymbol{\varepsilon}^s) + tr(\boldsymbol{\varepsilon}^w) = \boldsymbol{m}^T\boldsymbol{\varepsilon} = \begin{bmatrix}1 & 1 & 1 & 0 & 1 & 1 & 1\end{bmatrix}\begin{bmatrix}\varepsilon_r^s\\ \varepsilon_z^s\\ \varepsilon_\theta^s\\ \gamma_{rz}^s\\ \varepsilon_r^w\\ \varepsilon_z^w\\ \varepsilon_\theta^w\end{bmatrix}.$$

Thus, in Voigt notation:

$$\begin{aligned}\boldsymbol{\sigma} &= \boldsymbol{D}^{e*}\,\boldsymbol{\varepsilon} + Q\,\boldsymbol{m}^T\boldsymbol{\varepsilon}\;\;\boldsymbol{m} = (\boldsymbol{D}^{e*} + Q\,\boldsymbol{m}^T\boldsymbol{m})\,\boldsymbol{\varepsilon} = \boldsymbol{D}^{\boldsymbol{u-w}}\,\boldsymbol{\varepsilon}\\ &= \begin{bmatrix}\frac{\lambda(1-\nu)}{\nu}+Q & \lambda+Q & \lambda+Q & 0 & Q & Q & Q\\ \lambda+Q & \frac{\lambda(1-\nu)}{\nu}+Q & \lambda+Q & 0 & Q & Q & Q\\ \lambda+Q & \lambda+Q & \frac{\lambda(1-\nu)}{\nu}+Q & 0 & Q & Q & Q\\ 0 & 0 & 0 & \mu & 0 & 0 & 0\\ Q & Q & Q & 0 & Q & Q & Q\\ Q & Q & Q & 0 & Q & Q & Q\\ Q & Q & Q & 0 & Q & Q & Q\end{bmatrix}\begin{bmatrix}\varepsilon_r^s\\ \varepsilon_z^s\\ \varepsilon_\theta^s\\ \gamma_{rz}^s\\ \varepsilon_r^w\\ \varepsilon_z^w\\ \varepsilon_\theta^w\end{bmatrix}.\end{aligned}$$

If the problem needs a B-Bar based algorithm, it is necessary to calculate the average value of $\boldsymbol{\varepsilon}$. Thus, the main equation yields:

$$\overline{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} - \frac{1}{d}tr(\boldsymbol{\varepsilon}^s)\mathbf{I} + \frac{1}{d}[tr(\boldsymbol{\varepsilon}^s)]^p\mathbf{I} - \frac{1}{d}tr(\boldsymbol{\varepsilon}^w)\mathbf{I} + \frac{1}{d}[tr(\boldsymbol{\varepsilon}^w)]^p\mathbf{I}.$$

In Voigt notation the equation, the lth-component of the strain tensor is:

$$\overline{\varepsilon_l} = \varepsilon_l + \frac{1}{d}\left(-\varepsilon_{kk}\;m_l^s + \sum_{j=1}^{Nb}[\varepsilon_{kk}^{(j)}w^{(j)}]\;m_l^s - \varepsilon_{kk}^w\;m_l^w + \sum_{j=1}^{Nb}[\varepsilon_{kk}^{w(j)}w^{(j)}]\;m_l^w\right),$$

where

$$\varepsilon_{kk}^s = m_k^s\,\varepsilon_k = \begin{bmatrix}1 & 1 & 1 & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\varepsilon_r^s\\ \varepsilon_z^s\\ \varepsilon_\theta^s\\ \gamma_{rz}^s\\ \varepsilon_r^w\\ \varepsilon_z^w\\ \varepsilon_\theta^w\end{bmatrix},\ \varepsilon_{kk}^w = m_k^w\,\varepsilon_k = \begin{bmatrix}0 & 0 & 0 & 0 & 1 & 1 & 1\end{bmatrix}\begin{bmatrix}\varepsilon_r^s\\ \varepsilon_z^s\\ \varepsilon_\theta^s\\ \gamma_{rz}^s\\ \varepsilon_r^w\\ \varepsilon_z^w\\ \varepsilon_\theta^w\end{bmatrix}.$$

Additionally, we know that the strain tensor in Voigt notation is:

$$\varepsilon_l = S_{lj}u_j = S_{lj}N_{jk}u_k^h = B_{lk}u_k^h$$

where, in this case, yields:

$$\begin{bmatrix} \varepsilon_r \\ \varepsilon_z \\ \varepsilon_\theta \\ \gamma_{rz} \\ \varepsilon_r^w \\ \varepsilon_z^w \\ \varepsilon_\theta^w \end{bmatrix} = \left[\begin{array}{cccc|cccc|c} \frac{\partial N_1}{\partial r} & 0 & 0 & 0 & \frac{\partial N_2}{\partial r} & 0 & 0 & 0 & \\ 0 & \frac{\partial N_1}{\partial z} & 0 & 0 & 0 & \frac{\partial N_2}{\partial z} & 0 & 0 & \\ \frac{N_1}{r} & 0 & 0 & 0 & \frac{N_2}{r} & 0 & 0 & 0 & \\ \frac{\partial N_1}{\partial z} & \frac{\partial N_1}{\partial r} & 0 & 0 & \frac{\partial N_2}{\partial z} & \frac{\partial N_2}{\partial r} & 0 & 0 & \cdots \\ 0 & 0 & \frac{\partial N_1}{\partial r} & 0 & 0 & 0 & \frac{\partial N_2}{\partial r} & 0 & \\ 0 & 0 & 0 & \frac{\partial N1}{\partial z} & 0 & 0 & 0 & \frac{\partial N_2}{\partial z} & \\ 0 & 0 & \frac{N_1}{r} & 0 & 0 & 0 & \frac{N_2}{r} & 0 & \end{array}\right] \begin{bmatrix} u_r^{(1)} \\ u_z^{(1)} \\ w_r^{(1)} \\ w_z^{(1)} \\ u_r^{(2)} \\ u_z^{(2)} \\ w_r^{(2)} \\ w_z^{(2)} \\ \vdots \end{bmatrix}.$$

In order to calculate ε_{ll}, the above equation can be rearranged as follows:

$$\begin{aligned} \varepsilon_{ll}^s &= m_l^s \varepsilon_l = m_l B_{lk}\, u_k^h = T_k^s\, u_k^h, \\ \varepsilon_{ll}^w &= m_l^w \varepsilon_l = m_l B_{lk}\, u_k^h = T_k^w\, u_k^h \end{aligned}$$

where

$$T^s = \left[\, \frac{\partial N_1}{\partial r} + \frac{N_1}{r} \;\; \frac{\partial N_1}{\partial z} \;\; 0 \;\; 0 \middle| \frac{\partial N_2}{\partial r} + \frac{N_2}{r} \;\; \frac{\partial N_2}{\partial z} \;\; 0 \;\; 0 \middle| \cdots \right],$$

$$T^w = \left[0 \;\; 0 \;\; \frac{\partial N_1}{\partial r} + \frac{N_1}{r} \;\; \frac{\partial N_1}{\partial z} \middle| 0 \;\; 0 \;\; \frac{\partial N_2}{\partial r} + \frac{N_2}{r} \;\; \frac{\partial N_2}{\partial z} \middle| \cdots \right].$$

Thus, the final lth-component for the new strain tensor $\overline{\varepsilon}$ at a single integration point i in Voigt notation is calculated as:

$$\begin{aligned} \overline{\varepsilon_l}^{(i)} &= B_{lk}^{(i)} u_k^h - \frac{1}{d} m_l^s \left(T_k^{s(i)} u_k^h - \sum_{j=1}^{Nb} [T_k^{s(j)} w^{(j)}] u_k^h \right) \\ &\qquad - \frac{1}{d} m_l^w \left(T_k^{w(i)} u_k^h - \sum_{j=1}^{Nb} [T_k^{w(j)} w^{(j)}] u_k^h \right) \\ &= \left[B_{lk}^{(i)} - \frac{1}{d} m_l^s \left(T_k^{s(i)} - \sum_{j=1}^{Nb} [T_k^{s(j)} w^{(j)}] \right) \right. \\ &\qquad \left. - \frac{1}{d} m_i^w \left(T_k^{w(i)} - \sum_{j=1}^{Nb} [T_k^{w(j)} w^{(j)}] \right) \right] u_k^h \\ &\equiv \overline{B_{lk}}\, u_k^h. \end{aligned}$$

3 Application to Consolidation of Soils

As mentioned before, the settlement of saturated soils under loading is caused by a gradual interchange between pore pressure and effective stress. In this Section, we apply the above developed methodology for consolidation of soils at three different configurations: one dimensional case, radial consolidation and consolidation with sinks. Both static and dynamic scenarios are studied. The obtained solutions are compared with analytical or available numerical solutions.

3.1 Consolidation of a Column of Soil: Static, One Dimensional Case

In this case, the analytical solution for this problem is available, and thus it is compared with the solution proposed by the present method. Although the analytical solution is presented in non-dimensional terms, the geometry of the problem carried out is shown in Fig. 1. It consists of a column of 30 m of soil resting on an impermeable rigid base layer and loaded by a vertical, homogeneous loading at the top. The lateral displacements are restricted for both the solid and fluid phase. At the base layer, the solid phase is fixed, whereas the vertical movement of the fluid phase is

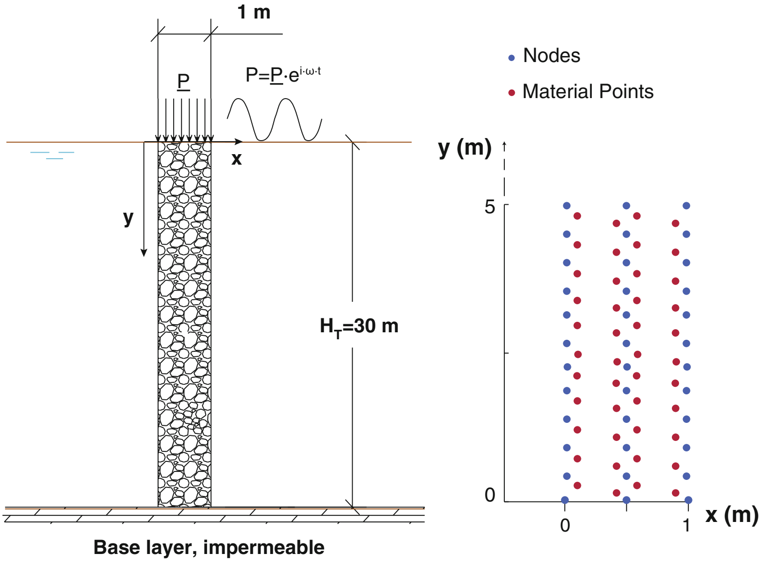
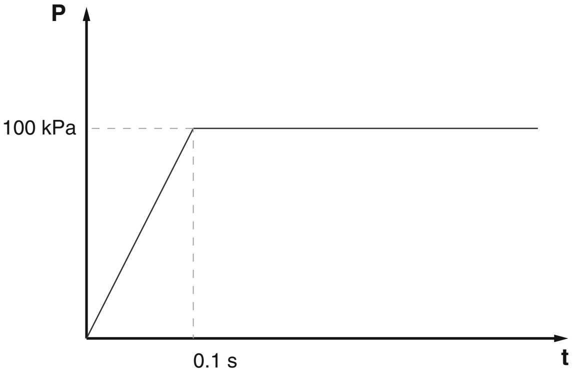


Fig. 1 Geometry, loading condition of the consolidation column, and the discretised nodes and material points (shown for the first 5 m only). The same geometry has been employed for both static and dynamic simulations

Fig. 2 Loading history in a monotonic problem



prevented. The column is discretised into 240 nodes and 183 material points. The external loading, in this case, is static at the end, but gradually applied as depicted in Fig. 2. The behaviour of the consolidation is led by the vertical consolidation coefficient, c_v, which is function of the vertical permeability coefficient, k_v:

$$c_v = \frac{k_v(1+e)}{\rho_w g a_v} = \frac{k_v}{\rho_w g} E_m = \frac{k_v}{\rho_w g m_v} \tag{27}$$

where a_v is the compressibility coefficient and m_v is the volumetric compressibility coefficient. The porous index, e, a measure of the porosity is expressed as follows:

$$e = \frac{n}{1-n}. \tag{28}$$

In Eq. (27), E_m is the oedometric modulus, related with the Young's modulus E according to the following equation:

$$E = E_m \left(1 - \frac{2\nu^2}{1-\nu}\right). \tag{29}$$

Typical values adopted for clays are 2 MPa for the Young's modulus and 0.33 for the Poisson's ratio.

The basic equation for one-dimensional consolidation derived by Terzaghi in 1923 [23] is

$$c_v \frac{\partial^2 \overline{u}}{\partial z^2} = \frac{\partial \overline{u}}{\partial t}. \tag{30}$$

The solution searched is a measure of the consolidation of the soil. It depends on the vertical time factor, T_v, defined by:

$$T_v = \frac{c_v t}{H^2}. \tag{31}$$

Adopting the degree of interstitial pressure dissipation, U_v, we can compare the analytical solution, given by the following equation, with the results obtained with the present research:

$$U_v(z) = 1 - \frac{u_e(z)}{u_{0e}} = 1 - \sum_{m=0}^{m=\infty} \frac{2}{M} \sin\left[M\left(1 - \frac{z}{H}\right)\right] \exp(-M^2 T_v) \quad (32)$$

where

$$M = \frac{\pi}{2}(2m + 1), \quad m = 0, 1, 2, \ldots, \infty. \quad (33)$$

In Fig. 3, it is given the comparison between the analytical and the numerical solution along the depth of the column of soil for different values of T_v. As it is seen, all the values for T_v are dimensionless. With this comparison, we consider the current model employed is sufficiently validated.

This solution has been also compared with PLAXIS, in order to have an idea on the accuracy of this commercial software, since it is going to be employed hereinafter for several other theoretical examples. The direct conclusion obtained in Fig. 4 is that the accuracy decreases at the final stages of the consolidation. For low values of T_v PLAXIS solution of U_v along the column of soil is similar to the calculated

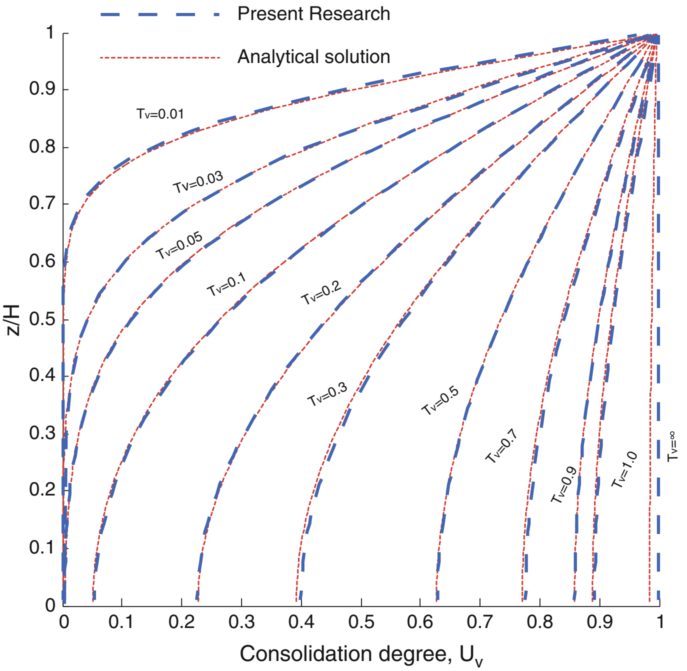


Fig. 3 Analytical and computational solution of U_v for different values of T_v

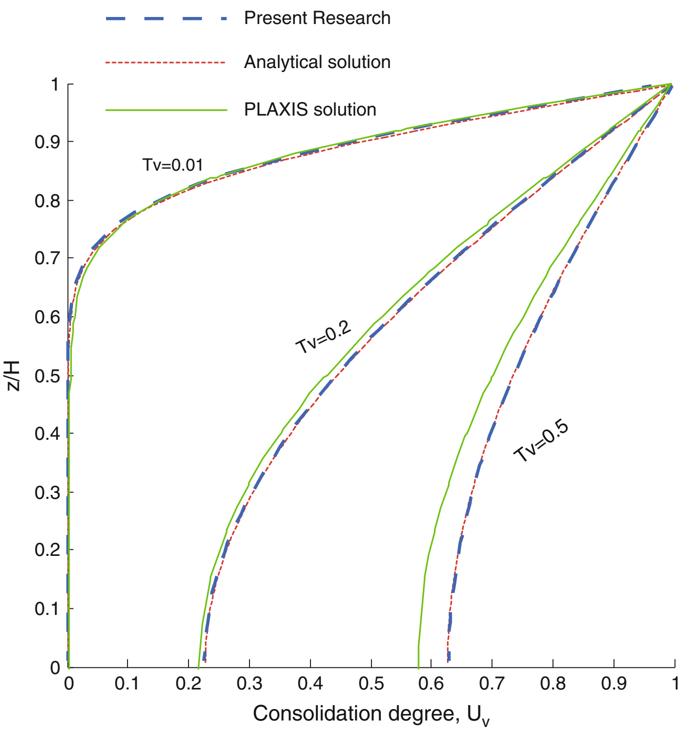


Fig. 4 Solutions of U_v for several values of T_v obtained in the current work compared with the analytical ones and those obtained with PLAXIS software

in this research, but is getting noticebly different for higher values of T_v. Therefore, PLAXIS solutions will provide a good approach of the trend of the pressure along the consolidation without an accurate degree of precision, at least, for one dimensional, static problems.

The solution for the monotonic loading at the upper side of the column shows the dissipation of the pore pressure along time. Figure 5 provides the comparison between the solution obtained with the present methodology, the $u - w$ solution obtained with a quadratic FEM model, and the one calculated using the software GeHoMadrid [25]. Hardly any perceptible difference can be noticed from these three solutions.

3.2 Consolidation of a Column of Soil: Dynamic, 1D Case

In this example, using the same geometry for the soil column, the consolidation of a soil column vertically subjected to harmonic pressure is obtained. This problem was first analytically solved by Zienkiewicz et al. [3] in 1980s, and recently by López-

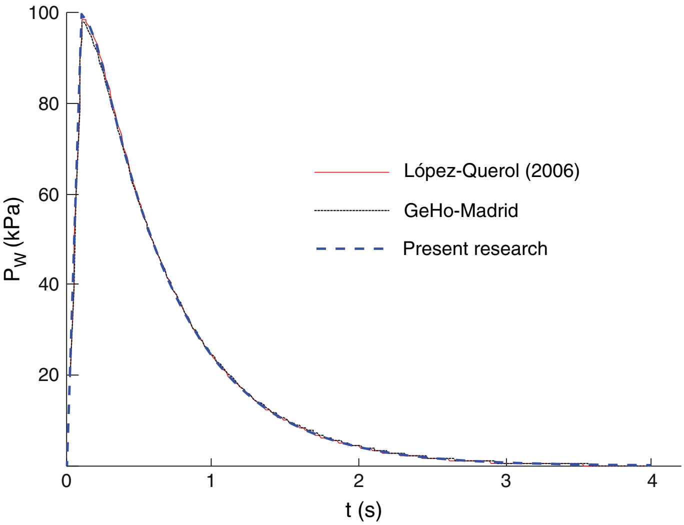


Fig. 5 Pore pressure evolution solutions at the *top* of the consolidation column

Table 1 Material parameters employed for the dynamic consolidation problem of a soil column

G (MPa)	ν (–)	n (–)	ρ (kg/m^3)	ρ_f (kg/m^3)	K_f (MPa)	ω (rad/s)	K_s (MPa)
312.5	0.2	0.333	3003	1000	10^3	3.379	10^{34}

Querol [26], among others, through employing quadratic finite element method. The parameters for the material are those presented in Table 1, where ω, is the applied loading frequency. A periodic surface load with the amplitude of 100 kPa, a frequency of 3.379 rad/s, is imposed. This dynamic load as well as all the material properties in Table 1 are chosen to be the same as those of Zienkiewicz et al. [3].

The variation of the pore pressure with depth is illustrated for different values of π_1, a dimensionless parameter defined as follows,

$$\pi_1 = k \frac{V_c^2}{g \, \frac{\rho_f}{\rho} \, \omega \, H_T^2}, \tag{34}$$

where H_T is the column height and V_c is the compressive wave velocity calculated as:

$$V_c = \sqrt{\left[D + \frac{K_f}{n} \right] \frac{1}{\rho}}, \quad D = \frac{2G(1-\nu)}{1-2\nu}, \tag{35}$$

where D the bulk modulus of the soil skeleton (dry mixture). Note that, for the given material properties and loading frequency, π_1 is proportional to the hydraulic conductivity k. Once k (thus π_1) is known, transient calculations can be carried out to obtain the envelop of the pore pressure history for different points along the column depth, or the isochrone.

