S. Rionero
G. Romano (Eds)

Trends and Applications of Mathematics to Mechanics

STAMM 2002

Springer

S. Rionero

G. Romano

Trends and Applications of Mathematics to Mechanics

S. Rionero (Editor)
G. Romano (Editor)

Trends and Applications of Mathematics to Mechanics

STAMM 2002

With 35 Figures

Springer

Salvatore Rionero
Dept. of Mathematics Renato Caccioppoli
University of Naples Federico II
Naples, Italy

Giovanni Romano
Dept. of Scienza delle Costruzioni
University of Naples Federico II
Naples, Italy

Library of Congress Control Number: 2004116220

ISBN 88-470-0269-9 Springer Milan Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in Italy

Cover-Design: Simona Colombo, Milan
Typesetting: PTP-Berlin Protago-TEX-Production Gmbh, Germany
Printing and Binding: Signum, Bollate (Mi)

Preface

The book offers a selection of papers most of which are revised and enriched versions of the contributions presented at the 12th Symposium on Trends of Applications of Mathematics to Mechanics (STAMM) which was sponsored by the International Society for the Interaction between Mathematics and Mechanics (ISIMM) and held in Maiori (Salerno), Italy, from September 29th to October 4th, 2002. The Symposium attracted leading researchers from around the world who are working at the interface between mathematics and mechanics. The importance of a close link between these two disciplines has long been recognized; each benefits from and is stimulated by open problems, methods and results emerging from the other. The book comprises 22 papers which report specialized investigations and which contribute broader presentations of linear and nonlinear problems.

It is with the deepest gratitude to the authors who have contributed to the volume and to the publisher, for its highly professional assistance, that the editors submit this book to the international mathematics and mechanics communities. The editors gratefully acknowledge the financial support of the following institutions: Gruppo Nazionale di Fisica Matematica (GNFM) of the Istituto Nazionale di Alta Matematica (INDAM), Università degli Studi di Napoli Federico II and Regione Campania.

Naples, Italy, October 2004

Salvatore Rionero
Giovanni Romano

Contents

List of Contributors

- **Florinda Capone,** Dept. of Mathematics Renato Caccioppoli, University of Naples Federico II, Naples, Italy
- **Bernardino D'Acunto,** Dept. of Mathematics Renato Caccioppoli, University of Naples Federico II, Naples, Italy
- **Fabio De Angelis,** Dept. of Scienza delle Costruzioni, University of Naples Federico II, Naples, Italy
- **Monica De Angelis,** Dept. of Mathematics Renato Caccioppoli, University of Naples Federico II, Naples, Italy
- **Marco Degiovanni,** Dept. of Mathematics and Physics, Università Cattolica del Sacro Cuore, Brescia, Italy
- **Marina Diaco,** Dept. of Scienza delle Costruzioni, University of Naples Federico II, Naples, Italy
- **James N. Flavin,** Dept. of Mathematical Physics, National University of Ireland, Galway, Ireland
- **Maurizio Gentile,** Dept. of Mathematics Renato Caccioppoli, University of Naples Federico II, Naples, Italy
- **Barry Gleeson,** Dept. of Geomatic Engineering, University College London, London U.K.
- **Yongzhong Huo,** Shanghai Institute of Ceramics, Shanghai
- **Mikhail Itskov,** Dept. of Applied Mechanics and Fluid Dynamics, Lehrstuhl für Technische Mechanik und Strömungsmechanik, University of Bayreuth, Bayreuth, Germany
- **Francesco Marotti de Sciarra,** Dept. of Scienza delle Costruzioni, University of Naples Federico II, Naples, Italy
- **Alfredo Marzocchi,** Dept. of Mathematics, University of Brescia, Brescia, Italy
- **Ingo Müller,** Technical University Berlin, Institute of Process Engineering, Thermodynamics, Berlin, Germany
- **Alessandro Musesti,** Dept. of Mathematics and Physics, Università Cattolica del Sacro Cuore, Brescia, Italy
- **AnnaMaria Monte,** Dept. of Mathematics Renato Caccioppoli, University of Naples Federico II, Naples, Italy
- **Daniel Reitebuch,** Technical University Berlin, Institute of Process Engineering, Thermodynamics, Berlin, Germany
- **Salvatore Rionero,** Dept. of Mathematics Renato Caccioppoli, University of Naples Federico II, Naples, Italy

- **Alessandra Romano,** Dept. of Scienza delle Costruzioni, University of Naples Federico II, Naples, Italy
- **Giovanni Romano,** Dept. of Scienza delle Costruzioni, University of Naples Federico II, Naples, Italy
- **Tommaso Ruggeri,** Research Center of Applied Mathematics, C.I.R.A.M., University of Bologna, Bologna, Italy
- **Giovanni Russo,** Dept. of Mathematics and Informatics, University of Catania, Catania, Italy
- **Carmen Sellitto,** Dept. of Scienza delle Costruzioni, University of Naples Federico II, Naples, Italy
- **Henning Struchtrup,** Dept. of Mechanical Engineering, University of Victoria, Victoria, Canada
- **Manuel Torrilhon,** Seminar for Applied Mathematics, ETH Zurich, Switzerland
- **Massimo Trovato,** Dept. of Mathematics and Informatics, University of Catania, Catania, Italy
- **Doretta Vivona,** Dept. of Mathematical Models for Applied Science, University of Rome La Sapienza, Rome, Italy
- **Krzysztof Wilmanski,** Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

On the instability of double diffusive convection in porous media under boundary data periodic in space

F. Capone, S. Rionero

Abstract. The linear instability analysis of the motionless state for a binary fluid mixture in a porous layer, under horizontal periodic temperature and concentration gradients, is performed.

1 Introduction

Let $Oxyz$ be a cartesian frame of reference with the z-axis vertically upwards and $S = \boldsymbol{R}^2 \times [0, d]$ a horizontal (infinite) porous layer filled by a binary fluid mixture. When S is strictly uniformly heated and salted from below, the onset of convection in S has been widely studied since it has many geophysical and technical applications [2–4,8–12]. Recently an analysis of this problem was carried out under the assumption of a linear spatial variation of temperature and concentration along the boundaries [6,7,13]. This assumption is more realistic than strictly uniform heating and salting but, since it implies infinite temperature and concentration on the horizontal planes at large spatial distances, it does not appear to be completely acceptable. In order to overcome this problem, a diffusion-convection model driven by horizontally periodic temperature and concentration gradients was considered in [5]. Specifically, in [5], only a condition of global nonlinear stability of the steady state has been obtained. In the present paper we reconsider the problem in order to determine an instability condition when $\varepsilon\, Le > 1$. The basic proof is based on the instability theorem obtained by Rionero in [14] for a general binary reaction-diffusion system of PDE.

The plan of the paper is as follows. Section 2 is devoted to the mathematical statement of the convection problem for a double diffusive fluid mixture modelled by the Darcy-Oberbeck-Boussinesq (DOB) equations. In Sect. 3, by applying the general instability theorem obtained in [14], we obtain an instability condition.

2 Basic equations and steady state solution

The Darcy-Oberbeck-Boussinesq equations governing the motion of a binary porous fluid mixture are [2–4]:

$$\begin{cases} \nabla p = -(\mu/k)\,\mathbf{v} + \rho_0[1 - \gamma_T(T - T_0) + \gamma_C(C - C_0)]\mathbf{g}, \\ \\ \nabla \cdot \mathbf{v} = 0, \\ \\ \dfrac{(\rho_0\, c)_m}{(\rho_0\, c_p)_f}\, T_t + \mathbf{v} \cdot \nabla T = k_T \Delta T, \\ \\ \varphi C_t + \mathbf{v} \cdot \nabla C = k_C \Delta C, \end{cases} \tag{1}$$

with:

γ_C = solute expansion coefficient,
γ_T = thermal expansion coefficient,
T_0 = reference temperature,
$\mathbf{v}$ = seepage velocity field,
μ = viscosity,
k_T = thermal diffusivity,
c = specific heat of the solid,
c_p = specific heat of fluid at constant pressure ,
p = pressure field,
φ = porosity of the medium,
C_0 = reference concentration,
C = concentration field,
T = temperature field,
k_C = salt diffusivity,
ρ_0 = fluid density at T_0,

and the subscripts m and f refer to the porous medium and the fluid, respectively. To (1) we append the boundary conditions

$$\begin{aligned} &T_L(x) = \beta_T^* \sin(x/d) + T_1, \;\; C_L(x) = \beta_C^* \sin(x/d) + 2C_1 \qquad \text{on } z = 0, \\ &T_U(x) = (\beta_T^*/e)\, \sin(x/d), \quad C_U(x) = (\beta_C^*/e)\, \sin(x/d) + C_1 \;\; \text{on } z = d, \end{aligned} \tag{2}$$

in which $T_1 > 0$, $C_1 > 0$; here β_C^* and β_T^* are two positive constants having, respectively, the dimensions of concentration and temperature such that

$$\beta_T^* = \beta_C^* \frac{\gamma_C}{\gamma_T}\, (> 0). \tag{3}$$

On assuming

$$0 < \beta_C^* < \min\left\{ \frac{e}{e-1} \frac{\gamma_T}{\gamma_C}\, T_1,\, C_1 \right\}, \tag{4}$$

we see that (1-3) admit the steady state solution (motionless state) [5]:

$$\mathbf{v}_s(x) = 0, \quad p_s(z) = -\rho_0 g[A_1\, z + (\gamma_T T_1 - \gamma_C C_1)/(2d)\, z^2], \tag{5}$$

$$T_s(x, z) = \beta_T^* e^{-z/d} \sin(x/d) + T_1(1 - z/d), \tag{6}$$

$$C_s(x, z) = \beta_C^* e^{-z/d} \sin(x/d) + C_1(2 - z/d), \tag{7}$$

where $A_1 = 1 + \gamma_C(2C_1 - C_0) - \gamma_T(T_1 - T_0)$ is a constant. The solution (5-7), because of (4), corresponds to the case of a porous fluid mixture layer heated and salted from below.[1]

[1] We observe that the temperature and concentration fields $T_s(x, z)$, $C_s(x, z)$ exhibit a bounded periodic behaviour in the unbounded x-direction.

Let $\mathbf{u} = (u, v, w)$, θ, Γ, π be the perturbations to the (seepage) velocity, temperature, concentration and (reduced) pressure fields, respectively, and introduce the non-dimensional quantities [4]:

$$\mathbf{x} = d\,\mathbf{x}^*\,, \qquad t = \frac{A\,d^2}{k_T}t^*\,, \qquad \mathbf{u} = \frac{k_T}{d}\mathbf{u}^*\,, \qquad \pi = \frac{\mu\,k_T}{k}\pi^*\,,\ \theta = \tilde{T}\,\theta^*\,,$$

$$\beta_T = \frac{\beta_T^*}{\sqrt{R}\,\tilde{T}}\,,\ N^2 = \frac{C_1\gamma_C}{T_1\gamma_T}, \qquad \beta_C = \frac{\beta_C^*}{\sqrt{\mathscr{C}\,Le}\,\tilde{C}}, \qquad Le = \frac{k_T}{k_C}, \qquad \Gamma = \tilde{C}\,\Gamma^*\,,$$

$$\varepsilon = \frac{\varphi}{A}\,, \qquad \tilde{T} = \sqrt{\frac{T_1\mu k_T}{g\gamma_T\rho_0 k d}}\,,\ \tilde{C} = \sqrt{\frac{C_1\mu k_T^2}{g\gamma_C\rho_0 k k_c d}}\,,\ A = \frac{(\rho_0 c)_m}{(\rho_0 c_p)_f}$$

$$R = \frac{\rho_0 g\gamma_T k d T_1}{\mu k_T} \quad \text{the vertical thermal Rayleigh number,}$$

$$\mathscr{C} = \frac{\rho_0 g\gamma_C k d C_1}{\mu k_T} \quad \text{the vertical solutal Rayleigh number}\,.$$

In particular, on taking into account (4), one obtains

$$0 < \beta_C < \frac{1}{Le}\min\left\{\frac{e}{e-1}\frac{1}{N^2},\ 1\right\}. \tag{8}$$

Dropping all asterisks, we see that, for all $(x, y, z) \in \boldsymbol{R}^2 \times [0, 1]$,

$$\begin{cases} \nabla\pi = -\mathbf{u} + (\sqrt{R}\,\theta - \sqrt{Le\,\mathscr{C}}\,\Gamma)\mathbf{k}, \\ \nabla\cdot\mathbf{u} = 0, \\ \theta_t + \ \mathbf{u}\cdot\nabla\theta = -\beta_T\sqrt{R}\,\mathbf{e}\cdot\mathbf{u} + \sqrt{R}w + \Delta\theta, \\ \varepsilon\,Le\,\Gamma_t + Le\,\mathbf{u}\cdot\nabla\Gamma = -\beta_C\,Le\sqrt{Le\,\mathscr{C}}\,\mathbf{e}\cdot\mathbf{u} + \sqrt{Le\,\mathscr{C}}\,w + \Delta\Gamma, \end{cases} \tag{9}$$

where $\mathbf{e} = (e^{-z}\cos x,\ 0,\ -e^{-z}\sin x)$. To (9) we append the boundary conditions

$$w = \theta = \Gamma = 0 \quad \text{on} \qquad z = 0, z = 1\,. \tag{10}$$

In the sequel we assume that the perturbation fields are periodic functions of x and y of periods $2\pi/a_x$, $2\pi/a_y$, respectively, [2] and we denote the periodicity cell by $\Omega = [0, 2\pi/a_x] \times [0, 2\pi/a_y] \times [0, 1]$. Finally, to ensure that the steady state (5 - 7) is unique, we assume that

$$\int_\Omega u\,d\Omega = \int_\Omega v\,d\Omega = 0\,.$$

[2] When $\alpha \neq 0$, $\tau = 2\pi$ is the only admissible period in the x-direction [5].

3 Instability analysis

Our aim, in this section, is to determine conditions guaranteeing the linear instability of the solution (5-7) when $N^2 < 1$ and $\varepsilon\, Le > 1$. To this end, we consider the linear version of (9), i.e.,

$$\begin{cases} \nabla \pi = -\mathbf{u} + (\sqrt{R}\,\theta - \sqrt{Le\,\mathscr{C}}\,\Gamma)\mathbf{k}, \\ \nabla \cdot \mathbf{u} = 0, \\ \theta_t = -\beta_T \sqrt{R}\,\mathbf{e}\cdot\mathbf{u} + \sqrt{R} w + \Delta\theta, \\ \varepsilon\, Le\, \Gamma_t = -\beta_C\, Le \sqrt{Le\,\mathscr{C}}\,\mathbf{e}\cdot\mathbf{u} + \sqrt{Le\,\mathscr{C}}\, w + \Delta\Gamma\,. \end{cases} \tag{11}$$

On considering the third component of the double curl of $(11)_1$, we obtain

$$\begin{cases} \Delta w = \Delta_1(\sqrt{R}\,\theta - \sqrt{Le\,\mathscr{C}}\,\Gamma), \\ \nabla \cdot \mathbf{u} = 0, \\ \theta_t = -\beta_T \sqrt{R}\,\mathbf{e}\cdot\mathbf{u} + \sqrt{R} w + \Delta\theta, \\ \varepsilon\, Le\, \Gamma_t = -\beta_C\, Le \sqrt{Le\,\mathscr{C}}\,\mathbf{e}\cdot\mathbf{u} + \sqrt{Le\,\mathscr{C}}\, w + \Delta\Gamma\,, \end{cases} \tag{12}$$

in which $\Delta_1 \cdot = \partial_x^2 \cdot + \partial_y^2 \cdot$.

In order to obtain conditions guaranteeing the linear instability of the steady state (5-7), following Remark 2 of [14], we consider the class of perturbations for which the first component of the seepage velocity $\mathbf{u} = (u, v, w)$ is zero, i.e.,

$$u = 0\,.$$

In this case, (12) reduces to

$$\begin{cases} \Delta w = \Delta_1(\sqrt{R}\,\theta - \sqrt{Le\,\mathscr{C}}\,\Gamma), \\ \theta_t = \beta_T \sqrt{R}\, e^{-z} \sin x\, w + \sqrt{R} w + \Delta\theta, \\ \varepsilon\, Le\, \Gamma_t = \beta_C\, Le \sqrt{Le\,\mathscr{C}}\; e^{-z} \sin x\, w + \sqrt{Le\,\mathscr{C}}\, w + \Delta\Gamma\,. \end{cases} \tag{13}$$

We define

$$\begin{aligned} \mathscr{H} = \{ & w, \theta, \Gamma \text{ regular in } \Omega, \text{ periodic in } x \text{ and } y, \text{ satisfying} \\ & (13)_1 \text{ and the boundary conditions (10)}\} \end{aligned} \tag{14}$$

the class of *kinematically admissible perturbations* and choose

$$\begin{cases} w = \alpha(\sqrt{R}\theta - \sqrt{Le\,\mathscr{C}}\;\Gamma), \\ \theta = \hat{\theta}(x, y, t)\, \sin(\pi z), \\ \Gamma = \hat{\Gamma}(x, y, t)\, \sin(\pi z)\,. \end{cases} \tag{15}$$

By requiring that $\hat{\theta}$, $\hat{\Gamma}$ satisfy the plan form equation $\Delta_1 \cdot = -a^2 \cdot$, according to the periodicity of θ and Γ in the x and y directions, from (15) we easily obtain

$$\Delta w = -\xi\, w\,, \qquad \Delta\theta = -\xi\,\theta\,, \qquad \Delta\Gamma = -\xi\,\Gamma\,, \tag{16}$$

with

$$a^2 = a_x^2 + a_y^2\,, \quad \xi = a^2 + \pi^2\,, \quad \alpha = \frac{a^2}{\xi}\,, \tag{17}$$

and hence, as is easily verified, (15) belongs to $\mathcal{H}$. On substituting (15) in $(13)_2 - (13)_3$, since

$$\beta_T = \beta_C\, Le\, N^2\,, \tag{18}$$

one obtains

$$\begin{cases} \theta_t = a_1(x,z)\theta + b_1(x,z)\Gamma, \\ \Gamma_t = c_1(x,z)\theta + d_1(x,z)\Gamma \end{cases} \tag{19}$$

with

$$\begin{cases} a_1(x,z) = \alpha(1 + \beta_C\, Le\, N^2\, e^{-z}\, \sin x)R - \xi, \\ b_1(x,z) = -\alpha\,\sqrt{Le}(1 + \beta_C\, Le\, N^2\, e^{-z}\, \sin x)\sqrt{R\,\mathscr{C}}, \\ c_1(x,z) = \dfrac{\alpha}{\sqrt{Le}\,\varepsilon}(1 + \beta_C\, Le\, e^{-z}\, \sin x)\sqrt{R\,\mathscr{C}}, \\ d_1(x,z) = -\dfrac{\alpha}{\varepsilon}(1 + \beta_C\, Le\, e^{-z}\, \sin x)\mathscr{C} - \dfrac{\xi}{\varepsilon\, Le}\,. \end{cases} \tag{20}$$

To obtain sufficient conditions for the instability of the zero solution of (19), we apply the following theorem [14].

Theorem 1. *Let $A_1(x,z) = a_1 d_1 - b_1 c_1$ and $I_1(x,z) = a_1 + d_1$. If $A_1(x,z) > 0$ and $I_1(x,z) > 0$ for all $(x,z) \in [0, 2\pi/a_x] \times [0,1]$, then $O \equiv (\theta \equiv \Gamma \equiv 0)$ is linearly unstable with respect to the L^2-norm.*

Starting from (20), we obtain

$$A_1(x,z) = \frac{\alpha\xi}{\varepsilon Le}\left[-(1 + \beta_C Le N^2 e^{-z}\sin x)R + Le(1 + \beta_C Le e^{-z}\sin x)\mathscr{C} + \frac{\xi}{\alpha}\right], \tag{21}$$

$$I_1(x,z) = \alpha\left[(1 - \beta_C Le N^2 e^{-z}\sin x)R - \frac{1 + \beta_C Le e^{-z}\sin x}{\varepsilon}\mathscr{C} - \frac{(\varepsilon\, Le + 1)\xi}{\varepsilon\, Le\alpha}\right]. \tag{22}$$

On defining

$$\begin{cases} A_1^* = -(1+\beta_C\, Le\, N^2)R + Le(1-\beta_C\, Le)\mathscr{C} + \dfrac{\xi}{\alpha}, \\ \\ I_1^* = (1-\beta_C\, Le\, N^2)R - \dfrac{1+\beta_C\, Le}{\varepsilon}\,\mathscr{C} - \dfrac{(\varepsilon\, Le+1)\xi}{\varepsilon\, Le\,\alpha}, \end{cases} \tag{23}$$

and taking (21) and (22) into account, one immediately obtains

$$\frac{\varepsilon\, Le}{\alpha\,\xi}\, A_1(x,z) \geq A_1^*\,, \qquad \frac{1}{\alpha}\, I_1(x,z) \geq I_1^* \quad \forall (x,z) \in [0, 2\pi/a_x] \times [0,1] \tag{24}$$

and hence

$$\begin{cases} A_1^* > 0 \\ \\ I_1^* > 0 \end{cases} \Longrightarrow \begin{cases} A_1(x,z) > 0 \\ \\ I_1(x,z) > 0\,. \end{cases} \quad \forall\,(x,z) \in [0, 2\pi/a_x] \times [0,1] \tag{25}$$

We observe that the system of inequalities (for at least one $a \in \boldsymbol{R}^+$)

$$\begin{cases} A_1^* > 0, \\ \\ I_1^* > 0\,, \end{cases} \tag{26}$$

in view of (23), is equivalent to the system

$$\begin{cases} -(1+\beta_C Le\, N^2)R + Le(1-\beta_C Le)\mathscr{C} + \dfrac{\xi}{\alpha} > 0, \\ \\ (1-\beta_C Le\, N^2)R - \dfrac{1+\beta_C Le}{\varepsilon}\,\mathscr{C} - \dfrac{(\varepsilon\, Le+1)\xi}{\varepsilon\, Le\,\alpha} > 0\,. \end{cases} \tag{27}$$

Setting

$$\begin{cases} g(\beta_C) = Le[Le^2N^2(\varepsilon\, Le-1)\beta_C^2 - Le(\varepsilon\, Le+1)(N^2+1)\beta_C + \varepsilon\, Le - 1], \\ \\ R_B = 4\pi^2, \\ \\ f(\beta_C) = Le\, N^2(2\varepsilon\, Le+1)\,\beta_C + 1 (>0)\,, \end{cases} \tag{28}$$

we see that $N^2 < 1$ implies $\beta_C \in \left(0,\ \dfrac{1}{Le}\right)$ and it follows immediately that

$$\varepsilon\, Le - 1 > 0 \implies \exists \beta_C^* < \frac{1}{Le} : \quad g(\beta_C) > 0 \quad \forall \beta_C \in (0, \beta_C^*) \tag{29}$$

and that (26) holds only if

$$\begin{cases} \varepsilon\, Le > 1, \\ \\ \mathscr{C} > \mathscr{C}^* = \dfrac{f(\beta_C)}{g(\beta_C)}\, R_B \quad \forall\, \beta_C \in (0, \beta_C^*)\,. \end{cases} \tag{30}$$

Theorem 2. *Assume that (30) holds. Then a critical number for the linear instability of the steady state (5-7) is given by*

$$R_C = \frac{1+\beta_C Le}{\varepsilon(1-\beta_C Le\, N^2)}\,\mathscr{C} + \frac{\varepsilon\, Le+1}{\varepsilon\, Le(1-\beta_C Le\, N^2)}\,R_B\,. \tag{31}$$

Proof. It is enough to prove that, for any $k \in \left(0,\ \dfrac{g(\beta_C)(\mathscr{C}-\mathscr{C}^*)}{(\varepsilon\, Le+1)(1+\beta_C Le\, N^2)}\right]$,

$$R = \frac{1+\beta_C Le}{\varepsilon(1-\beta_C Le\, N^2)}\,\mathscr{C} + \frac{\varepsilon\, Le+1}{\varepsilon\, Le(1-\beta_C Le\, N^2)}\,(R_B+k) \tag{32}$$

implies linear instability. To this end, setting

$$F(a^2) = \frac{\xi}{\alpha}, \tag{33}$$

we show that there exists a suitable $a^2 > 0$ such that (27) holds, i.e.,

$$\begin{cases} R - \dfrac{Le(1-\beta_C Le)}{1+\beta_C Le\, N^2}\mathscr{C} - \dfrac{1}{1+\beta_C Le\, N^2}F(a^2) < 0, \\[2ex] R - \dfrac{1+\beta_C Le}{\varepsilon(1-\beta_C Le\, N^2)}\mathscr{C} - \dfrac{\varepsilon\, Le+1}{\varepsilon\, Le(1-\beta_C Le\, N^2)}F(a^2) > 0\,. \end{cases} \tag{34}$$

By (28-30), (34) becomes

$$R_B - k_1 < F(a^2) < R_B + k \tag{35}$$

with

$$k_1 = \frac{g(\beta_C)(\mathscr{C}-\mathscr{C}^*) - (\varepsilon\, Le+1)(1+\beta_C Le\, N^2)k}{\varepsilon\, Le(1-\beta_C Le\, N^2)}\ (>0)\,. \tag{36}$$

Since

$$\begin{cases} F(a^2) \in C(\boldsymbol{R}^+), \\ F(a^2) \geq F(\bar{a}^2) = R_B, \\ \bar{a}^2 = \pi^2 \end{cases} \tag{37}$$

in the interval $]\bar{a}^2, a_*^2[$, with

$$F(a_*^2) = R_B + k\,, \tag{38}$$

there exists suitable a $a^2 > 0$ such that (35) holds. Then on taking (25) into account and applying Theorem 1, one obtains the linear instability of the critical point O.

Acknowledgements

This work was carried out under the auspices of the GNFM of INDAM and MIUR (PRIN): “Nonlinear mathematical problems of wave propagation and stability in models of continuous media".

References

[1] Flavin, J.N., Rionero, S. (1996): Qualitative estimates for partial differential equations. CRC Press, Boca Raton, FL
[2] Nield, D.A., Bejan, A. (1992): Convection in porous media. Springer, New York
[3] Joseph, D.D. (1976): Stability of fluid motions. I, II. Springer, New York
[4] Straughan, B. (2004): The energy method, stability, and nonlinear convection. 2nd ed. (Applied Mathematical Sciences, vol. 91). Springer, New York
[5] Capone, F., Rionero, S. (2004): On the onset of convection for a double difusive mixture in a porous medium under periodic in space boundary data. Proc. WASCOM 2003, to appear
[6] Kaloni, P.N., Qiao, Z.-C. (2000): Nonlinear convection induced by inclined thermal and solutal gradients with mass flow. Cont. Mech. Thermodyn. **12**, 185–194
[7] Guo, J., Kaloni, P.N. (1995): Nonlinear stability of convection induced by thermal and solutal gradients. Z. Angew. Math. Phys. **46**, 645–654
[8] Lombardo, S., Mulone, G., Rionero, S. (2000): Global stability of the Bénard problem for a mixture with superimposed plane parallel shear flows. Math. Methods Appl. Sci. **23**, 1447–1465
[9] Lombardo, S., Mulone, G., Rionero, S. (2001): Global nonlinear exponential stability of the conduction-diffusion solution for Schmidt numbers greater than Prandtl numbers. J. Math. Anal. Appl. **262**, 191–207
[10] Lombardo, S., Mulone, G., Straughan, B. (2001): Non-linear stability in the Bénard problem for a double-diffusive mixture in a porous medium. Math. Methods Appl. Sci. **24**, 1229–1246
[11] Mulone, G. (1994): On the nonlinear stability of a fluid layer of a mixture heated and salted from below. Contin Mech. Thermodyn. **6**, 161–184
[12] Mulone, G., Rionero, S. (1998): Unconditional nonlinear exponential stability in the Bénard problem for a mixture: necessary and sufficient conditions. Atti Accad. Naz. Lincei Cl. Sci. Fiz. Mat. Natur. Rend. Lincei (9) Mat. Appl. **9**, 221–236
[13] Nield, D.A., Manole, D.M., Lage, J.L. (1993): Convection induced by inclined thermal and solutal gradients in a shallow horizontal layer of a porous medium. J. Fluid Mech. **257**, 559–574
[14] Rionero, S. (2004): On instability sources in dynamical systems. In: Romano, G., Rionero, S. (eds.): Recent trends in the applications of mathematics to mechanics. Springer, Berlin, pp. 141–153

Modelling of a free piston problem

B. D'Acunto, A. Monte

Abstract. We give a qualitative analysis of the free boundary value problem which models the motion of a piston. A basic role is played by a third-order operator, whose properties are used for solving preliminary problems. The differential equations are transformed into a nonlinear Volterra system, which is dealt with by the fixed point theorem.