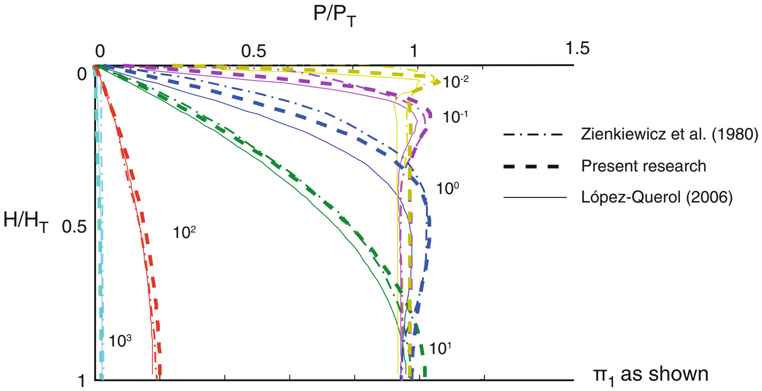


Fig. 6 Isochrones of the pressure in the whole column for different π_1 values: comparison for solutions taken from Zienkiewicz et al. [3], López-Querol [26] and those obtained in the present research. The depth is normalised by the column height, H_T, whereas the pore pressure is made non-dimensional by P_T, 100 kPa

After performing six different calculations for six different levels of π_1, from 10^{-2} to 10^3, we obtain the isochrones of the pore pressure depicted in Fig. 6. Additionally plotted are the results obtained by López-Querol [26] using quadratic finite elements formulating the problem in displacements as well, along with the analytical solutions provided by Zienkiewicz et al. [3]. It is noteworthy that the three different approaches achieve quite similar isochrone maps; nevertheless, whenever more scattering is observed, the current meshfree solution is closer to the analytical one.

3.3 Radial Consolidation: Static Axisymmetric Problems

The second problem carried out in this research is about radial consolidation. The physical equation which governs this problem is different from the Terzaghi's equation [23] shown in the previous section, i.e.:

$$c_h \left(\frac{\partial^2 u_r}{\partial r^2} + \frac{1}{r} \frac{\partial u_r}{\partial r} \right) = \frac{\partial u_r}{\partial t}, \tag{36}$$

where c_h is the horizontal consolidation coefficient, equivalent to the c_v coefficient in vertical consolidation:

$$c_h = \frac{k_h(1+e)}{\rho_w g a_v}. \tag{37}$$

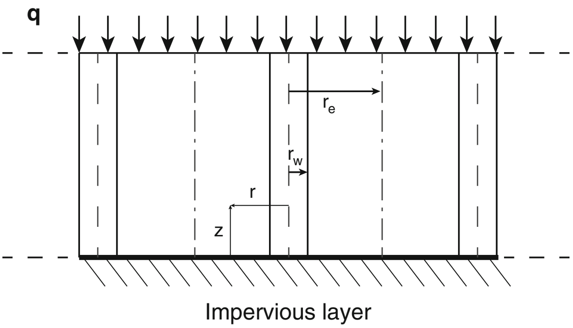


Fig. 7 Scheme of section of set of drains

In Fig. 7, a scheme of drains with induced radial flux is shown, where r and z directions are defined as depicted. There is an analytical solution for this problem given by Barron in 1948 [27] who defined the radial consolidation degree as a function of the non-dimensional time T_r:

$$U_r(T_r) = 1 - \exp\left[-\frac{2T_r}{F(n_r)}\right], \tag{38}$$

where

$$F(n_r) = \frac{n_r^2}{n_r^2 - 1}\log(n_r) - \frac{3n_r^2 - 1}{4n_r^2}. \tag{39}$$

The coefficient n_r depends on the relative extension of the drain in a particular geometry, for the section defined in Fig. 7, it is calculated as

$$n_r = \frac{r_e}{r_w}. \tag{40}$$

where r_w is the drain radius and r_e is the radius of influence for each type of problem. In this case a quadrangular net of drains shown in the scheme of Fig. 8 is studied.

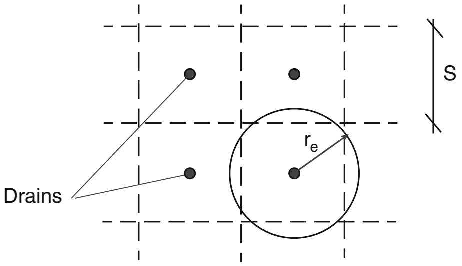


Fig. 8 Quadrangular net of drains ($r_e = 0.564\,S$)

Since radial consolidation equation involves a second term (tangential flow), it is not possible to solve within a plane strain formulation, as the one employed for vertical, one dimensional consolidation. Therefore, the axisymmetric framework, shown in previous sections, is employed instead and excellent results are obtained. In Fig. 9, several solutions of the radial consolidation degree, U_r, along the non-dimensional time T_r are shown. In addition, a comparison with a commercial software, PLAXIS, is given, even though this program does not allow us to implement a perfect radial consolidation due to the impossibility to neglect the vertical displacement throughout the domain. Two alternatives are proposed instead of the original problem, allowing for the vertical displacement: one assumes an impervious boundary condition on the top layer; the other one allows the flux of water through this boundary. Results in Fig. 10 offer a good trend in both cases but not the accuracy expected, as it occurs in the vertical consolidation too.

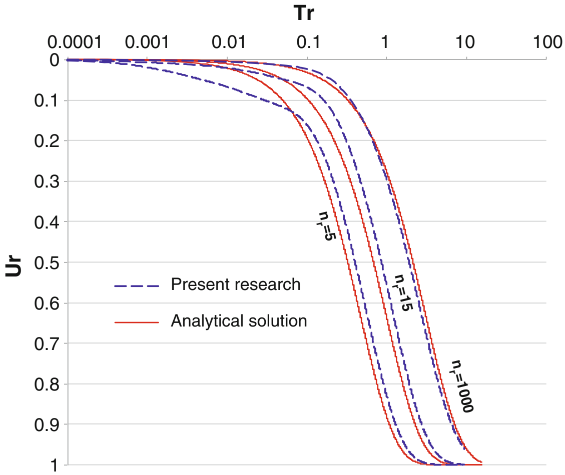


Fig. 9 Analytical and computational solutions of U_r along the non-dimensional time T_r

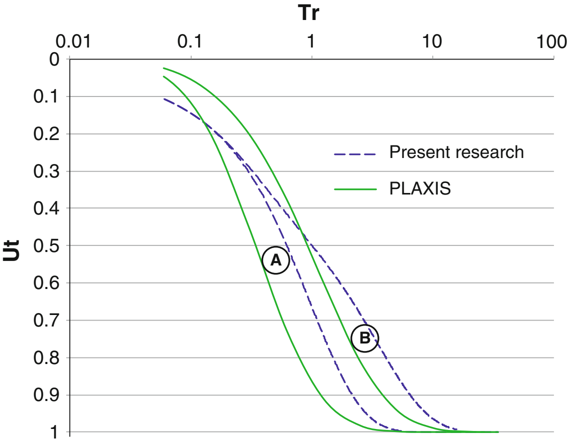


Fig. 10 Comparison between PLAXIS and present research solutions of $U_t = U_r + U_v$ along the non-dimensional time T_r for case A (permeable boundary) and B (impervious boundary)

3.4 Radial Consolidation: Dynamic Axisymmetric Problems

A dynamic loading on the surface has been applied, aiming to simulate its effect on the development of excess pore water pressure in the domain. The frequency of loading is the same as for the case of the soil column, while the amplitude is 50 kN. The geometry is the same as represented in Fig. 7. Vertical displacements of water in the entire domain have been prevented. Figure 11 represents the evolution of pore water pressure at the lateral, lowest corner, which clearly demonstrates the cyclic response of this result as well. From this figure it can be concluded that the generation and dissipation of excess pore water pressure are balanced in every cycle, and steady state is achieved from the very beginning of the loading. Additionally, Fig. 12 represents

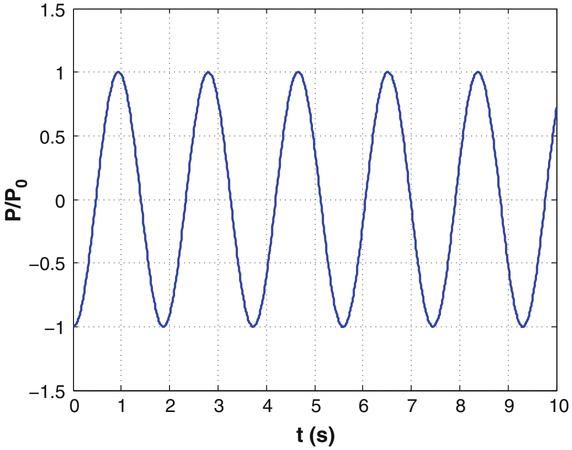


Fig. 11 Evolution of normalised excess pore water pressure during external cyclic loading (for the dynamic, radial consolidation problem)

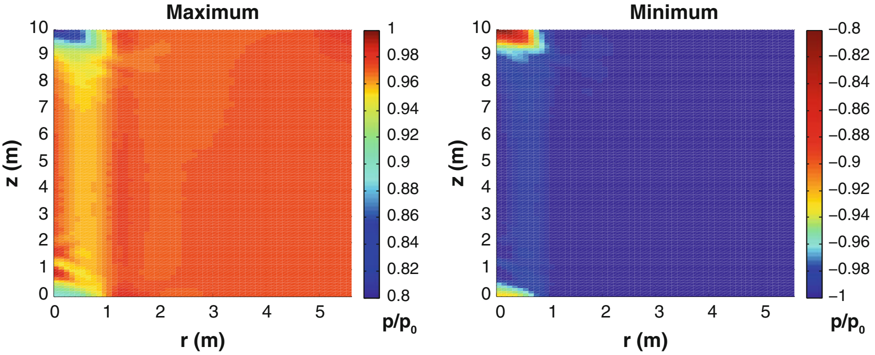


Fig. 12 Maximum and minimum normalised excess pore water pressures in dynamic, radial consolidation

maximum and minimum values of excess pore water pressure in the entire domain during the cyclic loading. This figure demonstrates the 2D nature of the problem, the drain being clearly displayed on the left hand side of the domain. Higher excess pore water pressures can be seen at the lower, left corner for both maximum and minimum cases. Moreover, the figure demonstrates the alternate positive and negative values for the pressures in the domain, as in Fig. 11.

3.5 Static Consolidation in a Soil with a Singular Point: Sink

The insertion of singular points inside the domain may vary the consolidation behaviour. The existence of a sink in the middle of horizontal soil layer is expected to accelerate the consolidation of the porous media, since this means an output of water at the permeable top boundary. To reproduce the sink, excess pore water pressure is not allowed to develop in several nodes in the centre of the domain (see Fig. 13). A square with one meter edge length is proposed for the study of this problem. Material properties are given in Table 2. Figures 14 and 15 show the comparison of results obtained with the present model as well as with PLAXIS for point (0, 0). Figure 14 provides the evolution in time of the degree of consolidation, U, at one of the corners at the bottom of the domain, whereas Fig. 15 represents the solutions in the whole domain after two seconds. In spite of slight differences in the final part of the evolution, it can be concluded that both results are fairly similar, demonstrating the good performance of the present formulation for this kind of problems.

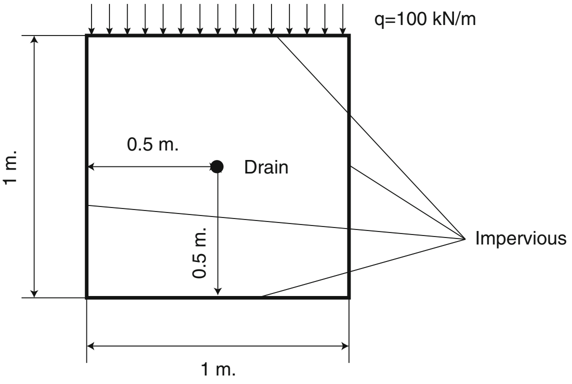


Fig. 13 Geometry for the problem of the sink

Table 2 Material parameters employed for the consolidation problem of a soil with a sink

E (MPa)	ν (–)	n (–)	ρ (kg/m^3)	ρ_f (kg/m^3)	K_f (MPa)	K_s (MPa)	k (m/s)	k_{sink} (m/s)
100	0.0	0.333	3003	1000	10^3	10^{34}	10^{-3}	10

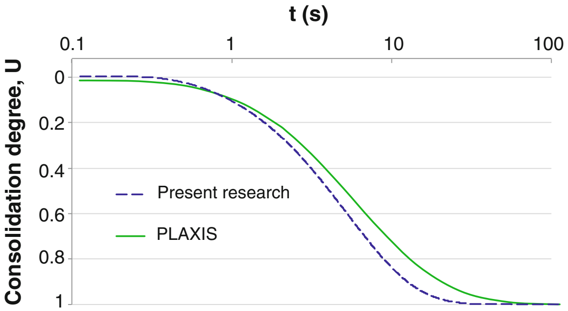


Fig. 14 Evolution of consolidation degree at the *bottom* corner for the sink problem. Present model versus PLAXIS

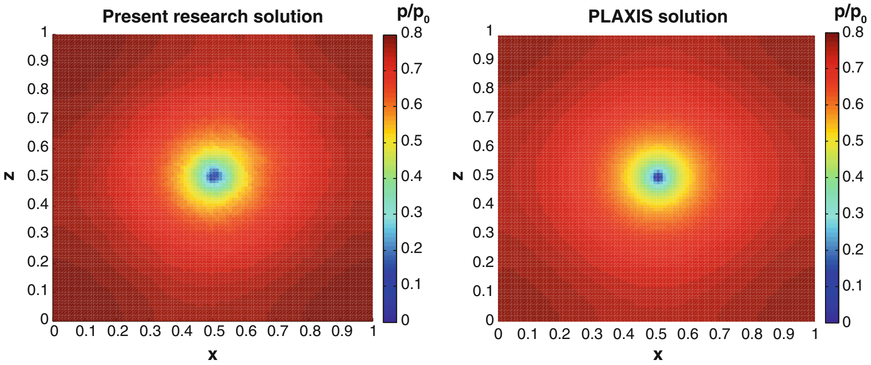


Fig. 15 Field of consolidation degrees in the domain after 2 s. Present model versus PLAXIS

3.6 *Dynamic Consolidation in a Soil with a Singular Point: Sink*

As for the case of radial consolidation, a dynamic simulation has been carried out. The geometry is the same as in Fig. 13. Figure 16 represents the evolution of normalised excess pore water pressure at a bottom corner, clearly showing the cyclic nature of the solution, which is steady from the beginning. Moreover, Fig. 17 shows maximum and minimum values in the whole domain. This figure clearly demonstrates the location of the sink, with zero water pressures in the centre of the domain for both cases. The alternate positive and negative values are also clear from the plot. Once again, these results demonstrate the suitability of the present formulation for dynamic consolidation problems in saturated soils.

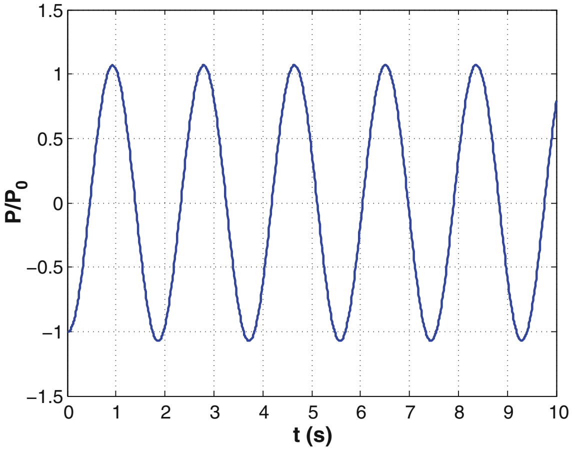


Fig. 16 Evolution of normalised excess pore water pressure during external cyclic loading (dynamic consolidation in a soil with a sink)

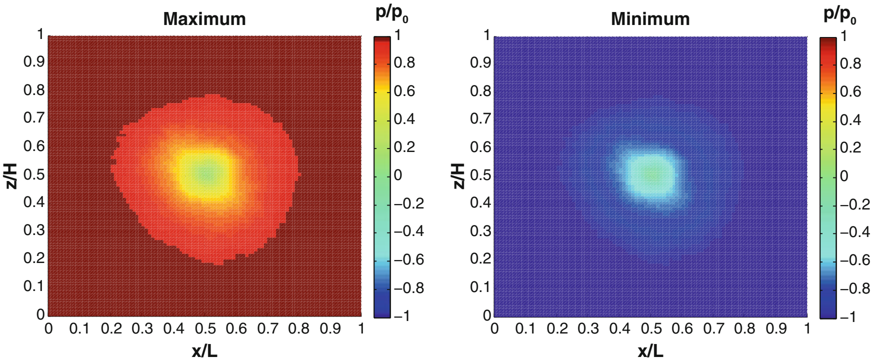


Fig. 17 Maximum and minimum normalised excess pore water pressures for dynamic soil consolidation with a sink

4 Conclusions

We have extended the previously developed B-bar based algorithm to meshfree numerical schemes in axisymmetric framework for both elastic and porous media. The methodology is applied to both static and dynamic consolidation problems in saturated soils. In particular, static and dynamic consolidation of a soil column, static and dynamic radial consolidation, static and dynamic consolidation with singular points (sinks), are carried out and compared with analytical solutions (whenever exist) or available finite element solutions. The feasibility of the current formulation in solving consolidation problems in saturated soils has been clearly demonstrated.

Acknowledgments This research has been partially funded by the Spanish Ministry of Economy and Competitiveness through the projects *BIA2012–31678* and *MAT2012–35416*. The first author also acknowledges the financial support via the fellowship No. *BES2013–063924*.

References

1. Biot, M. A. (1941). General theory of three-dimensional consolidation. *Journal of Applied Physics*, *12*(2), 155–164.
2. Biot, M. A. (1956). General solutions of the equations of elasticity and consolidation for a porous material. *Journal of Applied Mechanics*, 91–96.
3. Zienkiewicz, O. C., Chang, C. T., & Bettes, P. (1980). Drained, undrained, consolidating and dynamic behaviour assumptions in soils. *Géotechnique*, *30*(4), 385–395.
4. López-Querol, S., & Blazquez, R. (2006). Liquefaction and cyclic mobility model in saturated granular media. *International Journal for Numerical and Analytical Methods in Geomechanics*, *30*, 413–439.
5. Cividini, A., & Gioda, G. (2013). On the dynamic analysis of two-phase soils. In S. Pietruszczak & G. N. Pande (Eds.), *Proceedings of the third international symposium on computational geomechanics (ComGeo III)* (pp. 452–461).
6. Ortiz, A., Puso, M. A., & Sukumar, N. (2004). Construction of polygonal interpolants: A maximum entropy approach. *International Journal for Numerical Methods in Engineering*, *61*(12), 2159–2181.
7. Arroyo, M., & Ortiz, M. (2006). Local maximum-entropy approximation schemes: A seamless bridge between finite elements and meshfree methods. *International Journal for Numerical Methods in Engineering*, *65*(13), 2167–2202.
8. Li, B., Habbal, F., & Ortiz, M. (2010). Optimal transportation meshfree approximation schemes for fluid and plastic flows. *International Journal for Numerical Methods in Engineering*, *83*, 1541–1579.
9. Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308–313.
10. Hughes, T. J. R. (1980). Generalization of selective integration procedures to anisotropic and nonlinear media. *International Journal for Numerical Methods in Engineering*, *15*, 1413–1418.
11. Simo, J. C., & Rifai, M. S. (1990). A class of mixed assumed strain methods and the method of incompatible modes. *International Journal for Numerical Methods in Engineering*, *29*, 1595–1638.
12. Kasper, E. P., & Taylor, R. L. (2000). A mixed-enhanced strain method: Part I: Geometrically linear problems. *Computers and Structures*, *75*(3), 237–250.
13. De Souza Neto, E. A., Pires, F. M., & Owen, D. R. J. (1980). F-bar-based linear triangles and tetrahedra for finte strain analysis of nearly incompressible solids. Part I: Formulation and benchmarking. *International Journal for Numerical Methods in Engineering*, *62*, 353–383.
14. Bonet, J., & Burton, A. J. (1998). A simple average nodal pressure tetrahedral element for incompressible and nearly incompressible dynamic Explicit applications. *Communications in Numerical Methods in Engineering*, *14*(5), 437–449.
15. Hauret, P., Kuhl, E., & Ortiz, M. (2007). Diamond elements: A finite element/discrete-mechanics approximation scheme with guaranteed optimal convergene in incompressible elasticity. *International Journal for Numerical Methods in Engineering*, *73*, 253–294.
16. Elguedj, T., Bazilevs, Y., Calo, V. M., & Hughes, T. J. R. (2008). $\overline{B}$ and $\overline{F}$ projection methods for nearly incompressible linear and non-linear elasticity and plasticity using higher-order NURBS elements. *Computer Methods in Applied Mechanics and Engineering*, *197*(33–40), 2732–2762.
17. Artioli, E., Castellazzi, G., & Krysl, P. (2014). Assumed strain nodally integrated hexahedral finite element formulations for elastoplastic applications. *International Journal for Numerical Methods in Engineering*, *99*(11), 844–866.

18. Ortiz, A., Puso, M. A., & Sukumar, N. (2010). Maximum-entropy meshfree method for compressible and near-incompressible elasticity. *Computer Methods in Applied Mechanics and Engineering*, *199*, 1859–1871.
19. Navas, P., López-Querol, S., Yu, R. C., & Li, B. (2016). B-bar based algorithm applied to meshfree numerical schemes to solve unconfined seepage problems through porous media. *International Journal for Numerical and Analytical Methods in Geomechanics*, *40*, 962–984.
20. López-Querol, S., Navas, P., Peco, J., & Arias-Trujillo, J. (2011). Changing impermeability boundary conditions to obtain free surfaces in unconfined seepage problems. *Canadian Geotechnical Journal*, *48*, 841–845.
21. Navas, P., & López-Querol, S. (2013). Generalized unconfined seepage flow model using displacement based formulation. *Engineering Geology*, *166*, 140–141.
22. Biot, M. A. (1956). Theory of propagation of elastic waves in a fluid-saturated porous solid. I. Low-Frequency range. *Journal of the Acoustical Society of America*, *28*(2), 168–178.
23. Terzaghi, K. V. (1925). Principles of Soil Mechanics. *Engineering News-Record*, *95*, 19–27.
24. Zienkiewicz, O. C., & Taylor, R. L. (1994). *El método de los elementos finitos. Vol 1: Formulación básica y problemas lineales*. Barcelona: CIMNE.
25. Fernández Merodo, J. A., Mira, P., Pastor, M., & Li, T. (1999). *GeHoMadrid User Manual*. Madrid: CEDEX. Technical Report.
26. López-Querol, S. (2006). *Modelización geomecánica de los procesos de densificación, licuefacción y movilidad cíclica de suelos granulares sometidos a solicitaciones dinámicas*. PhD thesis, University of Castilla-La Mancha, Ciudad Real, Spain.
27. Barron, R. A. (1948). Consolidation of fine-grained soils by drain wells. *Transationc ASCE*, *113*, 718–754.

A Multiscale Microstructural Model of Permeability in Fractured Solids

Anna Pandolfi, Maria Laura De Bellis and Gabriele Della Vecchia

Abstract We discuss a microstructural model of permeability in fractured solids, where fractures are modeled as recursive families of parallel, equidistant faults. Faults are originated by the attainment of a resistance threshold under the action of a confinement pressure over an initially undamaged, fully elastic matrix, not necessarily isotropic. The initially undamaged matrix might possess a natural permeability, which is modified by the progressive damage of the rock. The particular organization of the micro-faults considered in the model allows to define analytically the equivalent permeability of the solid. The model is particularly appealing to describe the permeability of rock undergoing the process of fracking, in a form that has great computational advantages, since the approach does not track explicitly the formation of individual macro-faults.

1 Fluid Flow in Porous Media

Permeability is the property of materials that measures the ability for fluids (gas or liquid) to flow through a porous solid material, essentially related to the void topology and not to the properties of the fluid.

The void space of a porous medium, where liquid or gaseous fluids are allowed to flow, can be thought as composed of a spatial network of interconnected random passages, channels or tubes of varying length, cross section and orientation, and junctions, where channels meet [1]. Considering an elementary volume with a sufficiently large number of channels, experimentally it is observed that the network of channels produces average gradients of pressure, density, viscosity and solute con-

A. Pandolfi (✉) · M.L. De Bellis · G.D. Vecchia
Politecnico di Milano, Milano, Italy
e-mail: anna.pandolfi@polimi.it

M.L. De Bellis
e-mail: marialaura.debellis@polimi.it

G.D. Vecchia
e-mail: gabriele.dellavecchia@polimi.it

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_12

centration, which are independent of the geometry of a single channel. Contrariwise, local deviations from the average at points within the void space depend strongly on the geometry of the solid matrix.

A physical property of a material containing voids is the porosity, or void fraction n, expressing the ratio between the volume of the voids and the total volume V of solid V_S and voids V_V

$$n = \frac{V_V}{V} = \frac{V_V}{V_S + V_V}. \tag{1}$$

In a volume of space occupied by a multispecies fluid system, every species α has a different velocity $\mathbf{v}_\alpha$, and it is possible to introduce several definitions of average ($\mathbf{v}^*$ mass or $\mathbf{v}'$ volume) velocities. In general, the three velocities differ in direction and magnitude. This discrepancy disappears if the fluid contains homogeneous single species; in the following we consider a homogeneous fluid, and we denote its velocity with $\mathbf{v}$, of magnitude v.

According to the geometrical arrangement, fluids lose energy most in the channels and not in the junctions that can be disregarded. The energy of a fluid is traditionally measured in terms of total hydraulic head h, a simplified form of the Bernoulli's definition for incompressible fluids, i.e.,

$$h = \frac{p}{\rho g} + z + \frac{v^2}{2g}, \tag{2}$$

where p is the gauge pressure, ρ is the density of the fluid, and g is the gravitational acceleration. The first term, $p/\rho g$, is called pressure head, the second term, z, is called elevation head, and the third term, $v^2/2g$, is called velocity head. The pressure head is the equivalent gauge pressure of a column of water at the base of the piezometer. The elevation head is the relative potential energy in terms of an elevation. The velocity head is related to the kinetic energy of the fluid, and in most applications, where the fluid velocity is very small, is disregarded. In the following we will discard the kinetic contribution.

Fluid flow across packed porous media is in general characterized by laminar regime (Reynolds number $\mathrm{Re} \leq 1$) and a drop in the hydraulic head. The hydraulic gradient ∇h measures the change of the hydraulic head field and is related to the direction of the flow in the porous medium. In a cartesian orthogonal reference system it is defined as

$$\nabla h = \frac{\partial h}{\partial x_1}\mathbf{e}_1 + \frac{\partial h}{\partial x_2}\mathbf{e}_2 + \frac{\partial h}{\partial x_3}\mathbf{e}_3. \tag{3}$$

Analytical models of fluid flow in rocks use constitutive relations that link the average velocity of the fluid across the medium to the pressure drop. In particular, Darcy's law states that the discharge rate in a porous media is proportional to the hydraulic head gradient and inversely proportional to the fluid viscosity

$$\mathbf{q} = -\boldsymbol{\kappa}\frac{\rho g}{\mu}\nabla h, \tag{4}$$

where $\mathbf{q}$ is the discharge rate and μ the fluid viscosity, and $\boldsymbol{\kappa}$ the medium permeability.

In anisotropic media the permeability is described by a symmetric and positive definite second order tensor. Symmetry of $\boldsymbol{\kappa}$ is the consequence of the Onsager reciprocal relations, while positive definiteness follows from the fact that a fluid cannot flow against the pressure drop. These two properties render the tensor diagonalizable. Eigenvalues of the permeability tensor represent the principal permeabilities, and the corresponding eigenvectors define the principal directions of flow, i.e., the directions where flow is parallel to the pressure drop.

1.1 Analytical Models of Permeability

Permeability is an overall important physical property of porous media. In order to be used in applications, it has to be measured, either directly or through estimations using formulas derived empirically. Relationships between permeability and commonly measured physical variables of porous media, such as porosity and elastic wave velocity, are not easily established. Permeability is also very difficult to characterize theoretically. Difficulties derive from the complexity of the pore fabric geometry and connectivity, which introduces large uncertainty on the range of applicability and on the predictability of analytical models. However, for simple and structured models of porous media, permeability can be estimated through analytical relationships. Analytical models often are limited to the particular porous medium under investigation, and apply only under a narrow range of conditions.

The class of Kozeny-Carman type models collects simple relations that, under the assumption of laminar flow of the pore fluid, link the permeability to the microstructural characteristics of the porous medium. The original Kozeny-Carman relation [2–4] was derived by extending Poiseuille's law, valid for straight circular section pipes, to the flow of a fluid in a collection of curving passages and tubes embedded in a porous media. The original pipe radius of Poiseuille's law is replaced by the hydraulic radius, defined as the ratio of the pore volume to the solid-fluid interfacial area, a new concept introduced to bypass the definition of a representative radius, complex even in a natural homogeneous porous medium. The Kozeny-Carman equation reads

$$k = \frac{c}{8a_v^2\tau}\, n\left(\frac{n}{1-n}\right)^2, \tag{5}$$

where k is a scalar permeability, c an empirical geometric parameter, n the porosity, a_v the ratio of the exposed surface of the channels to the volume of the solids (also called specific internal surface area), and τ the tortuosity. In a simple manner, the tortuosity can be defined as

$$\tau = \left(\frac{L_a}{L}\right)^2, \tag{6}$$

where L_a is the average length of the channels and L the macroscopic length of the flow path. Relation (5) incorporates a characteristic microstructural length parameter similar to that used in other analyses of permeability [5].

The estimation of the shape coefficients appearing in the equation has been promoting an active research [6]. The specific internal surface area has been evaluated with several methods, for example by combining the contributions of plastic clayey and granular silty-sandy fractions [7], or scanning electron microscope images [8]. Simplified and more tractable tortuosity models have been introduced, for example considering pore channels of variable shape but constant cross-section [9] or by introducing alternative definitions of tortuosity [10]. By using the analog of networks of electrical resistors, τ can be linked empirically to the electrical conductivity of rocks and the brine (salt solution) saturation through a quantitative relation due to Archie [11]. Starting from Archie's law, permeability models have been able to include the pore connectedness in the correlation between permeability and local electric field [12–15].

By describing the connected pore space as a bundle of tortuous leaky hydraulic tubes, alternative permeability models have been proposed by Civan [16, 17]

$$k = \gamma\, n \left(\frac{n}{1-n}\right)^{\beta}, \tag{7}$$

where the parameter γ and the fractal exponent β have been originally determined by experimental data fitting, and successively derived analytically, showing how the two parameters γ and β can be linked to physically meaningful parameters of the porous media.

The scalar nature of variables and parameters used in this class of models leads to scalar definitions, and the correct tensor nature of the permeability is disregarded. Therefore, such models are not meaningful if applied to soils characterized by the presence of sedimentation layers or fissures. The complexity of the relationship between the permeability tensor and a scalar property such as the porosity in rocks has been clearly pointed out [18]. In particular, permeability depends not only on the actual stress and on the strain during the loading history, but also on the evolution of the crack patterns, which is anisotropic in nature. The Kozeny-Carman permeability models are primarily intended for applications to static porous materials, whose effective or conductive pore structure and properties remain unchanged during fluid flow. Hence, these models do not allow for the modification of the porous medium microstructure due to fluid-porous matrix interactions, or by the presence of a variable confining pressure.

1.2 Computational Methods for Permeability

The permeability's parameters can be evaluated using numerical approaches. Modified models of permeability are obtained by introducing the uncertainty deriving from particle size distribution, through first-order error analysis methods [19]. Neural networks and genetic algorithms are employed to elaborate wide samples of data [20] and to link, in modified models, the saturated permeability to the effective porosity. Lattice Boltzmann simulation of flow in simplified 2D or 3D porous media have been able to provide numerical evaluation of tortuosity [21] and permeability [22] for different values of the solid fraction, considering also the presence of spanning planar fractures in the matrix [22]. Numerical results illustrate the importance of matrix-fracture interactions, and prove the inadequacy of using simplified assumptions to predict permeability from porosity in fractured porous rock.

More recently, advances in imaging techniques gave impulse to the numerical simulations of physical processes occurring within sub volumes of rock samples, in order to characterize the rock permeability and correlate it to the microstructural features of the pores. Most of the research has been conducted using specialized research software, but recently applications use commercial software [23].

1.3 Mechanics and Permeability

Fractures and faults represent the most ubiquitous and efficient ways for flows in natural rock formations. The availability of fault and fracture mappings in reservoirs is an important recent achievement in geology, but the understanding of the influence of these structures on fluid flows is still far from being satisfactory, in particular when the mechanical coupling is significant. Difficulties emerge from the complexity of the topology and geometry of faults. Each group or class of faults is characterized by orientation, spacing, distribution, and connectivity, which affect the entrapment of fluids, limiting or advantaging the migration and flow of fluids in a given environment [24].

Clearly, the complexity of natural fracture networks is associated to the stress state history, which is hardly known. Furthermore, cracks and fracture can evolve due to the action of gravity, superposed localized pressures, and shear tractions resulting from the viscosity of the flowing fluids.

The observation that intact rocks contain distributed flaws and cracks, arranged between particles of various shape, has motivated the use of fracture mechanics to study their organization and the conditions that promote their growth [25]. The most widely used fracture mechanics models to describe the progressive microfracturing of rocks upon increasing loading are the open crack and the sliding crack models. However, standard approaches of fracture mechanics have limiting drawbacks related to the explicit treatment of cracking.

As an alternative to fracture mechanics, continuum damage mechanics considers the averaged effect of microstructural changes, following a phenomenological approach able to reproduce hydro-mechanical responses during rock progressive degeneration. One of the most interesting approaches treats rock masses containing a large number of discontinuities as homogeneous, anisotropic porous media [26]. The cracks in the medium are assumed to follow a probability distribution function $PDF\,(\mathbf{N}, L, \Delta)$ in terms of crack orientation $\mathbf{N}$, size L and opening Δ. By using an averaging procedure, a symmetric crack tensor associated to the permeability tensor of the cracked porous medium is derived. The principal directions of the permeability tensor are coaxial to the ones of the crack tensor. Thus, the first invariant of the crack tensor results to be proportional to the mean permeability, while the deviatoric part of the crack tensor is related to the anisotropic permeability.

Various methods to take into account the effects on the permeability of rocks of several factors, such as the coupled effect of flow, stress and deformation, the propagation of existing fractures, and the initiation of new fractures, have been developed in the framework of continuum mechanics.

A simplified coupled hydro-mechanical continuum approach, based on the Biot's theory of fluid saturated porous media and on brittle-elastic solids with residual strength, is considered in a finite element model that includes damage with elastic unloading/reloading [27]. Hydraulic anisotropy and internal state variables are not considered, thus the resulting permeability is treated as a scalar and directly dependent on the stress state.

A micromechanical point of view has been taken in [28] to assess the influence of local damage on the macroscopic hydro-mechanical response of porous media. The damage variables are related to the degradation of elastic properties and to the characteristics of the fracture network, thus the model has the possibility to describe the evolution of the porous network with deformation and its influence on permeability [29]. The model distinguishes between the natural pore network, sensitive to the deformation of the representative volume, and the crack network, enucleated in the damaged material, and assumes laminar flow. The natural pore network is characterized by a Pore Size Distribution (PSD) curve, updated with the state variables and with the evolution of the cracks, and linked to the permeability. In spite of the anisotropic nature of the crack pattern, a scalar value of hydraulic conductivity is defined by integration of the PSD.

Coupling between deformation and fluid flow has been accounted for in various manners. By considering the volumetric strain as an additional controlling parameter, in [30] numerical simulations of permeability reduction (increase) upon elastic contraction (dilation) in rocks are presented. Permeability is considered there as a scalar variable, although the approach accounts for flow in both matrix and fractures.

The dependence of rock permeability on material deformation has been enforced alternatively in terms of crack opening. A sophisticated coupled semi-empirical hydro-mechanical constitutive model accounting for anisotropic damage induced by cracks and modification in permeability in brittle rocks under deviatoric compressive stresses has been proposed in [31]. The rock is regarded as a porous medium with embedded microcracks. Upon homogenization, the cracked material is treated as an equivalent porous medium where the permeability tensor is decomposed additively into initial permeability and crack induced permeability tensors. Since the microcrack distribution is orientation-dependent, the crack permeability tensor has to be anisotropic in nature and it is regarded as a function of crack number, orientation, radius, and average opening. The mechanical model is formulated in terms of linear elasticity and the crack propagation conditions are based on linear elastic fracture mechanics, without the support of a thermodynamical framework.

The effects of damage on anisotropic permeability have been discussed in [32] by adopting a relationship between macroscopic and microscopic aspects of damage, and exploiting micro-level analyses of flow through randomly generated crack networks.

In agreement with the literature on the topic, material models that account for the progressive evolution of defects, flaws and cracks seem to possess the right characteristics to account for the variation of the permeability according to the evolution of the damage in the considered porous material.

In the next paragraph we recall a multi-scale brittle damage material model [33] that describes the fractures using a cohesive approach [34, 35]. The mechanical model has been discussed in the original paper; here we analyze the kinematic aspects of the model in view of deriving analytically the permeability.

2 Brittle Damage Model

The brittle damage model presented in [33] is characterized by a homogeneous matrix where nested microstructures characterized by different length scales are embedded. In each level k of the nested architecture, microstructures assume the form of families of equidistant cohesive faults, characterized by an orientation $\mathbf{N}_k$ and a spacing L_k. In line with general mathematical results pertaining to free discontinuity problems, the constitutive model is derived within a thermodynamical approach where we assume the existence of a free energy density which accounts for reversible and dissipative behaviors of the material.

The key of the model is given by the kinematic assumptions. We begin by considering the particular case of a single family of fault planes of normal $\mathbf{N}$ and spacing L, and later extend the behavior to nested families. The total deformation gradient $\mathbf{F}$ of the material is assumed to decompose multiplicatively into a part $\mathbf{F}^{\mathrm{m}}$ pertaining the uniform deformation of the matrix, and a second part $\mathbf{F}^{\mathrm{f}}$ pertaining the discontinuous kinematics of the cohesive faults, i.e.,

$$\mathbf{F} = \mathbf{F}^{\mathrm{m}}\mathbf{F}^{\mathrm{f}}. \tag{8}$$

The deformation gradient $\mathbf{F}^{\mathrm{f}}$ can be easily linked to the fault kinematic activity. Consider a material vector $d\mathbf{X}$, shorter than the system size but longer than the internal scale L, that spans two material points P and Q in the material configuration. The number of faults m traversed by the vector is

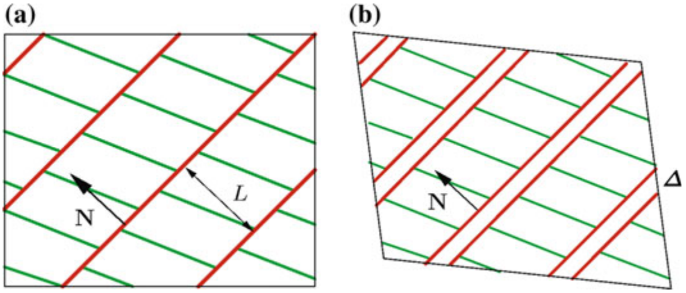


Fig. 1 Inelastic kinematics of the fault system. The opening displacement $\boldsymbol{\Delta}$ applied to all the faults at distance L leads to a deformed configuration characterized by the inelastic deformation gradient $\mathbf{F}^{\mathrm{f}}$. **a** Reference configuration. **b** Spatial configuration

$$m = \frac{1}{L} d\mathbf{X} \cdot \mathbf{N}.$$

Let us now apply an opening displacement $\boldsymbol{\Delta}$ to each fault, Fig. 1a. In the spatial configuration the two points P and Q are joined by the vector $d\mathbf{x}$ given by

$$d\mathbf{x} = d\mathbf{X} + m\boldsymbol{\Delta} = d\mathbf{X} + \frac{1}{L}(d\mathbf{X} \cdot \mathbf{N})\,\boldsymbol{\Delta} = (\mathbf{I} + \frac{1}{L}\boldsymbol{\Delta} \otimes \mathbf{N})\, d\mathbf{X},$$

where we set

$$\mathbf{F}^{\mathrm{f}} \equiv \mathbf{I} + \frac{1}{L}\boldsymbol{\Delta} \otimes \mathbf{N}. \tag{9}$$

Note that, once $\mathbf{N}$ and L are supplied, $\mathbf{F}^{\mathrm{f}}$ and $\boldsymbol{\Delta}$ are in one-to-one correspondence, and the opening displacements $\boldsymbol{\Delta}$ follow from $\mathbf{F}^{\mathrm{f}}$ through the relation

$$\boldsymbol{\Delta} = L\,(\mathbf{F}^{\mathrm{f}} - \mathbf{I})\,\mathbf{N}. \tag{10}$$

Note that the matrix may, in turn, accommodate a second fault family:

$$\mathbf{F} = \mathbf{F}^{\mathrm{m}1}\mathbf{F}^{\mathrm{f}^1}, \qquad \mathbf{F}^{\mathrm{m}1} = \mathbf{F}^{\mathrm{m}2}\mathbf{F}^{\mathrm{f}^2}. \tag{11}$$

This decomposition can be applied recursively for as many levels as necessary, assigning the innermost level a purely elastic behavior

$$\mathbf{F} = \mathbf{F}^{\mathrm{e}}\mathbf{F}^{\mathrm{f}^k}\mathbf{F}^{\mathrm{f}^{k-1}} \ldots \mathbf{F}^{\mathrm{f}^2}\mathbf{F}^{\mathrm{f}^1}. \tag{12}$$

The constitutive behavior of the material is thus obtained by introducing a free energy density that decomposes into the sum of two contributions

$$A(\mathbf{F}^{\mathrm{m}}, \boldsymbol{\Delta}, q) = W^{\mathrm{m}}(\mathbf{F}^{\mathrm{m}}) + \frac{1}{L}\Phi(\boldsymbol{\Delta}, q), \tag{13}$$

where W^{m} is the strain-energy density per unit volume of the matrix, Φ is the cohesive energy per unit fault surface, suitably divided by the scale length L to result in a specific energy per unit of volume, and q is an internal variable describing the maximum opening displacement ever experienced by the faults. Note that the separation of the variables excludes strong coupling between the two energies. The particular form of the energy densities W^{m} and Φ can be selected with a certain degree of freedom according to the particular material considered.