1 Introduction

Free boundary value problems occur in many fields of mechanics and are the objects of studies documented in an extensive bibliography; see, e.g., [1,2] and their references. Also, from this point of view, the motion of the piston in gas dynamics has been newly addressed in recent years [3–5]. In this context, we discuss a free boundary value problem related to the motions of a viscous isentropic gas in a cylinder. In the model analyzed in this note we consider aspects of the physical process not examined in previous papers. In fact, we assume that the cylinder is finite and delimited by a fixed wall and by a movable piston. In addition, we account for the friction, that necessarily arises during the motion, by means of a quite general force F applied on the piston head. Furthermore, the case of a forcing term f, depending on the gas speed, is also considered.

In the framework of the one-dimensional model, with ρ denoting the density, u the velocity, p the pressure, μ the viscosity coefficient, c_p (c_v) the specific heat at constant pressure (constant volume), the following equations apply:

$$\rho u_t + \rho u u_x = (4/3)\mu u_{xx} - p_x + f(u),$$

$$\rho_t + (\rho u)_x = 0,$$

$$p = A\rho^\gamma, \quad \gamma = c_p/c_v, \quad A > 0,$$

$$\dot{s} = u(s,t),$$

$$m\ddot{s} = \alpha[p(s,t) - (4/3)\mu u_x(s,t)] + F(s,\dot{s},t),$$

where $x = s(t)$ represents the piston path, α the surface of the piston head and m its mass. Moreover, a boundary condition at the fixed end must be given in order to consider the gas entering into the cylinder. Finally, arbitrary initial conditions are prescribed.

We emphasize that, although in this paper we consider a linear form of the equations, the problem itself remains nonlinear because of the free boundary. In

Sect. 2 we discuss the model and give the basic equations of the problem. In the next section we solve intermediate initial boundary value problems. Then, the problem is transformed into a nonlinear Volterra system, which can be analyzed by the fixed point theorem. This enables us to obtain a uniqueness and existence theorem for our model of the free piston problem.

2 Model and basic equations

Consider a cylinder filled with gas that can move between a fixed wall and a movable piston. We assume that a quite general force F_1 can act on its head; in particular, the friction is included in this force. Moreover, the gas motions can also be influenced by a forcing term f which depends on the gas speed. The one-dimensional motions of an isentropic viscous gas are governed by the equations:

$$\rho u_t + \rho u u_x = -p_x + (4/3)\mu u_{xx} + f(u), \tag{1}$$

$$\rho_t + (\rho u)_x = 0, \tag{2}$$

$$p = A\rho^\gamma, \quad \gamma = c_p/c_v, \quad A > 0, \tag{3}$$

where $u(x,t)$ represents the gas velocity at location x and time t, ρ the density, p the pressure, μ the viscosity coefficient and c_p (c_v) the specific heat at constant pressure (constant volume). The piston path is described by the unknown function $x = s(t)$, which represents the free boundary.

Of course, the gas elements in contact with the piston head move with the same velocity as that of the piston:

$$\dot{s} = u(s(t), t). \tag{4}$$

Finally, Newton's law for the piston motion is interpreted as

$$m\ddot{s} = \alpha_1[p(s,t) - (4/3)\mu u_x(s,t)] + F_1(s, \dot{s}, t), \tag{5}$$

where α_1 identifies the surface of the piston head and m its mass.

The cylinder is assumed to be insulated except at the fixed wall through which the gas enters. We take account of this by means of the boundary conditions at $x = 0$. In addition, suitable initial conditions must be given.

Now, we consider the linear case of equations (1), (2). However the free boundary value problem is nonlinear. We denote a reference density by ρ_0 and we use the condensation

$$\sigma = (\rho - \rho_0)/\rho_0, \tag{6}$$

for giving a linear form to $(3)_1$

$$p = A\rho_0^\gamma + c^2\rho_0\sigma, \tag{7}$$

where c denotes the speed of sound. The last result is used to eliminate the pressure from the motion equations which, in linearized form, become

$$u_t + c^2\sigma_x - \varepsilon u_{xx} + au = 0, \;\; 0 < x < s(t), \; 0 < t \le T, \tag{8}$$

$$\sigma_t + u_x = 0, \;\; 0 < x < s(t), \; 0 < t \le T, \tag{9}$$

where $\varepsilon = 4\mu/3\rho_0$ is the kinematical coefficient of viscosity and $a\rho_0 u$ is the linear term of f.

Now, we use (7) in the piston equation (5) and we obtain

$$\ddot{s} = \alpha[c^2\sigma(s,t) - \varepsilon u_x(s,t)] + F(s,\dot{s},t), \; 0 < t \le T, \tag{10}$$

where $\alpha = \alpha_1\rho_0/m$, and $F = (F_1 + \tilde{\alpha}\rho_0^{\gamma})/m$. Obviously, (4) holds:

$$\dot{s} = u(s(t),t), \; 0 < t \le T. \tag{11}$$

The mathematical problem is completed by arbitrary initial and boundary conditions:

$$u(0,t) = g(t), \; 0 < t \le T. \tag{12}$$

$$u(x,0) = u_0(x), \;\; \sigma(x,0) = \sigma_0(x), \; 0 < x < b_1, \tag{13}$$

$$s(0) = b_1, \; \dot{s}(0) = b_2. \tag{14}$$

For the functions g, u_0, σ_0 we assume that

$$g \in C^1([0,T]), \;\; u_0, \sigma_0 \in C^2([0,b_1]). \tag{15}$$

Remark 2.1. It is easy to verify that any function u, which is a solution of (8), (9), also satisfies the third-order equation

$$\varepsilon u_{xxt} + c^2 u_{xx} = u_{tt} + au_t. \tag{16}$$

The same equation is also satisfied by the function σ.

Equation (16) is a special case of a nonlinear equation recently discussed in a wide-ranging monograph [6]. We remark that the free boundary value problem studied here cannot be deduced from the results obtained in this book.

3 Discussion of intermediate problems

In this section we analyze preliminary initial-boundary value problems related to the system (8), (9). First we consider the Cauchy problem

$$v_t + c^2 z_x - \varepsilon v_{xx} + av = 0, \;\; x \in R, \;\; 0 < t \le T, \tag{17}$$

$$z_t + v_x = 0, \; x \in R, \; 0 < t \le T. \tag{18}$$

As initial conditions for this problem we use the initial data u_0, σ_0 defined in (13), after a smooth extension with compact support on R:

$$v(x,0) = u_0(x), \quad z(x,0) = \sigma_0(x), \quad u_0, \sigma_0 \in C^2(R). \tag{19}$$

This problem can be explicitly solved by using the fundamental solution K of Eq. (16), determined and discussed in [7]. We use the following expression of K for convenience:

$$K(x,t) = e^{-bt} \int_r^\infty \frac{e^{-z_1^2/4\varepsilon t}}{2\varepsilon\sqrt{\pi\varepsilon t}} B(z_1, r)\, dz_1, \ r = |x|, \tag{20}$$

with

$$B(z_1, r) = \int_r^{z_1} I_0\left(\beta\sqrt{u_1^2 - r^2}\right) I_0\left(2\delta\sqrt{u(z_1 - u_1)}\right)\, du_1, \tag{21}$$

where I_0 denotes the modified Bessel function of the first kind and

$$b = c^2/\varepsilon, \ k = \sqrt{1 - a/b}, \ \ \delta = (c/\varepsilon)\sqrt{k}, \ \beta = c(1-k)/\varepsilon. \tag{22}$$

From (20) it is apparent that K is never negative. Moreover, this function has other basic properties, proved in [7], which are similar to those of the fundamental solution of the heat operator. We use these properties later.

Now, we can write the solution of (17)-(19) as:

$$v(x,t) = \int_R u_0(\xi) K_t(x-\xi, t) d\xi - \int_R c^2 \sigma_0'(\xi) K(x-\xi, t) d\xi, \tag{23}$$

$$z(x,t) = \int_R \sigma_0(\xi) K_t(x-\xi, t) d\xi - \int_R (u_0' + \varepsilon\sigma_0'' - a\sigma_0)(\xi) K(x-\xi, t) d\xi. \tag{24}$$

Indeed, it is easy to verify that these functions satisfy (17), (18). Furthermore, the initial conditions are satisfied as

$$\lim_{t\downarrow 0} \int_R u_0(\xi) K_t(x-\xi, t) d\xi = u_0(x), \ \ \lim_{t\downarrow 0} \int_R \sigma_0(\xi) K_t(x-\xi, t) d\xi = \sigma_0(x),$$

whereas the other limits vanish [7].

Then, we discuss the initial-boundary value problem:

$$u_t + c^2\sigma_x - \varepsilon u_{xx} + au = 0, \ \ 0 < x < s(t), \ 0 < t \le T, \tag{25}$$

$$\sigma_t + u_x = 0, \ \ 0 < x < s(t), 0 < t \le T, \tag{26}$$

$$u(0,t) = g(t), \ 0 < t \le T. \tag{27}$$

$$u(s(t), t) = g_1(t), \ 0 < t \le T. \tag{28}$$

$$u(x,0) = u_0(x), \ \ \sigma(x,0) = \sigma_0(x), \ 0 < x < b_1, \tag{29}$$

where $s(t)$ is a given function. We examine this problem under hypotheses (15) and the following hypotheses:

$$g_1 \in C^1([0,T]),\ s \in C^1([0,T]), \tag{30}$$

$$g(0) = u_0(0),\ g_1(0) = u_0(b_1), \tag{31}$$

$$\Delta = \inf_{0\le t\le T} s(t) > 0. \tag{32}$$

In order to obtain the solution of the above problem, we extend the initial data exactly as we did for the Cauchy problem (17)-(19) and we express, as before, by v and z the functions which solve that problem. Then, we give a solution of (25)-(29) by means of two auxiliary functions $\varphi(t)$, $\psi(t)$ belonging to $C^1([0,T])$ such that $\varphi(0) = \psi(0) = 0$. First we provide the solution and afterwards we show how to find these functions. We apply a method already employed to analyze other free boundary value problems governed by third-order operators [5,8]. Thus, we put

$$K_1(x,t) = (\varepsilon\partial_t + c^2)K(x,t), \tag{33}$$

and we state that a solution of (25)-(29) is given by

$$u(x,t) = v(x,t) + 2\int_0^t \varphi(\tau)K_1(x,t-\tau)d\tau \tag{34}$$
$$+\ 2\int_0^t \psi(\tau)[\dot{s}(\tau)K_t(x-s(\tau),t-\tau) + K_1(x-s(\tau),t-\tau)]d\tau,$$
$$\sigma(x,t) = z(x,t) - 2\int_0^t \dot{\varphi}(\tau)K(x,t-\tau)d\tau \tag{35}$$
$$-\ 2\int_0^t [\dot{\psi}(\tau) + a\psi(\tau)]K(x-s(\tau),t-\tau)d\tau.$$

Indeed, it is easy to verify that these functions satisfy the system (25)-(26); moreover, since v, z are solutions of (17)-(19), the initial conditions are also satisfied. It remains to show that the boundary conditions (27), (28) are verified and the auxiliary functions can be uniquely found; these problems are solved together. First, letting $x \to 0$ and $x \to s(t)$ in (34), we get

$$\varphi(t) = v(0,t) - g(t) \tag{36}$$
$$+\ 2\int_0^t \psi(\tau)[\dot{s}(\tau)K_t(-s(\tau),t-\tau) + K_1(-s(\tau),t-\tau)]d\tau,$$
$$\psi(t) = g_1(t) - v(s(t),t) - 2\int_0^t \varphi(\tau)K_1(s(t),t-\tau)d\tau \tag{37}$$
$$-\ 2\int_0^t \psi(\tau)[\dot{s}(\tau)K_t(s(t)-s(\tau),t-\tau) + K_1(s(t)-s(\tau),t-\tau)]d\tau.$$

Then, we differentiate (34) with respect to t and let $x \to 0$ and thus obtain

$$\dot{\varphi}(t) = v_t(0,t) - \dot{g}(t) + 2\int_0^t \dot{\psi}(\tau)K_1(-s(\tau),t-\tau)d\tau \tag{38}$$

$$- 2a \int_0^t \dot{s}(\tau) K_t(-s(\tau), t-\tau) d\tau.$$

Similarly, by noting that

$$\dot{g}_1(t) = u_x(s(t), t)\dot{s}(t) + u_t(s(t), t),$$

we get

$$\dot{\psi}(t) = \dot{g}_1(t) - \dot{v}(s(t), t) - 2 \int_0^t \dot{\varphi}(\tau)[\dot{s}(t) K_t(s(t), t-\tau) + K_1(s(t), t-\tau)] d\tau \quad (39)$$
$$- 2a \int_0^t \dot{s}(t)\varphi(\tau) K_t(s(t), t-\tau) d\tau - 2a \int_0^t \dot{s}(t)\psi(\tau) K_t(s(t) - s(\tau), t-\tau) d\tau$$
$$- 2 \int_0^t \dot{\psi}(\tau)[\dot{s}(t) K_t(s(t) - s(\tau), t-\tau) + K_1(s(t) - s(\tau), t-\tau)] d\tau.$$

Now, we note that (36)-(39) provide a system of Volterra integral equations whose kernels are bounded by $C/\sqrt{t-\tau}$), where C is a constant depending on T [7]. Consequently, we can apply well-known results [9] in order to obtain the existence and uniqueness of the functions φ, $\psi \in C^1([0, T])$. This proves that a solution of problem (25)-(29) exists and that it is given by (23), (24). Moreover, this solution is also unique, as we next show.

To this end we denote by (u_1, σ_1) the solution of (25)-(29) such that

$$u_1(x, 0) = 0, \quad \sigma_1(x, 0) = 0, \quad u_1(0, t) = 0, \quad u_1(s(t), t) = 0. \quad (40)$$

Now, we put

$$u = u_1 \exp(-\beta t), \quad \sigma = \sigma_1 \exp(-\beta t), \quad (41)$$

where β is a positive constant, and we obtain

$$u_t + c^2 \sigma_x - \varepsilon u_{xx} + (\beta + a) u = 0, \quad 0 < x < s(t), \ 0 < t \leq T, \quad (42)$$

$$\sigma_t + \beta\sigma + u_x = 0, \quad 0 < x < s(t), \ 0 < t \leq T. \quad (43)$$

Next, we consider the energy functional

$$E(t) = \frac{1}{2} \int_0^s [u^2(x, t) + c^2 \sigma^2(x, t)] dx, \quad (44)$$

which we differentiate and evaluate along the solutions of (42)-(43):

$$\dot{E}(t) = \frac{c^2}{2}\dot{s}(t)\sigma(s(t),t) \tag{45}$$

$$-\int_0^s [(\beta+a)u^2(x,t) + c^2\beta\sigma^2(x,t) + \varepsilon u_x^2(x,t)]dx.$$

Finally, from (45) we get

$$E(t) = \frac{c^2}{2}\int_0^t \dot{s}(\tau)\sigma(s(\tau),\tau)d\tau \tag{46}$$

$$-\int_0^t d\tau \int_0^s [(\beta+a)u^2(x,\tau) + c^2\beta\sigma^2(x,\tau) + \varepsilon u_x^2(x,\tau)]dx.$$

Since β can be chosen sufficiently enough, we obtain $E(t) \leq 0$ and this proves the uniqueness.

4 Existence and uniqueness of the free boundary

First, we give the hypotheses under which the free boundary value problem (8)-(14) is discussed. We assume that the known force acting on the piston is a continuous bounded function which satisfies a Lipschitz condition

$$|F(\dot{s},s,t)| \leq C_D, \quad |F(\dot{s}_1,s_1,t) - F(\dot{s}_2,s_2,t)| \leq \quad L\{|\dot{s}_1-\dot{s}_2| + |s_1-s_2|\}. \tag{47}$$

In addition, since we use the results of the last section, all the assumptions introduced there are assumed to hold. Specifically,

$$g \in C^1([0,T]), \quad u_0, \sigma_0 \in C^2([0,b_1]), \tag{48}$$

$$g(0) = u_0(0), \ \dot{s}(0) = u_0(b_1). \tag{49}$$

Consider, now, solutions (34) and (35) depending on the unknown functions $s(t)$, $\varphi(t)$, $\psi(t)$. From the results of Sect. 3 it easily follows that u, σ satisfy the equations (8), (9) as well as the initial conditions (13). Therefore, we have to show that the cited unknown functions can be determined by the system (36)-(39), where the function g_1 is replaced by $\dot{s}$, and by the piston equation (10). However, this last equation must be suitably transformed. First we note that from (34), (35) it follows that

$$c^2\sigma(s(t),t) - \varepsilon u_x(s(t),t) = c^2 z(s(t),t) - \varepsilon v_x(s(t),t) \tag{50}$$

$$-2\int_0^t [a\varphi(\tau) + \dot{\varphi}(\tau)]K_1(s(t),t-\tau)d\tau$$

$$-2\int_0^t [a\psi(\tau) + \dot{\psi}(\tau)]K_1(s(t)-s(\tau),t-\tau)d\tau.$$

In addition, from (23), (24), we obtain immediately that

$$c^2 z(s(t),t) - \varepsilon v_x(s(t),t) = \int_R c^2[a\sigma_0''(\xi) + \sigma_0(\xi)]K(s(t)-\xi,t)d\xi \qquad (51)$$
$$- \int_R u_0'(\xi)(\varepsilon\partial_t + c^2)K(s(t)-\xi,t)d\xi.$$

Then we substitute (50) and (51) into the piston equation (10) and obtain

$$y = F(s,\dot{s},t) + \alpha\int_R [c^2 a\sigma_0''(\xi) + c^2\sigma_0(\xi) - u_0'(\xi)(\varepsilon\partial_t + c^2)]K(s(t)-\xi,t)d\xi \qquad (52)$$
$$- 2\alpha\int_0^t \{[a\varphi(\tau) + \dot{\varphi}(\tau)]K_1(s(t),t-\tau) + [a\psi(\tau) + \dot{\psi}(\tau)]K_1(s(t)-s(\tau),t-\tau)\}d\tau,$$

where we put

$$y = \ddot{s}. \qquad (53)$$

Consequently, we also have

$$\dot{s}(t) = b_2 + \int_0^t y(\tau)d\tau, \quad s(t) = b_1 + b_2 t + \int_0^t (t-\tau)y(\tau)d\tau. \qquad (54)$$

We note that Eq. (52), together with (36)-(39) (when the function g_1 is replaced by $\dot{s}$), gives rise to a nonlinear Volterra system. The existence and uniqueness of a continuous solution of this system can be proved by the fixed point theorem. Indeed, since the function K has properties analogous to those of the heat fundamental solution [7], we can apply the well-known method used for the classical Stefan problem [10,11].

We outline that the result proves only existence and uniqueness for small times. In fact, when the free boundary value problem is discussed, condition (32) cannot be assumed as hypothesis, and, of course, it is satisfied, in general, only for small T. However, conditions under which a solution exists for large intervals can be provided, [5].

We summarize the results in the following theorem.

Theorem 1. *Under the hypotheses (47)-(49) the free boundary value problem (8)-(14) admits a unique smooth solution.*

References

[1] Friedman, A., Reitich, F. (2001): On the existence of spatially patterned dormant malignancies in a model for the growth of non-necrotic vascular tumors. Math. Models Methods Appl. Sci. **11**, 601–625

[2] De Angelis, E., Preziosi, L. (2000): Advection-diffusion problems for solid tumour in vivo and related free boundary problem. Math. Models Meth. Appl. Sci. **10**, 379–408

[3] Takeno, S. (1995): Free piston problem for isentropic gas dynamics. Japan J. Indust. Appl. Math. **12**, 163–194

[4] Yanagi, S. (1996): Existence of uniform bounded solution to the piston problem for one-dimensional equations of compressible viscous gas. Adv. Math. Sci. Appl. **6**, 509–521

[5] D'Acunto, B., Rionero, S. (1999): A note on the existence and uniqueness of solutions to a free piston problem. Rend. Accad. Sci. Fis. Mat. Napoli (4) **66**, 75–84
[6] Maslov, V.P., Mosolov, P.P. (2000): Nonlinear wave equations perturbed by viscous terms. Walter de Gruyter, Berlin
[7] D'Acunto, B., De Angelis, M., Renno, P. (1997): Fundamental solution of a dissipative operator. Rend. Accad. Sci. Fis. Mat. Napoli (4) **64**, 295–314
[8] D'Acunto, B., De Angelis, M. (2002): A phase-change problem for an extended heat conduction model. Math. Comput. Modelling **35**, 709–717
[9] Cannon, J.R. (1984): The one-dimensional heat equation. Addison-Wesley, Reading, MA
[10] Friedman, A. (1959): Free boundary problems for parabolic equations. I. Melting of solids. J. Math. Mech. **8**, 499–517
[11] Rubinstein, L.I. (1971): The Stefan problem. (Translations of Mathematical Monographs, vol. 27). American Mathematical Society, Providence, RI

Reflections on frequently used viscoplastic constitutive models

F. De Angelis

Abstract. The constitutive problems of plasticity and viscoplasticity are considered in detail via an internal variable formulation. The treatment is set within the framework of the generalized standard material model and exploits the appropriate mathematical tools of convex analysis and subdifferential calculus. Furthermore two frequently used viscoplastic constitutive models are analyzed, the Perzyna viscoplastic model and the Duvaut-Lions viscoplastic model. In the existing literature these two models are frequently used as alternatives. In the sequel interesting relations between them are outlined and it is shown that, under particular hypotheses, the Duvaut-Lions model may be regarded as derived from the Perzyna model.

1 Introduction

In non-smooth plasticity and viscoplasticity the appropriate mathematical framework is determined by the tools and concepts of convex analysis and subdifferential calculus (Rockafellar [1], Hiriart-Urruty and Lemaréchal [2]), which are capable of dealing with convex non-differentiable functions and multivalued operators (see, e.g., Halphen and Nguyen [3], Moreau [4], Eve et al. [5], Romano et al. [6]). Within this framework the evolutive laws are here derived in a generalized form which naturally encompasses the flow law and the evolutive laws of the internal variables for non-smooth plasticity and viscoplasticity. The elasto/viscoplastic model is thus set in a unitary context [7].

According to this approach the evolutive equations in viscoplasticity are interpreted as optimality conditions of a convex optimization problem which naturally provides the elasto/viscoplastic model with a complete variational formulation (De Angelis [8]).

In the literature, in order to describe the viscoplastic behaviour of materials the Perzyna model [9] and the Duvaut-Lions model [10] are frequently used. In particular the Duvaut-Lions model is used as an alternative for the Perzyna model; see, e.g., Simo et al. [11] and Ju [12]. Useful clarifications regarding these two models in non-smooth viscoplasticity were also provided by Ristinmaa and Ottosen [13,14]. In the present paper, by means of a suitable application of the rules of convex analysis, it is shown that the flow law of the Perzyna model reduces to the flow law of the Duvaut-Lions model when, for the function present in the Perzyna model, a particular function of the excess stress is chosen, which represents the difference between the actual stress and its projection onto the elastic domain. This result was originally presented in De Angelis [7]. For this exposition it is necessary to use properties and definitions typical of convex analysis and ideas about projections on convex sets.

2 The continuum model

We consider a body $\mathcal{B}$ whose reference configuration is $\Omega \subset \Re^n$, with $1 \leq n \leq 3$. The time interval of interest is defined by $\mathcal{T} \subset \Re_+$. Let $\boldsymbol{V}$ be the space of displacements, $\boldsymbol{D}$ the strain space and $\boldsymbol{S}$ the dual stress space. We denote by $\mathbf{u} : \Omega \times \mathcal{T} \to \boldsymbol{V}$ the displacement of a particle at a point $\mathbf{x} \in \Omega$ and by $\boldsymbol{\sigma} : \Omega \times \mathcal{T} \to \boldsymbol{S}$ the stress tensor. The compatible strain tensor is expressed by $\boldsymbol{\varepsilon} = \nabla^s \mathbf{u} : \Omega \times \mathcal{T} \to \boldsymbol{D}$, where ∇^s is the symmetric part of the gradient.

A pair of conjugate convex potentials representing the elastic energy $\mathcal{W} : \boldsymbol{D} \to \Re$ and the complementary elastic energy $\mathcal{W}^* : \boldsymbol{S} \to \Re$ are introduced. For linear elasticity they are expressed in the form

$$\mathcal{W}(\boldsymbol{\varepsilon}^e) = \frac{1}{2}\langle \mathbf{E}\boldsymbol{\varepsilon}^e, \boldsymbol{\varepsilon}^e \rangle, \qquad \mathcal{W}^*(\boldsymbol{\sigma}) = \frac{1}{2}\langle \boldsymbol{\sigma}, \mathbf{E}^{-1}\boldsymbol{\sigma} \rangle, \tag{1}$$

where $\boldsymbol{\varepsilon}^e \in \boldsymbol{D}$ is the elastic strain, the symbol $\langle \cdot, \cdot \rangle$ is used to denote a non-degenerate bilinear form acting on dual spaces and $\mathbf{E}$ is the linear elastic stiffness.

The relations

$$\boldsymbol{\sigma} = \mathrm{d}\,\mathcal{W}(\boldsymbol{\varepsilon}^e) \iff \boldsymbol{\varepsilon}^e = \mathrm{d}\,\mathcal{W}^*(\boldsymbol{\sigma}) \tag{2}$$

may also be expressed in the equivalent Legendre form

$$\mathcal{W}(\boldsymbol{\varepsilon}^e) + \mathcal{W}^*(\boldsymbol{\sigma}) = \langle \boldsymbol{\sigma}, \boldsymbol{\varepsilon}^e \rangle \tag{3}$$

and hold true for pairs $\{\boldsymbol{\sigma}, \boldsymbol{\varepsilon}^e\}$ satysfying the elastic constitutive relation.

In the sequel we assume a linearized theory and a quasi-static formulation with an additive decomposition of the total strain into an elastic and an inelastic part. Following Naghdi and Murch [15], we consider the class of rate-sensitive materials and assume that viscous effects are exhibited beyond the elastic range (see also Skrzypek and Hetnarski [16] and Lemaitre and Chaboche [17]). Accordingly the inelastic strain is specified as a plastic strain $\boldsymbol{\varepsilon}^p$ for the rate independent material behavior and a viscoplastic strain $\boldsymbol{\varepsilon}^{vp}$ for the rate dependent material behavior, where combined viscous and plastic effects are represented.

The kinematic internal variable $\boldsymbol{\alpha} \in \Re^{n+1}$ and the corresponding static internal variable $\boldsymbol{\chi} \in \Re^{n+1}$ are defined as

$$\boldsymbol{\alpha} = \left\{ \begin{array}{c} \alpha_{iso} \\ \\ \boldsymbol{\alpha}_{kin} \end{array} \right\}, \qquad \boldsymbol{\chi} = \left\{ \begin{array}{c} \chi_{iso} \\ \\ \boldsymbol{\chi}_{kin} \end{array} \right\}, \tag{4}$$

where $\alpha_{iso} \in \Re$ and $\chi_{iso} \in \Re$ are introduced to model isotropic hardening and $\boldsymbol{\alpha}_{kin} \in \Re^n$ and $\boldsymbol{\chi}_{kin} \in \Re^n$ are introduced to model kinematic hardening.