In keeping with the irreversible nature of fracture, decohered faults permanently damage the material. This can be expressed by the internal variable q evolution constraint $\dot{q} \geq 0$. Another important constraint is the impenetrability of the faults surfaces upon closure, i.e., the component of the opening displacement along the fault normal cannot be negative, thus $\mathbf{N} \cdot \boldsymbol{\Delta} \geq 0$.

The behavior of irreversible materials can be characterized variationally by recourse to time discretization [33, 36], that requires to consider a process of deformation at distinct successive times $t_0, \ldots, t_{n+1} = t_n + \Delta t, \ldots$. We assume that the state of the material at time t_n is known and the deformation $\mathbf{F}_{n+1}$ at time t_{n+1} is assigned. The problem is to determine the state of the material at time t_{n+1}. We define an effective, incremental, strain-energy density taking the infimum of the constrained energy with respect to $\boldsymbol{\Delta}_{n+1}$ and q_{n+1}, as

$$W_n(\mathbf{F}_{n+1}) = \inf_{\substack{\boldsymbol{\Delta}_{n+1},\, q_{n+1} \\ \boldsymbol{\Delta}_{n+1} \cdot \mathbf{N} \geq 0 \\ q_{n+1} \geq q_n}} A(\mathbf{F}_{n+1}, \boldsymbol{\Delta}_{n+1}, q_{n+1}). \tag{14}$$

The subindex n used in W_n signifies the dependence on the initial state. The irreversibility and the impenetrability constraints render the effective strain-energy density W_n dependent on the initial conditions at time t_n, and account for all the inelastic behaviors, such as irreversibility, hysteresis, and path dependency. Thus, $W_n(\mathbf{F}_{n+1})$ acts as a potential for the first Piola-Kirchhoff stress tensor $\mathbf{P}_{n+1}$ at time t_{n+1} [36], i.e., as

$$\mathbf{P}_{n+1} = \frac{\partial W_n(\mathbf{F}_{n+1})}{\partial \mathbf{F}_{n+1}}. \tag{15}$$

The stable equilibrium configurations are the minimizers of the corresponding effective energy. The fault geometrical features $\mathbf{N}$ and L, until now considered as known, can be determined with the aid of the time-discretized variational formulation.

Energy optimization will ascertain whether the insertion of faults is energetically favorable, and the optimal orientation of the faults. For a given deformation $\mathbf{F}_{n+1}$, we test two end states of the material, one with faults and another without faults, and choose the end state which results in the lowest incremental energy density $W_n(\mathbf{F}_{n+1})$. The orientation of the faults $\mathbf{N}$ and the remaining state variables are obtained variationally from the extended minimum problem:

$$W_n(\mathbf{F}_{n+1}) = \inf_{\substack{\boldsymbol{\Delta}_{n+1},\, q_{n+1},\, \mathbf{N} \\ \boldsymbol{\Delta}_{n+1} \cdot \mathbf{N} \geq 0 \\ q_{n+1} \geq q_n \\ |\mathbf{N}|^2 = 1}} A(\mathbf{F}_{n+1}, \boldsymbol{\Delta}_{n+1}, q_{n+1}, \mathbf{N}). \tag{16}$$

The actual orientation of the faults is defined by the surrounding stress state. In particular, faults can be originated under tensile stress if the maximum tensile stress reaches the tensile resistance of the material. Under compressive stress, it is likely that the material would fail in shear.

The length L can be computed variationally by accounting for the misfit energy $E^{\mathrm{mis}}(\boldsymbol{\Delta}, L)$ contained in the boundary layers that forms where the faults meet a confining boundary. In fact, the compatibility between the faults and their container is only on average, and this gives rise to boundary layers that penetrate into the faulted region to a certain depth. The addition to the energy furnishes a selection mechanism among all possible microstructures leading to the relaxed energy [33]. The total free-energy density of the faulted region becomes

$$\begin{aligned} &A(\mathbf{F}_{n+1}, \boldsymbol{\Delta}_{n+1}, q_{n+1}, L_{n+1}) \\ &\quad = W^{\mathrm{m}}(\mathbf{F}^{\mathrm{m}}{}_{n+1}, L_{n+1}) + \frac{1}{L_{n+1}} \Phi(\boldsymbol{\Delta}_{n+1}, \mathbf{q}_{n+1}) + E^{\mathrm{mis}}(\boldsymbol{\Delta}_{n+1}, L_{n+1}). \end{aligned} \tag{17}$$

The variational update (14) now becomes

$$W_n(\mathbf{F}_{n+1}) = \inf_{\substack{\boldsymbol{\Delta}_{n+1},\, q_{n+1},\, L_{n+1} \\ \boldsymbol{\Delta}_{n+1} \cdot \mathbf{N} \geq 0 \\ q_{n+1} \geq q_n}} A(\mathbf{F}_{n+1}, \boldsymbol{\Delta}_{n+1}, q_{n+1}, L_{n+1}). \tag{18}$$

The optimal fault separation is determined by two competing demands. On one hand, the cohesive energy favors a large value of L resulting in fewer faults per unit volume. On the other hand, the misfit energy favors a small value of L resulting in a narrow boundary layer.

Internal friction is an important dissipation mechanism in brittle materials, especially in geological applications. We assume that friction operates at the faults concurrently with cohesion, but if faults loose cohesion completely upon the attainment of a critical opening displacement, friction may become the sole dissipation mechanism at the faults. In order to retain the variational structure of the model, we define an incremental strain energy density $W_n(\mathbf{F}_{n+1})$ that includes a dual kinetic potential, attending the frictional dissipation, with suitable convexity and regularity properties [37].

So far we have consider either an intact material or a single family of parallel faults. The material with a single fault family is referred to as rank-1 faulting pattern material. More complex microstructures, can effectively be generated by applying the previous construction recursively. In the first level of recursion, we simply replace the elastic strain-energy density $W(\mathbf{F}^{\mathrm{m}})$ of the matrix by $W_n(\mathbf{F}^{\mathrm{m}})$, i.e., by the effective strain-energy density of a rank-1 faulting pattern. This substitution can now be iterated, resulting in a recursive definition of $W_n(\mathbf{F}_{n+1})$. The recursion stops when the matrix between the faults remains elastic. The resulting microstructures are shown in Fig. 1a, and consist of faults within faults. The level of recursion is the rank of the microstructure.

According to the particular loading history, at the time t_n and at the generic point P the material is characterized by a particular opening displacement $\boldsymbol{\Delta}$, which will be respectful of the equilibrium and compatibility conditions. The model is therefore able to account for a variable opening of the faults.

3 Permeability of the Brittle Damage Model

Under the assumption of a perfectly impermeable matrix and considering the presence of a single fault family, the permeability tensor for the material fractured brittle damage model can be directly derived from the particular geometry of faults. The permeability of a particular geometry of parallel and equidistant faults has been examined by Irmay [38]. Snow [39, 40] and Parsons [41] obtained analytical expressions for the anisotropic permeability, similar to the one described here, by considering networks of parallel fissures.

According to Fig. 2, the opening displacement $\boldsymbol{\Delta}$ decomposes into a normal component Δ_N and a sliding component Δ_T computed as

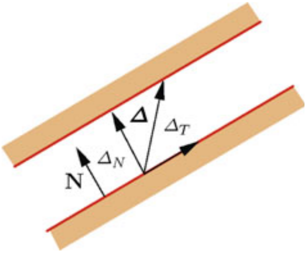
$$\Delta_N = \mathbf{N} \cdot \boldsymbol{\Delta}, \qquad \boldsymbol{\Delta}_T = (\mathbf{I} - \mathbf{N} \otimes \mathbf{N})\, \boldsymbol{\Delta}, \qquad \Delta_T = |\boldsymbol{\Delta}_T|, \tag{19}$$

The fluid fills layers of variable thickness Δ_N and the average fluid flow will take place in the direction of the layer.

We begin by considering the solution of the Navier-Stokes' equation for average velocity v_s along the direction s in a fault of constant width Δ_N:

$$v_s = \frac{\Delta_N^2}{12} \frac{\rho g}{\mu} \frac{\partial h}{\partial s}, \tag{20}$$

Fig. 2 Kinematics of the single fault, defined by an opening displacement $\boldsymbol{\Delta}$, with a component Δ_N along the normal and a component Δ_T along the fault



where $\partial h / \partial s$ is the hydraulic head gradient in the direction s. By considering a porous medium where pores are in the form of parallel faults, the discharge q_s in the direction of the flow is

$$q_s = n\, v_s = \frac{\Delta_N}{L} \frac{\Delta_N^2}{12} \frac{\rho g}{\mu} \frac{\partial h}{\partial s}, \tag{21}$$

where

$$n = \frac{\Delta_N}{L} \tag{22}$$

can be intended as a measure of the local porosity. The permeability in the direction of the fault plane becomes

$$k_s = \frac{\Delta_N}{L} \frac{\Delta_N^2}{12}. \tag{23}$$

Now we restate the above equations in vector form. By introducing the unit vector $\mathbf{d}$ in the direction of the fluid flow

$$\mathbf{d} = \frac{\partial x_1}{\partial s} \mathbf{e}_1 + \frac{\partial x_2}{\partial s} \mathbf{e}_2 + \frac{\partial x_3}{\partial s} \mathbf{e}_3 \tag{24}$$

and using Eq. (3), the directional gradient can be expressed as

$$\frac{\partial h}{\partial s} = \nabla h \cdot \mathbf{d} = \frac{\partial h}{\partial x_1} \frac{\partial x_1}{\partial s} + \frac{\partial h}{\partial x_2} \frac{\partial x_2}{\partial s} + \frac{\partial h}{\partial x_3} \frac{\partial x_3}{\partial s}. \tag{25}$$

The average velocity v_s (20) becomes

$$v_s = \frac{\Delta_N^2}{12} \frac{\rho g}{\mu} \nabla h \cdot \mathbf{d}, \tag{26}$$

and the average flow velocity vector $\mathbf{v}_s = v_s \mathbf{d}$ is

$$\mathbf{v}_s = \frac{\Delta_N^2}{12} \frac{\rho g}{\mu} (\nabla h \cdot \mathbf{d})\ \mathbf{d}. \tag{27}$$

Since the hydraulic discharge can be written as

$$\mathbf{q}_s = n\, \mathbf{v}_s = \frac{\Delta_N}{L} \frac{\Delta_N^2}{12} \mathbf{d} \otimes \mathbf{d} \frac{\rho g}{\mu} \nabla h, \tag{28}$$

the fault permeability tensor $\mathbf{k}^{\mathrm{f}}$ corresponds to

$$\mathbf{k}^{\mathrm{f}} = \frac{\Delta_N}{L} \frac{\Delta_N^2}{12} \mathbf{d} \otimes \mathbf{d}. \tag{29}$$

To account for a generic direction of the flow in the layer of normal $\mathbf{N}$, in (29) the tensor $\mathbf{d} \otimes \mathbf{d}$ must be replaced by the projection $(\mathbf{I} - \mathbf{N} \otimes \mathbf{N})$, leading to

$$\mathbf{k}^{\mathrm{f}} = \frac{\Delta_N}{L} \frac{\Delta_N^2}{12} \left(\mathbf{I} - \mathbf{N} \otimes \mathbf{N}\right). \tag{30}$$

If Q fault families are present in the porous medium, each of which characterized by a normal $\mathbf{N}^K$, a separation L^K and a normal opening displacement Δ_N^K, the equivalent permeability is given by the sum of the corresponding permeabilities:

$$\mathbf{k}^{\mathrm{f}} = \sum_{K=1}^{Q} \frac{\Delta_N^K}{L^K} \frac{\Delta_N^{K^2}}{12} \left(\mathbf{I} - \mathbf{N}^K \otimes \mathbf{N}^K\right). \tag{31}$$

In the brittle damage model the permeability is described by an anisotropic tensor variable from point to point.

The model does not exclude the presence of an initial permeability of the intact matrix. If this is the case, the resulting permeability will be given by the sum of the intact matrix and of the faults

$$\mathbf{k} = \mathbf{k}^{\mathrm{m}} + \mathbf{k}^{\mathrm{f}}. \tag{32}$$

4 Examples

We want to study the response of the brittle damage model to the action of external loadings and to analyze the correspondent variation of the permeability. For the sake of simplicity, we assume that the permeability of the intact matrix is null, thus the permeability will be exclusively related to the formations of the faults.

We specialize the strain energy density W to a standard neo-Hookean material extended to the compressible range, i.e.,

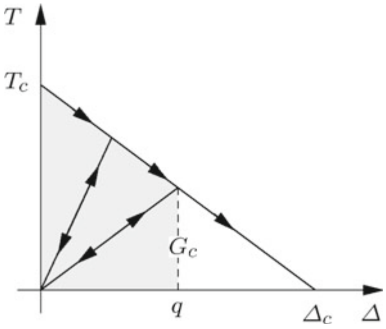
$$W(\mathbf{F}^{\mathrm{m}}) = \frac{1}{2}\lambda \log^2 J^{\mathrm{m}} + \frac{1}{2}\mu \left((\mathbf{F}^{\mathrm{m}T}\mathbf{F}^{\mathrm{m}}) : \mathbf{I} - 3 - 2\log J^{\mathrm{m}} \right), \tag{33}$$

where λ and μ are the Lamé coefficients, and $J^{\mathrm{m}} = \det \mathbf{F}^{\mathrm{m}}$ is the determinant of $\mathbf{F}^{\mathrm{m}}$. Following [34], the cohesive energy on a fault with orientation $\mathbf{N}$ is assumed to depend on an effective opening displacement Δ defined as

$$\Delta = \sqrt{(1 - \beta^2)\,(\boldsymbol{\Delta} \cdot \mathbf{N})^2 + \beta^2 |\boldsymbol{\Delta}|^2}, \tag{34}$$

where $|\boldsymbol{\Delta}|$ is the norm of the opening displacement and β a material parameter expressing the ratio between shear and tensile resistance of the material. In the first loading, the cohesive law follows a linearly decreasing law, i.e.,

Fig. 3 Irreversible linear decreasing cohesive law. The enclosed area represents the critical energy release rate G_c



$$\Phi(\boldsymbol{\Delta}, q) = \Phi(\Delta, q) = \begin{cases} T_c \Delta \left(1 - 0.5\, \Delta / \Delta_c\right) & \text{if } \Delta \leq \Delta_c \\ G_c = 0.5\, T_c \Delta_c & \text{otherwise} \end{cases}, \tag{35}$$

where G_c is the critical energy release rate of the material, T_c is the tensile resistance and Δ_c the maximum opening displacement associated to the presence of cohesive tractions, see Fig. 3. Irreversibility is enforced by recording the maximum ever attained effective opening displacement $q = \Delta_{\max}$, and assuming unloading and reloading to/from the origin, see Fig. 3, with the kinetic equation

$$\dot{q} = \begin{cases} \dot{\Delta} & \text{if} \quad \Delta = q \quad \text{and} \quad \dot{\Delta} \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

The cohesive tractions follows as [33]

$$\mathbf{T} = \frac{\partial \Phi}{\partial \boldsymbol{\Delta}} = \frac{T}{\Delta} \left[\left(1 - \beta^2\right) (\boldsymbol{\Delta} \cdot \mathbf{N})\, \mathbf{N} + \beta^2 \boldsymbol{\Delta} \right], \tag{37}$$

where

$$T = \frac{\partial \Phi}{\partial \Delta} = \sqrt{\left(1 - \beta^{-2}\right) (\mathbf{T} \cdot \mathbf{N})^2 + \beta^{-2} |\mathbf{T}|^2}. \tag{38}$$

The material is characterized by the constants reported in Table 1. We assume an intact material, with no pre-existent or natural faults, and limit our attention to the constitutive response. We feed the material with an assigned deformation gradient, whose significant components grow according to a prescribed history. The material is allowed to form up to three families of faults, which cannot have the same orientation $\mathbf{N}^K$.

Table 1 Rock material constants adopted in the examples

λ (MPa)	μ (MPa)	T_c (MPa)	G_c (N/mm)	β
2778	4167	10	0.01	1.0

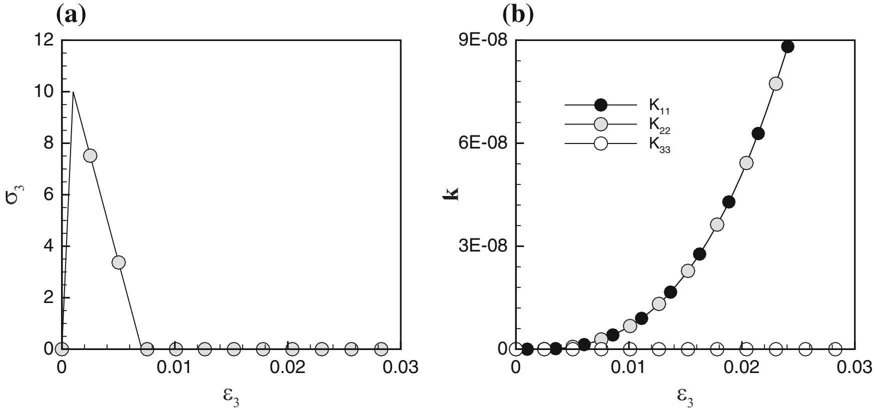


Fig. 4 **a** Uniaxial response σ_3 versus ε_3 of the material, stress in MPa. The drop in stiffness follows the linearly decreasing law of the cohesive layers. **b** Variation of the material permeability, expressed in m^2, in direction $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ as a function of the deformation ε_3 related to the opening of the faults. After the formation of the faults, the matrix is unstressed

A first example is a simple uniaxial tensile test in direction $\mathbf{e}_3$, with a maximum stretch $\lambda_3 = l/l_0 = 1.05$, where l is the final length of the specimen and l_0 the original length. The only non zero component of the stress tensor is in the direction of loading. The material begins behaving elastically, and fails at the attainment of the tensile strength, originating a single family of faults normal to the loading direction. Figure 4a shows the uniaxial response of the material in direction $\mathbf{e}_3$, and Fig. 4b shows the variation of the permeability tensor components k_{11}, k_{22} and k_{33} with the strain in direction $\mathbf{e}_3$. As expected from the orientation of the faults, the permeability in direction $\mathbf{e}_3$ is always null. Note that, after the formation of the faults, the matrix remains unstressed.

A second example is a multistage multiaxial test, that wants to mimic the variation of stress and permeability in the field due to a fracking job. The material is initially compressed isotropically by applying a uniform stretch $\lambda_1 = \lambda_2 = \lambda_3 = 0.99$, to induce a geostatic-like stress state. Then, the material undergoes an isotropic extension $\lambda_1 = \lambda_2 = \lambda_3 = 1.01$, as it may happen in terms of effective stress when a high-pressurized fluid is injected. Given the isotropy of the stress state, at the loading corresponding to the material strength the material fails three times in tension, creating in sequence three families of faults, with normal in direction $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$, respectively. The three families of faults differ because they are characterized by different spacings. The failed material is still able to sustain an overall compressive stress, since the interpenetration of the faults is controlled by a contact algorithm. In a last stage of loading, the material is compressed again with an anisotropic stretch. A $\lambda_1 = \lambda_2 = 0.97$ stretch is applied in direction $\mathbf{e}_1$ and $\mathbf{e}_2$, while the original geostatic-like stretch $\lambda_3 = 0.99$ is applied in direction $\mathbf{e}_3$. Figure 5a shows the mechanical response of the model in direction $\mathbf{e}_1$. The material initially undergoes a compression (black circles). The following extension induces a tensile state that

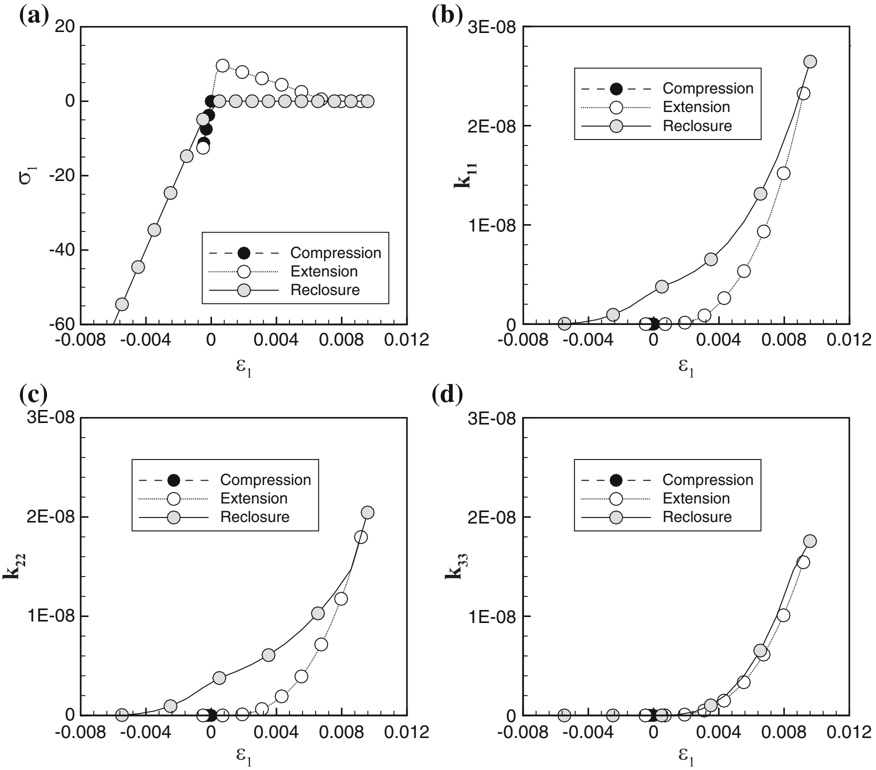


Fig. 5 Multi-stage multiaxial response of the material, undergoing an isotropic compression, followed by an isotropic extension, and by a final anisotropic compression. **a** Stress σ_1 (MPa) versus strain ε_1. Although the material fails, generating three nested families of faults, the material preserves its ability of sustaining load. **b** Permeability (m^2) in direction $\mathbf{e}_1$ as a function of the strain. **c** Permeability (m^2) in direction $\mathbf{e}_2$ as a function of the strain. **d** Permeability (m^2) in direction $\mathbf{e}_3$ as a function of the strain

reaches the strength of the material and causes triple tensile failure (open circles); note that the three stress components show the behavior observed in the uniaxial loading, cf. Fig. 4a. The final compressive stretch is characterized by a null stress until faults close completely. Afterwards, the contact algorithm provides the compressive tractions that guarantee the equilibrium of the system (grey circles). The resulting reduction of the stiffness of the material due to the damage is remarkable.

Figures 5b–d show the permeability in direction $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$, respectively. The permeability is null until the material fails (black circles). Then the permeability reaches a maximum corresponding to the maximum extension imposed to the material (open circles). The different values of the maximum permeability for the three directions is the combined result of the different spacing of the fault families and of the stress anisotropy deriving from the formation of faults. Upon faults closure, permeability decreases until it goes to zero (gray circles). Note that the anisotropy of the compressive loading reflects in the anisotropy of the permeability history. In particular, the permeability reduces more quickly in the direction $\mathbf{e}_3$, where no extra-confinement is applied. In fact, the over-compression in the two directions $\mathbf{e}_1$ and $\mathbf{e}_2$ closes the faults parallel to direction $\mathbf{e}_3$, while the flow is still allowed in the faults normal to $\mathbf{e}_3$.

5 Remarks on the Porous Brittle Damage Model and Possible Applications

The brittle damage model hereby introduced may be used in applications where porous brittle materials experience a stress history leading to the formation of damaged zones characterized by micro or macro cracks. In particular, the model has the potential to be successfully employed in numerical simulations of localized stimulation of oil or gas reservoirs by means of hydraulic fracture.

Brittle materials are sensitive and fail under tensile stresses or under non-isotropic compressive stresses characterized by high shear components. The model here illustrated accounts for the progressive damage that can occur in brittle porous materials failing under tensile or shear stress. As far as deep rock formations are concerned, while the shear critical case can occur in natural situations under non-isotropic stress states, the tensile critical state is more likely to happen under the action of a highly pressurized fluid acting on a localized area, which changes the compressive stress state in a tensile stress.

The mechanical features of the damage model have been described widely in the original paper [33]. More importantly, unlikely standard damage theory models, the brittle damage model is based on cohesive theories of fracture, therefore it describes the formation of cohesive surfaces within the medium in the correct physical way. Moreover, the model is characterized by multiple scales, represented by the L^K spacings between faults in each set, and is able to describe the microstructures that can be observed in layered geologic media. The model is characterized variationally, by introducing a generalized free energy density, so that stress and material tangent stiffness are obtained analytically, and preserve a symmetric structure that makes them appealing in numerical applications. Furthermore, inclusion of friction and other forms of dissipative phenomena is easily achieved through the expedient of introducing dual dissipation potential in the free energy density.

The porous version of the damage model retains the same mechanical features as the original model. Thus, the microstructures resulting from the damage evolution can be characterized with a permeability tensor, which is of fundamental importance for the evaluation of the permeation and of the flow of fluids within the porous material.

References

1. Bear, J. (1972). *Dynamics of fluids in porous media*. New York: American Elsevier Publishing Company Inc.
2. Kozeny, J. (1927). Über kapillare leitung des wassers im boden. *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften*, *136*(2a), 36.
3. Carman, P. C. (1937). Fluid flow through granular beds. *Transactions, Institution of Chemical Engineers*, *15*, 17.
4. Carman, P. C. (1956). *Flow of gases through porous media*. London.
5. Blair, S. C., Berge, P. A., & Berryman, J. G. (1996). Using two-point correlation functions to characterize microgeometry and estimate permeabilities of sandstones and porous glass. *Journal of Geophysical Research - Solid Earth*, *101*, 20359–20375.
6. Göktepe, A. B., & Sezer, A. (2010). Effect of particle shape on density and permeability of sands. *Proceedings of the Institution of Civil Engineers: Geotechnical Engineering*, *163*, 307–320.
7. Sanzeni, A., Colleselli, F., & Grazioli, D. (2013). Specific surface and hydraulic conductivity of fine-grained soils. *Journal of Geotechnical and Geoenvironmental Engineering*, *139*, 1828–1832.
8. Schaap, M. G., & Lebron, I. (2001). Using microscope observations of thin sections to estimate soil permeability with the Kozeny-Carman equation. *Journal of Hydrology*, *251*, 186–201.
9. Lasowska, A., Cieṡlicki, K., & Smolarski, A. Z. (2000). The influence of channel tortuousity on hydraulic resistance. *Proceedings of the Institution of Civil Engineers: Geotechnical Engineering*, *48*, 161–173.
10. Saripalli, K. P., Serne, R. J., Meyer, P. D., & McGrail, B. P. (2002). Prediction of diffusion coefficients in porous media using tortuosity factors based on interfacial areas. *Ground Water*, *40*, 346–352.
11. Archie, G. E. (2003). The electrical resistivity log as an aid in determining some reservoir characteristicsn. In *SPE reprint series* (pp. 9–16).
12. Katz, A. J., & Thompson, A. H. (1987). Prediction of rock electrical conductivity from mercury injection measurements. *Journal of Geophysical Research*, *92*, 599–607.
13. Schwartz, L. M., Sen, P. N., & Johnson, D. L. (1989). Novel geometrical effects in electrolytic conduction in porous media. *Physica A: Statistical Mechanics and its Applications*, *157*, 493–496.
14. Revil, A., & Cathles, L. M. (1999). Permeability of shaly sand. *Physica A: Statistical Mechanics and its Applications*, *35*, 651–662.
15. Cosentini, R. M., Della Vecchia, G., Foti, S., & Musso, G. (2012). Estimation of the hydraulic parameters of unsaturated samples by electrical resistivity tomography. *Géotechnique*, *62*(7), 583–594.
16. Civan, F. (2001). Scale effect on porosity and permeability: Kinetics, model, and correlation. *AIChE Journal*, *47*, 271–287.
17. Civan, F. (2002). Relating permeability to pore connectivity using a power-law flow unit equation. *Petrophysics*, *43*, 457–476.
18. Schulze, O., Popp, T., & Kern, H. (2001). Development of damage and permeability in deforming rock salt. *Engineering Geology*, *61*(2), 163–180.
19. Mishra Parker, J. C., & Singhal, S. N. (1989). Estimation of soil hydraulic properties and their uncertainty from particle size distribution data. *Journal of Hydrology*, *108*, 1–18.
20. Pachepsky, Y. A., Timlin, D. J., & Ahuja, L. R. (1999). Estimating saturated soil hydraulic conductivity using water retention data and neural networks. *Soil Science*, *164*, 552–560.
21. Latief, F. D. E., & Fauzi, U. (2012). Kozeny-Carman and empirical formula for the permeability of computer rock models. *International Journal of Rock Mechanics and Mining Sciences*, *50*, 117–123.
22. Dardis, O., & McCloskey, J. (1998). Permeability porosity relationships from numerical simulations of fluid flow. *Geophysical Research Letters*, *25*(9), 1471–1474.

23. Bird, M. B., Butler, S. L., Hawkes, C. D., & Kotzer, T. (2014). Numerical modeling of fluid and electrical currents through geometries based on synchrotron X-ray tomographic images of reservoir rocks using Avizo and COMSOL. *Computers and Geosciences*, *73*, 6–16.
24. Aydin, A. (2000). Fractures, faults, and hydrocarbon entrapment, migration and flow. *Marine and Petroleum Geology*, *17*(7), 797–814.
25. Yuan, S. C., & Harrison, J. P. (2006). A review of the state of the art in modelling progressive mechanical breakdown and associated fluid flow in intact heterogeneous rocks. *International Journal of Rock Mechanics and Mining Sciences*, *43*(7), 1001–1022.
26. Oda, M. (1985). Permeability tensor for discontinuous rock masses. *Geotechnique*, *35*(4), 483–495.
27. Tang, C. A., Tham, L. G., Lee, P. K. K., Yang, T. H., & Li, L. C. (2002). Coupled analysis of flow, stress and damage (fsd) in rock failure. *International Journal of Rock Mechanics and Mining Sciences*, *39*(4), 477–489.
28. Arson, C., & Gatmiri, B. (2008). On damage modelling in unsaturated clay rocks. *Physics and Chemistry of the Earth, Parts A/B/C*, *33*, S407–S415.
29. Arson, C., & Pereira, J.-M. (2013). Influence of damage on pore size distribution and permeability of rocks. *37*(8), 810–831.
30. Yuan, S. C., & Harrison, J. P. (2005). Development of a hydro-mechanical local degradation approach and its application to modelling fluid flow during progressive fracturing of heterogeneous rocks. *International Journal of Rock Mechanics and Mining Sciences*, *42*(7), 961–984.
31. Shao, J.-F., Zhou, H., & Chau, K. T. (2005). Coupling between anisotropic damage and permeability variation in brittle rocks. *International Journal for Numerical and Analytical Methods in Geomechanics*, *29*(12), 1231–1247.
32. Maleki, K., & Pouya, A. (2010). Numerical simulation of damage- permeability relationship in brittle geomaterials. *Computers and Geosciences*, *37*(5), 619–628.
33. Pandolfi, A., Conti, S., & Ortiz, M. (2006). A recursive-faulting model of distributed damage in confined brittle materials. *Journal of Mechanics and Physics of Solids*, *54*, 1972–2003.
34. Ortiz, M., & Pandolfi, A. (1999). Finite-deformation irreversible cohesive elements for three-dimensional crack propagation analysis. *International Journal for Numerical Methods in Engineering*, *44*, 1267–1282.
35. Pandolfi, A., & Ortiz, M. (2002). An efficient adaptive procedure for three-dimensional fragmentation simulations. *Engineering with Computers*, *18*(2), 148–159.
36. Ortiz, M., & Stainier, L. (1999). The variational formulation of viscoplastic constitutive updates. *Computer Methods in Applied Mechanics and Engineering*, *171*, 419–444.
37. Pandolfi, A., Kane, C., Marsden, J. E., & Ortiz, M. (2002). Time-discretized variational formulation of non- smooth frictional contact. *International Journal for Numerical Methods in Engineering*, *53*(8), 1801–1829.
38. Irmay, S. (1937). Flow of liquid through cracked media. *Bulletin of the Research Council of Istrael*, *1*(5A), 84.
39. Snow, D. T. (1965). A parallel plate model of fractured permeable media. PhD thesis, California: University of California at Berkeley.
40. Snow, D. T. (1969). Anisotropic permeability of fractured media. *Water Resources Research*, *5*(6), 1263–1279.
41. Parsons, R. C. (1966). Permeability of idealized fractured rocks. *Journal of the Society of Petroleum Engineers*, *6*, 126–136.

Thermal Diffusion in a Polymer Blend

Kerstin Weinberg, Stefan Schuß and Denis Anders

Abstract This contribution presents a thermodynamically sound approach to model temperature sensitive diffusion in multi-phase solids. In order to describe the phenomena of thermal diffusion (thermophoresis) and to simulate the effect numerically, an extended version of the Cahn-Hilliard phase-field model is combined with the heat-diffusion equation. The derived model is formulated consistently with the basic laws of thermodynamics. Its discretized version is embedded in a NURBS-based finite element framework. Numerical simulations and a comparison to experimental results show the effect of thermal diffusion, induced by non-uniform and non-steady temperature fields, on the microstructural evolution of a binary polymer blend consisting of polydimethylsiloxane and polyethylmethylsiloxane.

1 Introduction

Composed materials, such as metallic alloys, solid solutions and multiphase plastics, play an increasing role in industrial design. In particular, polymer blends became a matter of interest in recent years. In order to combine beneficial characteristics of single polymers, specific multiphase blends are composed. Such mixtures are subjected to a great variety of microstructural changes such as separation of phases and coarsening processes, cf. [6, 9, 13, 17, 22, 24, 25].

When a multicomponent system at a critical composition is quenched from an initially homogeneous state, usually at high temperature, the blend destabilizes and becomes susceptible to external perturbations. At a critical point the mixture decomposes into two phases of different composition—an effect which is known from

K. Weinberg (✉) · S. Schuß · D. Anders
Lehrstuhl für Festkörpermechanik, Fakultät IV, Universität Siegen, 57068 Siegen, Germany
e-mail: kerstin.weinberg@uni-siegen.de

S. Schuß
e-mail: stefan.schuss@uni-siegen.de

D. Anders
e-mail: denis.anders@uni-siegen.de

K. Weinberg and A. Pandolfi (eds.), *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, Lecture Notes in Applied and Computational Mechanics 81, DOI 10.1007/978-3-319-39022-2_13

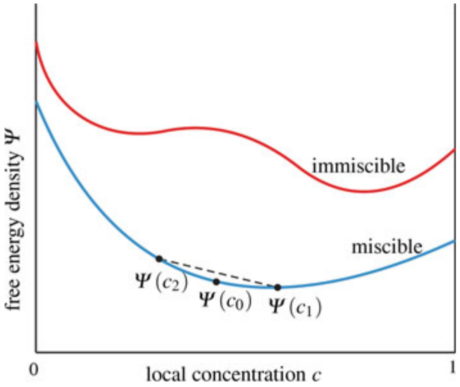


Fig. 1 Illustration of the free energy shape in an stable and unstable scenario

metallic alloys and corresponds to two minima in the free energy function, see Fig. 1. Polymer blends with a close-to-critical-point composition are very susceptible to external fields such as convective, strain, electric, thermal or just random fields. Such driven systems have been studied and simulated by the authors in [1, 3, 4]. External fields allow to control the microstructure evolution in the mixture.

In this text we will focus on the effect of thermal diffusion (Ludwig-Soret effect) induced by non-uniform temperature gradients on the microstructural evolution in a blend consisting of polydimethylsiloxane (PDMS) and polyethyl-methylsiloxane (PEMS). Despite of a wide range of applications, comparatively little studies have been performed for the phase behavior of such polymer blends subjected to non-uniform temperature fields. Lee et al. worked on spinodal decomposition in the presence of local temperature gradients in [19–21] and Köhler et al. published outstanding studies on periodically driven and thermally patterned polymer mixtures in [11, 25, 26]. The experimental setup to perform thermal patterning experiments is explained in detail by Voit in [24].

Motivated by these publications we will present here a robust numerical scheme to approximate the coupled heat-diffusion equations at hand by means of B-spline based finite element analyses. To do so, we first derive the required thermal diffusion equation and the resulting extended Cahn-Hilliard phase-field model. Then, the numerical discretization where we use B-splines to exactly fulfill the continuity requirements of the problem is presented. Finally, two- and three-dimensional simulations of a thermal patterning experiment in a polymer blend are illustrated.

2 Thermal Diffusion and the Ludwig-Soret Effect

To derive the thermal diffusion model the classical *Thermodynamics of Irreversible Processes* (TIP) in the sense of de Groot and Mazur [10] is employed. Point of departure is the entropy balance. In this context it is important to mention that a cornerstone of the classical TIP is the *local equilibrium hypothesis*, where thermodynamic state variables in non-equilibrium states are considered to be the same as in equilibrium. In the thermodynamic community it is still under debate that other variables, not found at equilibrium, are able to influence non-equilibrium processes. For detailed remarks on limitations of classical TIP the reader is referred to [18, pp. 63–65] or [8]. For thermal diffusion in solids and liquids, however, we have no reason to assume additional state variables and so the TIP framework is fully acceptable.

At equilibrium the entropy per unit mass s is a well-defined function depending on state variables such as the internal energy u, the specific volume v and the mass fractions c_k of a multicomponent system with n components. In such a situation the total differential of s in equilibrium is given by the common Gibbs relation

$$\mathrm{d}s = \frac{1}{T}\left(\mathrm{d}u + p\mathrm{d}v - \sum_{k=1}^{n} \mu_k \mathrm{d}c_k\right) \tag{1}$$

where p is the equilibrium pressure, T the absolute temperature, μ_k the chemical potential and c_k the concentration of component k.