A hardening potential $\mathcal{H}(\boldsymbol{\alpha})$ and its conjugate $\mathcal{H}^*(\boldsymbol{\chi})$, the complementary hardening potential, are also introduced. The relations

$$\boldsymbol{\chi} = \mathrm{d}\,\mathcal{H}(\boldsymbol{\alpha}) \iff \boldsymbol{\alpha} = \mathrm{d}\,\mathcal{H}^*(\boldsymbol{\chi}) \tag{5}$$

are equivalent to the relation in Legendre form

$$\mathscr{H}(\boldsymbol{\alpha}) + \mathscr{H}^*(\boldsymbol{\chi}) = \langle \boldsymbol{\chi}, \boldsymbol{\alpha} \rangle \tag{6}$$

and hold true for conjugate pairs $\{\boldsymbol{\chi}, \boldsymbol{\alpha}\}$. The hardening potential and the complementary hardening potential are assumed to be in decoupled form which takes isotropic and kinematic hardening into account separately:

$$\begin{aligned} \mathscr{H}(\boldsymbol{\alpha}) &= \mathscr{H}_{iso}(\alpha_{iso}) + \mathscr{H}_{kin}(\boldsymbol{\alpha}_{kin}), \\ \mathscr{H}^*(\boldsymbol{\chi}) &= \mathscr{H}^*_{iso}(\chi_{iso}) + \mathscr{H}^*_{kin}(\boldsymbol{\chi}_{kin}). \end{aligned} \tag{7}$$

The Helmholtz free energy in decoupled form is expressed as

$$\Psi(\boldsymbol{\varepsilon}^e, \boldsymbol{\alpha}) = \mathscr{W}(\boldsymbol{\varepsilon}^e) + \mathscr{H}(\boldsymbol{\alpha}). \tag{8}$$

The potential $\Psi^*(\boldsymbol{\sigma}, \boldsymbol{\chi})$, the conjugate of $\Psi(\boldsymbol{\varepsilon}^e, \boldsymbol{\alpha})$, represents the complementary free energy. In decoupled form it is expressed as

$$\Psi^*(\boldsymbol{\sigma}, \boldsymbol{\chi}) = \mathscr{W}^*(\boldsymbol{\sigma}) + \mathscr{H}^*(\boldsymbol{\chi}). \tag{9}$$

For linear hardening behavior, static and kinematic internal variables are linked by the relations

$$\boldsymbol{\chi} = \mathbf{H}\boldsymbol{\alpha}, \qquad \boldsymbol{\alpha} = \mathbf{H}^{-1}\boldsymbol{\chi}, \tag{10}$$

where $\mathbf{H}$ denotes the hardening matrix

$$\mathbf{H} = \begin{bmatrix} H_{iso} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{H}_{kin} \end{bmatrix}, \tag{11}$$

and H_{iso} and $\mathbf{H}_{kin}$ respectively denote the isotropic and kinematic hardening moduli. Accordingly the hardening potential and the complementary hardening potential can be expressed as

$$\begin{aligned} \mathscr{H}(\boldsymbol{\alpha}) &= \frac{1}{2} H_{iso}\alpha_{iso}^2 + \frac{1}{2}\mathbf{H}_{kin}\boldsymbol{\alpha}_{kin} \cdot \boldsymbol{\alpha}_{kin}, \\ \mathscr{H}^*(\boldsymbol{\chi}) &= \frac{1}{2} H_{iso}^{-1}\chi_{iso}^2 + \frac{1}{2}\boldsymbol{\chi}_{kin} \cdot \mathbf{H}_{kin}^{-1}\boldsymbol{\chi}_{kin}, \end{aligned} \tag{12}$$

which gives $\chi_{iso} = H_{iso}\alpha_{iso}$ and $\boldsymbol{\chi}_{kin} = \mathbf{H}_{kin}\boldsymbol{\alpha}_{kin}$.

2.1 The generalized standard material model

Following the generalized standard material model (Halphen and Nguyen [3]), we introduce generalized strains and stresses

$$\tilde{\boldsymbol{\varepsilon}} = \begin{bmatrix} \boldsymbol{\varepsilon} \\ \mathbf{o} \end{bmatrix}, \quad \tilde{\boldsymbol{\varepsilon}}^e = \begin{bmatrix} \boldsymbol{\varepsilon}^e \\ \boldsymbol{\alpha} \end{bmatrix}, \quad \tilde{\boldsymbol{\varepsilon}}^p = \begin{bmatrix} \boldsymbol{\varepsilon}^p \\ -\boldsymbol{\alpha} \end{bmatrix}, \quad \tilde{\boldsymbol{\varepsilon}}^{vp} = \begin{bmatrix} \boldsymbol{\varepsilon}^{vp} \\ -\boldsymbol{\alpha} \end{bmatrix}, \quad \tilde{\boldsymbol{\sigma}} = \begin{bmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\chi} \end{bmatrix}, \tag{13}$$

in order to take into account actual strains and stresses and also kinematic and static internal variables. Generalized variables are defined in product spaces, respectively $\tilde{\boldsymbol{D}} = \boldsymbol{D} \times \Re^{n+1}$ and $\tilde{\boldsymbol{S}} = \boldsymbol{S} \times \Re^{n+1}$, and they are often represented by the notation $\tilde{\boldsymbol{\varepsilon}} = (\boldsymbol{\varepsilon}, \mathbf{o})$ and $\tilde{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}, \boldsymbol{\chi})$.

The admissibility condition on the generalized stress is determined by a generalized convex elastic domain $\tilde{\boldsymbol{C}} \subseteq \tilde{\boldsymbol{S}}$, defined as

$$\tilde{\boldsymbol{C}} \stackrel{\text{def}}{=} \{\tilde{\boldsymbol{\sigma}} \in \tilde{\boldsymbol{S}} : \tilde{f}(\tilde{\boldsymbol{\sigma}}) \leq 0\},$$

where $\tilde{f} : \tilde{\boldsymbol{S}} \to \Re$ is a convex generalized material function. The convex sets $\tilde{\boldsymbol{C}}_{\boldsymbol{\sigma}} \subseteq \boldsymbol{S}$ and $\tilde{\boldsymbol{C}}_{\boldsymbol{\chi}} \subseteq \Re^{n+1}$, defined as

$$\tilde{\boldsymbol{C}}_{\boldsymbol{\sigma}} = \{\boldsymbol{\sigma} \in \boldsymbol{S} : (\boldsymbol{\sigma}, \boldsymbol{\chi}) \in \tilde{\boldsymbol{C}}\},$$

$$\tilde{\boldsymbol{C}}_{\boldsymbol{\chi}} = \{\boldsymbol{\chi} \in \Re^{n+1} : (\boldsymbol{\sigma}, \boldsymbol{\chi}) \in \tilde{\boldsymbol{C}}\},$$

represent sections of the generalized elastic domain respectively at the constant $\boldsymbol{\chi}$ and the constant $\boldsymbol{\sigma}$.

Consequently the duality product between generalized variables is defined as

$$\langle \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\varepsilon}} \rangle = \langle \boldsymbol{\sigma}, \boldsymbol{\varepsilon} \rangle, \qquad \langle \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\varepsilon}}^e \rangle = \langle \boldsymbol{\sigma}, \boldsymbol{\varepsilon}^e \rangle + \langle \boldsymbol{\chi}, \boldsymbol{\alpha} \rangle,$$

$$\langle \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\varepsilon}}^p \rangle = \langle \boldsymbol{\sigma}, \boldsymbol{\varepsilon}^p \rangle - \langle \boldsymbol{\chi}, \boldsymbol{\alpha} \rangle, \qquad \langle \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\varepsilon}}^{vp} \rangle = \langle \boldsymbol{\sigma}, \boldsymbol{\varepsilon}^{vp} \rangle - \langle \boldsymbol{\chi}, \boldsymbol{\alpha} \rangle,$$

and is induced by the duality products between the corresponding elements of $\boldsymbol{D}$ and $\boldsymbol{S}$ and between the corresponding elements of $\Re^{n+1}$ and $\Re^{n+1}$.

3 The constitutive model in plasticity

The maximum plastic dissipation principle (Hill [18]),

$$\mathscr{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^p) = \sup_{\tilde{\boldsymbol{\tau}} \in \tilde{\boldsymbol{C}}} \{\langle \tilde{\boldsymbol{\tau}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^p \rangle\} = \sup_{(\boldsymbol{\tau}, \mathbf{q}) \in \tilde{\boldsymbol{C}}} \{\langle \boldsymbol{\tau}, \dot{\boldsymbol{\varepsilon}}^p \rangle - \langle \mathbf{q}, \dot{\boldsymbol{\alpha}} \rangle\}, \tag{14}$$

plays a fundamental role in plasticity, since it implies normality of the flow law, normality of the evolutive law for the internal variables and the definition of the loading/unloading conditions in the Kuhn-Tucker complementarity form.

For a given generalized plastic strain rate $\dot{\tilde{\boldsymbol{\varepsilon}}}^p$, the Lagrangian of the plastic constitutive problem with hardening is defined as

$$\begin{aligned} \tilde{\mathscr{L}}^p(\tilde{\boldsymbol{\tau}}, \dot{\delta}) &\stackrel{\text{def}}{=} -\langle \tilde{\boldsymbol{\tau}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^p \rangle + \dot{\delta} \tilde{f}(\tilde{\boldsymbol{\tau}}) - \sqcup_{\Re^+}(\dot{\delta}) \\ &= -\langle \boldsymbol{\tau}, \dot{\boldsymbol{\varepsilon}}^p \rangle + \langle \mathbf{q}, \dot{\boldsymbol{\alpha}} \rangle + \dot{\delta} \tilde{f}(\boldsymbol{\tau}, \mathbf{q}) - \sqcup_{\Re^+}(\dot{\delta}), \end{aligned} \tag{15}$$

where $\sqcup_{\Re^+}(\dot{\delta})$ is the convex indicator function [1,2] of the set of non-negative real numbers $\Re^+$, namely,

$$\sqcup_{\Re^+}(\dot{\delta}) = \begin{cases} 0 & \text{if} \quad \dot{\delta} \geq 0, \\ +\infty & \text{if} \quad \dot{\delta} < 0. \end{cases} \tag{16}$$

The generic generalized stress is denoted here by $\tilde{\boldsymbol{\tau}} = (\boldsymbol{\tau}, \mathbf{q}) \in \tilde{\boldsymbol{S}}$, while the value at the solution is denoted by $\tilde{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}, \boldsymbol{\chi})$. Similarly, the generic Lagrange multiplier is denoted by $\dot{\delta}$, while $\dot{\gamma}$ is used to denote the value at the solution, whose significance is that of a plastic multiplier.

The solution of problem (14) is given by the point $(\tilde{\boldsymbol{\sigma}}, \dot{\gamma}) \in \tilde{\boldsymbol{S}} \times \Re^+$ which satisfies the Kuhn-Tucker optimality conditions [19]:

$$0 \in \left[\partial_{\tilde{\boldsymbol{\tau}}} \tilde{\mathscr{L}}^p(\tilde{\boldsymbol{\tau}}, \dot{\delta})\right]_{(\tilde{\boldsymbol{\sigma}}, \dot{\gamma})} \qquad \Leftrightarrow \qquad \dot{\tilde{\boldsymbol{\varepsilon}}}^p \in \dot{\gamma} \partial \tilde{f}(\tilde{\boldsymbol{\sigma}}), \tag{17}$$

$$0 \in \left[\partial_{\dot{\delta}} \tilde{\mathscr{L}}^p(\tilde{\boldsymbol{\tau}}, \dot{\delta})\right]_{(\tilde{\boldsymbol{\sigma}}, \dot{\gamma})} \qquad \Leftrightarrow \qquad \tilde{f}(\tilde{\boldsymbol{\sigma}}) \in \partial \sqcup_{\Re^+}(\dot{\delta}). \tag{18}$$

By making explicit the terms related to the generalized variables the solution is given by the conditions:

$$0 \in \left[\partial_{\boldsymbol{\tau}} \tilde{\mathscr{L}}^p(\boldsymbol{\tau}, \mathbf{q}, \dot{\delta})\right]_{(\boldsymbol{\sigma}, \boldsymbol{\chi}, \dot{\gamma})} \qquad \Leftrightarrow \qquad \dot{\boldsymbol{\varepsilon}}^p \in \dot{\gamma} \partial_{\boldsymbol{\sigma}} \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}), \tag{19}$$

$$0 \in \left[\partial_{\mathbf{q}} \tilde{\mathscr{L}}^p(\boldsymbol{\tau}, \mathbf{q}, \dot{\delta})\right]_{(\boldsymbol{\sigma}, \boldsymbol{\chi}, \dot{\gamma})} \qquad \Leftrightarrow \qquad -\dot{\boldsymbol{\alpha}} \in \dot{\gamma} \partial_{\boldsymbol{\chi}} \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}), \tag{20}$$

$$0 \in \left[\partial_{\dot{\delta}} \tilde{\mathscr{L}}^p(\boldsymbol{\tau}, \mathbf{q}, \dot{\delta})\right]_{(\boldsymbol{\sigma}, \boldsymbol{\chi}, \dot{\gamma})} \qquad \Leftrightarrow \qquad \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}) \in \partial \sqcup_{\Re^+}(\dot{\delta}). \tag{21}$$

Relation (17) gives the normality law of the plastic flow for the model problem with hardening. We explicitly note that the term $\partial \tilde{f}(\tilde{\boldsymbol{\sigma}})$ has the significance of a subdifferential [1,2] of the function $\tilde{f}(\tilde{\boldsymbol{\sigma}})$ at $\tilde{\boldsymbol{\sigma}}$ and therefore turns out to be a multivalued operator.

The relation (17) expresses the flow law (19) and the evolutive law (20) for the internal variables. Relation (18) or relation (21) give the loading/unloading conditions for the model problem with hardening and they may be written equivalently in the well-known complementarity form (see, e.g., [7]):

$$\tilde{f}(\tilde{\boldsymbol{\sigma}}) = \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}) \leq 0, \qquad \dot{\gamma} \geq 0, \qquad \dot{\gamma} \tilde{f}(\tilde{\boldsymbol{\sigma}}) = \dot{\gamma} \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}) = 0. \tag{22}$$

The principle of maximum plastic dissipation may be expressed in the form

$$\langle (\tilde{\boldsymbol{\tau}} - \tilde{\boldsymbol{\sigma}}), \dot{\tilde{\boldsymbol{\varepsilon}}}^p \rangle \leq 0 \qquad \forall \tilde{\boldsymbol{\tau}} \in \tilde{\boldsymbol{C}}. \tag{23}$$

By recalling the definition of normal cone to a convex set [1,2], we see that

$$\dot{\tilde{\boldsymbol{\varepsilon}}}^p \in \mathscr{N}_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}}), \tag{24}$$

which expresses the normality law for the plastic model with hardening and it ensures that the well-known property for the generalized plastic strain rate to belong to the normal cone to the generalized convex domain $\tilde{\boldsymbol{C}}$ at $\tilde{\boldsymbol{\sigma}}$, is satisfied.

The normality law (24) may be expressed in subdifferential terms by observing [7] that $\mathscr{N}_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}})$ coincides with the subdifferential of the indicator function $\sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}})$ of the convex set $\tilde{\boldsymbol{C}}$,

$$\mathscr{N}_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}}) = \partial \sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}}), \tag{25}$$

where the indicator function of the generalized convex elastic domain $\tilde{\boldsymbol{C}}$ is defined [1,2] as

$$\sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}}) = \begin{cases} 0 & \text{if} \quad \tilde{\boldsymbol{\sigma}} \in \tilde{\boldsymbol{C}}, \\ \\ +\infty & \text{if} \quad \tilde{\boldsymbol{\sigma}} \notin \tilde{\boldsymbol{C}}. \end{cases} \tag{26}$$

The normality law for the generalized standard material model may therefore be expressed in the subdifferential form

$$\dot{\tilde{\boldsymbol{\varepsilon}}}^p \in \partial \sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}}), \tag{27}$$

which expresses an equivalent form of the normality law for the plastic model with hardening. This law may be expressed in component variables as

$$\dot{\tilde{\boldsymbol{\varepsilon}}}^p \in \partial \sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}}) \iff \begin{cases} \dot{\boldsymbol{\varepsilon}}^p \in \partial \sqcup_{\tilde{\boldsymbol{C}}_{\boldsymbol{\sigma}}}(\boldsymbol{\sigma}), \\ \\ -\dot{\boldsymbol{\alpha}} \in \partial \sqcup_{\tilde{\boldsymbol{C}}_{\boldsymbol{\chi}}}(\boldsymbol{\chi}). \end{cases} \tag{28}$$

The function $\sqcup^*_{\tilde{\boldsymbol{C}}}(\dot{\tilde{\boldsymbol{\varepsilon}}}^p)$, the conjugate [1,2] of the function $\sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\tau}})$, is defined as

$$\sqcup^*_{\tilde{\boldsymbol{C}}}(\dot{\tilde{\boldsymbol{\varepsilon}}}^p) = \sup_{\tilde{\boldsymbol{\tau}} \in \tilde{\boldsymbol{S}}} \{ \langle \tilde{\boldsymbol{\tau}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^p \rangle - \sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\tau}}) \} = \sup_{\tilde{\boldsymbol{\tau}} \in \tilde{\boldsymbol{C}}} \langle \tilde{\boldsymbol{\tau}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^p \rangle \tag{29}$$

and has the mechanical significance of plastic dissipation, and therefore in the sequel is denoted by $\mathcal{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^p)$.

The pairs $(\tilde{\boldsymbol{\sigma}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^p)$ which satisfy the normality law in its subdifferential form are said to be conjugate. The normality law, in its subdifferential form (27), is equivalent to the inverse subdifferential relation

$$\tilde{\boldsymbol{\sigma}} \in \partial \mathcal{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^p). \tag{30}$$

Relations (27) and (30) may be expressed equivalently in Fenchel's form

$$\sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}}) + \mathcal{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^p) = \langle \tilde{\boldsymbol{\sigma}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^p \rangle, \tag{31}$$

that is, in components,

$$\sqcup_{\tilde{\boldsymbol{C}}}(\boldsymbol{\sigma}, \boldsymbol{\chi}) + \mathcal{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^p, -\dot{\boldsymbol{\alpha}}) = \langle \boldsymbol{\sigma}, \dot{\boldsymbol{\varepsilon}}^p \rangle - \langle \boldsymbol{\chi}, \dot{\boldsymbol{\alpha}} \rangle,$$

holding for conjugate pairs $(\tilde{\boldsymbol{\sigma}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^p)$, that is, for conjugate pairs $(\boldsymbol{\sigma}, \boldsymbol{\chi})$ and $(\dot{\boldsymbol{\varepsilon}}^p, -\dot{\boldsymbol{\alpha}})$.

4 The constitutive model in viscoplasticity

For a given generalized viscoplastic strain rate $\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} = (\dot{\boldsymbol{\varepsilon}}^{vp}, -\dot{\boldsymbol{\alpha}})$, the potential function of the evolutive viscoplastic constitutive problem with hardening is defined as (see, e.g., De Angelis [7])

$$\tilde{\mathscr{L}}_{vp}(\tilde{\boldsymbol{\tau}}) \stackrel{\text{def}}{=} -\langle \tilde{\boldsymbol{\tau}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} \rangle + \Pi^*(\tilde{\boldsymbol{\tau}}) = -\langle \boldsymbol{\tau}, \dot{\boldsymbol{\varepsilon}}^{vp} \rangle + \langle \mathbf{q}, \dot{\boldsymbol{\alpha}} \rangle + \Pi^*(\boldsymbol{\tau}, \mathbf{q}), \tag{32}$$

where $\Pi^*(\tilde{\boldsymbol{\tau}})$ is a convex viscoplastic potential. The potential $\tilde{\mathscr{L}}_{vp}(\tilde{\boldsymbol{\tau}})$ turns out to be convex in $\tilde{\boldsymbol{\tau}}$.

The solution of the viscoplastic constitutive problem is given by the value $\tilde{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}, \boldsymbol{\chi}) \in \tilde{\boldsymbol{S}}$ which satisfies the stationarity condition

$$0 \in \left[\partial_{\tilde{\boldsymbol{\tau}}} \tilde{\mathscr{L}}_{vp}(\tilde{\boldsymbol{\tau}})\right]_{(\tilde{\boldsymbol{\sigma}})} \qquad \Leftrightarrow \qquad \dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} \in \partial \Pi^*(\tilde{\boldsymbol{\sigma}}), \tag{33}$$

that is, in components,

$$\begin{cases} 0 \in \left[\partial_{\boldsymbol{\tau}} \tilde{\mathscr{L}}_{vp}(\boldsymbol{\tau}, \mathbf{q})\right]_{(\boldsymbol{\sigma}, \boldsymbol{\chi})} \\ \\ 0 \in \left[\partial_{\mathbf{q}} \tilde{\mathscr{L}}_{vp}(\boldsymbol{\tau}, \mathbf{q})\right]_{(\boldsymbol{\sigma}, \boldsymbol{\chi})} \end{cases} \Longleftrightarrow \begin{cases} \dot{\boldsymbol{\varepsilon}}^{vp} \in \partial_{\boldsymbol{\sigma}} \Pi^*(\boldsymbol{\sigma}, \boldsymbol{\chi}), \\ \\ -\dot{\boldsymbol{\alpha}} \in \partial_{\boldsymbol{\chi}} \Pi^*(\boldsymbol{\sigma}, \boldsymbol{\chi}), \end{cases} \tag{34}$$

which represent the flow law and the evolutive law for the internal variables of the viscoplastic problem with hardening expressed in subdifferential form.

The function $\Pi(\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp})$, the conjugate of the function $\Pi^*(\tilde{\boldsymbol{\tau}})$, is by definition expressed as

$$\begin{aligned} \Pi(\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp}) &= \sup_{\tilde{\boldsymbol{\tau}} \in \tilde{\boldsymbol{S}}} \{\langle \tilde{\boldsymbol{\tau}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} \rangle - \Pi^*(\tilde{\boldsymbol{\tau}})\} \\ &= \sup_{(\boldsymbol{\tau}, \mathbf{q}) \in \boldsymbol{S} \times \Re^{n+1}} \{\langle \boldsymbol{\tau}, \dot{\boldsymbol{\varepsilon}}^{vp} \rangle - \langle \mathbf{q}, \dot{\boldsymbol{\alpha}} \rangle - \Pi^*(\boldsymbol{\tau}, \mathbf{q})\}, \end{aligned} \tag{35}$$

and has the significance of a viscoplastic dissipation $\mathscr{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp})$.

The evolutive law, expressed in the subdifferential form (33), may be written equivalently as an inverse subdifferential relation

$$\tilde{\boldsymbol{\sigma}} \in \partial \mathscr{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp}). \tag{36}$$

The relations (33) and (36) may be written equivalently in Fenchel's form

$$\Pi^*(\tilde{\boldsymbol{\sigma}}) + \mathscr{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp}) = \langle \tilde{\boldsymbol{\sigma}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} \rangle, \tag{37}$$

that is, in components,

$$\Pi^*(\boldsymbol{\sigma}, \boldsymbol{\chi}) + \mathscr{D}(\dot{\boldsymbol{\varepsilon}}^{vp}, -\dot{\boldsymbol{\alpha}}) = \langle \boldsymbol{\sigma}, \dot{\boldsymbol{\varepsilon}}^{vp} \rangle - \langle \boldsymbol{\chi}, \dot{\boldsymbol{\alpha}} \rangle, \tag{38}$$

which holds for conjugate pairs $(\boldsymbol{\sigma}, \boldsymbol{\chi})$ and $(\dot{\boldsymbol{\varepsilon}}^{vp}, -\dot{\boldsymbol{\alpha}})$.

In the literature viscoplasticity is often presented (Yosida [20]) as a regularization process of plasticity (see, e.g., Simo et al. [11]). Constitutive equations of the rate-independent model result from the optimality conditions of maximum plastic dissipation. Similarly constitutive equations in viscoplasticity may be considered as optimality conditions of a properly regularized function representing maximum viscoplastic dissipation.

Consequently in viscoplasticity we now introduce the regularized potential function

$$\tilde{\mathscr{L}}_{\eta}^{vp}(\tilde{\boldsymbol{\tau}}) \stackrel{\text{def}}{=} -\langle \tilde{\boldsymbol{\tau}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} \rangle + \frac{1}{\boldsymbol{\eta}} g^{+}(\tilde{f}(\tilde{\boldsymbol{\tau}})) = -\langle \boldsymbol{\tau}, \dot{\boldsymbol{\varepsilon}}^{vp} \rangle + \langle \mathbf{q}, \dot{\boldsymbol{\alpha}} \rangle + \frac{1}{\boldsymbol{\eta}} g^{+}(\tilde{f}(\boldsymbol{\tau}, \mathbf{q})). \quad (39)$$

The regularized potential function is obtained by appending to the objective function of the plastic problem (14) a penalty function $g^{+} : \Re \to \Re^{+}$ of the constraint $\tilde{f}(\tilde{\boldsymbol{\sigma}}) \leq 0$, amplified by a penalty parameter $1/\boldsymbol{\eta}$. The penalty function $g^{+}(x)$ must be of class $\mathscr{C}^1$, defined in $\Re$, non-negative and such that $g^{+}(x) = 0$ if and only if $x \leq 0$. The parameter $\boldsymbol{\eta} \in (0, +\infty)$ in viscoplasticity represents a viscosity coefficient.

Consequently, a set of unconstrained problems is obtained,

$$\inf_{\tilde{\boldsymbol{\tau}} \in \tilde{\boldsymbol{S}}} \tilde{\mathscr{L}}_{\eta}^{vp}(\tilde{\boldsymbol{\tau}}), \quad (40)$$

as penalty regularization of the constrained plastic problem. The solution $\tilde{\boldsymbol{\sigma}}_{\eta}$ of the regularized problem tends to the solution $\tilde{\boldsymbol{\sigma}}$ of the constrained problem as $\boldsymbol{\eta} \to 0$ (Luenberger [19]).

A regularized form of the viscoplastic dissipation may therefore be expressed as

$$\begin{aligned} \mathscr{D}(\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp}) &= \sup_{\tilde{\boldsymbol{\tau}} \in \tilde{\boldsymbol{S}}} \{\langle \tilde{\boldsymbol{\tau}}, \dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} \rangle - \frac{1}{\boldsymbol{\eta}} g^{+}(\tilde{f}(\tilde{\boldsymbol{\tau}}))\} \\ &= \sup_{(\boldsymbol{\tau}, \mathbf{q}) \in \boldsymbol{S} \times \Re^{n+1}} \{\langle \boldsymbol{\tau}, \dot{\boldsymbol{\varepsilon}}^{vp} \rangle - \langle \mathbf{q}, \dot{\boldsymbol{\alpha}} \rangle - \frac{1}{\boldsymbol{\eta}} g^{+}(\tilde{f}(\tilde{\boldsymbol{\tau}}))\}. \end{aligned} \quad (41)$$

The expression (39) for the regularized Lagrangian is equivalent to the assumption that the convex viscoplastic potential $\boldsymbol{\Pi}^{*}(\tilde{\boldsymbol{\tau}})$ has the expression of a composed function $\mathscr{G}(\tilde{f}(\tilde{\boldsymbol{\tau}}))$, where $\mathscr{G}(x)$ depends upon a penalty function $g^{+}(x)$. By a suitable specialization of the function $\mathscr{G}(x)$, or, equivalently, of the penalty function, it is possible to show (see, e.g., De Angelis [7]) that this expression of the regularized viscoplastic potential is capable of reproducing different viscoplastic constitutive models such as the Odqvist law, the Norton law and the Perzyna law.

It may be shown that the viscoplastic constitutive law expressed in the form $(33)_2$ acquires a general relevance. In fact, when the function $\mathscr{G}(x)$ is assumed to be the indicator function of the non-positive real numbers [1,2],

$$\sqcup_{\Re^{-}}(x) = \begin{cases} 0 & \text{if} \quad x \leq 0, \\ +\infty & \text{if} \quad x > 0, \end{cases} \quad (42)$$

it follows [7] that

$$\boldsymbol{\Pi}^{*}(\tilde{\boldsymbol{\tau}}) = \mathscr{G}(\tilde{f}(\tilde{\boldsymbol{\tau}})) = \sqcup_{\Re^{-}}(\tilde{f}(\tilde{\boldsymbol{\sigma}})) = \sqcup_{\tilde{\boldsymbol{C}}}(\tilde{\boldsymbol{\sigma}}), \quad (43)$$

and thus the viscoplastic flow law $(33)_2$ reduces to the plastic flow law (27). Consequently the flow law $(33)_2$ can be specialized in order to represent different models of viscoplastic behavior and, at the limit, it gives the flow law of the plastic problem.

In the sequel we consider two viscoplastic constitutive models frequently used in the literature, namely, the Perzyna model and the Duvaut-Lions model. Some interesting relations between them are outlined and it is shown that, under particular hypotheses, the Duvaut-Lions model may be regarded as derived from the Perzyna model.

5 The Perzyna viscoplastic model

Different expressions of the viscoplastic constitutive relations are obtained by specializing the penalty function suitably [7]. For instance, by choosing the penalty function in the form

$$g^{+} \stackrel{\text{def}}{=} \begin{cases} \dfrac{1}{2}x^2 & \text{for} \quad x > 0, \\ \\ 0 & \text{for} \quad x \leq 0, \end{cases} \tag{44}$$

it follows that $dg^{+}(x)/dx =< x >$, where the McCauley brackets are defined as $< x >= (x + |x|)/2$. These positions ensure that the adopted function satisfies the conditions necessary to be considered as a penalty function (Luenberger [19]).