2.1 Balances

We follow the local equilibrium hypothesis and make use of the substantial time derivative $\mathrm{d}\left(\bullet\right)/\mathrm{d}t = \partial\left(\bullet\right)/\partial t + \mathbf{v}\cdot\nabla\left(\bullet\right)$ with the barycentric velocity field $\mathbf{v}$ and the spatial gradient ∇ to rewrite the balance (1) in the form

$$T\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}u}{\mathrm{d}t} + p\frac{\mathrm{d}v}{\mathrm{d}t} - \sum_{k=1}^{n} \mu_k \frac{\mathrm{d}c_k}{\mathrm{d}t}. \tag{2}$$

We see here that Eq. (2) already includes the mass balance of the kth component by $\mathrm{d}c_k/\mathrm{d}t$ and the energy balance by $\mathrm{d}u/\mathrm{d}t$ and, thus, compliance with the basic laws of thermodynamics is inherently guaranteed. Introducing the total mass density of our multicomponent system by

$$\rho = \sum_{k=1}^{n} \rho_k \quad \text{with } c_k = \frac{\rho_k}{\rho}, \tag{3}$$

the conservation of mass can also be expressed as

$$\rho\frac{\mathrm{d}c_k}{\mathrm{d}t} = -\operatorname{div}\mathbf{J}_k + \sum_{j=1}^{r} v_{kj}\hat{k}_j, \quad (k = 1, 2, \ldots, n)\,, \tag{4}$$

where $\mathbf{J}_k$ represents the diffusive mass current, $\hat{k}_j$ is a chemical reaction rate of reaction j and ν_{kj} denotes a parameter which is proportional to the stoichiometric coefficient that weights the contribution of component k in the chemical reaction j. In this way we also account for mass production due to chemical reactions; the parameter r denotes their total number. Since mass is conserved in each separate chemical reaction, it holds

$$\sum_{k=1}^{n} \nu_{kj} = 0, \quad (j = 1, 2, \ldots, r)\,. \tag{5}$$

Now we need to deduce an appropriate formulation for the rate of the specific internal energy $\mathrm{d}u/\mathrm{d}t$ implying the conservation of the total specific energy. To this end we start with the equation of motion

$$\rho \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\operatorname{div} \mathscr{P} + \sum_{k=1}^{n} \rho_k \mathbf{F}_k, \tag{6}$$

where $\mathscr{P}$ is a microscopic generalized pressure tensor due to mechanical load and $\mathbf{F}_k$ are vectorial contributions arising from external forces acting on the system. In this study we restrict the discussion to the consideration of conservative forces, which are derived from a stationary potential ψ_k with

$$\mathbf{F}_k = -\nabla \psi_k, \quad \frac{\partial \psi_k}{\partial t} = 0. \tag{7}$$

Multiplying Eq. (6) by the barycentric velocity field $\mathbf{v}$ and making use of the relation $\operatorname{div}(\mathscr{P}\mathbf{v}) = \operatorname{div} \mathscr{P} \cdot \mathbf{v} + \mathscr{P} : \nabla \mathbf{v}$, we obtain the balance equation for the kinetic energy

$$\rho \frac{\mathrm{d}\frac{1}{2}\mathbf{v}^2}{\mathrm{d}t} = -\operatorname{div}(\mathscr{P}\mathbf{v}) + \mathscr{P} : \nabla \mathbf{v} + \sum_{k=1}^{n} \rho_k \mathbf{F}_k \cdot \mathbf{v}. \tag{8}$$

Accordingly, from the local form of the continuity mass equation $\mathrm{d}\rho/\mathrm{d}t = -\rho \operatorname{div} \mathbf{v}$ we get for an arbitrary scalar field Φ the identity

$$\rho \frac{\mathrm{d}\Phi}{\mathrm{d}t} = \frac{\partial (\rho \Phi)}{\partial t} + \operatorname{div}(\rho \Phi \mathbf{v})\,, \tag{9}$$

which is now used for a reformulation of Eq. (8) into

$$\frac{\partial \frac{1}{2}\rho \mathbf{v}^2}{\partial t} = -\operatorname{div}\left(\frac{1}{2}\rho \mathbf{v}^2 \cdot \mathbf{v} + \mathscr{P}\mathbf{v}\right) + \mathscr{P} : \nabla \mathbf{v} + \sum_{k=1}^{n} \rho_k \mathbf{F}_k \cdot \mathbf{v}. \tag{10}$$

In the next step we reformulate Eq. (4) by means of identity (9) into

$$\frac{\partial \rho_k}{\partial t} = -\,\mathrm{div}\,(\mathbf{J}_k + \rho_k \mathbf{v}) + \sum_{j=1}^{r} \nu_{kj} \hat{k}_j, \quad (k = 1, 2, \ldots, n)\,. \tag{11}$$

By combining Eqs. (11) and (7) with the differential formulae (see [16])

$$\begin{aligned} \mathrm{div}\,(\mathbf{J}_k \psi_k) &= \psi_k\,\mathrm{div}\,\mathbf{J}_k + \mathbf{J}_k \cdot \nabla \psi_k \\ \mathrm{div}\,(\psi_k \rho_k \mathbf{v}) &= \psi_k\,\mathrm{div}\,(\rho_k \mathbf{v}) + \rho_k \mathbf{v} \cdot \nabla \psi_k \end{aligned} \tag{12}$$

we establish an equation for the rate of change of the potential energy density $\rho\psi \equiv \sum_k \rho_k \psi_k$ which is of the form

$$\frac{\partial \rho\psi}{\partial t} = -\,\mathrm{div}\left(\rho\psi\mathbf{v} + \sum_{k=1}^{n} \psi_k \mathbf{J}_k\right) - \sum_{k=1}^{n} \rho_k \mathbf{F}_k \cdot \mathbf{v} - \sum_{k=1}^{n} \mathbf{J}_k \cdot \mathbf{F}_k + \sum_{k=1}^{n} \sum_{j=1}^{r} \psi_k \nu_{kj} \hat{k}_j\,. \tag{13}$$

Since we consider only chemical reactions where the potential energy is conserved, for each reaction it holds

$$\sum_{k=1}^{n} \psi_k \nu_{kj} = 0, \quad (j = 1, 2, \ldots, r)\,, \tag{14}$$

and the last term of Eq. (13) vanishes. As a result we obtain an equation for the rate of change of the mechanical energy as a sum of kinetic energy $\frac{1}{2}\rho\mathbf{v}^2$ and potential energy $\rho\psi$,

$$\begin{aligned} \frac{\partial \rho \left(\frac{1}{2}\mathbf{v}^2 + \psi\right)}{\partial t} &= -\,\mathrm{div}\left(\rho\left(\frac{1}{2}\mathbf{v}^2 + \psi\right)\mathbf{v} + \mathscr{P}\mathbf{v} + \sum_{k=1}^{n} \psi_k \mathbf{J}_k\right) \\ &\quad \underline{+\ \mathscr{P} : \nabla\mathbf{v} - \sum_{k=1}^{n} \mathbf{J}_k \cdot \mathbf{F}_k}\,. \end{aligned} \tag{15}$$

Clearly, since a source term appears at the right-hand side of the equation (underlined in 15), the mechanical energy is not a conserved quantity.

According to the first law of thermodynamics (conservation of energy) the total energy within an arbitrary control volume Ω in the system can only change due to energy fluxes $\mathbf{J}_e$ through the boundary $\partial\Omega$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \rho e\,\mathrm{d}\Omega = \int_{\Omega} \frac{\partial \rho e}{\partial t}\,\mathrm{d}\Omega = -\int_{\partial\Omega} \mathbf{J}_e \cdot \mathbf{n}\,\mathrm{d}\Omega\,. \tag{16}$$

Here e is the total specific energy and $\mathbf{n}$ denotes the unit outward normal on $\partial\Omega$. An application of Gauss' theorem provides the local form of energy conservation

$$\frac{\partial \rho e}{\partial t} = -\operatorname{div} \mathbf{J}_e. \tag{17}$$

In general the total energy flux $\mathbf{J}_e$ includes a convective term $\rho e \mathbf{v}$, an energy flux $\mathscr{P}\mathbf{v}$ due to mechanical work performed on the system, a potential energy flux $\sum_k \psi_k \mathbf{J}_k$ due to diffusion and finally a heat flow $\mathbf{J}_\theta$

$$\mathbf{J}_e = \rho e \mathbf{v} + \mathscr{P}\mathbf{v} + \sum_{k=1}^{n} \psi_k \mathbf{J}_k + \mathbf{J}_\theta. \tag{18}$$

If we keep in mind that the total specific energy is defined as the sum of specific kinetic energy $\frac{1}{2}\mathbf{v}^2$, the specific potential energy ψ and the specific internal energy u, we can subtract Eq. (15) from Eq. (17) and use the formula for the energy flux (18) to obtain the balance equation for the internal energy

$$\frac{\partial \rho u}{\partial t} = -\operatorname{div}\left(\rho u \mathbf{v} + \mathbf{J}_\theta\right) \underline{- \mathscr{P} : \nabla \mathbf{v} + \sum_{k=1}^{n} \mathbf{J}_k \cdot \mathbf{F}_k}. \tag{19}$$

This equation shows that also the internal energy u is not a conserved quantity. Again there is a source term, which is equal but of opposite sign to the source term in the balance equation (15) of the mechanical energy. Therefore, such a formulation of the balance equation for the internal energy inherently guarantees the conservation of total energy. In our notation it is convenient to write Eq. (19) in an alternative form. For this purpose we split the total pressure tensor $\mathscr{P}$ into a spherical/hydrostatic part $p\mathbf{I} = \operatorname{tr}(\mathscr{P})/3\mathbf{I}$ and a deviatoric part $\bar{\mathbf{S}}$,

$$\mathscr{P} = p\mathbf{I} + \bar{\mathbf{S}} \tag{20}$$

where $\mathbf{I}$ is the identity tensor. With relations (20) and (9), Eq. (19) becomes

$$\rho \frac{\mathrm{d}u}{\mathrm{d}t} = -\operatorname{div} \mathbf{J}_\theta - p \operatorname{div} \mathbf{v} - \bar{\mathbf{S}} : \nabla \mathbf{v} + \sum_{k=1}^{n} \mathbf{J}_k \cdot \mathbf{F}_k. \tag{21}$$

Here we have used that $\mathbf{I} : \nabla \mathbf{v} = \operatorname{tr}(\nabla \mathbf{v}) = \operatorname{div} \mathbf{v}$. Another version of the mass continuity equation in terms of the specific volume $\nu \equiv \rho^{-1}$

$$\rho \frac{\mathrm{d}\nu}{\mathrm{d}t} = \operatorname{div} \mathbf{v} \tag{22}$$

is now used to formulate

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\nu\,\mathrm{div}\,\mathbf{J}_\theta - p\frac{\mathrm{d}\nu}{\mathrm{d}t} - \nu\bar{\mathbf{S}} : \nabla\mathbf{v} + \nu\sum_{k=1}^{n}\mathbf{J}_k\cdot\mathbf{F}_k. \tag{23}$$

In order to find an explicit form of the entropy balance equation we have to insert the expressions for $\mathrm{d}u/\mathrm{d}t$ (23) and $\mathrm{d}c_k/\mathrm{d}t$ (4) into Eq. (2), which becomes

$$\rho\frac{\mathrm{d}s}{\mathrm{d}t} = -\frac{1}{T}\,\mathrm{div}\,\mathbf{J}_\theta - \frac{1}{T}\bar{\mathbf{S}} : \nabla\mathbf{v} + \frac{1}{T}\sum_{k=1}^{n}\mathbf{J}_k\cdot\mathbf{F}_k + \frac{1}{T}\sum_{k=1}^{n}\mu_k\,\mathrm{div}\,\mathbf{J}_k - \frac{1}{T}\sum_{j=1}^{r}A_j\hat{k}_j. \tag{24}$$

Here we have introduced the so-called chemical affinity A_j of the j-th reaction defined by

$$A_j = \sum_{k=1}^{n}\nu_{kj}\mu_k, \quad (j = 1, 2, \ldots, r)\,. \tag{25}$$

Now we intend to bring Eq. (24) into the typical structure of a balance equation

$$\rho\frac{\mathrm{d}s}{\mathrm{d}t} = -\,\mathrm{div}\,\mathbf{J}_s + \pi_s, \tag{26}$$

where $\mathbf{J}_s$ is a general entropy flux and π_s is an entropy source strength. According to the second law of thermodynamics the entropy source π_s vanishes for reversible (or equilibrium) thermodynamic processes and it holds $\pi_s > 0$ for irreversible thermodynamic transformations. Consequently, it must hold $\pi_s \geq 0$ for a general thermodynamic process.

By means of the relations (12) we obtain the entropy balance equation in the required form

$$\begin{aligned}\rho\frac{\mathrm{d}s}{\mathrm{d}t} = &-\,\mathrm{div}\left(\frac{\mathbf{J}_\theta - \sum\limits_{k=1}^{n}\mu_k\mathbf{J}_k}{T}\right) - \frac{1}{T^2}\mathbf{J}_\theta\nabla T \\ &- \frac{1}{T}\sum_{k=1}^{n}\mathbf{J}_k\cdot\left(T\nabla\left(\frac{\mu_k}{T}\right) - \mathbf{F}_k\right) - \frac{1}{T}\bar{\mathbf{S}} : \nabla\mathbf{v} - \frac{1}{T}\sum_{j=1}^{r}A_j\hat{k}_j.\end{aligned} \tag{27}$$

From comparison with (26) it is possible to identify the entropy flux and the entropy source term as

$$\mathbf{J}_s = \frac{1}{T}\left(\mathbf{J}_\theta - \sum_{k=1}^{n}\mu_k\mathbf{J}_k\right), \tag{28}$$

$$\pi_s = -\frac{1}{T^2}\mathbf{J}_\theta \nabla T - \frac{1}{T}\sum_{k=1}^{n}\mathbf{J}_k \cdot \left(T\nabla\left(\frac{\mu_k}{T}\right) - \mathbf{F}_k\right) - \frac{1}{T}\bar{\mathbf{S}} : \nabla\mathbf{v} - \frac{1}{T}\sum_{j=1}^{r} A_j \hat{k}_j \geq 0. \tag{29}$$

At the first glance the applied separation into flux quantities and entropy source contribution seems to be arbitrary, but $\mathbf{J}_s$ and π_s have to satisfy a number of requirements which determine this separation uniquely; for a discussion we refer to [2].

2.2 *Sources and Fluxes*

Let us now have a closer look at the expressions for the entropy flux $\mathbf{J}_s$ (28) and the entropy source (29). Expression (28) indicates that the entropy flow consists of a reduced heat flow $\mathbf{J}_\theta/T$ and a current due to diffusion. The entropy source expression (29) demonstrates that it can be divided into four contributions which all are connected to a flow quantity. The first term of (29) arises from heat conduction and is connected to heat flow $\mathbf{J}_\theta$, the second is a weighted mass diffusion flow $\mathbf{J}_k$, the third one connects the momentum flow/viscous pressure $\bar{\mathbf{S}}$ to gradients of the velocity field and the fourth term is a sum of chemical rates J_j multiplied by their affinities A_j. In miscellaneous physical applications the terms in the entropy source are classified into *thermodynamic fluxes*, and quantities which multiply the fluxes are called *thermodynamic forces* or *affinities*.

In order to derive the model for thermal diffusion based on our entropy source formulation (29), it is convenient to split off all thermodynamic forces proportional to the temperature gradient which multiply the diffusive flux $\mathbf{J}_k$. By means of the thermodynamic relation

$$\mathrm{d}\left(\frac{\mu_k}{T}\right) = \frac{1}{T}\left(\mathrm{d}\mu_k\right)_{T=\text{const.}} - \frac{h_k}{T^2}\mathrm{d}T, \tag{30}$$

where $h_k := \mu_k - T\partial\mu_k/\partial T$ is the partial specific enthalpy of component k. Please note that relation (30) takes into account that the chemical potentials μ_k depend on a spatially non-uniform temperature field. Now we introduce a generalized heat flux $\mathbf{J}'_\theta$ as

$$\mathbf{J}'_\theta = \mathbf{J}_\theta - \sum_{k=1}^{n} h_k \mathbf{J}_k \tag{31}$$

which is conjugate to the temperature gradient to reformulate the entropy source

$$\pi_s = -\frac{1}{T^2}\mathbf{J}'_\theta \nabla T - \frac{1}{T}\sum_{k=1}^{n}\mathbf{J}_k \cdot \left((\nabla\mu_k)_{T=\text{const.}} - \mathbf{F}_k\right) - \frac{1}{T}\bar{\mathbf{S}} : \nabla\mathbf{v} - \frac{1}{T}\sum_{j=1}^{r} A_j \hat{k}_j. \tag{32}$$

The generalized heat flux $\mathbf{J}'_\theta$ involves with the term $\sum_k h_k \mathbf{J}_k$ a transfer of heat due to diffusion. Therefore the quantity $\mathbf{J}'_\theta$ can be interpreted as an irreversible heat flow. An alternative form of the conservation of mass accounts for the fact, that the sum of all mass fluxes is zero,

$$\sum_{k=1}^{n} \mathbf{J}_k = 0. \tag{33}$$

Thus, we can eliminate $\mathbf{J}_n$ from Eq. (32)

$$\begin{aligned} \pi_s = &-\frac{1}{T^2}\mathbf{J}'_\theta \nabla T - \frac{1}{T}\bar{\mathbf{S}} : \nabla \mathbf{v} - \frac{1}{T}\sum_{j=1}^{r} A_j \hat{k}_j \\ &- \frac{1}{T}\sum_{k=1}^{n-1} \mathbf{J}_k \cdot \left[(\nabla (\mu_k - \mu_n))_{T=\text{const.}} - \mathbf{F}_k + \mathbf{F}_n \right]. \end{aligned} \tag{34}$$

Now we will leave the general framework in order to reduce the general form of the entropy source (34) to a multicomponent isotropic mixture where the concentrations and the temperature are non-uniformly distributed over the system. In this text we consider only isobaric, mechanically equilibrated systems where no external forces and chemical reactions are supposed to be present. In this case the entropy source reduces to the simple equation

$$\pi_s = -\frac{1}{T^2}\mathbf{J}'_\theta \nabla T - \frac{1}{T}\sum_{k=1}^{n-1} \mathbf{J}_k \cdot [\nabla (\mu_k - \mu_n)]_{T,p=\text{const.}} \,. \tag{35}$$

An application of the Gibbs-Duhem relation

$$\sum_{k=1}^{n} \rho_k \delta \mu_k = -\rho s \delta T + \delta p \tag{36}$$

for constant temperature and pressure provides

$$\sum_{k=1}^{n} \rho_k \nabla \mu_k = 0. \tag{37}$$

In this context $\delta(\bullet)$ denotes the variation with respect to spatial coordinates. With the help of Eq. (37) we can eliminate μ_n from (35). This gives

$$\pi_s = -\frac{1}{T^2}\mathbf{J}'_\theta \nabla T - \frac{1}{T}\sum_{k=1}^{n-1} \mathbf{J}_k \cdot \mathscr{A}_{km} [\nabla \mu_m]_{T,p=\text{const.}} \,, \tag{38}$$

where the matrix components $\mathscr{A}_{km}$ are defined as

$$\mathscr{A}_{km} = \delta_{km} + \frac{c_m}{c_n}, \quad (k, m = 1, 2, \ldots, n-1). \tag{39}$$

We regard π_s as a linear combination of thermodynamic fluxes $\mathbf{J}_\bullet$ multiplied by their corresponding affinities $\mathfrak{X}_\bullet$,

$$\pi_s = \mathbf{J}'_\theta \mathfrak{X}_\theta + \sum_{k=1}^{n-1} \mathbf{J}_k \mathfrak{X}_k, \tag{40}$$

and assume a linear dependency between the fluxes and affinities in form of:

$$\mathbf{J}_\bullet = L_{\bullet\theta} \mathfrak{X}_\bullet + \sum_{k=1}^{n-1} L_{\bullet k} \mathfrak{X}_k. \tag{41}$$

This leads to a quadratic expression for the entropy source strength. To guarantee the positive definiteness of the entropy source strength π_s it is therefore appropriate to take the following choice of the phenomenological equations for the thermodynamical fluxes $\mathbf{J}'_\theta$ and $\mathbf{J}_i$:

$$\mathbf{J}'_\theta = -\frac{L_{\theta\theta}}{T^2} \nabla T - \frac{1}{T} \sum_{k,m=1}^{n-1} L_{\theta k} \mathscr{A}_{km} \left[\nabla \mu_m\right]_{T,p=\text{const.}}, \tag{42}$$

$$\mathbf{J}_i = -\frac{L_{i\theta}}{T^2} \nabla T - \frac{1}{T} \sum_{k,m=1}^{n-1} L_{ik} \mathscr{A}_{km} \left[\nabla \mu_m\right]_{T,p=\text{const.}}. \tag{43}$$

Since we study here isotropic systems, the phenomenological coefficients $L_{\theta\theta}$ and L_{ik} are scalars with symmetry in their indices as a consequence of Onsager's reciprocal relations. Due to the positive definiteness of the entropy source strength, additionally it holds

$$L_{\theta\theta} \geq 0, \quad L_{ii} \geq 0, \quad L_{ii} L_{kk} \geq \frac{1}{4} \left(L_{ik} + L_{ki}\right)^2. \tag{44}$$

2.3 Mixtures of a Thermophobic and a Thermophilic Component

Let us finally focus on the case of a binary mixture ($n = 2$). In this situation the coefficients $\mathscr{A}_{km}$ reduce to

$$\mathscr{A}_{11} = 1 + \frac{c_1}{c_2} = \frac{1}{c_2} \tag{45}$$

and the phenomenological equations for the thermodynamic fluxes become

$$\mathbf{J}'_\theta = -\frac{L_{\theta\theta}}{T^2}\nabla T - \frac{L_{\theta 1}}{Tc_2}\,[\nabla\mu_1]_{T,p=\text{const.}}\,, \tag{46}$$

$$\mathbf{J}_1 = -\frac{L_{1\theta}}{T^2}\nabla T - \frac{L_{11}}{Tc_2}\,[\nabla\mu_1]_{T,p=\text{const.}}\,. \tag{47}$$

The coupling coefficients $L_{1\theta}$ characterize the phenomenon of *thermal diffusion* or *thermophoresis* (Ludwig-Soret effect), where a mass diffusion current is caused by temperature gradients. The reciprocal phenomenon, where a heat flow is caused by concentration gradients, depends on the coupling coefficients $L_{\theta i}$. This phenomenon is referred to as Dufour effect. In detail we introduce instead of the phenomenological coefficients in (46) and (47) the following set of coefficients:

$$\Lambda = \frac{L_{\theta\theta}}{T^2} \qquad \text{thermal conductivity} \tag{48}$$

$$D_\theta = \frac{L_{\theta 1}}{\rho c_1 c_2 T^2} \qquad \text{Dufour coefficient} \tag{49}$$

$$D_T = \frac{L_{1\theta}}{\rho c_1 c_2 T^2} \qquad \text{thermal diffusion coefficient} \tag{50}$$

$$M = \frac{L_{11}}{\rho T c_2} \qquad \text{mobility coefficient} \tag{51}$$

The Onsager reciprocal relations imply equality between the thermal mass diffusivity D_T and the Dufour coefficient D_θ which leads to thermodynamical fluxes in the form:

$$\mathbf{J}'_\theta = -\Lambda\nabla T - \rho_1 D_\theta T\,[\nabla\mu]_{T,p=\text{const.}}\,, \tag{52}$$

$$\mathbf{J}_1 = \mathbf{J} = -\rho D_T c\,(1-c)\,\nabla T - \rho M\,[\nabla\mu]_{T,p=\text{const.}}\,. \tag{53}$$

For clarity we made here the choice $c_1 := c$, $\mu_1 := \mu$ and therefore $c_2 := 1 - c$. The indices $T, p = $ const. will be omitted. Equations (52) and (53) enable us now to study diffusion phenomena which arise in a binary mixture where both the temperature and the concentration are non-uniform within the system. Furthermore if the concentration gradients are moderate we may consider the overall density ρ as roughly uniform.