By imposing the stationarity condition of the regularized viscoplastic Lagrangian (39), it follows that

$$0 \in \left[\partial_{\tilde{\boldsymbol{\tau}}} \tilde{\mathcal{L}}_{\eta}^{vp}(\tilde{\boldsymbol{\tau}})\right]_{(\tilde{\boldsymbol{\sigma}})} \qquad \Leftrightarrow \qquad \dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} \in \frac{1}{\eta} < \tilde{f}(\tilde{\boldsymbol{\sigma}}) > \partial \tilde{f}(\tilde{\boldsymbol{\sigma}}), \tag{45}$$

that is, in components,

$$\begin{cases} 0 \in \left[\partial_{\boldsymbol{\tau}} \tilde{\mathcal{L}}_{\eta}^{vp}(\boldsymbol{\tau}, \mathbf{q})\right]_{(\boldsymbol{\sigma}, \boldsymbol{\chi})} \\ \\ 0 \in \left[\partial_{\mathbf{q}} \tilde{\mathcal{L}}_{\eta}^{vp}(\boldsymbol{\tau}, \mathbf{q})\right]_{(\boldsymbol{\sigma}, \boldsymbol{\chi})} \end{cases} \Longleftrightarrow \begin{cases} \dot{\boldsymbol{\varepsilon}}^{vp} \in \dfrac{1}{\eta} < \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}) > \partial_{\boldsymbol{\sigma}} \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}), \\ \\ -\dot{\boldsymbol{\alpha}} \in \dfrac{1}{\eta} < \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}) > \partial_{\boldsymbol{\chi}} \tilde{f}(\boldsymbol{\sigma}, \boldsymbol{\chi}). \end{cases} \tag{46}$$

Equations $(46)_1$ and $(46)_2$ represent the normality law and the internal variable evolutive law for the viscoplastic constitutive model of Perzyna-type [9] with linear viscous effects. The constitutive equations are reported here in subdifferential form [7], suitable for dealing properly with the singularities characterizing non-smooth problems.

6 The Duvaut-Lions viscoplastic model

It is well-known that the viscoplastic model presents a substantial difference from the plastic model. In fact, in the plastic model, the generalized stress $\tilde{\boldsymbol{\sigma}}$ is constrained

to belong to the closure of the elastic domain $\tilde{\boldsymbol{C}}$. On the contrary, in the viscoplastic model, generalized stress states external to the elastic domain are admissible. When the viscosity parameter $\boldsymbol{\eta}$ tends to zero the behavior of the rate-dependent model tends to the behavior of the rate-independent model and the solution in terms of generalized stresses tends to the solution of the plastic problem.

The constitutive model examined in this paragraph refers to the treatment originally proposed by Duvaut and Lions [10] and subsequently exploited by various authors; among others see Simo et al. [11] and Ju [12].

The Duvaut-Lions viscoplastic constitutive model is expressed in the form

$$\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} = \begin{cases} \dfrac{1}{\eta}\mathbf{G}^{-1}(\tilde{\boldsymbol{\sigma}} - \overline{\tilde{\boldsymbol{\sigma}}}) & \text{if} \quad f(\tilde{\boldsymbol{\sigma}}) > 0, \\ \\ \mathbf{0} & \text{if} \quad f(\tilde{\boldsymbol{\sigma}}) \leq 0, \end{cases} \tag{47}$$

where $\mathbf{G}^{-1}$ is defined as

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{E}^{-1} & \mathbf{0} \\ \\ \mathbf{0} & \mathbf{H}^{-1} \end{bmatrix}, \tag{48}$$

$\tilde{\boldsymbol{\sigma}}$ is the generalized actual stress and $\overline{\tilde{\boldsymbol{\sigma}}} \in \partial\tilde{\boldsymbol{C}}$ is defined as

$$\overline{\tilde{\boldsymbol{\sigma}}} = \arg \min_{\tilde{\boldsymbol{\tau}} \in \tilde{\boldsymbol{C}}} \|\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\tau}}\|_{\mathbf{G}^{-1}}. \tag{49}$$

The term $\overline{\tilde{\boldsymbol{\sigma}}} = (\overline{\boldsymbol{\sigma}}, \overline{\boldsymbol{\chi}})$ is the closest-point-projection (in the metric induced by $\mathbf{G}^{-1}$) of the generalized actual stress $\tilde{\boldsymbol{\sigma}}$ onto the elastic domain $\tilde{\boldsymbol{C}}$. Since $\tilde{\boldsymbol{C}}$ is closed and convex, the solution of the problem (49) exists and is unique for any generalized actual stress $\tilde{\boldsymbol{\sigma}} \in \tilde{\boldsymbol{S}}$.

The projection operator

$$\mathbf{P}(\tilde{\boldsymbol{\sigma}}) = \begin{cases} \tilde{\boldsymbol{\sigma}} & \text{if} \quad \tilde{\boldsymbol{\sigma}} \in \text{int}\, \tilde{\boldsymbol{C}}, \\ \\ \overline{\tilde{\boldsymbol{\sigma}}} & \text{if} \quad \tilde{\boldsymbol{\sigma}} \in \text{ext}\, \tilde{\boldsymbol{C}}, \end{cases} \tag{50}$$

satisfies the condition $\mathbf{P} \circ \mathbf{P} = \mathbf{P}$. Equation (49) may therefore be written as

$$\overline{\tilde{\boldsymbol{\sigma}}} = \mathbf{P}(\tilde{\boldsymbol{\sigma}}). \tag{51}$$

The Duvaut and Lions viscoplastic constitutive model is therefore formulated as

$$\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} = \frac{1}{\eta}\mathbf{G}^{-1}(\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})), \tag{52}$$

which is expressed in components as

$$\begin{cases} \dot{\boldsymbol{\varepsilon}}^{vp} = \dfrac{1}{\eta}\mathbf{E}^{-1}(\boldsymbol{\sigma} - \mathbf{P}(\boldsymbol{\sigma})), \\ \\ -\dot{\boldsymbol{\alpha}} = \dfrac{1}{\eta}\mathbf{H}^{-1}(\boldsymbol{\chi} - \mathbf{P}(\boldsymbol{\chi})). \end{cases} \tag{53}$$

It may be shown [7] that the Duvaut-Lions viscoplastic model (52) can be derived from the Perzyna viscoplastic model $(45)_2$ when, for the function $\tilde{f}(\tilde{\boldsymbol{\sigma}})$, one choses the function representative of the complementary energy norm of the generalized excess stress $\tilde{\boldsymbol{\sigma}}_{ex}$, which represents the difference between the generalized actual stress $\tilde{\boldsymbol{\sigma}}$ and its projection onto the elastic domain $\mathbf{P}(\tilde{\boldsymbol{\sigma}})$. In this regard we recall that the complementary energy norm of a generalized stress $\tilde{\boldsymbol{\sigma}}$ is expressed as

$$\|\tilde{\boldsymbol{\sigma}}\|_{\mathbf{G}^{-1}} = \langle \tilde{\boldsymbol{\sigma}}, \mathbf{G}^{-1}\tilde{\boldsymbol{\sigma}} \rangle = \langle \boldsymbol{\sigma}, \mathbf{E}^{-1}\boldsymbol{\sigma} \rangle + \langle \boldsymbol{\chi}, \mathbf{H}^{-1}\boldsymbol{\chi} \rangle. \tag{54}$$

In fact, if, in Eq. $(45)_2$, we assume the function $\tilde{f}(\tilde{\boldsymbol{\sigma}})$ to be the complementary energy norm of the generalized excess stress, i.e.

$$\tilde{f}(\tilde{\boldsymbol{\sigma}}) = \|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{G}^{-1}}, \tag{55}$$

then

$$< \tilde{f}(\tilde{\boldsymbol{\sigma}}) >= \|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{G}^{-1}} \tag{56}$$

and therefore relation $(45)_2$ may be written as

$$\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} = \frac{1}{\eta}\|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{G}^{-1}}\partial_{\tilde{\boldsymbol{\sigma}}}\|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{G}^{-1}}. \tag{57}$$

We explicitly note that the function $\|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{G}^{-1}}$ is both non-linear and non-differentiable and therefore in (57) it is necessary to consider the subdifferential $\partial_{\tilde{\boldsymbol{\sigma}}}\|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{G}^{-1}}$.

We now consider the function

$$\varphi(\tilde{\boldsymbol{\sigma}}) = \frac{1}{2}\|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|^2_{\mathbf{G}^{-1}}. \tag{58}$$

For the subdifferential rule of the composed functions it follows that

$$\partial\varphi(\tilde{\boldsymbol{\sigma}}) = \|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{G}^{-1}}\partial_{\tilde{\boldsymbol{\sigma}}}\|\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{G}^{-1}}. \tag{59}$$

Relation (57) may therefore be written as

$$\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} = \frac{1}{\eta}\partial\varphi(\tilde{\boldsymbol{\sigma}}), \tag{60}$$

where $\partial\varphi(\tilde{\boldsymbol{\sigma}})$ represents the subdifferential of $\varphi(\tilde{\boldsymbol{\sigma}})$.

A result related to projection problems on convex sets (Moreau [21], Zarantonello [22], Romano and Romano [23]) ensures that the function (58) is a non-linear function, but differentiable in $\tilde{\boldsymbol{S}}$, and it is

$$\mathrm{d}\,\varphi(\tilde{\boldsymbol{\sigma}}) = \mathbf{G}^{-1}\left[\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\right]. \tag{61}$$

Equation (60) may therefore be expressed as

$$\dot{\tilde{\boldsymbol{\varepsilon}}}^{vp} = \frac{1}{\eta}\mathbf{G}^{-1}\left[\tilde{\boldsymbol{\sigma}} - \mathbf{P}(\tilde{\boldsymbol{\sigma}})\right], \tag{62}$$

which represents the viscoplastic constitutive relation (52) for the Duvaut-Lions model.

Consequently the Duvaut-Lions viscoplastic model, expressed in the form (52), may be considered as derived from the Perzyna viscoplastic model if we assume that the function $\tilde{f}(\tilde{\boldsymbol{\sigma}})$ is the complementary energy norm of the generalized excess stress.

A different demonstration of the same result is also possible; see De Angelis [24].

References

[1] Rockafellar, R.T. (1970): Convex analysis. Princeton University Press, Princeton

[2] Hiriart-Urruty, J.-B., Lemaréchal, C. (1993): Convex analysis and minimization algorithms. Vols. I, II. Springer, Berlin

[3] Halphen, B., Nguyen, Q.S. (1975): Sur les matériaux standards généralisés. J. Mécanique **14**, 39–63

[4] Moreau, J.-J. (1976): Application of convex analysis to the treatment of elastoplastic systems. In: Germain, P., Nayroles, B. (eds.): Applications of methods of functional analysis to problems in mechanics. Springer, Berlin, pp. 56–89

[5] Eve, R.A., Reddy, B.D., Rockafellar, R.T. (1990): An internal variable theory of elastoplasticity based on the maximum plastic work inequality. Quart. Appl. Math. **48**, 59–83

[6] Romano, G., Rosati, L., Marotti de Sciarra, F. (1993): Variational formulations of nonlinear and nonsmooth structural problems. Internat. J. Non-Linear Mech. **28**, 195–208

[7] De Angelis, F. (1998): Constitutive models and computational algorithms in elastoviscoplasticity. (Italian) Ph.D. Thesis. Università di Napoli Federico II., Naples

[8] De Angelis, F. (2000): An internal variable variational formulation of viscoplasticity. Comput. Methods Appl. Mech. Engrg. **190**, 35–54

[9] Perzyna, P. (1963): The constitutive equations for rate sensitive plastic materials. Quart. Appl. Math. **20**, 321–332

[10] Duvaut, G., Lions, J.-L. (1972): Les inéquations en mécanique et en physique. Dunod, Paris

[11] Simo, J.C., Kennedy, J.J., Govindjee, S. (1988): Nonsmooth multisurface plasticity and viscoplasticity. Loading/unloading conditions and numerical algorithms. Internat. J. Numer. Methods Engrg. **26**, 2161–2185

[12] Ju, J.W. (1990): Consistent tangent moduli for a class of viscoplasticity. J. Engrg. Mech. **116**, 1764–1779

[13] Ristinmaa, M., Ottosen, N.S. (1998): Viscoplasticity based on an additive split of the conjugated forces. Eur. J. Mech. A Solids **17**, 207–235

[14] Ristinmaa, M., Ottosen, N.S. (2000): Consequences of dynamic yield surface in viscoplasticity. Internat. J. Solids Structures **37**, 4601–4622

[15] Naghdi, P.M., Murch, S.A. (1963): On the mechanical behaviour of viscoelastic/plastic solids. Trans. ASME Ser. E J. Appl. Mech. **30**, 321–328

[16] Skrzypek, J.J., Hetnarski, R.B. (1993): Plasticity and creep. CRC Press, Boca Raton, FL

[17] Lemaitre, J., Chaboche, J.L. (1990): Mechanics of solid materials. Cambridge University Press, Cambridge

[18] Hill, R. (1950): The mathematical theory of plasticity. Clarendon Press, Oxford

[19] Luenberger, D.G. (1973): Introduction to linear and nonlinear programming. Addison-Wesley, Reading, MA

[20] Yosida, K. (1980): Functional Analysis. 6th edition. Springer, Berlin
[21] Moreau, J.-J. (1965): Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France **93**, 273–299
[22] Zarantonello, E.H. (1971): Projections on convex sets in Hilbert space and spectral theory. I.,II. In: Zarantonello, E.H. (ed.): Contributions to nonlinear functional analysis. Academic Press, New York, pp. 237–424
[23] Romano, G., Romano, M. (1985): Elastostatics of structures with unilateral conditions on stress and displacement fields. In: Unilateral problems in structural analysis. (CISM Courses and Lectures, no. **288**). Springer, Vienna, pp. 315–338
[24] De Angelis, F. (2003): Relation between the Duvaut-Lions model and the Perzyna model in viscoplasticity. (Italian). In: 16th AIMETA Congress of Theoretical and Applied Mechanics. Ferrara, Sept. 9-12, 2003.

On hereditary models of polymers

M. De Angelis

Abstract. An equivalence between an integro-differential operator $\mathcal{M}$ and an evolution operator $\mathcal{L}_n$ is determined. From this equivalence the fundamental solution of $\mathcal{L}_n$ is estimated in terms of the fundamental solution related to the third-order operator $\mathcal{L}_1$ whose behavior is now available. Moreover, properties typical of wave hierarchies can be applied to polymeric materials. As an example the case $n = 2$ is considered and results are applied to the Rouse model and the reptation model which describe different aspects of polymer chains.

1 Statement of the problem

The creep and relaxation processes related to the viscoelastic behavior of many polymeric materials are specified by means of memory functions of the form:

$$g_n(t) = \sum_{h=1}^{n} B_h \, e^{-\beta_h t}, \tag{1}$$

where n, B_h and β_h depend on the polymer physics and are determined so as to fit the experimental curves for $g_n(t)$ to a given approximation [1–4].

Let $\mathcal{B}$ be a linear, isotropic, homogeneous system and let $\underline{u}(x,t)\underline{i}$ be the displacement field from an underformed homogeneous reference configuration $\mathcal{B}_0$. If ρ_0 denotes the mass density in $\mathcal{B}_0$, and $\underline{\mathbf{f}} = f\underline{\mathbf{i}}$ is the known body force, the one-dimensional linear motions of $\mathcal{B}$ are described by the higher order equation [5]

$$\mathcal{L}_n u = \sum_{k=0}^{n} a_k \, \partial_t^{(k)} \, (u_{tt} - c_k^2 \, u_{xx}) = F, \tag{2}$$

where

$$c_k = \alpha_k/\rho_0 a_k, \quad F = (1/\rho_0) \, \sum_{k=0}^{n} a_k \, \partial_t^k f. \tag{3}$$

In (2) the constants c_k are the characterized speeds depending on the material properties of the medium and in many physical problems $c_0^2 < c_1^2 .. < c_{n-1}^2 < c_n^2$ and so the equation is typical of *wave hierarchies* [6].

When $n = 1$, (2) turns into a strictly hyperbolic third-order equation which models the evolution of the standard linear solid [7] and its behavior was discussed in [8]: the fundamental solution $\mathcal{E}_1$ was explicitly determined, together with maximum theorems and boundary layer estimates.

Moreover, the behavior of most viscoelastic media is also fairly well modelled by linear hereditary equations of the form

$$\varepsilon(t) = J(0)\sigma(t) + \int_{-\infty}^{t} \dot{J}(t-\tau)\sigma(\tau)d\tau \tag{4}$$

where $J(t)$ denotes the creep-compliance and σ, ε are the only non-vanishing components of the stress and the strain tensors such that $\rho_0\, u_{tt} = \sigma_x + f, \quad \varepsilon = u_x,$

According to *fading memory* hypotheses [9,10], $\dot{J}(t)$ is a positive fast decreasing function and, for many real materials such as polymers, rubbers and bitumens, which can be representated by means of chains of S.L.S. elements in series or parallel [1,2], one has

$$\dot{J}_n(t) = J_n(0)g_n(t), \tag{5}$$

where n is the number of elements in the chain, $J_n(0)$ denotes the elastic compliances and constants B_k and frequencies β_k satisfy

$$0 < \beta_1 < \beta_2 < \ldots < \beta_n \quad and \quad B_k > 0 \quad \forall k = 1, 2 \ldots n. \tag{6}$$

The well-known creep representation of one-dimensional linear motions of $\mathcal{B}$ is given by [11] as

$$\mathcal{M}u = c^2 u_{xx} - u_{tt} - \int_0^t g(t-\tau)u_{\tau\tau}d\tau = -F_*(x,t), \tag{7}$$

where

$$c^2 = [\rho_0 J_n(0)]^{-1}, \;\; F_* = c^2[J_n(0)f + \int_{-\infty}^{0} \dot{J}_n(t-\tau)\sigma_x(\tau)d\tau + \int_0^t \dot{J}(t-\tau)f(\tau)d\tau]. \tag{8}$$

For all n, the fundamental solution E_n of the operator $\mathcal{M}$ has been explicitly determined [11,12]. Moreover, let E_1 be the fundamental solution related to an appropriate S.L.S. $\mathcal{B}_1^*$ defined by

$$g_1 = b\, e^{-\beta_1 t} \quad \text{with} \;\; \mathrm{b} = \beta_1 \sum_{1}^{\mathrm{n}} \frac{\mathrm{B_k}}{\beta_{\mathrm{k}}}\,; \tag{9}$$

the following theorem shows that the fundamental solution E_n can be rigorously estimated by means of E_1.

In fact, if Γ is the open forward characteristic cone $\{(t, x) : t > 0 \; |x| < ct\}$, and $\chi_n = \prod_{k=2}^{n} (\frac{B_k}{\beta_1})^2$, then the following theorem holds.

Theorem 1.1. *If the memory function is given by* (1) and (6), *then the fundamental solution E_n of $\mathcal{M}$ is a nonnegative $C^\infty(\Gamma)$ function and it satisfies the estimate*

$$0 < E_n(\beta_1..\beta_n, B_1..B_n) < \chi_n \; E_1(\beta_1, b) \tag{10}$$

everywhere in Γ and for all n. □

In this paper a *conditional* equivalence between (2) and the integro-differential equation (7) is established.

With this equivalence and Theorem 1.1, the fundamental solution $\mathscr{E}_n$ of (2) can be determined and an estimate in terms of $\mathscr{E}_1$ is obtained for all n.

Moreover, properties typical of wave hierarchies can be applied to polymeric materials. As an example the case $n = 2$ is considered and results are applied to the *Rouse model* and the *reptation model* which describe different aspects of polymer chains.

2 Conditioned equivalence between operators $\mathscr{L}_n$ and $\mathcal{M}$

Suppose that the initial data related to (2) and (7) are null and let

$$\mu_k = a_k/a_n, \;\; \lambda_k = a_k c_k / a_n c_n \quad (k = 0, \ldots, n). \tag{1}$$

Applying the Laplace transform and the polynomial identity one has $c_n = c^2$ and

$$\begin{cases} \lambda_0 = \beta_1\beta_2\cdots\beta_n \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \lambda_{n-2} = \beta_1\beta_2 + \beta_1\beta_3 + \cdots + \beta_{n-1}\beta_n \\ \lambda_{n-1} = \beta_1 + \cdots + \beta_n \end{cases} \tag{2}$$

By (6), the λ_k are positive. Further,

$$\begin{cases} \mu_0 = \lambda_0 + B_1(\beta_2 \ldots \beta_n) + \cdots + B_n(\beta_1 \cdots \beta_{n-1}) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \mu_{n-2} = \lambda_{n-2} + B_1(\beta_2 + \cdots + \beta_n) + \cdots + B_n(\beta_1 + \cdots + \beta_{n-1}) \\ \mu_{n-1} = \lambda_{n-1} + B_1 + \cdots + B_n \end{cases} \tag{3}$$

and (6), (3) imply that $0 < \lambda_k < \mu_k \;\; (k = 0, \ldots, n-1)$. Consequently,

$$0 < c_k < c_n = c^2 \quad (k = 0, \ldots, n-1). \tag{4}$$

Lastly, by (2), (3), it follows that $\frac{\lambda_0}{\mu_0} < \frac{\lambda_1}{\mu_1} < \frac{\lambda_{n-1}}{\mu n-1}$ and so

$$0 < c_0 < c_1 < \ldots < c_n. \tag{5}$$

Thus, the following property holds.

Property 2.1. *Hypotheses of fading memory* (1), (6) *imply that the differential operator* (2) *is typical of wave hierarchies.* □

Conversely, when the differential equation (2) is given, to obtain the dual hereditary equation (7) with a memory function $g(t)$ satisfying (5), (6), appropriate restrictions on the constants a_k, c_k must be imposed.

Example 2.1. When $n = 2$, then $c^2 = c_2$, $B_0 = 1$, and β_1, β_2 are real if and only if

$$\omega^2 = (a_1c_1)^2 - 4(a_0c_0)(a_2c_2) > 0. \tag{6}$$

Then,

$$\beta_1 = \frac{1}{2a_2c_2}(a_1c_1 - \omega), \quad \beta_2 = \frac{1}{2a_2c_2}(a_1c_1 + \omega) \tag{7}$$

so that $0 < \beta_1 < \beta_2$. Further,

$$B_i = \frac{(-1)^{i-1}}{\omega}[a_0(c_2 - c_0) - a_1\beta_i(c_2 - c_1)] \quad (i = 1, 2). \tag{8}$$

Thus, $B_1 > 0$, $B_2 > 0$ if and only if

$$\beta_1 < \frac{a_o}{a_1}\ \frac{c_2 - c_0}{c_2 - c_1} < \beta_2. \tag{9}$$

Therefore, the fourth-order operator

$$a_2(u_{tt} - c_2u_{xx})_{tt} + a_1(u_{tt} - c_1u_{xx})_t + a_0(u_{tt} - c_0u_{xx}) \tag{10}$$

can be analyzed by (5)-(7) when the constants a_k, c_k satisfy (6) and (9). □

3 Polymeric materials

Polymeric materials such as rubber are very flexible and are easily formed into fibres, thin films, etc. Moreover, the liquid state composed only of polymers (polymer melt) is an important state for industrial uses where polymeric materials are processed into various plastic products such as gaskets, seals, flexible joints, vehicle tires, etc. Also, durability is a requirement imposed on polymers and polymeric composites so the development of these materials has thus become an increasingly important part of engineering studies. In fact, a large literature deals with polymer physics. As for viscoelastic theories, two models which describe different aspects of polymer chains, have met with reasonable success: the *Rouse model* and the *reptation model* [4,3].

In both cases the memory function $g(t)$ assumes the form (1) of Sect. 1.

In fact in the the Rouse model the function $g(t)$ is given by

$$g(t) = k_1 \sum_{h=1}^{n} e^{2h^2 \frac{t}{\tau_1}}, \tag{1}$$

where the relaxation time τ_1 can be calculated by means of experimental results [4].

When the viscoelastic behavior is represented by the reptation model, the stress function decreases with a relaxation time τ_d, as times increases, and one has

$$g(t) = k \sum_{h=0}^{n} \frac{1}{(h)^2} \; e^{-h^2 \frac{t}{\tau_d}}, \tag{2}$$

where h ranges over the odd integers, the constant k depends on the polymer physics and the value of the reptation time τ_d can be fixed according to elasticity experiments [2].

From the first two steps in the reptation model, $B_1 = k, \;\; B_2 = B_1/9, \;\; \beta_1 = 1/\tau_d, \;\; \beta_2 = 9\beta_1$. Consequently the operator (10) is characterized by constants:

$$\begin{cases} c_0 = c^2 \, \frac{81}{81+82k\tau_d}, \; c_1 = c^2 \, \frac{9}{9+k\tau_d}, \qquad c_2 = c^2 \\ \\ a_0 = 1 + \frac{82}{81} \, k\tau_d, \;\; a_1 = \frac{10\tau_d^2}{9}(\frac{1}{\tau_d} + \frac{k}{9}), \; a_2 = \frac{\tau_d^2}{9} \end{cases} \tag{3}$$

Analogously, in the Rouse model, as $B_1 = B_2 = k_1, \;\; \beta_1 = 2/\tau_1 \; and \; \beta_2 = 4\beta_1$, one has:

$$\begin{cases} c_0 = c^2 \frac{8}{8+5k_1\tau_1}, \; c_1 = c^2 \frac{5}{5+k_1\tau_1}, \qquad c_2 = c^2 \\ \\ a_0 = 1 + \frac{5k_1}{8}\tau_1, \; a_1 = \frac{\tau_1^2}{16}(2k_1 + \frac{10}{\tau_1}), \; a_2 = \frac{\tau_1^2}{16} \end{cases} \tag{4}$$

The wave hierarchies defined by (3) or (4) are governed by the operator $\mathscr{L}_1^*$ of the standard linear solid defined, respectively, by:

$$\begin{cases} c_0 = c^2 \, \frac{81}{81+82k\tau_d}, \; c_1 = c^2, \; a_0 = 1 + \frac{82}{81} \, k\tau_d, \; \eta = \frac{81\tau_d}{81+82k\tau_d}, \\ \\ c_0 = c^2 \frac{8}{8+5k_1\tau_1}, \quad c_1 = c^2, \; a_0 = 1 + \frac{5}{8}k_1\tau_1, \;\; \eta = \frac{4\tau_1}{8+5k_1\tau_1}. \end{cases} \tag{5}$$

Remark 3.1. As shown, the memory function $g_n(t)$ can depend on h^2. Thus, the approximation to the two first terms appears to be reasonable. However, in many articles the model is limited to a single relaxation time (see, e.g., [13]). □

References

[1] Hunter, S.C. (1976): Mechanics of continuous media. Wiley, New York
[2] Ferry, J.D. (1961): Viscoelastic properties of polymers. Wiley, New York
[3] Doi, M. (1996): Introduction to polymer physics. Clarendon Press, Oxford
[4] Doi, M., Edwards, S.F. (1986): The theory of polymer dynamics. Clarendon Press. Oxford
[5] R.M. Christensen (1971): Theory of viscolasticity. Academic Press, N.Y. and London
[6] Whitham, G.B. (1974): Linear and nonlinear waves. Wiley, New York
[7] Haupt, P. (2000): Continuum mechanics and theory of materials. Springer, Berlin
[8] Renno, P. (1984): On a wave theory for the operator $\varepsilon\partial_t(\partial_t^2 - c_1^2\Delta_n) + \partial_t^2 - c_0^2\Delta_n$. Ann. Mat. Pura e Appl. (4) **136**, 355–389

[9] Graffi, D. (1983): On the fading memory. Applicable Anal. **15**, 295–311
[10] Graffi, D. (1982): Mathematical models and waves in linear viscoelasticity. In: Mainardi, F. (ed.): Wave propagation in viscoelastic media. (vol. 52). Research Notes in Math. Pitman, Boston, pp. 1–27
[11] Renno, P. (1983): On the Cauchy problem in linear viscoelasticity. Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8) **75**, 195–204
[12] Renno, P. (1983): On some viscoelastic models. Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8) **75**, 339–348
[13] Ianniruberto, G., Marrucci, G. (2001): A simple constitutive equation for entangled polymers with chain stretch. J. Rheol. **45**, 1305–1318

Edge contact forces in continuous media

M. Degiovanni, A. Marzocchi, A. Musesti

1 Introduction and preliminaries

In this note we present results contained in [2] concerning integral properties of second-order powers. More precisely, we introduce the power expended on a subbody by a virtual velocity field, in the spirit of Germain [6,7], but in an axiomatic way similar to that exploited for first-order powers in [11], in which the power is regarded as a function of the subbody and of the velocity field.