If the temperature gradient plays the dominant role within the irreversible heat flux $\mathbf{J}'_\theta$ and the contribution from the Dufour cross-phenomenon can be neglected, it is possible to derive an approximation of Eq. (52).

Let us now thoroughly elaborate on the diffusion flow $\mathbf{J}$. The concept of thermal diffusion expresses that different particle types move differently under a temperature gradient. Therefore the applied temperature gradient induces a diffusive mass flux $\mathbf{J}_T$. In the introduced notation $\mathbf{J}_T$ takes the form

$$\mathbf{J}_T = -\rho D_T c\,(1-c)\,\nabla T. \tag{54}$$

In a single phase it is not possible to achieve thermal diffusion, since $\mathbf{J}_T$ vanishes for $c = 0$ and $c = 1$. Usually the thermally activated diffusive mass current may occur in either direction, dependent on the materials involved. Thermophilic substances diffuse in the direction of the temperature gradient. Thermophobic materials diffuse in the direction opposite to the temperature gradient. Typically the heavier/larger species in a mixture exhibits a thermophobic behavior while the lighter/smaller species exhibit thermophilic behavior. In addition to the sizes of the various types of particles and the steepness of the temperature gradient, the heat conductivity and heat absorption of the particles play a significant role in thermal diffusion. However $\mathbf{J}_T$ leads to a buildup of a concentration gradient, which is accompanied by a generalized Fickean type mass diffusion current,

$$\mathbf{J}_D = -\rho \mathbf{M} \nabla \mu. \tag{55}$$

Consequently, the entire diffuse mass flux is then $\mathbf{J} = \mathbf{J}_D + \mathbf{J}_T$. Note that the mobility coefficient M may be replaced by a tensor valued mobility to account for anisotropic effects.

At this point it should be clear that the Ludwig-Soret effect can be used in various technical applications to influence the microstructure of materials. Thermal diffusion is a powerful tool in pharmacology to discover and design new types of drugs [5]. As well it is employed as a technique for manipulating single biological macromolecules such as DNA in polymer micro- and nanochannels, cf. [23, 27].

3 Phase Decomposition and Coarsening in a PDMS-PEMS Blend

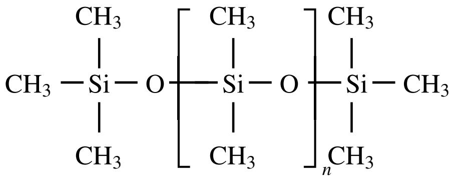
The evolution of multiphase polymer mixtures can be deduced from a variation of a thermodynamic energy functional with respect to a set of order parameters $\boldsymbol{\phi} = [\phi_1, \phi_2, \ldots, \phi_m]$ under the constraint of mass conservation. In general the energy functional has the following structure

$$E(\boldsymbol{\phi}) = \frac{k_B T}{v} \int_{\Omega} \left(\Psi^{\text{con}}(\boldsymbol{\phi}, T) + \Psi^{\text{int}}(\boldsymbol{\phi}, T) \right) \, \mathrm{d}\mathbf{x}, \tag{56}$$

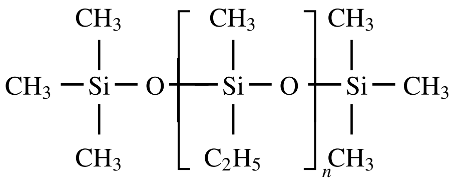
where Ψ^{con} is the configurational energy density and $\Psi^{\text{int}}(\boldsymbol{\phi}, T) = \sum_{i=1}^{m} \frac{\kappa_i(T)}{2} \|\nabla \phi_i\|^2$ characterizes the contribution of interfacial energy. The temperature dependent parameters κ_i are related to surface energy density γ_i and length l_i of the transition regions between the domains of each phase. In this context T is the temperature of the system, k_B denotes the Boltzmann constant and v characterizes a corresponding unit volume. In (56) the free energy density $\Psi = \Psi^{\text{con}} + \Psi^{\text{int}}$ is a dimensionless quantity which is multiplied by the coefficient $k_B T/v$ to obtain an energy density $\left(\mathrm{J/m^3}\right)$.

The corresponding unit volume v comprises an average volume that a monomer with Kuhn length (persistence length) l undergoes within the scenario of a random walk.

In the following we consider specifically a PDMS-PEMS mixture. Both components are polymeric organosilica with a wide range of technical applications, such as material for contact lenses, adhesives, coating implementations, silicone based lubricants, and even as a food additive in defoaming agents. The structural formulas of PDMS and PEMS show their similar molecular structure. Both are terminated with trimethylsiloxane endgroups, one methyl respectively ethyl group in the repeating unit makes the only difference.



(PDMS)



(PEMS)

We choose the mass concentration field c as order parameter, whereby it simply holds $c_{\text{PDMS}} = 1 - c_{\text{PEMS}} =: c$. The densities of PDMS $\left(0.969\,\text{g/cm}^3\right)$ and PEMS $\left(0.977\,\text{g/cm}^3\right)$ are very similar, so that mass and volume fraction are basically the same. According to Flory-Huggins thermodynamics of mixing [12, 15], for a binary polymer blend the configurational energy density can be written as

$$\Psi^{\text{con}}(c) = g_A c + g_B (1 - c) + \frac{c}{N_{\text{PDMS}}} \ln(c) + \frac{(1 - c)}{N_{\text{PEMS}}} \ln(1 - c) + \chi c (1 - c) . \tag{57}$$

Here N_{PDMS} and N_{PEMS} represent the degrees of polymerization, and χ is a temperature dependent material parameter characterizing the chemical interaction between the constituents of the mixture. The value of χ is usually approximated by a relation of the form $\chi = \mathfrak{a} + \mathfrak{b} T^{-1}$, where $\mathfrak{a}$ and $\mathfrak{b}$ are experimentally obtained fitting parameters. The terms $g_A c$ and $g_B (1 - c)$ quantify the free energy of the individual components.

In order to investigate the evolution of the concentration field $c(\mathbf{x}, t)$ and the corresponding temperature field $T(\mathbf{x}, t)$ in a representative domain Ω within the time $\bar{t}$ we make use of the diffusion equation, which expresses the conservation of mass in the system:

$$\frac{\partial c}{\partial t} = -\nabla \cdot \mathbf{j} + \xi(\mathbf{x}, t) . \tag{58}$$

The flux $\mathbf{j}$ is given by flux (53) divided by the average density ρ, $\xi\,(\mathbf{x}, t)$ is a random variable arising from thermal fluctuations. The chemical potential is related to $\Psi = \Psi^{\text{con}} + \Psi^{\text{int}}$ by the variational derivative:

$$\mu = \delta_c \Psi = \partial_c \Psi^{\text{con}} - \nabla \cdot \left(\partial_{\nabla c}\left(\Psi^{\text{int}}\right)\right) = \partial_c \Psi^{\text{con}} - \kappa \Delta c. \tag{59}$$

Using this relation, Eq. (58) results in a modified Cahn-Hilliard equation ([7], cf. [2]) coupled with a heat diffusion equation with specific heat capacity c_p, conductivity k and heat source q. Thus, in the strong form the problem reads: Find $c : \Omega \times \left[0, \bar{t}\right] \to \mathbb{R}$ and $T : \Omega \times \left[0, \bar{t}\right] \to \mathbb{R}$ such that

$$\frac{\partial c}{\partial t} = \nabla \cdot (\mathbf{M}\nabla\mu + D_T c(1-c)\nabla T)\,, \qquad \text{in } \Omega \times [0, \bar{t}] \tag{60}$$

$$c_p \rho \frac{\partial T}{\partial t} = \nabla \cdot (k\nabla T) + q, \qquad \text{in } \Omega \times [0, \bar{t}]. \tag{61}$$

Initial concentration and temperature are given by

$$c(\mathbf{x}, 0) = c_0(\mathbf{x}), \quad T(\mathbf{x}, 0) = T_0(\mathbf{x}) \quad \text{in } \Omega, \tag{62}$$

along with the boundary conditions

$$\begin{aligned} (\mathbf{M}\nabla\mu + D_T c(1-c)\nabla T) \cdot \mathbf{n} = 0 \quad & \text{on } \partial\Omega \times [0, \bar{t}], \\ \nabla c \cdot \mathbf{n} = 0 \quad & \text{on } \partial\Omega \times [0, \bar{t}], \\ T = \tilde{T} \quad & \text{on } \partial\Omega_e \times [0, \bar{t}], \\ \nabla T \cdot \mathbf{n} = 0 \quad & \text{on } \partial\Omega_n \times [0, \bar{t}], \end{aligned} \tag{63}$$

with properties $\partial\Omega = \partial\Omega_e \cup \partial\Omega_n$ and $\partial\Omega_e \cap \partial\Omega_n = \emptyset$. Vector $\mathbf{n}$ denotes the unit outward normal to $\partial\Omega$. The specific heat capacity c_p and thermal conductivity k are assumed to have the form

$$\begin{aligned} c_p(c, T) &= c P_{\text{PDMS}}(T) + (1-c) P_{\text{PEMS}}(T), \\ k(c, T) &= c Q_{\text{PDMS}}(T) + (1-c) Q_{\text{PEMS}}(T), \end{aligned} \tag{64}$$

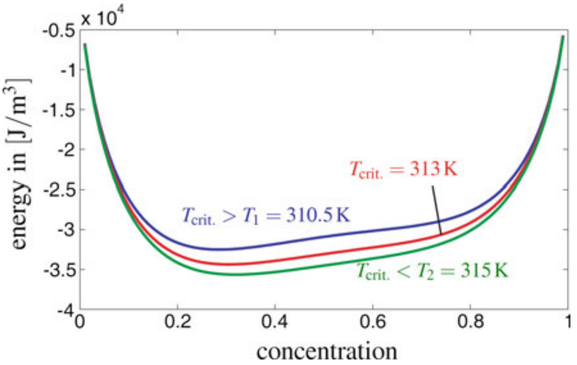
where P_{PDMS}, P_{PEMS}, Q_{PDMS} and Q_{PEMS} are polynomial functions of degree 3 which fit the specific heat capacity and thermal conductivity of the two species for $T \in [50\,\text{K}, 340\,\text{K}]$ and $T \in [230\,\text{K}, 410\,\text{K}]$, respectively. We use an Onsager coefficient of $M = 1.815 \times 10^{-18}\,\text{m}^5/\,(\text{J s})$ and a thermal diffusivity of $D_T = -2 \times 10^{-13}\,\text{m}^2/\,(\text{s K})$. Additional parameters are summarized in Table 1. The material data as well as the experimental results are obtained from [17, 24–26].

Since PDMS is a thermophilic polymer its thermal diffusion coefficient has a negative sign. The corresponding unit volume is assumed as $v = 4\pi\,(\sigma_A/2)^3\,/3$. The critical temperature of the considered polymer blend is $T_{\text{crit.}} = 313\,\text{K}$ and the critical interaction parameter $\chi_{\text{crit.}}$ evaluates to $\chi_{\text{crit.}} = 0.0084318$. At this point it is

Table 1 Material parameters for PDMS (A) and PEMS (B)

N_A (–)	N_B (–)	$\mathfrak{a}$ (–)	$\mathfrak{b}$ (K)	κ	σ_A (nm)	σ_B (nm)
219.4	257.25	-1.86×10^{-3}	3.22	9×10^{-9}	0.583	0.64

Fig. 2 Shapes of the configurational energy density at different temperatures



important to mention that the chosen value for $\mathfrak{a} = -1.8557 \times 10^{-3}$ significantly differs from the value given in [25]. There the authors suggest $\mathfrak{a} = 2.9 \times 10^{-3}$, which is obviously erroneous, because it does not fit $\chi_{\text{crit.}}$. The approximation procedure for the values of the interaction parameter χ prescribes negative values for $\mathfrak{a}$ and positive values for $\mathfrak{b}$ in such a manner that it holds

$$\chi\,(T) = \begin{cases} \chi_{\text{crit.}}, & \text{for } T = T_{\text{crit.}} \\ \chi\,(T) < \chi_{\text{crit.}}, & \text{for } T > T_{\text{crit.}} \\ \chi\,(T) > \chi_{\text{crit.}}, & \text{for } T < T_{\text{crit.}}. \end{cases} \tag{65}$$

Our value for $\mathfrak{a}$ perfectly fits these conditions. The corresponding plot of the configurational energy density is presented in Fig. 2.

4 Numerical Approximation

For finite element analysis we reformulate our coupled diffusion model in a variational form. Note that the mass diffusion equation involves spatial derivatives of fourth order. Thus, we define the spaces of admissible test functions $\mathscr{V}^c = \{\delta c \in \mathscr{H}^2(\Omega) \mid \nabla \delta c \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \times (0, \tilde{t})\}$ and $\mathscr{V}^T = \{\delta T \in \mathscr{H}^1(\Omega) \mid \delta T = 0 \text{ on } \partial\Omega_e \times (0, \tilde{t})\}$ where $H^1\,(\Omega)$ and $H^2\,(\Omega)$ are the Sobolev space of square integrable functions with square integrable derivatives of first and of second order, respectively. The weak forms of the boundary-value problems (60) and (61) follows as

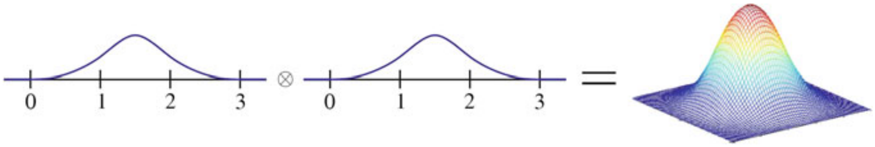


Fig. 3 Two-dimensional B-spline of order $p = 2$

$$\begin{aligned}
&\rho \int_\Omega c_p \frac{\partial T}{\partial t} \delta T \, \mathrm{d}V + \int_\Omega k \nabla T \cdot \nabla \delta T \, \mathrm{d}V - \alpha \int_\Omega I \delta T \, \mathrm{d}V = 0, \\
&\int_\Omega \frac{\partial c}{\partial t} \delta c \, \mathrm{d}V + M \int_\Omega \nabla \partial_c \Psi^{\mathrm{con}} \cdot \nabla \delta c \, \mathrm{d}V + D_T \int_\Omega c(1-c) \nabla T \cdot \nabla \delta c \, \mathrm{d}V \\
&\qquad + \lambda M \int_\Omega \Delta c \Delta \delta c \, \mathrm{d}V = 0,
\end{aligned} \tag{66}$$

for all $\delta c \in \mathscr{V}^c$, $\delta T \in \mathscr{V}^T$. Clearly, the variational formulation of the problem requires approximation functions which are piecewise smooth and globally C^1-continuous. For this reason we decide to employ B-splines as finite element basis (Fig. 3).

A multivariate B-spline basis of degree $\mathbf{p} = [p_1, \ldots, p_d]$ and dimension $d \in \mathbb{N}$ is defined by the tensor product $\Theta_1 \otimes \cdots \otimes \Theta_d$ of knot vectors, built by a sequence of knots $\Theta_l = [\xi_1^l \le \xi_2^l \le \cdots \le \xi_{\mathfrak{n}_l + p_l + 1}^l]$, $l \in \{1, \ldots, d\}$. In the absence of repeated knots, the partition $[\xi_{i_1}^1, \xi_{i_1+1}^1] \times \cdots \times [\xi_{i_d}^d, \xi_{i_d+1}^d]$ forms an element of the mesh in the parametric domain. A single multivariate B-spline B^A, $A \in [1, \ldots, \mathfrak{n}]$, $\mathfrak{n} := \mathfrak{n}_1 \ldots \mathfrak{n}_d$, is then defined by

$$B^A = B_{\mathbf{p}}^{\mathbf{i}}(\boldsymbol{\xi}) = B_{\mathbf{p}}^{\mathbf{i}}(\xi^1, \ldots, \xi^d) = \prod_{l=1}^{d} N_{i_l, p_l}(\xi^l), \tag{67}$$

with multi-index $\mathbf{i}$=$[i_1, \ldots, i_d]$ and $\mathrm{supp}(B^A) = [\xi_{i_1}^1, \xi_{i_1+p_1+1}^1] \times \cdots \times [\xi_{i_d}^d, \xi_{i_d+p_d+1}^d]$. It provides the necessary support for the required continuity. The recursive definition of a univariate B-spline is given as follows

$$N_{i_l, p_l}(\xi) = \frac{\xi - \xi_{i_l}^l}{\xi_{i_l+p_l}^l - \xi_{i_l}^l} N_{i_l, p_l - 1}(\xi) + \frac{\xi_{i_l+p_l+1}^l - \xi}{\xi_{i_l+p_l+1}^l - \xi_{i_l+1}^l} N_{i_l+1, p_l-1}(\xi), \tag{68}$$

starting with piecewise constant functions

$$N_{i_l,0}(\xi) = \begin{cases} 1 & \text{if } \xi_{i_l}^l \le \xi < \xi_{i_l+1}^l \\ 0 & \text{otherwise} \end{cases} . \tag{69}$$

Linear independence as a fundamental property of finite element basis as well as local support are given by a B-spline basis. Moreover, smoothness is related to knot multiplicity, i.e., the number of repetitions in Θ at node $\mathbf{i}$. Unfortunately, the tensor product structure in (67) impedes standard local refinement strategies which motivated us to introduce a specific hierarchical refinement strategy in [14].

For temporal discretization the considered time interval $[0, \bar{t}]$ is divided into n_t subintervals. The first order time derivative is approximated by finite differences

$$\frac{\partial c}{\partial t} = \frac{c_{n+1} - c_n}{\Delta t} \tag{70}$$

with time step $\Delta t = t_{n+1} - t_n$. The time integration is performed by an implicit Crank-Nicholson scheme, known to be second-order accurate. Now the fully discretized problem reads

$$\begin{aligned}
&\frac{\rho}{\Delta t}\int\limits_{\Omega} c_{p,n+1/2}(T_{n+1} - T_n)B^A \,\mathrm{d}V + \int\limits_{\Omega} k_{n+1/2}\nabla T_{n+1/2}\cdot\nabla B^A \,\mathrm{d}V - \alpha\int\limits_{\Omega} I B^A \,\mathrm{d}V = 0,\\
&\frac{1}{\Delta t}\int\limits_{\Omega}(c_{n+1} - c_n)B^A \,\mathrm{d}V + M\int\limits_{\Omega}(\partial_c^2 \Psi^{\mathrm{con}}_{n+1/2}\nabla c_{n+1/2} + \partial_c\partial_T\nabla T_{n+1/2})\cdot\nabla B^A \,\mathrm{d}V\\
&\qquad + D_T\int\limits_{\Omega} c_{n+1/2}(1 - c_{n+1/2})\nabla T_{n+1/2}\cdot\nabla B^A \,\mathrm{d}V + \lambda M\int\limits_{\Omega}\Delta c_{n+1/2}\Delta B^A \,\mathrm{d}V = 0,
\end{aligned} \tag{71}$$

for all $A \in \{1, \ldots, \mathfrak{n}\}$.