As already shown by Dell'Isola and Seppecher [3] and Di Carlo and Tatone [4], higher order powers can be used to describe edge effects, in a way that seems to be simpler than the use of edge interactions (see Noll and Virga [12] and Forte and Vianello [5]).

Here we investigate the above subject by paying attention to the regularity of the stress (or hyper-stress) fields, as well as of the subbodies on which the stresses act. In doing this, we first obtain results for finite perimeter subbodies and fields with divergence measure in order to represent a contact power as a surface integral; secondly, since the power is of order two, a further integration by parts is formally possible, leading to subsets of codimension 2, i.e., edges. To this end, we introduce a subclass of the sets of finite perimeter, called *sets with curvature measure*, where such an integral representation can be obtained. The result is that edge effects are seen as surface integrals involving curvature and/or density which is singular with respect to the area.

Finally, we find, as in previous papers [1,10,11], that powers are uniquely determined by their properties on n-intervals.

For the proofs of all results cited below, the reader is referred to [2].

In the sequel, $\mathscr{L}^n$ denotes the n-dimensional Lebesgue outer measure and $\mathscr{H}^k$ the k-dimensional Hausdorff outer measure on $\mathbb{R}^n$. Given a Borel subset $\Omega \subseteq \mathbb{R}^n$, we denote by $\mathfrak{B}(\Omega)$ the collection of all Borel subsets of Ω.

The topological closure, interior and boundary of $E \subseteq \mathbb{R}^n$ are denoted as usual by $\operatorname{cl} E$, $\operatorname{int} E$ and $\operatorname{bd} E$, respectively. Denoting by $B_r(x)$ the open ball with radius r centered at x, we introduce the *measure-theoretic interior* of E

$$E_* = \left\{ x \in \mathbb{R}^n : \lim_{r \to 0^+} \left(r^{-n} \mathscr{L}^n(B_r(x) \setminus E) \right) = 0 \right\}$$

and the *measure-theoretic boundary* of E

$$\partial_* E = \mathbb{R}^n \setminus \left(E_* \cup (\mathbb{R}^n \setminus E)_* \right),$$

which are both Borel subsets of $\mathbb{R}^n$. We say that $E \subseteq \mathbb{R}^n$ is *normalized* if $E_* = E$.

Let Ω be a Borel subset of $\mathbb{R}^n$. We denote by $\mathfrak{M}(\Omega)$ the set of Borel measures $\mu : \mathfrak{B}(\Omega) \to [0, +\infty]$ finite on compact subsets of Ω and by $\mathscr{L}^p_{loc,+}(\Omega)$, $p \in [1, +\infty]$, the set of Borel functions $h : \Omega \to [0, +\infty]$ such that

$$\int_K h^p \, d\mathscr{L}^n < +\infty \quad (p < +\infty), \qquad \operatorname*{ess\,sup}_K h < +\infty \quad (p = +\infty)$$

for every compact subset $K \subseteq \Omega$.

Definition 1. A *full grid* G is an ordered triple

$$G = \left(x_0, (e_1, \dots, e_n), \widehat{G}\right),$$

where $x_0 \in \mathbb{R}^n$, $(e_1, \dots, e_n)$ is a positively oriented orthonormal basis of $\mathbb{R}^n$ and $\widehat{G}$ is a Borel subset of $\mathbb{R}$ with $\mathscr{L}^1(\mathbb{R} \setminus \widehat{G}) = 0$.

If G_1, G_2 are two full grids, we write $G_1 \subseteq G_2$ if $\widehat{G}_1 \subseteq \widehat{G}_2$ and they share the point x_0 and the list $(e_1, \dots, e_n)$.

Definition 2. We denote by Sym_2 the finite-dimensional linear space of all symmetric bilinear forms on $\mathbb{R}^n$ and by Sym_3 the linear space of all symmetric 3-linear forms on $\mathbb{R}^n$.

Definition 3. We denote by $\mathscr{R}$ the class of open n-intervals I such that $\mathrm{cl}\, I \subseteq \Omega$.

Definition 4. Let $G = \left(x_0, (e_1, \dots, e_n), \widehat{G}\right)$ be a full grid. A subset M of $\mathbb{R}^n$ is said to be a *G-interval* if

$$M = \left\{x \in \mathbb{R}^n : \, a_j < (x - x_0) \cdot e_j < b_j \ \forall j = 1, \dots, n\right\}$$

for some $a_1, b_1, \dots, a_n, b_n \in \widehat{G}$. We set

$$\mathscr{M}_G = \{M \subseteq \mathbb{R}^n : \, M \text{ is a } G\text{-interval with } \mathrm{cl}\, M \subseteq \Omega\} \, .$$

Definition 5. Let $\mathscr{A} \subseteq \mathscr{R}$. We say that $\mathscr{A}$ *contains almost all of* $\mathscr{R}$ if, for every $x_0 \in \mathbb{R}^n$ and every positively oriented orthonormal basis $(e_1, \dots, e_n)$ in $\mathbb{R}^n$, there exists a full grid

$$G = \left(x_0, (e_1, \dots, e_n), \widehat{G}\right)$$

such that $\mathscr{M}_G \subseteq \mathscr{A}$.

2 Second-order powers

We now give our main definition.

Definition 6. Let $\mathscr{A}$ be a subset of $\mathscr{R}$ containing almost all of $\mathscr{R}$. We say that a function $P : \mathscr{A} \times C^\infty(\Omega) \to \mathbb{R}$ is a *second-order power* if the following properties hold:

(a) for every $v \in C^\infty(\Omega)$, $P(\,\cdot\,, v)$ is countably $*$-additive, i.e.,

$$P\left(\left(\bigcup_{i\in\mathbb{N}} M_i\right)_*, v\right) = \sum_{i\in\mathbb{N}} P(M_i, v)$$

for every disjoint sequence $(M_i) \in \mathscr{A}$ such that $\left(\bigcup_{i\in\mathbb{N}} M_i\right)_* \in \mathscr{A}$;

(b) for every $M \in \mathscr{A}$, $P(M, \,\cdot\,)$ is linear;

(c) there exist $\mu_0, \mu_1, \mu_2 \in \mathfrak{M}(\Omega)$ such that, for every $M \in \mathscr{A}$, $v \in C^\infty(\Omega)$,

$$|P(M, v)| \leqslant \int_M |v(x)|\, d\mu_0(x) + \int_M |\nabla v(x)|\, d\mu_1(x) + \int_M |\nabla\nabla v(x)|\, d\mu_2(x).$$

Definition 7. We call a second-order power with $\mu_2 = 0$ a *first-order power*, and a first-order power with $\mu_1 = 0$ a *power with order 0*.

Remark 1. Let $M \in \mathscr{R}$; then it is easy to prove that for every full grid G there exists a disjoint sequence $(M_i) \subseteq \mathscr{M}_G$ such that

$$\left(\bigcup_{i\in\mathbb{N}} M_i\right)_* = M.$$

Moreover, one can replace (a) by the following weaker assumption:

(a') for every $v \in C^\infty(\Omega)$ and for every full grid G,

$$P\left(\left(\bigcup_{i\in\mathbb{N}} M_i\right)_*, v\right) = \sum_{i\in\mathbb{N}} P(M_i, v)$$

whenever $(M_i) \in \mathscr{A} \cap \mathscr{M}_G$ is a disjoint sequence such that $\left(\bigcup_{i\in\mathbb{N}} M_i\right)_* \in \mathscr{A}$.

Remark 2. One can also consider powers $P(M, \boldsymbol{v})$, where $\boldsymbol{v}$ takes values in $\mathbb{R}^N$, $N \geqslant 1$, and define the corresponding power by linearity.

Our first goal is to establish a representation formula for a second-order power. This is not a matter of routine, since $P(M, v)$ does not depend only on v and hence is not merely a linear functional on the velocity field.

Theorem 1. *Let P be a second-order power.*

Then there exist bounded Borel maps $A_0 : \Omega \to \mathbb{R}$, $A_1 : \Omega \to (\mathbb{R}^n)^*$, $A_2 : \Omega \to \mathrm{Sym}_2$ such that, for every $M \in \mathscr{A}$, $v \in C^\infty(\Omega)$,

$$P(M, v) = \int_M A_0(x)v(x)\, d\mu_0(x) + \int_M \langle A_1(x), \nabla v(x)\rangle\, d\mu_1(x) + \\ + \int_M \langle A_2(x), \nabla\nabla v(x)\rangle\, d\mu_2(x). \quad (1)$$

Moreover, each A_j is uniquely determined μ_j-a.e.

The following is a form of converse of the previous theorem.

Proposition 1. *Let $\mu_0, \mu_1, \mu_2 \in \mathfrak{M}(\Omega)$ and A_0, A_1, A_2 as above be Borel and bounded.*

Then there exists a set $\mathscr{A} \subseteq \mathscr{R}$ containing almost all of $\mathscr{R}$ such that the function $P : \mathscr{A} \times C^\infty(\Omega) \to \mathbb{R}$ defined as

$$P(M, v) = \int_M A_0(x)v(x)\, d\mu_0(x) + \int_M \langle A_1(x), \nabla v(x)\rangle\, d\mu_1(x) + \\ + \int_M \langle A_2(x), \nabla\nabla v(x)\rangle\, d\mu_2(x)$$

is a second-order power.

Now we turn to a similar representation formula on Borel subsets of Ω.

Definition 8. Let $\eta \in \mathfrak{M}(\Omega)$. We set

$$\mathscr{B}_\eta = \{M \subseteq \mathbb{R}^n : M = M_*\,, \operatorname{cl} M \subseteq \Omega,\ \eta(\partial_* M) = 0\}\,.$$

Theorem 2. *Let P be a second-order power. Let A_j, $j = 0, 1, 2$, be as in Theorem 1.*

Then there exists $\eta \in \mathfrak{M}(\Omega)$ such that the function $\tilde{P} : \mathscr{B}_\eta \times C^\infty(\Omega) \to \mathbb{R}$ defined as

$$\tilde{P}(M, v) = \int_M A_0(x)v(x)\, d\mu_0(x) + \int_M \langle A_1(x), \nabla v(x)\rangle\, d\mu_1(x) + \\ + \int_M \langle A_2(x), \nabla\nabla v(x)\rangle\, d\mu_2(x)$$

is an extension of P which satisfies (a), (b) and (c) of Definition 6 on $\mathscr{B}_\eta$.

3 Decomposition of powers

Up to this point, the definitions and assumptions made imply that the power P behaves as an integral on the subbodies, but they do not imply, for example, that the power can be represented as a surface integral, as is often the case in continuum mechanics. Our next definition makes these features precise.

Definition 9. A second-order power P is said to be *weakly balanced* if there exists $\nu \in \mathfrak{M}(\Omega)$ such that,

$$\forall M \in \mathscr{A},\ \forall v \in C_c^\infty(M)\,, \quad |P(M, v)| \leqslant \int_M |v|\, d\nu\,.$$

In particular, P is said to be a *contact power* if,

$$\forall M \in \mathscr{A},\ \forall v \in C_c^\infty(M)\,, \quad P(M, v) = 0\,,$$

namely, if it is weakly balanced with $\nu = 0$.

A power P of order 0 is said to be a *body power*.

Note that a body power is always weakly balanced, as can be seen by choosing trivially $\nu = \mu_0$.

Theorem 3. *Let P be a weakly balanced second-order power and let A_j, $j = 0, 1, 2$, be as in Theorem 1.*

Then the following facts hold:

(a) there exists a bounded Borel function $B : \Omega \to \mathbb{R}$ such that, for every $v \in C_c^\infty(\Omega)$,

$$\int_\Omega A_0(x)v(x)\,d\mu_0(x) + \int_\Omega \langle A_1(x), \nabla v(x)\rangle\,d\mu_1(x) + \\ + \int_\Omega \langle A_2(x), \nabla\nabla v(x)\rangle\,d\mu_2(x) = \int_\Omega B(x)v(x)\,d\nu(x)\,; \quad (2)$$

moreover, B is uniquely determined ν-a.e.;

(b)

$$\forall M \in \mathcal{A},\ \forall v \in C_c^\infty(M)\,, \quad P(M, v) = \int_M B(x)v(x)\,d\nu(x)\,.$$

Now let P be a weakly balanced second-order power, let μ_j, A_j, $0 \leqslant j \leqslant k$, be as in Theorem 1 and let ν, B be as in Theorem 3. According to Proposition 1 we can define, for a suitable class $\mathcal{A}$ containing almost all of $\mathcal{R}$, two powers $P_b, P_c : \mathcal{A} \times C^\infty(\Omega) \to \mathbb{R}$ by

$$P_b(M, v) := \int_M B(x)v(x)\,d\nu(x)\,,$$
$$P_c(M, v) := P(M, v) - \int_M B(x)v(x)\,d\nu(x)\,.$$

It is readily seen that P_b is a body power and P_c a second-order contact power. Of course, we have $P = P_b + P_c$.

Definition 10. P_b is said to be the *body part* of P and P_c the *contact part* of P.

4 First-order contact powers

Let P be a first-order contact power such that (c) of Definition 6 holds with μ_1 absolutely continuous with respect to the Lebesgue measure. We set $\eta = \mu_0$.

According to Theorems 1 and 3, there exist a bounded Borel function $a : \Omega \to \mathbb{R}$ and $T \in L^1_{loc}(\Omega; \mathbb{R}^n)$ such that,

$$\forall M \in \mathcal{B}_\eta, \forall v \in C^\infty(\Omega)\,, \quad P(M, v) = \int_M av\,d\eta + \int_M T \cdot \nabla v\,d\mathcal{L}^n\,, \quad (3)$$

$$\forall v \in C_c^\infty(\Omega)\,, \quad \int_\Omega av\,d\eta + \int_\Omega T \cdot \nabla v\,d\mathcal{L}^n = 0\,. \quad (4)$$

Moreover, a is uniquely determined η-a.e. and T is uniquely determined $\mathcal{L}^n$-a.e.

We now briefly recall the concept of *outer normal* to the measure-theoretic boundary of a set. Let $M \subseteq \mathbb{R}^n$ and $x \in \partial_* M$. We denote by $\boldsymbol{n}^M(x) \in \mathbb{R}^n$ a unit vector such that

$$\mathcal{L}^n\left(\{\xi \in B_r(x) \cap M : (\xi - x) \cdot \boldsymbol{n}^M(x) > 0\}\right) \Big/ r^n \to 0,$$

$$\mathcal{L}^n\left(\{\xi \in B_r(x) \setminus M : (\xi - x) \cdot \boldsymbol{n}^M(x) < 0\}\right) \Big/ r^n \to 0$$

as $r \to 0^+$. At most one such vector can exist. Setting $\boldsymbol{n}^M(x) = 0$ elsewhere, we can consider the map $\boldsymbol{n}^M : \partial_* M \to \mathbb{R}^n$, which is called the *unit outer normal* to M. It turns out that $\boldsymbol{n}^M$ is Borel and bounded.

Whenever $\mathcal{H}^{n-1}(\partial_* M) < +\infty$, we say that M is a *set with finite perimeter*. In that case it is well-known that $\boldsymbol{n}^M(x) \neq 0$ for $\mathcal{H}^{n-1}$-a.e. $x \in \partial_* M$ and the Gauss-Green theorem holds for Lipschitz functions.

Now we define a suitable subclass of $\mathcal{B}_\eta$ which allows us to give a representation formula for a first-order contact power involving only the measure-theoretic boundary of the subbodies. We refer to [14,1] for a discussion of this class.

Definition 11. For $h \in \mathcal{L}^1_{loc,+}(\Omega)$ we set

$$\mathcal{M}_{h\eta} = \left\{ M \in \mathcal{B}_\eta : \mathcal{H}^{n-1}(\partial_* M) < +\infty,\ \int_{\partial_* M} h \, d\mathcal{H}^{n-1} < +\infty \right\}.$$

We are now in a position to state the boundary representation formula for first-order contact powers. We refer to it as the Cauchy's Stress Theorem, since it states the linearity of the stress with respect to the normal.

Theorem 4 (Cauchy's Stress Theorem). *There exists $h \in \mathcal{L}^1_{loc,+}(\Omega)$ such that,*

$$\forall M \in \mathcal{M}_{h\eta}, \forall v \in C^\infty(\Omega), \quad P(M, v) = \int_{\partial_* M} v\, T \cdot \boldsymbol{n}^M \, d\mathcal{H}^{n-1}.$$

5 Second-order contact powers

Now we study second-order contact powers in more detail and the possibility of representing them as surface integrals.

Throughout this section, we assume that P is a second-order contact power such that (c) of Definition 6 holds with μ_1 and μ_2 absolutely continuous with respect to the Lebesgue measure. We set $\eta = \mu_0$.

According to Theorems 1 and 3, there exist a bounded Borel function $a : \Omega \to \mathbb{R}$, $B \in L^1_{loc}(\Omega; \mathbb{R}^n)$ and $\mathsf{C} \in L^1_{loc}(\Omega; \mathrm{Sym}_2)$ such that,

$$\forall M \in \mathcal{B}_\eta, \forall v \in C^\infty(\Omega), \quad P(M, v) = \int_M a v \, d\eta +$$

$$+ \int_M B \cdot \nabla v \, d\mathcal{L}^n + \int_M \mathsf{C} \cdot \nabla\nabla v \, d\mathcal{L}^n, \tag{5}$$

$$\forall v \in C^\infty_c(\Omega), \quad \int_\Omega a v \, d\eta + \int_\Omega B \cdot \nabla v \, d\mathcal{L}^n + \int_\Omega \mathsf{C} \cdot \nabla\nabla v \, d\mathcal{L}^n = 0. \tag{6}$$

Moreover, a is uniquely determined η-a.e. and B, C are uniquely determined $\mathscr{L}^n$-a.e. When the distribution $\operatorname{div}\mathsf{C}$ is a function, we have a representation of $P(M, v)$ as a surface integral.

Theorem 5. *Assume that* $\operatorname{div}\mathsf{C} \in L^1_{loc}(\Omega, \mathbb{R}^n)$. *Then there exists* $h \in \mathscr{L}^1_{loc,+}(\Omega)$ *such that*

$$P(M, v) = \int_{\partial_* M} [v(B - \operatorname{div}\mathsf{C}) \cdot \boldsymbol{n}^M + \nabla v \cdot \mathsf{C}\boldsymbol{n}^M]\, d\mathscr{H}^{n-1} \tag{7}$$

for every $v \in C^\infty(\Omega)$ *and* $M \in \mathscr{M}_{h\eta}$.

A remarkable feature of our approach is that the condition $\operatorname{div}\mathsf{C} \in L^1_{loc}(\Omega, \mathbb{R}^n)$, mentioned in the above theorem, has a counterpart in terms of the power P, as we show in Theorem 6 below.

This is quite interesting, since assumptions made on P are in general more 'physical' than those made on its densities.

To state this, we need a definition and a proposition.

Definition 12. Let $G = \big(x_0, (e_1, \dots, e_n), \widehat{G}\big)$ be a full grid and $M \in \mathscr{M}_G$ of the form

$$M = \{x \in \mathbb{R}^n : a_j < (x - x_0) \cdot e_j < b_j,\ j = 1, \dots, n\}, \tag{8}$$

where $a_1, b_1, \dots, a_n, b_n \in \widehat{G}$. Whenever $1 \leqslant j \leqslant n$ and $a_j \leqslant \alpha < \beta \leqslant b_j$, we set

$$M^{(j)}_{\alpha,\beta} = \{x \in \mathbb{R}^n : \alpha < (x - x_0) \cdot e_j < \beta,\ a_i < (x - x_0) \cdot e_i < b_i \quad \forall i \neq j\}.$$

We simply write $M^{(j)}_{\beta}$ in the case $\alpha = a_j$.

Proposition 2. *Let* $\mathscr{M}_G \subseteq \mathscr{A}$, $M \in \mathscr{M}_G$ *be represented as in* (8), $v \in C^\infty_c(M)$ *and* $1 \leqslant j \leqslant n$. *Then* $M^{(j)}_{\beta} \in \mathscr{M}_G$ *for* $\mathscr{L}^1$*-a.e.* $\beta \in (a_j, b_j]$ *and the map*

$$\big\{\beta \mapsto P(M^{(j)}_{\beta}, v)\big\}$$

belongs to $L^\infty(a_j, b_j)$ *for every* $v \in C^\infty_c(\Omega)$.

At this point we may state the following result.

Theorem 6. *The distribution* $\operatorname{div}\mathsf{C} \in L^1_{loc}(\Omega, \mathbb{R}^n)$ *if and only if there exists* $h \in \mathscr{L}^1_{loc,+}(\Omega)$ *such that*

$$\left| \int_{a_j}^{b_j} P(M^{(j)}_{\beta}, v)\, d\beta \right| \leqslant \int_M |v| h\, d\mathscr{L}^n \tag{9}$$

for every $\mathscr{M}_G \subseteq \mathscr{A}$, $M \in \mathscr{M}_G$, $v \in C^\infty_c(M)$ *and* $j = 1, \dots, n$. *In this case, we have* $|B - 2\operatorname{div}\mathsf{C}| \leqslant h$ *on* $\mathscr{L}^n$*-a.a. of* Ω.

6 Boundary representation with edges

Now we come to the most interesting application of second-order powers, namely, the possibility of having a representation formula on edges, or simply sets with non-smooth normal. Roughly speaking, this is made possible by the gradient term in (7), which allows a further integration by parts.

To do this, we need to introduce a new class of sets.

Definition 13. Let M be a normalized set with finite perimeter. We say that M is a *set with curvature measure* if there exist $\lambda_M \in \mathfrak{M}(\partial_* M)$ with $\lambda_M(\partial_* M) < +\infty$ and a Borel tensor field $\mathsf{U} : \partial_* M \to \mathrm{Sym}_2$ with $|\mathsf{U}(x)| = 1$ for λ_M-a.e. $x \in \partial_* M$ such that

$$-\int_{\partial_* M} [-(\operatorname{div} \mathsf{C}) \cdot \boldsymbol{n}^M + ((\nabla \mathsf{C})\boldsymbol{n}^M \boldsymbol{n}^M) \cdot \boldsymbol{n}^M]\, d\mathscr{H}^{n-1} = \int_{\partial_* M} \mathsf{C} \cdot \mathsf{U}\, d\lambda_M$$

for every $\mathsf{C} \in C_c^\infty(\mathbb{R}^n; \mathrm{Sym}_2)$. It turns out that λ_M is uniquely determined and U is uniquely determined λ_M-a.e.

For $h \in \mathscr{L}^1_{loc,+}(\Omega)$ we set

$$\mathscr{C}_{h\eta} = \Big\{ M \in \mathscr{M}_{h\eta} : \ M \text{ has curvature measure and } \int_{\partial_* M} h\, d\lambda_M < +\infty \Big\}.$$

Remark 3. One can prove that the elements of $\mathscr{R}$ are sets with curvature measure. Indeed, since on each face the term $[-(\operatorname{div} \mathsf{C}) \cdot \boldsymbol{n}^M + ((\nabla \mathsf{C})\boldsymbol{n}^M \boldsymbol{n}^M) \cdot \boldsymbol{n}^M]$ is a surface divergence, it turns out that λ_M is the Hausdorff measure $\mathscr{H}^{n-2}$ restricted to the edges, and $\mathsf{U} = \boldsymbol{n}^M \otimes \boldsymbol{N} + \boldsymbol{N} \otimes \boldsymbol{n}^M$, where $\boldsymbol{N}$ is the normal to the edge in the hyperplane of the surface.

We are now ready to perform the last integration by parts. In doing this, however, we remark that the normal derivative of v cannot be dropped, since it corresponds to a field of doublets assigned on the boundary.

We also want to let line integrals appear as surface integrals, with respect to a singular measure.

Since the formal integration by parts involves the symmetric gradient of a tensor, we briefly recall its definition.

Definition 14. Let $\mathsf{C} \in L^1_{loc}(\Omega, \mathrm{Sym}_2)$. We define the *symmetric gradient* of C by setting $(\nabla^s \mathsf{C})uvw$ on Ω as

$$\langle (\nabla^s \mathsf{C})uvw, \varphi \rangle = \frac{1}{3} \int_\Omega [(\mathsf{C}v \cdot w)(\nabla\varphi \cdot u) + (\mathsf{C}u \cdot w)(\nabla\varphi \cdot v) + (\mathsf{C}u \cdot v)(\nabla\varphi \cdot w)]\, d\mathscr{L}^n$$

for every $\varphi \in C_c^\infty(\Omega)$. The function $\big\{(u, v, w) \mapsto (\nabla^s \mathsf{C})uvw\big\}$ is 3-linear and symmetric; moreover, $((\nabla \mathsf{C})uu) \cdot u = (\nabla^s \mathsf{C})uuu$ for every $u \in \mathbb{R}^n$.

The following theorem gives us our final goal, provided $\operatorname{div} \mathsf{C} \in L^1_{loc}(\Omega, \mathbb{R}^n)$ and $\nabla^s \mathsf{C} \in L^1_{loc}(\Omega, \mathrm{Sym}_3)$, that is, the corresponding distributions are represented by locally integrable functions.

Theorem 7. *Let P be a contact power of order 2 such that (c) of Definition 6 holds with $\mu_1 \ll \mathscr{L}^n$ and $\mu_2 \ll \mathscr{L}^n$ and let $\eta = \mu_0$. Assume further that $\operatorname{div} \mathsf{C} \in L^1_{loc}(\Omega, \mathbb{R}^n)$ and $\nabla^s \mathsf{C} \in L^1_{loc}(\Omega, \operatorname{Sym}_3)$.*

Then there exists $h \in \mathscr{L}^1_{loc,+}(\Omega)$ such that

$$P(M, v) = \int_{\partial_* M} v \Big[(B - 2 \operatorname{div} \mathsf{C}) \cdot \boldsymbol{n}^M + (\nabla^s \mathsf{C}) \boldsymbol{n}^M \boldsymbol{n}^M \boldsymbol{n}^M \Big] \, d\mathscr{H}^{n-1} + \int_{\partial_* M} \frac{\partial v}{\partial n} (\mathsf{C} \boldsymbol{n}^M \cdot \boldsymbol{n}^M) \, d\mathscr{H}^{n-1} + \int_{\partial_* M} v \, \mathsf{C} \cdot \mathsf{U} \, d\lambda_M \quad (10)$$

for every $M \in \mathscr{C}_{h\eta}$ and $v \in C^\infty(\Omega)$.

In the same spirit as above, we show that the condition $\nabla^s \mathsf{C} \in L^1_{loc}(\Omega, \operatorname{Sym}_3)$ has a counterpart in terms of P.

Theorem 8. *The function $\nabla^s \mathsf{C} \in L^1_{loc}(\Omega, \operatorname{Sym}_3)$ if and only if there exists $h \in \mathscr{L}^1_{loc,+}(\Omega)$ such that*

$$\left| \int_\alpha^\beta P(M_s^{(j)}, v) \, ds \right| \leqslant \int_{M^{(j)}_{\alpha,\beta}} \left(|v| + \left| \frac{\partial v}{\partial e_j} \right| \right) h \, d\mathscr{L}^n \quad (11)$$

for every $\mathscr{M}_G \subseteq \mathscr{A}$, $M \in \mathscr{M}_G$, $v \in C^\infty_c(M)$, $j = 1, \ldots, n$ and $a_j < \alpha < \beta < b_j$. In this case, we have $|\nabla^s \mathsf{C}| \leqslant \frac{3}{2}(h + |B - 2 \operatorname{div} \mathsf{C}|)$ on $\mathscr{L}^n$-a.a. of Ω.

Acknowledgements

The research of the first author was partially supported by the MIUR project "Variational and topological methods in the study of nonlinear phenomena" (COFIN 2003) and by Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (INdAM).

The research of the second and third author was partially supported by the MIUR project "Modelli matematici per la scienza dei materiali" (COFIN 2002) and by Gruppo Nazionale per la Fisica Matematica (INdAM).

References

[1] Degiovanni, M., Marzocchi, A., Musesti, A. (1999): Cauchy fluxes associated with tensor fields having divergence measure. Arch. Ration. Mech. Anal. **147**, 197–223

[2] Degiovanni, M., Marzocchi, A., Musesti, A. (2004): Edge force densities and second order powers, to appear in Ann. Mat. Puro Appl.

[3] Dell'Isola, F., Seppecher, P. (1997): Edge contact forces and quasi-balanced power. Meccanica **32**, 33–52

[4] Di Carlo, A., Tatone, A. (2001): (Iper-)tensioni equi-potenza. 15^{th} AIMETA Congress of Theoretical and Applied Mechanics. Taormina, Sept. 26–29, 2001

[5] Forte, S., Vianello, M. (1988): On surface stresses and edge forces, Rend. Mat. Appl. (7) **8**, 409–426

[6] Germain, P. (1973): La méthode des puissances virtuelles en mécanique des milieux continus. I. Théorie du second gradient. J. Mécanique **12**, 235–274
[7] Germain, P. (1973): The method of virtual power in continuum mechanics. II. Microstructure. SIAM J. Appl. Math. **25**, 556–575
[8] Gurtin, M.E., Martins, L.C. (1976): Cauchy's theorem in classical physics. Arch. Ration. Mech. Anal. **60**, 305–324
[9] Gurtin, M.E., Williams, W.O., Ziemer, W.P. (1986): Geometric measure theory and the axioms of continuum thermodynamics. Arch. Ration. Mech. Anal. **92**, 1–22
[10] Marzocchi, A., Musesti, A. (2001): Decomposition and integral representation of Cauchy interactions associated with measures. Contin. Mech. Thermodyn. **13**, 149–169
[11] Marzocchi, A., Musesti, A. (2003): Balanced powers in continuum mechanics. Meccanica **38**, 369–389
[12] Noll, W., Virga, E.G. (1990): On edge interactions and surface tension. Arch. Ration. Mech. Anal. **111**, 1–31
[13] Šilhavý, M. (1985): The existence of the flux vector and the divergence theorem for general Cauchy fluxes. Arch. Ration. Mech. Anal. **90**, 195–212
[14] Šilhavý, M. (1991): Cauchy's stress theorem and tensor fields with divergences in L^p. Arch. Ration. Mech. Anal. **116**, 223–255
[15] Ziemer, W.P. (1983): Cauchy flux and sets of finite perimeter. Arch. Ration. Mech. Anal. **84**, 189–201

Tangent stiffness of a Timoshenko beam undergoing large displacements

M. Diaco, A. Romano, C. Sellitto

Abstract. The polar model of an elastic Timoshenko beam undergoing large displacements is investigated in detail. Special emphasis is given to the problems involved in the evaluation of the tangent stiffness to provide a complete answer to the question of whether or not tangent stiffness is tensorial and symmetric.

1 Introduction

The paper deals with the analysis of the Timoshenko beam model (i.e., a shear deformable beam) undergoing large displacements and deformations during an evolution process in the elastic range. The Timoshenko beam model provides a significant example of a continuum whose configuration space is an infinite-dimensional nonlinear manifold modeled on a Banach space [1,2,6]. Indeed the rotations of the cross sections of the beam are primary kinematic parameters ranging over the special orthogonal group which is a three-dimensional non-linear compact manifold. Due to the nonlinearity of the configuration manifold, it is compelling to apply the rules of calculus on manifolds in the evaluation of the tangent stiffness. The constitutive tangent stiffness is provided by the Hessian of the elastic potential defined as the second covariant derivative according to a chosen connection on the rotation manifold. The Hessian operator is the difference between the second directional derivative along trial and test fields and the first derivative in the direction of the covariant derivative of the test field in the direction of trial fields. The evaluation of the Hessian requires the extension of the virtual displacement to a vector field on the configuration manifold, but the result is tensorial as it is independent of this arbitrary choice. We show that the geometric tangent stiffness is symmetric if the torsion of the connection on the manifold vanishes. Moreover we explain that the lack of symmetry found in early calculations based on a canonical extension of the virtual displacement is due to the implicit assumption of a non-symmetric connection on the configuration manifold. The constitutive tangent stiffness plays an essential role both from the theoretical and the computational points of view. The dependence of its expression on the connection and on the extension of the virtual displacement, recently stressed in [10,11], and investigated in detail in the present paper, seems to have been overlooked in previous treatments [3–5,7].

2 The beam model

We develop the treatment of the Timoshenko beam model along the general guidelines set forth in the companion paper [11]. We also refer to that paper for more general definitions and results that are presented here in the special context of one-dimensional polar beam theory.

As usual let E^3 denote three-dimensional affine euclidean space, with translation linear space V^3, and SO(3) the special orthogonal group of rotations, which is a three-dimensional compact nonlinear manifold.

A placement of a Timoshenko beam is described by a regular curve in E^3, named the axis of the beam, and by a field of rotations $\mathbf{Q} \in \mathrm{SO}(3)$, attached at each point of the beam axis, which simulates the rigid body kinematics of the cross sections of the beam. The ambient space $\mathbb{S}$ is the fiber bundle defined by the projection $\boldsymbol{\pi}_{\mathbb{S}} : E^3 \times \mathrm{SO}(3) \mapsto E^3$.

This is a trivial fiber bundle defined by the cartesian product of the euclidean space (the base manifold) and the non-linear three-dimensional compact group of rotations (the fiber).

The material body is a set of particles which can be put in one-to-one correspondence with an interval $\mathcal{B}$ of the real line $\mathcal{R}$.

We consider an evolution process of the beam in a time observation interval $I = [t_o, t_f]$ and a reference base placement of the beam at time $s \in I$, which is a closed interval $\mathbb{B}_s$ of a regular curve in E^3. We denote the curvilinear abscissa along $\mathbb{B}_s$ by $\lambda \in \mathcal{R}$.

A configuration of the beam at time $t \in I$: $\mathbf{u}_t : \mathcal{B} \mapsto E^3 \times \mathrm{SO}(3)$ maps a particle $\mathbf{p} \in \mathcal{B}$ into a pair $\{\mathbf{r}_t(\mathbf{p}), \mathbf{Q}_t(\mathbf{p})\} \subset E^3 \times \mathrm{SO}(3)$ defining the position $\mathbf{r}_t(\mathbf{p}) \in E^3$ of the beam axis and the rotation $\mathbf{Q}_t(\mathbf{p}) \in \mathrm{SO}(3)$ of the corresponding cross section of the beam. The image of a configuration is called a placement of the beam. The map $\mathbf{r}_t : \mathcal{B} \mapsto E^3$ is named the base configuration of the beam at time $t \in I$ and its image is the base placement of the beam $\mathbb{B}_t$ at time $t \in I$.

The *polar structure* of the beam $\mathbf{s}_t : \mathbb{B}_t \mapsto \mathbb{S}$ is the map from the base placement at time $t \in I$ onto the placement $\mathbb{P}_t = \mathbf{s}_t(\mathbb{B}_t) \subset \mathbb{S}$ defined by

$$\mathbf{s}_t \circ \mathbf{r}_t = \mathbf{u}_t := \{\mathbf{r}_t, \mathbf{Q}_t\} \quad \forall \mathbf{r}_t \in \mathbb{B}_t ; \tag{1}$$

it has the property

$$(\boldsymbol{\pi}_{\mathbb{S}} \circ \mathbf{s}_t)(\mathbf{x}) = \mathbf{x} \qquad \forall \mathbf{x} \in \mathbb{B}_t \subset E^3 . \tag{2}$$

The *change of base configuration* from $\mathbf{r}_s$ to $\mathbf{r}_t$ is the diffeomorphism $\mathbf{r}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s ; \mathbb{B}_t)$ defined by

$$\mathbf{r}_{t,s} \circ \mathbf{r}_s = \mathbf{r}_t , \tag{3}$$

where the index k denotes a suitable integer.

The *change of configuration* from $\mathbf{u}_s$ to $\mathbf{u}_t$ is the map $\mathbf{u}_{t,s} := \{\mathbf{r}_{t,s}, \mathbf{Q}_{t,s}\} : \mathbf{u}_s(\mathcal{B}) \mapsto \mathbf{u}_t(\mathcal{B}) \subset \mathbb{S}$ defined by

$$\mathbf{r}_{t,s}(\mathbf{r}_s) = \mathbf{r}_t , \qquad \mathbf{Q}_{t,s}\, \mathbf{Q}_s = \mathbf{Q}_t . \tag{4}$$

The beam axis displacement field is given by $\mathbf{d}_{t,s} = \mathbf{r}_t - \mathbf{r}_s$ so that

$$\mathbf{r}_{t,s}(\mathbf{r}_s) = \mathbf{r}_s + \mathbf{d}_{t,s}\,, \tag{5}$$

$$\mathbf{r}_{t,s} \circ \mathbf{r}_s = \mathbf{r}_t\,. \tag{6}$$

The composition rules are given by

$$\mathbf{r}_{\tau,t} \circ \mathbf{r}_{t,s} = \mathbf{r}_{\tau,s}\,, \qquad \mathbf{Q}_{\tau,t} \circ \mathbf{Q}_{t,s} = \mathbf{Q}_{\tau,s}\,. \tag{7}$$

Since $\mathbf{r}_{s,s} \in \mathrm{C}^k(\mathbb{B}_s\,;\mathbb{B}_s)$ and $\mathbf{Q}_{s,s} : \mathbf{u}_s(\mathcal{B}) \mapsto \mathbf{u}_s(\mathcal{B})$ are identities, the maps $\mathbf{r}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s\,;\mathbb{B}_t)$ and $\mathbf{Q}_{t,s} : \mathbf{u}_s(\mathcal{B}) \mapsto \mathbf{u}_t(\mathcal{B})$ are invertible and their inverses are given by

$$(\mathbf{r}_{t,s})^{-1} = \mathbf{r}_{s,t}\,, \qquad (\mathbf{Q}_{t,s})^{-1} = \mathbf{Q}_{s,t}\,. \tag{8}$$

The Timoshenko beam undergoes a rigid displacement in the passage from the configuration $\mathbf{u}_s = \{\mathbf{r}_s\,, \mathbf{Q}_s\}$ to the configuration $\mathbf{u}_t = \{\mathbf{r}_t\,, \mathbf{Q}_t\}$ if and only if the relative rotation $\mathbf{Q}_{t,s}$ between the cross sections of the beam is uniform and the positions of the axis are given by

$$\mathbf{r}_t = \mathbf{Q}_{t,s}\,\mathbf{r}_s + \mathbf{c}\,, \tag{9}$$

where $\mathbf{c} \in V^3$ is a constant vector field.

The next proposition provides a measure of finite deformation for the Timoshenko beam which satisfies the requirements of consistency and nonredundancy illustrated in [11].

Proposition 3 (Deformation measures). *The finite deformation measure*

$$\mathfrak{D}(\mathbf{u}_{t,s}) := \left| \begin{array}{c} \boldsymbol{\delta}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s}) \\ \mathbf{C}(\mathbf{Q}_{t,s}) \end{array} \right|, \tag{10}$$

vanishes if and only if the Timoshenko beam undergoes a rigid displacement. Here

$$\begin{aligned} \boldsymbol{\delta}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s}) &:= \mathbf{Q}_{t,s}^T\,\mathbf{r}_t' - \mathbf{r}_s' : \mathbb{B}_s \mapsto V^3\,, && \textit{sliding vector field}\,, \\ \mathbf{C}(\mathbf{Q}_{t,s}) &:= \mathbf{Q}_{t,s}^T\,\mathbf{Q}_{t,s}' : \mathbb{B}_s \mapsto L(V^3\,;V^3)\,, && \textit{curvature tensor field}\,. \end{aligned} \tag{11}$$

The prime $(\cdot)'$ denotes the derivative with respect to the curvilinear abscissa along the beam axis in the configuration at the initial time $s \in I$.

Proof. Under a rigid transformation, the relative rotation field $\mathbf{Q}_{t,s}$ is uniform so that $\mathbf{Q}_{t,s}' = \mathbf{O}$ and hence $\mathbf{C}(\mathbf{Q}_{t,s}) = \mathbf{O}$. By differentiating the expression

$$\mathbf{r}_t = \mathbf{Q}_{t,s}\,\mathbf{r}_s + \mathbf{c}\,, \tag{12}$$

with respect to the curvilinear abscissa ξ along the beam axis at time $s \in I$, we see that

$$\mathbf{r}_t' = \mathbf{Q}_{t,s}\,\mathbf{r}_s' \quad \Longleftrightarrow \quad \boldsymbol{\delta}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s}) = \mathbf{o}\,. \tag{13}$$

Vice versa, if $\mathfrak{D}(\mathbf{r}_{t,s}) = \{\mathbf{o}, \mathbf{O}\}$, the condition $\mathbf{C}(\mathbf{Q}_{t,s}) = \mathbf{O}$ implies that the rotation $\mathbf{Q}_{t,s}$ is uniform. The condition $\boldsymbol{\delta}(\mathbf{r}_{t,s}) = \mathbf{o}$ then implies that $\mathbf{r}'_t = \mathbf{Q}_{t,s}\,\mathbf{r}'_s$. Integrating with respect to λ we get the relation

$$\mathbf{r}_t = \mathbf{Q}_{t,s}\,\mathbf{r}_s + \mathbf{c}\,, \tag{14}$$

which is characteristic of a rigid transformation.

The deformation measure of Proposition 3 satisfies the *consistency condition*

$$\boldsymbol{\delta}(\mathbf{r}_{\tau,s}, \mathbf{Q}_{\tau,s}) = \mathbf{Q}_{t,s}^T\,\boldsymbol{\delta}(\mathbf{r}_{\tau,t}, \mathbf{Q}_{\tau,t})\,\frac{d\xi_t}{d\xi_s} + \boldsymbol{\delta}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s})\,, \tag{15}$$

$$\mathbf{C}(\mathbf{Q}_{\tau,s}) = \mathbf{Q}_{t,s}^T\,\mathbf{C}(\mathbf{Q}_{\tau,t})\mathbf{Q}_{t,s}\frac{d\xi_t}{d\xi_s} + \mathbf{C}(\mathbf{Q}_{t,s}) \tag{16}$$

for any $s, t, \tau \in I$. In particular, if the relative deformations $\boldsymbol{\delta}(\mathbf{r}_{\tau,t}, \mathbf{Q}_{\tau,t})$ and $\mathbf{C}(\mathbf{Q}_{\tau,t})$ vanish, we infer that

$$\boldsymbol{\delta}(\mathbf{r}_{\tau,s}, \mathbf{Q}_{\tau,s}) = \boldsymbol{\delta}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s})\,, \tag{17}$$

$$\mathbf{C}(\mathbf{Q}_{\tau,s}) = \mathbf{C}(\mathbf{Q}_{t,s})\,. \tag{18}$$

Regarding the rigid transformation $\{\mathbf{r}_{\tau,t}, \mathbf{Q}_{\tau,t}\}$ as a change of observer, we may conclude that the deformation measure is frame indifferent.

3 Elastic equilibrium

The *space of configuration changes* from a reference placement (in short, the *configuration space*) is the differentiable manifold $\mathbb{M} := \mathrm{C}^k(\mathbb{B}_s\,;\mathbb{S})$ modeled on the Banach space $\mathrm{C}^k(\mathbb{B}_s\,;\mathcal{R}^d)$, $d = \dim\mathbb{S}$ (see [11]).

We now consider an elastic behavior defined by an elastic potential φ which is assumed to be a differentiable convex function of the finite deformation $\mathfrak{D}(\mathbf{u}_{t,s})$ corresponding to the change of configuration $\mathbf{u}_{t,s} \in \mathbb{M}$ evaluated from a natural configuration $\mathbf{u}_s : \mathcal{B} \mapsto \mathbb{S}$. Since the deformation measure is *frame indifferent*, the value of the elastic potential is independent of the observer and the *principle of material indifference* is satisfied.

The elastic law is imposed pointwise in the reference placement $\mathbb{B}_s$ by assuming that the local stress $\mathfrak{S} = \{\mathbf{F}_o\,, \mathbf{M}_o\}$, conjugate to the deformation measure $\mathfrak{D} = \{\boldsymbol{\delta}\,, \mathbf{C}\}$, is the gradient of the potential φ, according to the local formula

$$\mathfrak{S}_{\mathbf{x}} = \partial\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}}(\mathbf{u})) \quad \forall\,\mathbf{x} \in \mathbb{B}_s\,. \tag{19}$$

Since the deformation measure satisfies the *consistency condition*, the equilibrium of the elastic continuum at the configuration $\mathbf{u}_t : \mathcal{B} \mapsto \mathbb{S}$ may be enforced in terms of fields defined in the reference placement $\mathbb{B}_s$, by means of the variational condition

$$\int_{\mathbb{B}_s} \partial(\varphi_{\mathbf{x}} \circ \mathfrak{D}_{\mathbf{x}})(\mathbf{u}_{t,s}) \cdot \delta\mathbf{u}_{t,s}(\mathbf{x})d\mu \;= \langle\, \mathbf{G}(\mathbf{u}_{t,s}) \cdot \mathbf{f}_t, \delta\mathbf{u}_{t,s} \rangle\,,$$

which must be satisfied for any virtual displacement

$$\delta\mathbf{u}_{t,s}(\mathbf{x}) \in \mathbb{T}_{\mathbb{S}}(\mathbf{u}_{t,s}(\mathbf{x})) \quad \forall\, \mathbf{x} \in \mathbb{B}_s\,. \tag{20}$$

Here and in the sequel a dot denotes a linear dependence on the subsequent argument.

The *global elastic potential* $\phi \in \mathrm{C}^2(\mathrm{C}^k(\mathbb{B}_s, \mathbb{S})\,; \mathcal{R})$ provides the elastic energy associated with the configuration change $\mathbf{u} \in \mathbb{M} = \mathrm{C}^k(\mathbb{B}_s, \mathbb{S})$ and is given by

$$\phi(\mathbf{u}) := (\varphi \circ \mathfrak{D})(\mathbf{u}) = \int_{\mathbb{B}} (\varphi_{\mathbf{x}} \circ \mathfrak{D}_{\mathbf{x}})(\mathbf{u})\, d\mu\,. \tag{21}$$

Setting $\mathbf{u} = \mathbf{u}_{t,s}$ we denote by $\mathbb{T}_{\mathbb{M}}(\mathbf{u})$ the linear space of tangent vectors to the manifold $\mathbb{M} := \mathrm{C}^k(\mathbb{B}_s\,; \mathbb{S})$ at the configuration $\mathbf{u} \in \mathbb{M}$ and by $\mathbb{T}^*_{\mathbb{M}}(\mathbf{u}) = BL\,(\mathbb{T}_{\mathbb{M}}(\mathbf{u})\,; \mathcal{R})$ the dual space of continuous linear forms on $\mathbb{T}_{\mathbb{M}}(\mathbf{u})$.

The referential equilibrium of the body at time $t \in I$ is then expressed by

$$\langle\, \partial\phi(\mathbf{u}), \delta\mathbf{u} \rangle \;=\; \langle\, \mathbf{G}(\mathbf{u}) \cdot \mathbf{f}, \delta\mathbf{u} \rangle \qquad\qquad \forall\, \delta\mathbf{u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})\,. \tag{22}$$

The bounded linear functionals $\mathbf{G}(\mathbf{u}) \cdot \mathbf{f} \in \mathbb{T}^*_{\mathbb{M}}(\mathbf{u})$ and $\partial\phi(\mathbf{u}) \in \mathbb{T}^*_{\mathbb{M}}(\mathbf{u})$ respectively provide the referential applied load and the referential elastic response of the body (see [11]).

4 Tangent stiffness

The condition of incremental elastic equilibrium of a Timoshenko beam at the configuration $\mathbf{u}_{t,s} \in \mathbb{M} = \mathrm{C}^k(\mathbb{B}_s, \mathbb{S})$ is obtained by linearizing the equilibrium condition.

In performing the linearization we must recall that the kinematic parameters of the beam are fields whose values belong to the non-linear differential manifold $E^3 \times \mathrm{SO}(3)$.

With $\mathbf{G}_{\mathbf{f}}(\mathbf{u}) := \mathbf{G}(\mathbf{u}) \cdot \mathbf{f}$, the incremental equilibrium is imposed by carrying out the total time derivative of the non-linear condition along the equilibrium path

$$\frac{d}{dt}\Big[(\partial\phi - \mathbf{G}_{\mathbf{f}})(\mathbf{u})\Big] \;=\; 0\,. \tag{23}$$

Since both the configuration change $\mathbf{u}$ and the force $\mathbf{f}$ depend on $t \in I$, the incremental equilibrium condition may be rewritten as

$$\partial_{\dot{\mathbf{u}}}(\partial\phi - \mathbf{G}_{\mathbf{f}})(\mathbf{u}) \;=\; \mathbf{G}(\mathbf{u}) \cdot \dot{\mathbf{f}}\,, \tag{24}$$

where a superimposed dot denotes the time derivative and $\dot{\mathbf{u}} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$.

The *total tangent stiffness* of the body is the directional derivative

$$\mathbf{K}(\mathbf{u}) := \partial(\partial\phi - \mathbf{G}_{\mathbf{f}})(\mathbf{u})\,, \tag{25}$$

and the incremental equilibrium is accordingly written as

$$\mathbf{K}(\mathbf{u}) \cdot \dot{\mathbf{u}} = \mathbf{G}(\mathbf{u}) \cdot \dot{\mathbf{f}}\,. \tag{26}$$

However, as illustrated in detail in [11], when dealing with a polar continuum, such as a Timoshenko beam, the directional derivative of the covector field $(\partial\phi - \mathbf{G_f}) \in \mathrm{C}^k(\mathbb{M};\mathbb{T}^*_{\mathbb{M}})$ at a configuration $\mathbf{u} \in \mathbb{M}$ must be taken according to a connection defined on the configuration space, the non-linear manifold of maps $\mathbb{M} = \mathrm{C}^k(\mathbb{B}_s;\mathbb{S})$.

Once a connection $\nabla^{\mathbb{S}}$ is defined on the ambient space $\mathbb{S}$ a corresponding connection $\nabla^{\mathbb{M}}$ is induced on the configuration manifold $\mathbb{M}$ (see [11]). The total tangent stiffness is then computed by taking covariant derivatives instead of directional derivatives to get the expression

$$\mathbf{K}(\mathbf{u}) := \nabla^{\mathbb{M}}(\partial\phi - \mathbf{G_f})(\mathbf{u}) = (\nabla^{\mathbb{M}}\boldsymbol{\alpha})(\mathbf{u})\,, \tag{27}$$

where

$$\boldsymbol{\alpha} = \partial\phi - \mathbf{G_f} \tag{28}$$

is the equilibrium gap resulting from the difference between the elastic response $\partial\phi \in \mathrm{C}^k(\mathbb{M};\mathbb{T}^*_{\mathbb{M}})$ and the referential load $\mathbf{G_f} \in \mathrm{C}^k(\mathbb{M};\mathbb{T}^*_{\mathbb{M}})$.

The covariant derivative of the covector field $\boldsymbol{\alpha} \in \mathrm{C}^k(\mathbb{M};\mathbb{T}^*_{\mathbb{M}})$ is the linear form $\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}}\boldsymbol{\alpha} \in \mathrm{C}^k(\mathbb{M};\mathbb{T}^*_{\mathbb{M}})$ defined by

$$(\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}}\boldsymbol{\alpha})(\mathbf{u}) \cdot \delta\mathbf{u} := \partial_{\dot{\mathbf{u}}}\,(\boldsymbol{\alpha} \cdot \hat{\delta\mathbf{u}})(\mathbf{u}) - \boldsymbol{\alpha}(\mathbf{u}) \cdot (\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}}\,\hat{\delta\mathbf{u}})(\mathbf{u})\,, \tag{29}$$

where $\hat{\delta\mathbf{u}}$ is an extension of the virtual displacement $\delta\mathbf{u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$ to a neighborhood $U(\mathbf{u}) \subset \mathbb{M}$ of the configuration $\mathbf{u} \in \mathbb{M}$ and $(\boldsymbol{\alpha} \cdot \hat{\delta\mathbf{u}}) \in \mathrm{C}^k(\mathbb{M};\mathscr{R})$ is the scalar field defined by

$$(\boldsymbol{\alpha} \cdot \hat{\delta\mathbf{u}})(\mathbf{u}) := \boldsymbol{\alpha}(\mathbf{u}) \cdot \hat{\delta\mathbf{u}}(\mathbf{u})\,. \tag{30}$$

It can be shown that the Hessian is independent of the choice of the extension [8,11]. Hereafter the suffices $^{\mathbb{S}}$ and $^{\mathbb{M}}$ are dropped unless necessary.

We consider two tangent vector fields at the configuration $\mathbf{u}_{t,s} \in \mathbb{M}$,

$$\begin{matrix} \dot{\mathbf{u}}_{t,s}(\mathbf{x}) \ \in \mathbb{T}_{\mathbb{S}}(\mathbf{u}_{t,s}(\mathbf{x})) \\ \delta\mathbf{u}_{t,s}(\mathbf{x}) \in \mathbb{T}_{\mathbb{S}}(\mathbf{u}_{t,s}(\mathbf{x})) \end{matrix} \qquad \forall\,\mathbf{x} \in \mathbb{B}_s\,. \tag{31}$$

To simplify the exposition we set $\mathbf{m} = \mathbf{u}_t(\mathbf{x})$ and adopt the shorthand notation

$$\mathbf{Q} = \mathbf{Q}_{t,s}\,, \qquad \mathbf{r} = \mathbf{r}_{t,s} \tag{32}$$

$$\mathbf{X_m} = \dot{\mathbf{u}}_{t,s}(\mathbf{x}) = \{\dot{\mathbf{r}}_{\mathbf{X}}\,,\dot{\mathbf{Q}}_{\mathbf{X}}\}\,, \quad \mathbf{Y_m} = \delta\mathbf{u}_{t,s}(\mathbf{x}) = \{\dot{\mathbf{r}}_{\mathbf{Y}}\,,\dot{\mathbf{Q}}_{\mathbf{Y}}\}\,. \tag{33}$$

The field $\mathbf{X_m} = \{\dot{\mathbf{r}}_{\mathbf{X}}\,,\dot{\mathbf{Q}}_{\mathbf{X}}\}$ is the unknown velocity along the equilibrium path and the field $\mathbf{Y_m} = \{\dot{\mathbf{r}}_{\mathbf{Y}}\,,\dot{\mathbf{Q}}_{\mathbf{Y}}\}$ is a virtual displacement which plays the role of test field in the variational condition of equilibrium.

In the sequel, in common with earlier treatments in the literature (see [4,7]), we analize only the tangent stiffness stemming from the constitutive response of the

beam. The constitutive stiffness is the Hessian of the elastic potential $\phi = \varphi \circ \mathfrak{D} \in \mathrm{C}^k(\mathbb{M}\,;\mathcal{R})$, a twice covariant tensor field on the configuration manifold $\mathbb{M}$ defined as the covariant derivative of the linear form $\boldsymbol{\alpha} = \partial\,\phi \in \mathrm{C}^k(\mathbb{M}\,;\mathbb{T}^*_{\mathbb{M}})$:

$$\mathbf{K}(\mathbf{u}) = (\nabla^2_{\dot{\mathbf{u}}\,\delta\mathbf{u}}\phi)(\mathbf{u}) := (\nabla_{\dot{\mathbf{u}}}\partial\,\phi)(\mathbf{u})\cdot\delta\mathbf{u} = \partial_{\dot{\mathbf{u}}}\,(\partial_{\hat{\delta\mathbf{u}}}\,\phi)(\mathbf{u}) - \partial_{\left(\nabla_{\dot{\mathbf{u}}}\,\hat{\delta\mathbf{u}}\right)}\phi(\mathbf{u})\,. \quad (34)$$

The explicit expression of the Hessian is given by

$$\mathbf{K}(\mathbf{u}) = \int_{\mathbb{B}_s} \nabla^2_{\mathbf{X_m}\mathbf{Y_m}}(\varphi_{\mathbf{x}} \circ \mathfrak{D}_{\mathbf{x}})(\mathbf{u})\,d\mu\,.$$

Accordingly the tangent stiffness can be decomposed in the sum of two terms:

- the *elastic tangent stiffness*

$$\int_{\mathbb{B}_s} \partial^2\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}}(\mathbf{u}))\cdot(\partial_{\mathbf{Y_m}}\mathfrak{D}_{\mathbf{x}})(\mathbf{u})\cdot(\partial_{\mathbf{X_m}}\mathfrak{D}_{\mathbf{x}})(\mathbf{u})\,d\mu\,, \quad (35)$$

which is a two-times covariant symmetric tensor as it is the second directional derivative of the scalar function $\varphi_{\mathbf{x}}$ in the linear space $V^3 \times \mathrm{L}(V^3\,;V^3)$;
- the *geometric tangent stiffness*

$$\int_{\mathbb{B}_s} \partial\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}}(\mathbf{u}))\cdot(\nabla^2_{\mathbf{X_m}\mathbf{Y_m}}\,\mathfrak{D}_{\mathbf{x}})(\mathbf{u})\;d\mu\,, \quad (36)$$

where

$$(\nabla^2_{\mathbf{X_m}\mathbf{Y_m}}\,\mathfrak{D}_{\mathbf{x}})(\mathbf{u}) = \partial_{\mathbf{X_m}}(\partial_{\hat{\mathbf{Y}}}\,\mathfrak{D}_{\mathbf{x}})(\mathbf{u}) - \partial_{\left(\nabla_{\mathbf{X_m}}\hat{\mathbf{Y}}\right)}\mathfrak{D}_{\mathbf{x}}(\mathbf{u}) \quad (37)$$

is the Hessian of the deformation measure.