5 Simulation Results

In our numerical simulations phase separation induced by laser light absorption is studied. The setting for simulation is arranged in such a manner that the polymer blend is homogeneously quenched from the one-phase regime and simultaneously is heated. In order to integrate a focused laser spot into the model, a time independent heat source of the form $q(\mathbf{x}) = \alpha\, I(\mathbf{x})$ with intensity

$$I(\mathbf{x}) = I_0(\mathbf{x})\sum_{i,j=1}^{\infty} a_{ij}\sin\left(\frac{i\pi x_1}{l_{x_1}}\right)\sin\left(\frac{j\pi x_2}{l_{x_2}}\right) \tag{72}$$

is used. The laser spot points at the middle of the upper surface of an aged cuboid of side length $l_x = l_y = 1.5\,\mu$m and hight $0.75\,\mu$m, see Fig. 4, with the diameter of the spot being about 3/5 of the domain size in the two-dimensional and 1/5 in the tree-dimensional model. The laser intensity and the optical absorption coefficient are given by $I_0 = 3.75 \times 10^7\,\mathrm{W\,m^{-2}}$ and $\alpha = 500\,\mathrm{m^{-1}}$, respectively. In both settings we consider a critical PDMS-PEMS blend with a composition of 55 % PDMS and

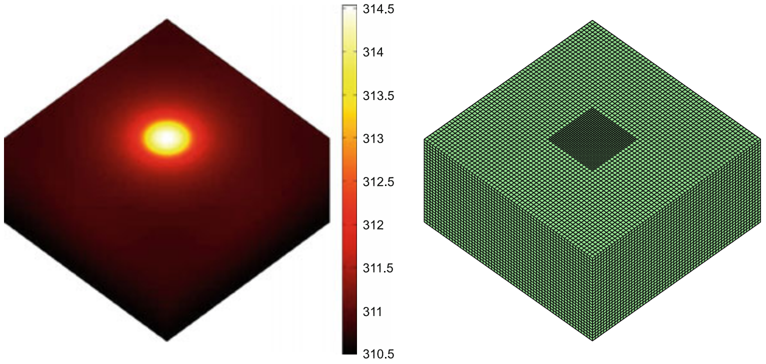


Fig. 4 Temperature field with laser intensity (72) in K (*left*) and computational mesh (*right*) for the three-dimensional simulation

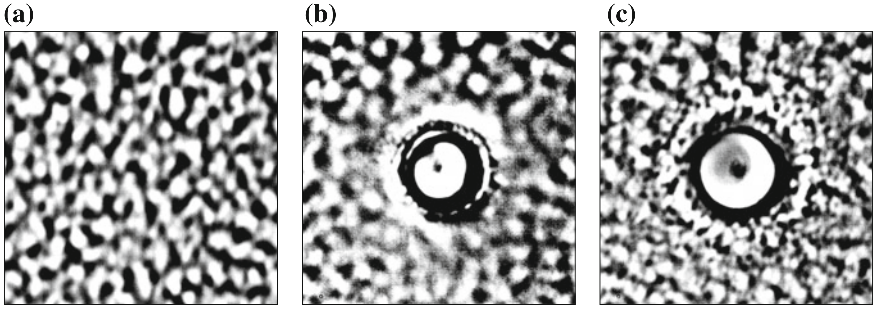


Fig. 5 Micrographs of a PDMS-PEMS polymer blend exposed to a focused laser spot. The specimen size is about 100 μm × 100 μm. Subfigures are adapted from [24]

45 % PEMS. Therefore, the initial concentration is set $c_0\left(\mathbf{x}\right) = 0.55$ including slight randomly perturbed inhomegeneities (±1 %).

Köhler et al. studied the microstructural evolution of a PDMS-PEMS blend subjected to a focused laser spot in [24, 25]. In their experimental set-up the blend under consideration was initially quenched into the spinodal regime for approximately 120 min before exposing it to the laser, see Fig. 5a. At this moment the blend has already completed phase decomposition and reached an intermediate state of coarsening. The recorded experimental observations indicate the following scenario. Starting from a homogeneous one-phase configuration the early stages of microstructural evolution are driven by spinodal decomposition of phases. After the sample is exposed to laser illumination, thermal forcing asserts itself gradually against the suppression of phase decomposition. Over time the inhomogeneous temperature field increasingly coins the microstructure. Figure 5b, c show the sample after an exposure of 100 and 200 s, respectively. Due to the local heating, a circle of PDMS evolves in the middle of the micrograph. Since a PDMS enrichment takes place at the heated domain within the sample an outer ring of PEMS-rich material surrounds the concentric PDMS circle. The structures slowly propagate through the sample like spherical waves in a medium. Outside the heated spot a more irregular spinodal pattern unfolds due to the decreasing impact of thermal diffusion, see Fig. 5 and [24, 25].

The simulation setting is also arranged in such a manner that the polymer blend is at first homogeneously quenched from the one-phase regime, see Fig. 6. Here and below the reddish areas denote the PDMS-rich α-phase and the blue domains represent the PEMS-rich β-phase. Heating with the laser spot starts at a system time $t = 0.03$ s. Similar to the experimental observations a spherical PDMS-rich phase emerges in the center of the domain, where the laser induces an enrichment of the thermophilic PDMS. The PDMS is attracted from the surrounding area, and—like in the experiments—a second ring of a PEMS-rich phase evolves around the heated spot. It seems that the PEMS-ring spreads like a spherical wave starting at the spot in the middle of the domain Ω. In the outer part of Ω the thermal effect decreases and more irregular spinodal pattern of phases become prevalent. Note that since the relaxation into thermal equilibrium occurs on a time scale that is by a factor of 1000 faster then the time scale at which diffusion takes place, we simplified the two-dimensional model by assuming quasistatic equilibrium for the heat equation.

As we continue in time, a typical coarsening process dominates the evolution of phases outside the laser spot. The microstructure within the exposure of the laser spot is characterized by a PDMS-rich concentric ring structure, which grows towards the center of the laser illumination, see Fig. 7. After the PDMS enrichment within the

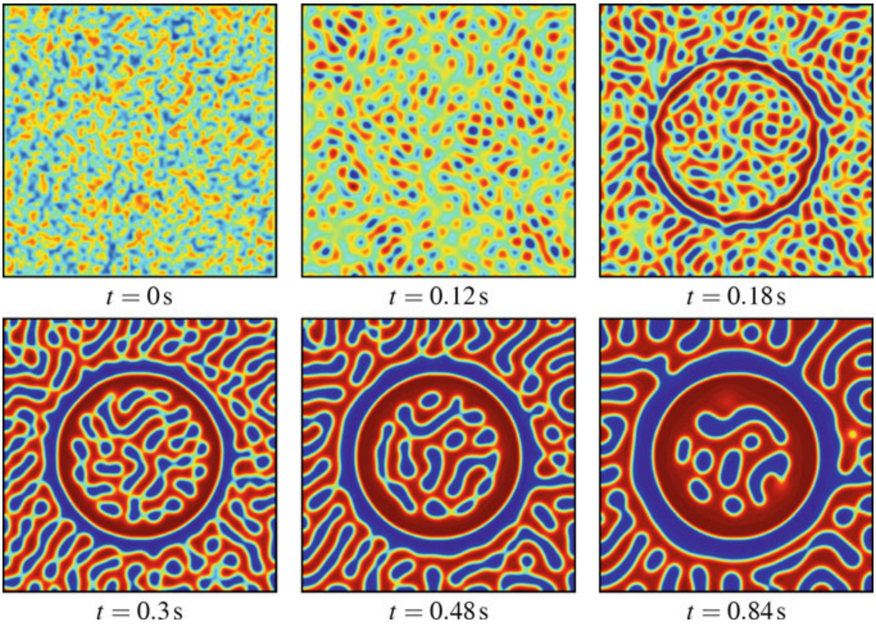


Fig. 6 Phase decomposition and early stages of coarsening for a critical PDMS-PEMS polymer blend subjected to a focused laser spot

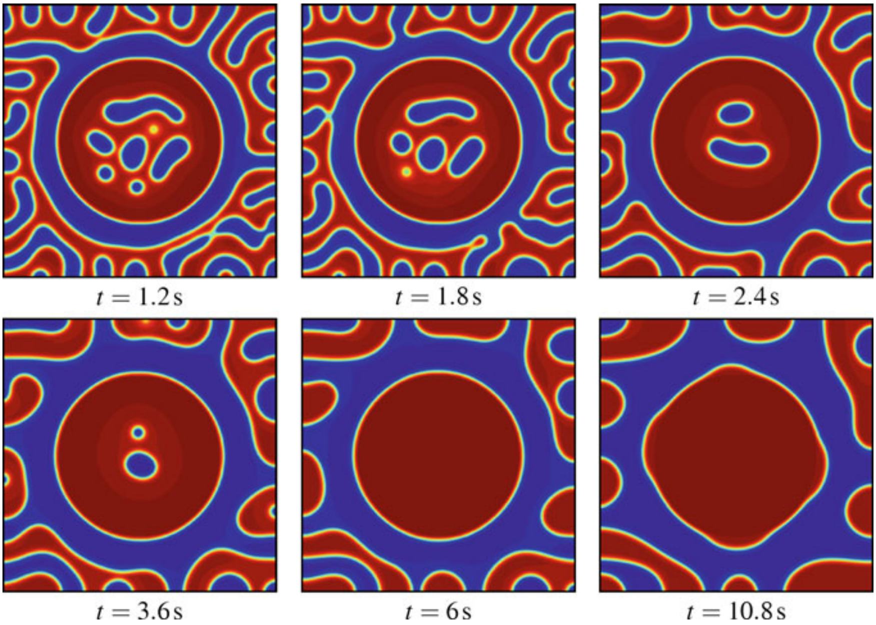


Fig. 7 Phase coarsening of the PDMS-PEMS polymer blend subjected to a focused laser spot (*red* PDMS, *blue* PEMS)

heated spot is finished and the PEMS material disappeared from the center of the spot at $t = 6$ s, the phase coarsening process takes the leading role in microstructural evolution. In the very late stages additional PDMS material from the surroundings is deposited at the PDMS circle.

In order to integrate the focused laser spot into the three-dimensional simulation, a laser intensity $I_0(\mathbf{x}) = 3.2 \times 10^{11}$ W m^{-2} is focused on a domain $\bar{\Omega}$ which is defined by

$$\bar{\Omega} := \left\{ \mathbf{x} \mid \left(x_1 - \frac{3}{4}\,\mu\text{m} \right)^2 + \left(x_2 - \frac{3}{4}\,\mu\text{m} \right)^2 \leq \frac{3}{20}\,\mu\text{m},\ \frac{29}{40}\,\mu\text{m} \leq x_3 \leq \frac{3}{4}\,\mu\text{m} \right\}.$$

Outside the domain $\bar{\Omega}$ the laser intensity is zero.

In order to locally refine the computational mesh around the domain $\bar{\Omega}$, a hierarchical refinement scheme based on B-spline subdivision is used, see [14] for details. The resulting mesh is shown in Fig. 4. Periodic boundary conditions on the left and right, and front and rear surfaces of the computational domain are applied.

At the beginning of the simulation the laser intensity $I_0(\mathbf{x})$ is set for approximately 2.5 s to zero so that a typical coarsening scenario can take place, see the pictures in the upper row of Fig. 8. Starting at time $t = 2.4691$ s the middle of the upper surface, i.e., the region $\bar{\Omega}$, is heated by the laser spot. As can be seen in the second row of Fig. 8,

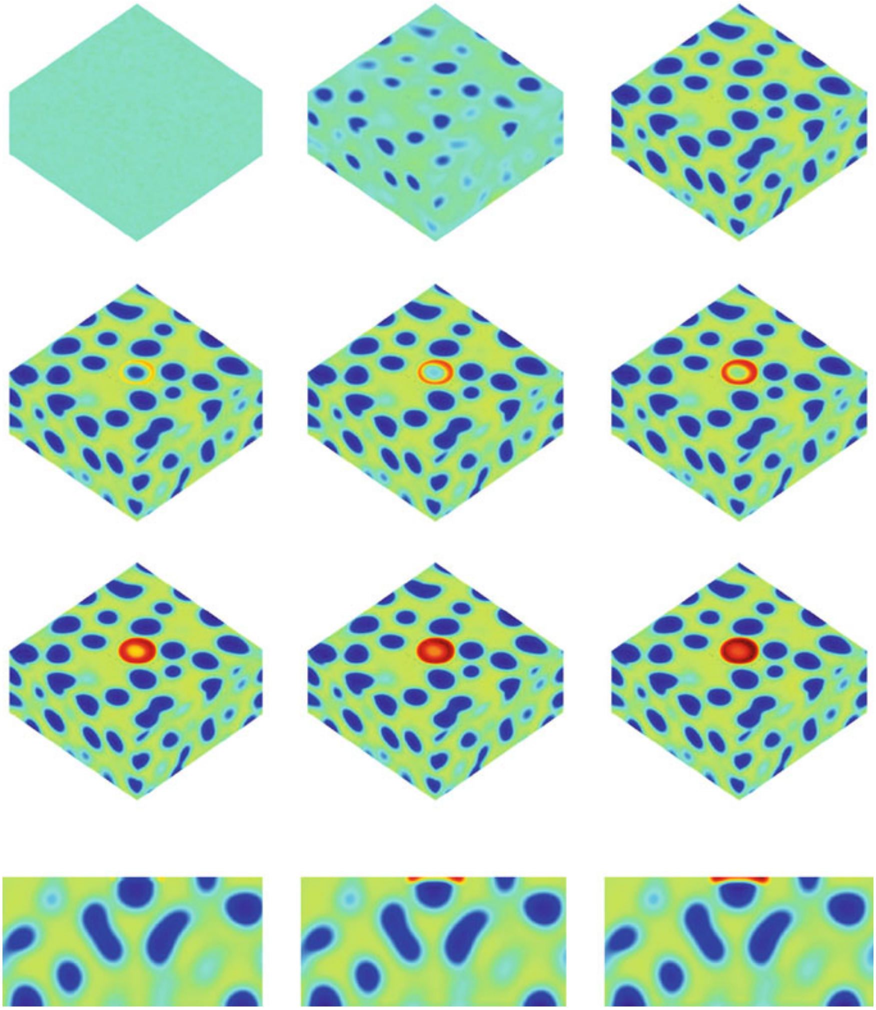


Fig. 8 Decomposition, phase coarsening and thermophoresis in a cuboid of critical PDMS-PEMS polymer blend, (*red* PDMS, *blue* PEMS). *First row* phase evolution without heat supply at start $t = 0$ and $t = 1.75$ s, $t = 2.4691$ s; *second* and *third row* results with a focused laser spot at times $t = [2.4697\,\text{s}, 2.4703\,\text{s}, 2.4709\,\text{s}, 2.4715\,\text{s}, 2.4721\,\text{s}, 2.4727\,\text{s}]$; *fourth row* profile of the heated region at times $t = 2.4697$ s, $t = 2.4715$ s and $t = 2.4727$ s

a ring of PDMS forms around the heated region. The subsequent pictures show an enrichment of PDMS in the heated center whereas the microstructural evolution in the remaining region is mainly driven by spinodal decomposition of phases. The last row of Fig. 8 illustrates the distribution across the depth where we see that thermophoresis mainly takes place at the heated surface, whereas the decomposition of phases in the remaining domain is a slower effect.

6 Conclusions

In this contribution we have presented an extended version of the commonly known Cahn-Hilliard phase-field model in order to capture diffusion phenomena induced by local non-uniform temperature gradients. Our diffusion model was formulated consistently with the basic laws of thermodynamics. Its discrete version was embedded into the isogeometric finite element concept in order to perform numerical simulations. The simulations were compared to experimental studies of PDMS-PEMS blends by the application of realistic material parameters.

Acknowledgments The authors gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) under the grants WE2525/2-3, WE2525/4-1 and WE2525/8-1.

References

1. Anders, D., & Weinberg, K. (2011). A variational approach to the decomposition of unstable viscous fluids and its consistent numerical approximation. *ZAMM—Zeitschrift für angewandte Mathematik und Mechanik*, *91*(8), 609–629.
2. Anders, D., & Weinberg, W. (2012). Thermophoresis in binary blends. *Mechanics of Materials*, *47*, 33–50.
3. Anders, D., Reichert, R., & Weinberg, K. (2011). Isogeometric analysis of thermal diffusion in binary blends. *Computational Materials Science*, *52*(1), 182–188.
4. Anders, D., Hoffmann, A., Scheffler, H.-P., & Weinberg, K. (2011). Application of operator-scaling anisotropic random fields to binary mixtures. *Philosophical Magazine*, *91*(29), 3766–3792.
5. Baaske, P., Wienken, C. J., Reineck, P., Duhr, S., & Braun, D. (2010). Optical thermophoresis for quantifying the buffer dependence of aptamer binding. *Angewandte Chemie International Edition*, *49*, 2238–2241.
6. Baumgärtner, A., & Heermann, D. W. (1986). Spinodal decomposition of polymer films. *Polymer*, *27*(11), 1777–1780.
7. Cahn, J. W. (1961). On spinodal decomposition. *Acta Metallurgica*, *9*(9), 795–801.
8. Cimmelli, V. A., Jou, D., Ruggeri, T., & Ván, P. (2014). Entropy principle and recent results in non-equilibrium theories. *Entropy*, *16*(3), 1756.
9. de Gennes, P. G. (1980). Dynamics of fluctuations and spinodal decomposition in polymer blends. *Journal of Chemical Physics*, *72*(9), 4756–4763.
10. de Groot, S. R., & Mazur, P. (1962). *Non-equilibrium thermodynamics*. Amsterdam: North-Holland.
11. Enge, W., & Köhler, W. (2004). Thermal diffusion in a critical polymer blend. *Physical Chemistry Chemical Physics*, *6*, 2373–2378.
12. Flory, P. J. (1942). Thermodynamics of high polymer solutions. *Journal of Chemical Physics*, *10*(1), 51–61.
13. Hashimoto, T., Kumaki, J., & Kawai, H. (1983). Time-resolved light scattering studies on kinetics of phase separation and phase dissolution of polymer blends. 1. Kinetics of phase separation of a binary mixture of polystyrene and poly(vinyl methyl ether). *Macromolecules*, *16*(4), 641–648.
14. Hesch, C., Schuß, S., Dittmann, M., Franke, M., & Weinberg, K. (2016). Isogeometric analysis and hierarchical refinement for higher-order phase-field models. *Computer Methods in Applied Mechanics and Engineering*, *303*, 185–207.

15. Huggins, M. L. (1942). Theory of solutions of high polymers. *Journal of the American Chemical Society*, *64*(7), 1712–1719.
16. Itskov, M. (2007). *Tensor algebra and tensor analysis for engineers (with applications to continuum mechanics)*. Berlin: Springer.
17. Krekhov, A. P., & Kramer, L. (2004). Phase separation in the presence of spatially periodic forcing. *Physical Review E*, *70*, 061801.
18. Lebon, G., Jou, D., & Casas-Vázquez, J. (2007). Understanding non-equilibrium thermodynamics. In *Foundations, Applications, Frontiers*. Springer: Berlin.
19. Lee, K.-W. D., Chan, P. K., & Feng, X. (2002). A computational study of the thermal-induced phase separation phenomenon in polymer solutions under a temperature gradient. *Macromolecular Theory and Simulations*, *11*, 996–1005.
20. Lee, K.-W. D., Chan, P. K., & Feng, X. (2003). A computational study of the polymerization-induced phase separation phenomenon in polymer solutions under a temperature gradient. *Macromolecular Theory and Simulations*, *12*(6), 413–424.
21. Lee, K.-W. D., Chan, P. K., & Feng, X. (2004). Morphology development and characterization of the phase-separated structure resulting from the thermal-induced phase separation phenomenon in polymer solutions under a temperature gradient. *Chemical Engineering Science*, *59*(7), 1491–1504.
22. Strobl, G. R. (1985). Structure evolution during spinodal decomposition of polymer blends. *Macromolecules*, *18*(3), 558–563.
23. Thamdrup, L. H., Larsen, N. B., & Kristensen, A. (2010). Light-induced local heating for thermophoretic manipulation of DNA in polymer micro- and nanochannels. *Nano Letters*, *10*(2), 826–832.
24. Voit, A. (2007). *Photothermische Strukturierung binärer Polymermischungen*. Ph.D. thesis, University of Bayreuth.
25. Voit, A., Krekhov, A., Enge, W., Kramer, L., & Köhler, W. (2005). Thermal patterning of a critical polymer blend. *Physical Review Letters*, *94*, 214501.
26. Voit, A., Krekhov, A., & Köhler, W. (2007). Quenching a UCST polymer blend into phase separation by local heating. *Macromolecules*, *40*, 9–11.
27. Würger, A. (2007). Das Salz in der DNA-Suppe. *Physik Journal*, *7*(22–24), 2014.