The geometric tangent stiffness is then a two-times covariant tensor which is symmetric if the Hessian $(\nabla^2_{\mathbf{X_m}\mathbf{Y_m}}\,\mathfrak{D}_{\mathbf{x}})(\mathbf{u})$ is so. In turn this Hessian is symmetric if the torsion of the connection ∇ vanishes.

The vector field $\hat{\mathbf{Y}}$ is an extension of the virtual displacement $\mathbf{Y} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$ to a neighborhood $U(\mathbf{u}) \subset \mathbb{M}$ of the configuration $\mathbf{u} \in \mathbb{M}$.

This extension pertains only to the component of the virtual displacement in SO(3) since the component in E^3 is trivially extended as a constant field.

We remark that any vector $\dot{\mathbf{Q}}$ tangent to the manifold SO(3) at a point $\mathbf{Q}$ can be represented by $\dot{\mathbf{Q}} = \mathbf{W}\,\mathbf{Q}$ where $\mathbf{W} \in \mathrm{so}(3)$ is a skew-symmetric tensor. By taking constant the tensor $\mathbf{W} \in \mathrm{so}(3)$, this formula provides a simple way to extend a vector tangent at a point of SO(3) to a vector field on SO(3). This extension, which is referred to as the *canonical extension*, was the one adopted by Simo and Vu-Quoc in [4].

By differentiating the expression of the finite deformation

$$\mathfrak{D}_{\mathbf{x}}(\mathbf{r},\mathbf{Q}) := \left| \begin{matrix} \boldsymbol{\delta}(\mathbf{r},\mathbf{Q}) \\ \mathbf{C}(\mathbf{Q}) \end{matrix} \right| = \left| \begin{matrix} \mathbf{Q}^T\,\mathbf{r}'_t - \mathbf{r}'_s \\ \mathbf{Q}^T\,\mathbf{Q}' \end{matrix} \right| , \tag{38}$$

we see that

$$\partial\mathfrak{D}_{\mathbf{x}}(\{\,\mathbf{r},\mathbf{Q}\,\})\cdot\{\,\dot{\mathbf{r}}_{\mathbf{Y}},\dot{\mathbf{Q}}_{\mathbf{Y}}\,\} = \left| \begin{matrix} \dot{\mathbf{Q}}^T_{\mathbf{Y}}\,\mathbf{r}'_t + \mathbf{Q}^T\,\dot{\mathbf{r}}'_{\mathbf{Y}} \\ \dot{\mathbf{Q}}^T_{\mathbf{Y}}\mathbf{Q}' + \mathbf{Q}^T\dot{\mathbf{Q}}'_{\mathbf{Y}} \end{matrix} \right| , \tag{39}$$

and, setting $\dot{\mathbf{Q}}_{\mathbf{Y}} = \mathbf{W}_{\mathbf{Y}}\,\mathbf{Q}$, we get

$$\partial\mathfrak{D}_{\mathbf{x}}(\{\,\mathbf{r},\mathbf{Q}\,\})\cdot\{\,\dot{\mathbf{r}}_{\mathbf{Y}},\mathbf{W}_{\mathbf{Y}}\,\mathbf{Q}\,\} = \left| \begin{matrix} \mathbf{Q}^T\,(\dot{\mathbf{r}}'_{\mathbf{Y}} - \mathbf{W}_{\mathbf{Y}}\,\mathbf{r}'_t) \\ \mathbf{Q}^T\mathbf{W}'_{\mathbf{Y}}\mathbf{Q} \end{matrix} \right| . \tag{40}$$

The second derivative $\partial_{\mathbf{X_m}}(\partial_{\mathbf{Y}}\,\mathfrak{D}_{\mathbf{x}})(\mathbf{u})$ is evaluated by putting $\dot{\mathbf{Q}}_{\mathbf{X}} = \mathbf{W}_{\mathbf{X}}\mathbf{Q}$ in $\mathbf{X_m} = \{\dot{\mathbf{r}}_{\mathbf{X}}\,,\dot{\mathbf{Q}}_{\mathbf{X}}\}$.

Then, as $\dot{\mathbf{Q}}^T_{\mathbf{X}} = -\mathbf{Q}^T\mathbf{W}_{\mathbf{X}}$, we see that

$$\partial(\mathbf{Q}^T\,(\dot{\mathbf{r}}'_{\mathbf{Y}} - \mathbf{W}_{\mathbf{Y}}\,\mathbf{r}'_t))[\dot{\mathbf{Q}}_{\mathbf{X}}] = \mathbf{Q}^T\left[\mathbf{W}_{\mathbf{X}}\,\mathbf{W}_{\mathbf{Y}}\,\mathbf{r}'_t - (\mathbf{W}_{\mathbf{X}}\,\dot{\mathbf{r}}'_{\mathbf{Y}} + \mathbf{W}_{\mathbf{Y}}\,\dot{\mathbf{r}}'_{\mathbf{X}})\right], \tag{41}$$

$$\begin{aligned} \partial(\mathbf{Q}^T\mathbf{W}'_{\mathbf{Y}}\mathbf{Q})\,[\dot{\mathbf{Q}}_{\mathbf{X}}] &= \dot{\mathbf{Q}}^T_{\mathbf{X}}\mathbf{W}'_{\mathbf{Y}}\mathbf{Q} + \mathbf{Q}^T\mathbf{W}'_{\mathbf{Y}}\dot{\mathbf{Q}}_{\mathbf{X}} = \\ &= -\mathbf{Q}^T\mathbf{W}_{\mathbf{X}}\mathbf{W}'_{\mathbf{Y}}\mathbf{Q} + \mathbf{Q}^T\mathbf{W}'_{\mathbf{Y}}\mathbf{W}_{\mathbf{X}}\mathbf{Q} = \\ &= \mathbf{Q}^T\left[-\mathbf{W}_{\mathbf{X}}\mathbf{W}'_{\mathbf{Y}} + \mathbf{W}'_{\mathbf{Y}}\mathbf{W}_{\mathbf{X}}\right]\mathbf{Q} = \mathbf{Q}^T\left[\mathbf{W}'_{\mathbf{Y}},\mathbf{W}_{\mathbf{X}}\right]\mathbf{Q}\,, \end{aligned} \tag{42}$$

where $\left[\mathbf{A},\mathbf{B}\right]$ is the commutator of two tensors in $\mathrm{L}(V^3\,;V^3)$ defined by

$$\left[\mathbf{A},\mathbf{B}\right] := \mathbf{AB} - \mathbf{BA}\,. \tag{43}$$

It is apparent that the commutator of two skew-symmetric tensors is a skew-symmetric tensor. In conclusion the second directional derivative of the deformation measure has the expression

$$\partial_{\mathbf{X_m}}(\partial_{\hat{\mathbf{Y}}}\,\mathfrak{D}_{\mathbf{x}})(\mathbf{u}) \;= \left| \begin{matrix} \mathbf{Q}^T\left[\mathbf{W}_{\mathbf{X}}\,\mathbf{W}_{\mathbf{Y}}\,\mathbf{r}'_t - (\mathbf{W}_{\mathbf{X}}\,\dot{\mathbf{r}}'_{\mathbf{Y}} + \mathbf{W}_{\mathbf{Y}}\,\dot{\mathbf{r}}'_{\mathbf{X}})\right] \\ \mathbf{Q}^T\left[\mathbf{W}'_{\mathbf{Y}},\mathbf{W}_{\mathbf{X}}\right]\mathbf{Q} \end{matrix} \right| , \tag{44}$$

which is apparently non-symmetric with respect to an exchange of the vectors

$$\mathbf{X_m} = \{\dot{\mathbf{r}}_{\mathbf{X}}\,,\dot{\mathbf{Q}}_{\mathbf{X}}\} = \{\dot{\mathbf{r}}_{\mathbf{X}}\,,\mathbf{W}_{\mathbf{X}}\,\mathbf{Q}\}\,, \qquad \mathbf{Y_m} = \{\dot{\mathbf{r}}_{\mathbf{Y}}\,,\dot{\mathbf{Q}}_{\mathbf{Y}}\} = \{\dot{\mathbf{r}}_{\mathbf{Y}}\,,\mathbf{W}_{\mathbf{Y}}\,\mathbf{Q}\}\,. \tag{45}$$

This is the expression that was evaluated in [4].

Lastly, the directional derivative $\partial\mathfrak{D}_{\mathbf{x}}(\mathbf{u})\cdot(\nabla_{\mathbf{X_m}}\hat{\mathbf{Y}})$ must be evaluated. To this end it is necessary to equip the manifold SO(3) with a connection. The Levi–Civita connection corresponding to the Riemannian metric is the natural candidate.

This metric is simply that induced on SO(3) by the euclidean metric in the ambient linear space $\mathrm{L}(V^3;V^3)$. The Levi–Civita connection may then be defined either by the general Koszul formula or as the orthogonal projection on SO(3) of the directional derivative in $\mathrm{L}(V^3;V^3)$ [8,9].

The former route was followed in [7]. We follow the latter route which is by far the simpler. Indeed, we consider the tensor field $\hat{\mathbf{T}}_{\mathbf{Y}}:\mathrm{L}(V^3;V^3)\mapsto\mathrm{L}(V^3;V^3)$ defined by

$$\hat{\mathbf{T}}_{\mathbf{Y}}(\mathbf{B}):=\mathbf{W}_{\mathbf{Y}}\,\mathbf{B}\qquad\forall\,\mathbf{B}\in\mathrm{L}(V^3;V^3)\,, \tag{46}$$

so that

$$\partial_{\mathbf{A}}\hat{\mathbf{T}}_{\mathbf{Y}}(\mathbf{B})=\mathbf{W}_{\mathbf{Y}}\,\mathbf{A}\quad\forall\,\mathbf{A}\in\mathrm{L}(V^3;V^3)\,. \tag{47}$$

Then, setting $\mathbf{T}_{\mathbf{X}}=\hat{\mathbf{T}}_{\mathbf{X}}(\mathbf{Q})\in\mathbb{T}_{\mathrm{SO(3)}}(\mathbf{Q})=\mathrm{so}(3)\,\mathbf{Q}$, we have

$$(\partial_{\mathbf{T}_{\mathbf{X}}}\hat{\mathbf{T}}_{\mathbf{Y}})(\mathbf{Q})=\mathbf{W}_{\mathbf{Y}}\,\mathbf{W}_{\mathbf{X}}\,\mathbf{Q}\,,\qquad\mathbf{Q}\in\mathrm{SO(3)}\,. \tag{48}$$

The orthogonal projection on the subspace $\mathrm{so}(3)\,\mathbf{Q}$, which is tangent to the manifold SO(3) at $\mathbf{Q}$, provides the expression of the covariant derivative

$$(\nabla_{\mathbf{T}_{\mathbf{X}}}\hat{\mathbf{T}}_{\mathbf{Y}})(\mathbf{Q})=\big(\mathrm{emi}\,(\mathbf{W}_{\mathbf{Y}}\,\mathbf{W}_{\mathbf{X}})\big)\,\mathbf{Q}=-\frac{1}{2}\,[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}]\,\mathbf{Q}\,, \tag{49}$$

where the commutator $[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}]$ is a skew-symmetric tensor.

We can now evaluate the directional derivative $\partial\mathfrak{D}_{\mathbf{x}}(\mathbf{u})\cdot(\nabla_{\mathbf{X_m}}\hat{\mathbf{Y}})$. To this end we observe that the directional derivatives of the curvature and of the sliding along $\dot{\mathbf{Q}}$ are respectively given by

$$\begin{aligned}\partial(\mathbf{Q}^T\,\mathbf{Q}')\,[\dot{\mathbf{Q}}]&=\dot{\mathbf{Q}}^T\,\mathbf{Q}'+\mathbf{Q}^T\,\dot{\mathbf{Q}}'\,,\\ \partial(\mathbf{Q}^T\,\mathbf{r}'_t-\mathbf{r}'_s)\,[\dot{\mathbf{Q}}]&=\dot{\mathbf{Q}}^T\,\mathbf{r}'_t\,.\end{aligned} \tag{50}$$

By setting

$$\dot{\mathbf{Q}}=(\nabla_{\mathbf{T}_{\mathbf{X}}}\hat{\mathbf{T}}_{\mathbf{Y}})(\mathbf{Q})=-\frac{1}{2}\,[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}]\,\mathbf{Q}\,, \tag{51}$$

the directional derivatives of the curvature and of the sliding become

$$\begin{aligned}\dot{\mathbf{Q}}^T\,\mathbf{Q}'+\mathbf{Q}^T\,\dot{\mathbf{Q}}'&=\tfrac{1}{2}\,\mathbf{Q}^T\,[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}]\,\mathbf{Q}'-\tfrac{1}{2}\mathbf{Q}^T\,[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}]'\,\mathbf{Q}+\\ &\quad-\tfrac{1}{2}\,\mathbf{Q}^T\,[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}]\,\mathbf{Q}'=-\tfrac{1}{2}\,\mathbf{Q}^T\,[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}]'\,\mathbf{Q}\,,\\ \dot{\mathbf{Q}}^T\,\mathbf{r}'_t&=\tfrac{1}{2}\,\mathbf{Q}^T\,[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}]\,\mathbf{r}'_t\,,\end{aligned} \tag{52}$$

and hence we get

$$\partial\mathfrak{D}_{\mathbf{x}}(\mathbf{u}) \cdot (\nabla_{\mathbf{X_m}} \hat{\mathbf{Y}}) = \frac{1}{2} \begin{vmatrix} \mathbf{Q}^T [\mathbf{W_X}, \mathbf{W_Y}] \mathbf{r}'_t \\ -\mathbf{Q}^T [\mathbf{W_X}, \mathbf{W_Y}]' \mathbf{Q} \end{vmatrix} . \tag{53}$$

Then, from the formula

$$(\nabla^2_{\mathbf{X_m Y_m}} \mathfrak{D}_{\mathbf{x}})(\mathbf{u}) = \partial_{\mathbf{X_m}} (\partial_{\hat{\mathbf{Y}}} \mathfrak{D}_{\mathbf{x}})(\mathbf{u}) - \partial_{\left(\nabla_{\mathbf{X_m}} \hat{\mathbf{Y}}\right)} \mathfrak{D}_{\mathbf{x}}(\mathbf{u}) , \tag{54}$$

we infer the final result

$$(\nabla^2_{\mathbf{X_m Y_m}} \mathfrak{D}_{\mathbf{x}})(\mathbf{u}) = \frac{1}{2} \begin{vmatrix} \mathbf{Q}^T \left[(\mathbf{W_X W_Y} + \mathbf{W_Y W_X}) \mathbf{r}'_t - 2(\mathbf{W_X} \dot{\mathbf{r}}'_{\mathbf{Y}} + \mathbf{W_Y} \dot{\mathbf{r}}'_{\mathbf{X}}) \right] \\ \mathbf{Q}^T ([\mathbf{W}'_{\mathbf{X}}, \mathbf{W_Y}] + [\mathbf{W}'_{\mathbf{Y}}, \mathbf{W_X}]) \mathbf{Q} \end{vmatrix} \tag{55}$$

which is clearly symmetric with respect to the exchange of the vectors $\mathbf{X_m}, \mathbf{Y_m} \in V^3 \times \mathrm{so}(3)\,\mathbf{Q}$.

The symmetry of the constitutive tangent stiffness is a direct consequence of the symmetry of the Levi–Civita connection. Moreover, since the covariant derivative $(\nabla_{\mathbf{T_X}} \hat{\mathbf{T}}_{\mathbf{Y}})(\mathbf{Q})$ is skew-symmetric, the second directional derivative

$$\partial_{\mathbf{X_m}} (\partial_{\hat{\mathbf{Y}}} \mathfrak{D}_{\mathbf{x}})(\mathbf{u}) \tag{56}$$

is the sum of the symmetric bilinear form $(\nabla^2_{\mathbf{X_m Y_m}} \mathfrak{D}_{\mathbf{x}})(\mathbf{u})$ and the skew-symmetric bilinear form $\partial\mathfrak{D}_{\mathbf{x}}(\mathbf{u}) \cdot (\nabla_{\mathbf{X_m}} \hat{\mathbf{Y}})$.

It is then apparent that the statement claimed in [4,7] that the symmetric constitutive tangent stiffness can be obtained simply by taking the symmetric part of the classical Hessian of the elastic potential (the second directional derivative), is not a general rule but a direct special consequence of the special assumptions concerning the connection on $\mathrm{SO}(3)$ (the Levi–Civita connection) and the extension of the virtual displacement (the canonical extension).

On the other hand complete understanding of the reason why the non-symmetric classical Hessian of the elastic potential turns out to be tensorial can be had by observing that it coincides with the second covariant derivative taken according to the connection induced by the parallel transport defined by the canonical extension of a vector tangent at a point of $\mathrm{SO}(3)$ to a vector field on $\mathrm{SO}(3)$. As proved in the Appendix, the torsion of this connection does not vanish. As a consequence the Hessian is not symmetric but is tensorial.

5 Matrix form of the tangent stiffness

It is convenient to rewrite the expression of the stiffness in terms of axial vectors associated to the skew-symmeric tensors $\mathbf{W_X}$ and $\mathbf{W_Y}$. To rewrite the deformation

measure $\mathfrak{D}$ in terms of axial vectors we recall the formulas

$$\begin{aligned} &\text{axial}\,(\mathbf{Q}\,\mathbf{W}_{\mathbf{X}}\,\mathbf{Q}^T) = \mathbf{Q}\,\text{axial}\,\mathbf{W}_{\mathbf{X}} \qquad\qquad \forall\,\mathbf{Q}\in \text{SO(3)}\,,\\ &\text{axial}\,\left[\mathbf{W}_{\mathbf{X}},\,\mathbf{W}_{\mathbf{Y}}\right] = (\text{axial}\,\mathbf{W}_{\mathbf{X}})\times(\text{axial}\,\mathbf{W}_{\mathbf{Y}})\,, \end{aligned} \tag{57}$$

and set

$$\boldsymbol{\omega}_{\mathbf{X}} = \text{axial}\,\mathbf{W}_{\mathbf{X}}\,, \quad \boldsymbol{\omega}_{\mathbf{Y}} = \text{axial}\,\mathbf{W}_{\mathbf{Y}}\,. \tag{58}$$

5.1 Elastic tangent stiffness

We consider the vector form of the deformation measure

$$\mathfrak{D}(\mathbf{r},\mathbf{Q}) := \left|\begin{matrix} \mathbf{Q}^T\,\mathbf{r}'_t - \mathbf{r}'_s \\ \mathbf{c}(\mathbf{Q}) \end{matrix}\right|, \tag{59}$$

with $\mathbf{c}(\mathbf{Q}) := \text{axial}\,(\mathbf{Q}^T\,\mathbf{Q}')$. By recalling that

$$\boldsymbol{\delta}(\mathbf{r},\mathbf{Q}) = \dot{\mathbf{r}}' - \mathbf{W}\,\mathbf{r}'_t = \dot{\mathbf{r}}' + \mathbf{r}'_t\times\boldsymbol{\omega}\,, \qquad \dot{\mathbf{c}}(\mathbf{Q}) = \text{axial}\,(\mathbf{Q}^T\mathbf{W}'\mathbf{Q}) = \mathbf{Q}^T\boldsymbol{\omega}'\,, \tag{60}$$

and by defining the operators

$$\mathbb{Q} := \begin{bmatrix} \mathbf{Q} & \mathbf{O} \\ \mathbf{O} & \mathbf{Q} \end{bmatrix}, \qquad \boldsymbol{\Xi}^T := \begin{bmatrix} \frac{\partial}{\partial\lambda}\mathbf{I} & \mathbf{r}'_t\times \\ \mathbf{O} & \frac{\partial}{\partial\lambda}\mathbf{I} \end{bmatrix}, \tag{61}$$

we see that

$$\begin{aligned} \partial\,\mathfrak{D}(\mathbf{r},\mathbf{Q})\cdot(\{\dot{\mathbf{r}}_{\mathbf{Y}},\mathbf{W}_{\mathbf{Y}}\,\mathbf{Q}\}) &= \left|\begin{matrix} \mathbf{Q}^T\,(\dot{\mathbf{r}}'_{\mathbf{Y}} + \mathbf{r}'_t\times\boldsymbol{\omega}_{\mathbf{Y}}) \\ \mathbf{Q}^T\boldsymbol{\omega}'_{\mathbf{Y}} \end{matrix}\right| \\ &= \mathbb{Q}^T\left|\begin{matrix} \dot{\mathbf{r}}'_{\mathbf{Y}} + \mathbf{r}'_t\times\boldsymbol{\omega}_{\mathbf{Y}} \\ \boldsymbol{\omega}'_{\mathbf{Y}} \end{matrix}\right| = \mathbb{Q}^T\boldsymbol{\Xi}^T\left|\begin{matrix} \dot{\mathbf{r}}_{\mathbf{Y}} \\ \boldsymbol{\omega}_{\mathbf{Y}} \end{matrix}\right|. \end{aligned} \tag{62}$$

The constitutive elastic stiffness is represented by the two-times covariant symmetric tensor

$$\mathfrak{E}(\mathbf{r},\mathbf{Q}) = \partial^2\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}})(\mathbf{r},\mathbf{Q}) = \begin{bmatrix} \mathfrak{E}_{11} & \mathfrak{E}_{12} \\ \mathfrak{E}_{21} & \mathfrak{E}_{22} \end{bmatrix}. \tag{63}$$

The bilinear form of the elastic tangent stiffness then assumes the expression

$$\begin{aligned} &\int_{\mathbb{B}_s} \partial^2\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}}(\mathbf{u}))\cdot(\partial_{\mathbf{Y}_{\mathbf{m}}}\mathfrak{D}_{\mathbf{x}})(\mathbf{u})\cdot(\partial_{\mathbf{X}_{\mathbf{m}}}\mathfrak{D}_{\mathbf{x}})(\mathbf{u})\,d\mu \\ &= \int_{\mathbb{B}_s} \mathfrak{E}(\mathbf{r},\mathbf{Q})\cdot\mathbb{Q}^T\boldsymbol{\Xi}^T\left|\begin{matrix} \dot{\mathbf{r}}_{\mathbf{Y}} \\ \boldsymbol{\omega}_{\mathbf{Y}} \end{matrix}\right|\cdot\mathbb{Q}^T\boldsymbol{\Xi}^T\left|\begin{matrix} \dot{\mathbf{r}}_{\mathbf{X}} \\ \boldsymbol{\omega}_{\mathbf{X}} \end{matrix}\right| d\mu\,. \end{aligned}$$

5.2 Geometric tangent stiffness

In an analogous way, for the geometric tangent stiffness we get

$$
\begin{aligned}
&(\nabla^2_{\mathbf{X_m Y_m}}\,\mathfrak{D}_{\mathbf{x}})(\mathbf{r},\mathbf{Q}) = \\
&= \frac{1}{2}\begin{vmatrix} \mathbf{Q}^T\left[(\mathbf{W_X}\,\mathbf{W_Y}+\mathbf{W_Y}\,\mathbf{W_X})\,\mathbf{r}'_t - 2\,(\mathbf{W_X}\,\dot{\mathbf{r}}'_{\mathbf{Y}}+\mathbf{W_Y}\,\dot{\mathbf{r}}'_{\mathbf{X}})\right] \\ \mathbf{Q}^T\,\text{axial}\,([\mathbf{W}'_{\mathbf{X}},\mathbf{W_Y}]+[\mathbf{W}'_{\mathbf{Y}},\mathbf{W_X}]) \end{vmatrix} = \\
&= \frac{1}{2}\begin{vmatrix} \mathbf{Q}^T\left[\boldsymbol{\omega}_{\mathbf{X}}\times(\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{r}'_t)+\boldsymbol{\omega}_{\mathbf{Y}}\times(\boldsymbol{\omega}_{\mathbf{X}}\times\mathbf{r}'_t)-\boldsymbol{\omega}_{\mathbf{X}}\times\dot{\mathbf{r}}'_{\mathbf{Y}}-\boldsymbol{\omega}_{\mathbf{Y}}\times\dot{\mathbf{r}}'_{\mathbf{X}}\right] \\ \mathbf{Q}^T\,(\boldsymbol{\omega}'_{\mathbf{X}}\times\boldsymbol{\omega}_{\mathbf{Y}}+\boldsymbol{\omega}'_{\mathbf{Y}}\times\boldsymbol{\omega}_{\mathbf{X}}) \end{vmatrix} = \\
&= -\frac{1}{2}\,\mathbb{Q}^T\begin{vmatrix} \boldsymbol{\omega}_{\mathbf{X}}\times(\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{r}'_t)+\boldsymbol{\omega}_{\mathbf{Y}}\times(\boldsymbol{\omega}_{\mathbf{X}}\times\mathbf{r}'_t)-\boldsymbol{\omega}_{\mathbf{X}}\times\dot{\mathbf{r}}'_{\mathbf{Y}}-\boldsymbol{\omega}_{\mathbf{Y}}\times\dot{\mathbf{r}}'_{\mathbf{X}} \\ \boldsymbol{\omega}'_{\mathbf{X}}\times\boldsymbol{\omega}_{\mathbf{Y}}+\boldsymbol{\omega}'_{\mathbf{Y}}\times\boldsymbol{\omega}_{\mathbf{X}} \end{vmatrix}.
\end{aligned}
\tag{64}
$$

The two-times covariant symmetric tensor

$$
\int_{\mathbb{B}_s} \partial\boldsymbol{\varphi}_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}}(\mathbf{u}))\cdot(\nabla^2_{\mathbf{X_m Y_m}}\,\mathfrak{D}_{\mathbf{x}})(\mathbf{u})\;d\boldsymbol{\mu}\,,
\tag{65}
$$

provides the geometric tangent stiffness. By setting

$$
\mathbf{F}=\mathbf{Q}\,\mathbf{F}_o\,,\qquad \mathbf{M}=\mathbf{Q}\,\text{axial}\,\mathbf{M}_o=\text{axial}\,(\mathbf{Q}\,\mathbf{M}_o\,\mathbf{Q}^T)\,,
\tag{66}
$$

it follows from the constitutive relation

$$
\mathfrak{S}_{\mathbf{x}}=\{\mathbf{F}_o\,,\mathbf{M}_o\}=\partial\boldsymbol{\varphi}_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}}(\mathbf{u}))\,,
\tag{67}
$$

that

$$
\{\mathbf{F}\,,\mathbf{M}_{\mathbf{Q}}\}=\mathbb{Q}\;\partial\boldsymbol{\varphi}_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}}(\mathbf{u}))\,\overline{\mathbb{Q}}^T\,,\qquad \{\mathbf{F}\,,\mathbf{M}\}=\mathbb{Q}\;\partial\boldsymbol{\varphi}_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}}(\mathbf{u}))\,,
\tag{68}
$$

where

$$
\mathbb{Q}:=\begin{bmatrix}\mathbf{Q} & \mathbf{O}\\ \mathbf{O} & \mathbf{Q}\end{bmatrix}\,,\qquad \overline{\mathbb{Q}}=\begin{bmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \mathbf{Q}\end{bmatrix}\,.
\tag{69}
$$

Observing that

$$
\begin{aligned}
\mathbf{F}\cdot(\mathbf{W_X}\,\mathbf{r}'_{\mathbf{Y}}) &= \mathbf{F}\cdot(\boldsymbol{\omega}_{\mathbf{X}}\times\mathbf{r}'_{\mathbf{Y}})=(\mathbf{F}\times\boldsymbol{\omega}_{\mathbf{X}})\cdot\mathbf{r}'_{\mathbf{Y}}\,,\\
\mathbf{F}\cdot(\mathbf{W_X}\,\mathbf{W_Y}\,\mathbf{r}'_t) &= (\mathbf{W_X}\,\mathbf{W_Y})\,:\,(\mathbf{F}\otimes\mathbf{r}'_t)=\mathbf{F}\cdot\left[\boldsymbol{\omega}_{\mathbf{X}}\times(\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{r}'_t)\right]=\\
&= (\mathbf{F}\cdot\boldsymbol{\omega}_{\mathbf{Y}})\,(\mathbf{r}'_t\cdot\boldsymbol{\omega}_{\mathbf{X}})-(\mathbf{F}\cdot\mathbf{r}'_t)\,(\boldsymbol{\omega}_{\mathbf{X}}\cdot\boldsymbol{\omega}_{\mathbf{Y}})=\\
&= \left[(\mathbf{F}\otimes\mathbf{r}'_t)\,\boldsymbol{\omega}_{\mathbf{X}}\right]\cdot\boldsymbol{\omega}_{\mathbf{Y}}-(\mathbf{F}\cdot\mathbf{r}'_t)\,(\boldsymbol{\omega}_{\mathbf{X}}\cdot\boldsymbol{\omega}_{\mathbf{Y}})\,,
\end{aligned}
\tag{70}
$$

we can write

$$\begin{aligned}
&\partial\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}})(\mathbf{r},\mathbf{Q})\cdot\nabla^2_{\mathbf{X_m Y_m}}\,\mathfrak{D}_{\mathbf{x}}(\mathbf{r},\mathbf{Q}) = \\
&= \mathbf{F}\cdot\frac{1}{2}\left[\boldsymbol{\omega}_{\mathbf{X}}\times(\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{r}'_t)+\boldsymbol{\omega}_{\mathbf{Y}}\times(\boldsymbol{\omega}_{\mathbf{X}}\times\mathbf{r}'_t)-2\,(\boldsymbol{\omega}_{\mathbf{X}}\times\dot{\mathbf{r}}'_{\mathbf{Y}}+\boldsymbol{\omega}_{\mathbf{Y}}\times\dot{\mathbf{r}}'_{\mathbf{X}})\right] \\
&+\mathbf{M}\cdot\frac{1}{2}\left[\boldsymbol{\omega}'_{\mathbf{X}}\times\boldsymbol{\omega}_{\mathbf{Y}}+\boldsymbol{\omega}'_{\mathbf{Y}}\times\boldsymbol{\omega}_{\mathbf{X}}\right].
\end{aligned} \tag{71}$$

Then

$$\begin{aligned}
&\partial\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}})(\mathbf{r},\mathbf{Q})\cdot\nabla^2{}_{\mathbf{X_m Y_m}}\,\mathfrak{D}_{\mathbf{x}}(\mathbf{r},\mathbf{Q}) = \\
&= \left[\mathrm{sym}\,(\mathbf{F}\otimes\mathbf{r}'_t)\,\boldsymbol{\omega}_{\mathbf{X}}\right]\cdot\boldsymbol{\omega}_{\mathbf{Y}}-(\mathbf{F}\cdot\mathbf{r}'_t)\,(\boldsymbol{\omega}_{\mathbf{X}}\cdot\boldsymbol{\omega}_{\mathbf{Y}}) \\
&-(\mathbf{F}\times\boldsymbol{\omega}_{\mathbf{X}})\cdot\mathbf{r}'_{\mathbf{Y}}+(\mathbf{F}\times\mathbf{r}'_{\mathbf{X}})\cdot\boldsymbol{\omega}_{\mathbf{Y}}+ \\
&+\tfrac{1}{2}\left[(\mathbf{M}\times\boldsymbol{\omega}'_{\mathbf{X}})\cdot\boldsymbol{\omega}_{\mathbf{Y}}-(\mathbf{M}\times\boldsymbol{\omega}_{\mathbf{X}})\cdot\boldsymbol{\omega}'_{\mathbf{Y}}\right].
\end{aligned} \tag{72}$$

To provide a matrix form of the first member of the incremental equilibrium condition, we introduce the linear differential operator

$$\mathfrak{L}=\begin{bmatrix}\frac{\partial}{\partial\lambda}\mathbf{I} & \mathbf{O} & \mathbf{O}\\ \frac{\partial}{\partial\lambda}\mathbf{I} & \mathbf{O} & \mathbf{I}\end{bmatrix}, \tag{73}$$

so that

$$\mathfrak{L}\begin{vmatrix}\dot{\mathbf{r}}\\ \boldsymbol{\omega}\end{vmatrix}=\begin{vmatrix}\dot{\mathbf{r}}'\\ \boldsymbol{\omega}'\\ \boldsymbol{\omega}\end{vmatrix}. \tag{74}$$

The geometric tangent stiffness is then expressed as

$$\int_{\mathbb{B}_s}\partial\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}})(\mathbf{r},\mathbf{Q})\cdot\nabla^2{}_{\mathbf{X_m Y_m}}\,\mathfrak{D}_{\mathbf{x}}(\mathbf{r},\mathbf{Q})\,d\mu \tag{75}$$

$$=\int_{\mathbb{B}_s}\mathfrak{B}\begin{vmatrix}\dot{\mathbf{r}}'_{\mathbf{X}}\\ \boldsymbol{\omega}'_{\mathbf{X}}\\ \boldsymbol{\omega}_{\mathbf{X}}\end{vmatrix}\cdot\begin{vmatrix}\dot{\mathbf{r}}'_{\mathbf{Y}}\\ \boldsymbol{\omega}'_{\mathbf{Y}}\\ \boldsymbol{\omega}_{\mathbf{Y}}\end{vmatrix}d\mu=\int_{\mathbb{B}_s}\mathfrak{B}\;\mathfrak{L}\begin{vmatrix}\dot{\mathbf{r}}_{\mathbf{X}}\\ \boldsymbol{\omega}_{\mathbf{X}}\end{vmatrix}\cdot\mathfrak{L}\begin{vmatrix}\dot{\mathbf{r}}_{\mathbf{Y}}\\ \boldsymbol{\omega}_{\mathbf{Y}}\end{vmatrix}d\mu, \tag{76}$$

where the symmetric operator of geometric stiffness $\mathfrak{B}$ is defined by

$$\mathfrak{B}=\begin{bmatrix}\mathbf{O} & \mathbf{O} & -\mathbf{F}\times\\ \mathbf{O} & \mathbf{O} & -\frac{1}{2}\mathbf{M}\times\\ \mathbf{F}\times & \frac{1}{2}\mathbf{M}\times & \frac{1}{2}(\mathbf{F}\otimes\mathbf{r}'_t+\mathbf{r}'_t\otimes\mathbf{F})-(\mathbf{F}\cdot\mathbf{r}'_t)\,\mathbf{I}\end{bmatrix}. \tag{77}$$

5.3 Constitutive tangent stiffness

The previous analysis provides the matrix expression of the constitutive tangent stiffness, defined by the symmetric bilinear form

$$
\begin{aligned}
&\int_{\mathbb{B}_s} \nabla^2_{\mathbf{XY}}(\varphi_{\mathbf{x}} \circ \mathfrak{D}_{\mathbf{x}})(\mathbf{r},\mathbf{Q})\; d\mu \\
&= \int_{\mathbb{B}_s} \partial\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}})(\mathbf{r},\mathbf{Q}) \cdot \nabla^2{}_{\mathbf{X_m Y_m}}\, \mathfrak{D}_{\mathbf{x}}(\mathbf{r},\mathbf{Q})\, d\mu \\
&+ \int_{\mathbb{B}_s} \partial^2\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}})(\mathbf{r},\mathbf{Q}) \cdot \partial\mathfrak{D}_{\mathbf{x}}(\mathbf{r},\mathbf{Q})[\mathbf{Y_m}] \cdot \partial\mathfrak{D}_{\mathbf{x}}(\mathbf{r},\mathbf{Q})[\mathbf{X_m}]\, d\mu \\
&= \int_{\mathbb{B}_s} \left[\mathfrak{B}\; \mathfrak{L} \begin{vmatrix} \dot{\mathbf{r}}_{\mathbf{X}} \\ \boldsymbol{\omega}_{\mathbf{X}} \end{vmatrix} \cdot \mathfrak{L} \begin{vmatrix} \dot{\mathbf{r}}_{\mathbf{Y}} \\ \boldsymbol{\omega}_{\mathbf{Y}} \end{vmatrix} d\mu + \mathfrak{E}(\mathbf{r},\mathbf{Q}) \cdot \mathbb{Q}^T \boldsymbol{\Xi}^T \begin{vmatrix} \dot{\mathbf{r}}_{\mathbf{X}} \\ \boldsymbol{\omega}_{\mathbf{X}} \end{vmatrix} \cdot \mathbb{Q}^T \boldsymbol{\Xi}^T \begin{vmatrix} \dot{\mathbf{r}}_{\mathbf{Y}} \\ \boldsymbol{\omega}_{\mathbf{Y}} \end{vmatrix} \right] d\mu\,,
\end{aligned}
$$

where

$$
\mathbb{Q} := \begin{bmatrix} \mathbf{Q} & \mathbf{O} \\ \mathbf{O} & \mathbf{Q} \end{bmatrix}, \tag{78}
$$

$$
\boldsymbol{\Xi}^T := \begin{bmatrix} \frac{\partial}{\partial\lambda}\mathbf{I} & \mathbf{r}'_t\times \\ \mathbf{O} & \frac{\partial}{\partial\lambda}\mathbf{I} \end{bmatrix}, \tag{79}
$$

$$
\mathfrak{L} = \begin{bmatrix} \frac{\partial}{\partial\lambda}\mathbf{I} & \mathbf{O} & \mathbf{O} \\ \frac{\partial}{\partial\lambda}\mathbf{I} & \mathbf{O} & \mathbf{I} \end{bmatrix}, \tag{80}
$$

$$
\mathfrak{B} = \begin{bmatrix} \mathbf{O} & \mathbf{O} & -\mathbf{F}\times \\ \mathbf{O} & \mathbf{O} & -\frac{1}{2}\mathbf{M}\times \\ \mathbf{F}\times & \frac{1}{2}\mathbf{M}\times & \frac{1}{2}(\mathbf{F}\otimes\mathbf{r}'_t + \mathbf{r}'_t\otimes\mathbf{F}) - (\mathbf{F}\cdot\mathbf{r}'_t)\,\mathbf{I} \end{bmatrix}. \tag{81}
$$

6 Conclusion

The constitutive tangent stiffness of a Timoshenko beam model is composed of an elastic and a geometric part. The elastic part is always a symmetric bilinear form as it is the second directional derivative of the deformation measure field whose values belong to a linear space. The geometric part is a bilinear form which turns out to be symmetric if the torsion of the connection vanishes. We have shown that the calculations performed in early treatments of the Timoshenko beam model [4,7], in which the geometric stiffness is computed by taking the second directional derivative of the deformation measure according to a canonical extension of the virtual displacement, are equivalent to evaluating the second covariant derivative of the deformation measure according to a non-symmetric connection on the non-linear rotation manifold. This result explains why the tangent stiffness so evaluated is tensorial but

non-symmetric. By adopting the natural symmetric Levi–Civita connection the tangent stiffness turns out to be the symmetric part of the non-symmetric one classically evaluated on the basis of the canonical extension of the virtual displacement.

Appendix

Let SO(3) be the special orthogonal group of proper rotations, with tangent bundle $\mathbb{T}_{\mathrm{SO(3)}} \subset BL\,(V^3\,;V^3)$ and let so(3) $\subset BL\,(V^3\,;V^3)$ be the linear subspace of skew-symmetric tensors. We consider the trivial fiber bundle $\boldsymbol{\pi}$: SO(3) $\times$ so(3) $\mapsto$ SO(3) and a section $\hat{\mathbf{W}}$: SO(3) $\mapsto$ SO(3) $\times$ so(3) of this bundle, defined by

$$\hat{\mathbf{W}}(\mathbf{R}) = \{\,\mathbf{R}, \mathbf{W}_{\mathbf{R}}\,\}\,, \qquad \mathbf{R} \in \mathrm{SO(3)}\,,\, \mathbf{W}_{\mathbf{R}} \in \mathrm{so(3)}\,. \tag{82}$$

The composition $\hat{\mathbf{T}} \circ \hat{\mathbf{W}}$ with the map $\hat{\mathbf{T}}$: SO(3) $\times$ so(3) $\mapsto \mathbb{T}_{\mathrm{SO(3)}}$ defined by

$$\hat{\mathbf{T}}(\{\,\mathbf{R}, \mathbf{W}_{\mathbf{R}}\,\}) = \{\,\mathbf{R}, \mathbf{W}_{\mathbf{R}}\,\mathbf{R}\,\}\,, \tag{83}$$

yields a vector field $\hat{\mathbf{T}} \circ \hat{\mathbf{W}}$: SO(3) $\mapsto \mathbb{T}_{\mathrm{SO(3)}}$ on the tangent bundle $\mathbb{T}_{\mathrm{SO(3)}}$, according to the relation

$$(\hat{\mathbf{T}} \circ \hat{\mathbf{W}})(\mathbf{R}) = \{\,\mathbf{R}, \mathbf{W}_{\mathbf{R}}\,\mathbf{R}\,\}\,. \tag{84}$$

Let $\hat{\mathbf{W}}_{\mathbf{X}}, \hat{\mathbf{W}}_{\mathbf{Y}}$: SO(3) $\mapsto$ SO(3) $\times$ so(3) be sections of the fiber bundle $\boldsymbol{\pi}$: SO(3) $\times$ so(3) $\mapsto$ SO(3). The corresponding vector fields $\hat{\mathbf{X}}, \hat{\mathbf{Y}}$: SO(3) $\mapsto \mathbb{T}_{\mathrm{SO(3)}}$ are defined by the compositions

$$\hat{\mathbf{X}} := \hat{\mathbf{T}} \circ \hat{\mathbf{W}}_{\mathbf{X}}\,, \quad \hat{\mathbf{Y}} := \hat{\mathbf{T}} \circ \hat{\mathbf{W}}_{\mathbf{Y}}\,. \tag{85}$$

The directional derivative of $\hat{\mathbf{X}}$ at $\mathbf{Q}$ along $\mathbf{Y} = \hat{\mathbf{Y}}(\mathbf{Q}) \in \mathbb{T}_{\mathrm{SO(3)}}(\mathbf{Q})$ is given by

$$(\partial_{\mathbf{Y}}\hat{\mathbf{X}})(\mathbf{Q}) = [\partial_{\mathbf{Y}}\hat{\mathbf{W}}_{\mathbf{X}}(\mathbf{Q})]\mathbf{Q} + \hat{\mathbf{W}}_{\mathbf{X}}(\mathbf{Q})\partial_{\mathbf{Y}}\mathbf{Q}\,. \tag{86}$$

To simplify the notation we set $\mathbf{W}_{\mathbf{X}} = \hat{\mathbf{W}}_{\mathbf{X}}(\mathbf{Q})$ and $\mathbf{W}_{\mathbf{Y}} = \hat{\mathbf{W}}_{\mathbf{Y}}(\mathbf{Q})$.

We remark that the directional derivative $\partial_{\mathbf{Y}}\mathbf{Q}$ is defined by considering the canonical injection $\hat{J}$: SO(3) $\mapsto BL\,(V^3\,;V^3)$:

$$\hat{J}(\mathbf{Q}) := \mathbf{Q} \in BL\,(V^3\,;V^3) \quad \forall\,\mathbf{Q} \in \mathrm{SO(3)}\,, \tag{87}$$

and by setting

$$\partial_{\mathbf{Y}}\mathbf{Q} := (\partial_{\mathbf{Y}}\hat{J})(\mathbf{Q}) = \mathbf{Y} = \mathbf{W}_{\mathbf{Y}}\,\mathbf{Q} \in BL\,(V^3\,;V^3)\,. \tag{88}$$

The directional derivative of the vector field $\hat{\mathbf{X}}(\mathbf{R})$ at the point $\mathbf{Q}$ along $\mathbf{Y} = \hat{\mathbf{Y}}(\mathbf{Q}) \in \mathbb{T}_{\mathrm{SO(3)}}(\mathbf{Q})$ is given by

$$(\partial_{\mathbf{Y}}\hat{\mathbf{X}})(\mathbf{Q}) = (\partial_{\mathbf{Y}}\hat{\mathbf{W}}_{\mathbf{X}})\mathbf{Q} + \mathbf{W}_{\mathbf{X}}\mathbf{W}_{\mathbf{Y}}\mathbf{Q}\,. \tag{89}$$

Let $\{\mathbf{Q}_s,\ s \in I\}$ be the parametric equation of a curve on SO(3) passing through $\mathbf{Q}$ at time $t \in I$ so that $\mathbf{Q}_t = \mathbf{Q}$.

A covariant differentiation on the manifold SO(3) is uniquely defined once a parallel transport is chosen along the curves on the manifold. The covariant derivative of the vector field $\hat{\mathbf{X}}$ along the tangent vector

$$\mathbf{Y} = \frac{\partial}{\partial s}\bigg|_{s=t} \mathbf{Q}_s \tag{90}$$

is defined by

$$(\nabla_{\mathbf{Y}}\hat{\mathbf{X}})(\mathbf{Q}_t) = \frac{\partial}{\partial s}\bigg|_{s=t} \mathbf{S}_{t,s}\hat{\mathbf{X}}(\mathbf{Q}_s)\,, \tag{91}$$

where $\mathbf{S}_{t,s}$ denotes the parallel transport along the curve $\{\mathbf{Q}_s,\ s \in I\}$ from the point $\mathbf{Q}_s$ to the point $\mathbf{Q}_t$. We now define the parallel transport according to the relation

$$\mathbf{S}_{t,s}\hat{\mathbf{X}}(\mathbf{Q}_s) := \hat{\mathbf{W}}_{\mathbf{X}}(\mathbf{Q}_s)\mathbf{Q}_t\,. \tag{92}$$

By the formula above the covariant derivative is then given by

$$(\nabla_{\mathbf{Y}}\hat{\mathbf{X}})(\mathbf{Q}_t) = (\partial_{\mathbf{Y}}\hat{\mathbf{W}}_{\mathbf{X}})(\mathbf{Q}_t)\,\mathbf{Q}_t\,, \tag{93}$$

so that

$$(\partial_{\mathbf{Y}}\hat{\mathbf{X}})(\mathbf{Q}_t) = (\nabla_{\mathbf{Y}}\hat{\mathbf{X}})(\mathbf{Q}_t) + \mathbf{W}_{\mathbf{X}}\mathbf{W}_{\mathbf{Y}}\mathbf{Q}_t\,. \tag{94}$$

The torsion of a connection on the manifold SO(3) is the third-order tensor field which provides the vector measure of the lack of symmetry of the Hessian of a scalar function $f \in \mathrm{C}^2(\mathbb{M};\mathcal{R})$:

$$\mathbf{TOR}(\mathbf{X},\mathbf{Y})\,f := (\nabla^2_{\mathbf{XY}} - \nabla^2_{\mathbf{YX}})\,f = (\nabla_{\mathbf{X}}\hat{\mathbf{Y}} - \nabla_{\mathbf{Y}}\hat{\mathbf{X}})\,f - [\hat{\mathbf{X}},\hat{\mathbf{Y}}]\,f\,, \tag{95}$$

where $[\hat{\mathbf{X}},\hat{\mathbf{Y}}]$ is the Lie bracket given by

$$[\hat{\mathbf{X}},\hat{\mathbf{Y}}]\,f = (\mathscr{L}_{\hat{\mathbf{X}}}\hat{\mathbf{Y}})\,f = (\partial_{\mathbf{X}}\partial_{\hat{\mathbf{Y}}} - \partial_{\mathbf{Y}}\partial_{\hat{\mathbf{X}}})\,f\,, \tag{96}$$

where the Lie derivative is defined by

$$(\mathscr{L}_{\hat{\mathbf{X}}}\hat{\mathbf{Y}})(\mathbf{Q}_t) := \frac{\partial}{\partial s}\bigg|_{s=t} \boldsymbol{\varphi}_{t,s_*}(\hat{\mathbf{X}}(\mathbf{Q}_s))\,, \tag{97}$$

and $\boldsymbol{\varphi}_{t,s_*}$ is the differential of the flow $\boldsymbol{\varphi}_{t,s}$ on the manifold SO(3) associated with the vector field $\hat{\mathbf{X}}$ via the differential equation

$$\frac{d}{dt}\bigg|_{t=s} \boldsymbol{\varphi}_{t,s} = \hat{\mathbf{X}}\,. \tag{98}$$

It is worth noting that the Lie derivative is a first-order derivative which is the difference between two second directional derivatives. Although both terms of the right-hand side in the expression of the torsion are not tensorial, the expression of the torsion as a whole is in fact tensorial in its arguments, i.e., it depends only on the point values of the vector fields.

We now compute the explicit expression of the torsion. The two covariant derivatives are given by

$$(\nabla_{\mathbf{X}}\hat{\mathbf{Y}})\,f = \big[(\partial_{\mathbf{X}}\hat{\mathbf{W}}_{\mathbf{Y}})\mathbf{Q}\big]\,f\,, \tag{99}$$

$$(\nabla_{\mathbf{Y}}\hat{\mathbf{X}})\,f = \big[(\partial_{\mathbf{Y}}\hat{\mathbf{W}}_{\mathbf{X}})\mathbf{Q}\big]\,f\,. \tag{100}$$

The evaluation of the second directional derivatives yields

$$\partial_{\mathbf{X}}\partial_{\hat{\mathbf{Y}}}f = \partial^2_{\mathbf{XY}}f + (\partial_{\mathbf{X}}\hat{\mathbf{Y}})f = \partial^2_{\mathbf{XY}}f + \big[(\partial_{\mathbf{X}}\hat{\mathbf{W}}_{\mathbf{Y}})\mathbf{Q} + \mathbf{W}_{\mathbf{Y}}\mathbf{W}_{\mathbf{X}}\,\mathbf{Q}\big]\,f, \tag{101}$$

$$\partial_{\mathbf{Y}}\partial_{\hat{\mathbf{X}}}f = \partial^2_{\mathbf{YX}}f + (\partial_{\mathbf{Y}}\hat{\mathbf{X}})f = \partial^2_{\mathbf{YX}}f + \big[(\partial_{\mathbf{Y}}\hat{\mathbf{W}}_{\mathbf{X}})\mathbf{Q} + \mathbf{W}_{\mathbf{X}}\mathbf{W}_{\mathbf{Y}}\,\mathbf{Q}\big]\,f\,. \tag{102}$$

In the linear ambient space $\mathrm{L}(V^3\,;V^3)$ we have $\partial^2_{\mathbf{XY}}f = \partial^2_{\mathbf{YX}}f$ and hence the final expression of the torsion is given by

$$\begin{aligned}\mathbf{TOR}(\mathbf{X},\mathbf{Y})\,f = &+\big[(\partial_{\mathbf{X}}\hat{\mathbf{W}}_{\mathbf{Y}})\mathbf{Q}\big]\,f - \big[(\partial_{\mathbf{Y}}\mathbf{W}_{\mathbf{X}})\mathbf{Q}\big]\,f\\ &- \partial^2_{\mathbf{XY}}f - \big[(\partial_{\mathbf{X}}\hat{\mathbf{W}}_{\mathbf{Y}})\mathbf{Q} + \mathbf{W}_{\mathbf{Y}}\mathbf{W}_{\mathbf{X}}\mathbf{Q}\big]\,f\\ &+ \partial^2_{\mathbf{YX}}f + \big[(\partial_{\mathbf{Y}}\hat{\mathbf{W}}_{\mathbf{X}})\mathbf{Q} + \mathbf{W}_{\mathbf{X}}\mathbf{W}_{\mathbf{Y}}\mathbf{Q}\big]\,f\\ &= \big[\mathbf{W}_{\mathbf{X}}\mathbf{W}_{\mathbf{Y}} - \mathbf{W}_{\mathbf{Y}}\mathbf{W}_{\mathbf{X}}\big]\mathbf{Q}\,f = \big[\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}\big]\mathbf{Q}\,f\,.\end{aligned} \tag{103}$$

We may conlude that the torsion tensor at a point $\mathbf{Q}\in\mathrm{SO(3)}$ is a non-vanishing bilinear form. Moreover, as predicted by the theory, its expression is independent of the special functional form of the vector fields $\hat{\mathbf{X}},\hat{\mathbf{Y}}:\mathrm{SO(3)}\mapsto\mathbb{T}_{\mathrm{SO(3)}}$ since, in its expression, only the point values $\mathbf{W}_{\mathbf{X}},\mathbf{W}_{\mathbf{Y}}$ at $\mathbf{Q}$ of the sections $\hat{\mathbf{W}}_{\mathbf{X}},\hat{\mathbf{W}}_{\mathbf{Y}}:\mathrm{SO(3)}\mapsto\mathrm{SO(3)}\times\mathrm{so(3)}$ appear.

The canonical extension of a vector tangent at a point of $\mathrm{SO(3)}$ is obtained by setting

$$\hat{\mathbf{W}}(\mathbf{R}) = \{\,\mathbf{R},\mathbf{W}\,\}\,,\qquad \mathbf{R}\in\mathrm{SO(3)}\,, \tag{104}$$

with $\mathbf{W}\in\mathrm{so(3)}$ a fixed skew-symmetric tensor. In this case the covariant derivative vanishes since

$$(\nabla_{\mathbf{Y}}\hat{\mathbf{X}})(\mathbf{Q}) = (\partial_{\mathbf{Y}}\hat{\mathbf{W}}_{\mathbf{X}})(\mathbf{Q})\,\mathbf{Q} = 0\qquad \forall\,\mathbf{Y}=\hat{\mathbf{Y}}(\mathbf{Q})\in\mathbb{T}_{\mathrm{SO(3)}}(\mathbf{Q})\,. \tag{105}$$

Hence, from the formula

$$(\nabla^2_{\mathbf{X}_{\mathrm{m}}\mathbf{Y}_{\mathrm{m}}}\mathfrak{D}_{\mathbf{x}})(\mathbf{u}) = \partial_{\mathbf{X}_{\mathrm{m}}}(\partial_{\hat{\mathbf{Y}}}\mathfrak{D}_{\mathbf{x}})(\mathbf{u}) - \partial_{\left(\nabla_{\mathbf{X}_{\mathrm{m}}}\hat{\mathbf{Y}}\right)}\mathfrak{D}_{\mathbf{x}}(\mathbf{u})\,, \tag{106}$$

we infer that

$$(\nabla^2_{\mathbf{X_m Y_m}}\, \mathfrak{D}_{\mathbf{x}})(\mathbf{u}) = \partial_{\mathbf{X_m}}(\partial_{\hat{\mathbf{Y}}}\, \mathfrak{D}_{\mathbf{x}})(\mathbf{u})\,, \tag{107}$$

so that the second covariant derivative of the deformation measure coincides with the second directional derivative.

Acknowledgements

The financial support of the Italian Ministry for University and Scientific Research (MIUR) is gratefully acknowledged.

References

[1] Spivak, M. (1979): A comprehensive introduction to differential geometry. Vols. I-V. Publish or Perish, Wilmington, DE
[2] Marsden, J. E., Hughes, T.J.R. (1983): Mathematical foundations of elasticity. Prentice-Hall, Englewood Cliffs, NJ
[3] Simo, J.C.: A finite strain beam formulation. The three-dimensional-dynamic problem. I. Comput. Meth. Appl. Mech. Engrg. **49**, 55–70
[4] Simo, J.C., Vu-Quoc, L. (1986): A three-dimensional finite-strain rod model. II. Computational aspects. Comput. Meth. Appl. Mech. Engrg. **58**, 79–116
[5] Simo, J.C., Vu-Quoc, L. (1988): On the dynamics in space of rods undergoing large motions–a geometrically exact approach. Comput. Meth. Appl. Mech. Engrg. **66**, 125–161
[6] Abraham, R., Marsden, J.E., Ratiu, T. (1988): Manifolds, tensor analysis, and applications. 2nd edition. Springer, New York
[7] Simo, J.C. (1992): The (symmetric) Hessian for geometrically nonlinear models in solid mechanics: intrinsic definition and geometric interpretation. Comput. Meth. Appl. Mech. Engrg. **96**, 189–200
[8] Petersen, P. (1998): Riemannian geometry. Springer, New York
[9] Romano, G. (2002–2003): Scienza delle costruzioni. Tomi I-II. Hevelius, Benevento
[10] Romano, G., Diaco, M., Romano, A., Sellitto, C. (2003): When and why a nonsymmetric tangent stiffness may occur. 16th AIMETA Congress of Theoretical and Applied Mechanics. Ferrara, Italy, Sept. 9–12, 2003
[11] Romano, G., Diaco, M., Sellitto, C. (2004): Tangent stiffness of elastic continua on manifolds. In: Romano, G., Rionero, S. (eds.): Recent trends in the applications of mathematics to mechanics. Springer, Berlin, pp. 155–184

Qualitative estimates for cross-sectional measures in elasticity

J.N. Flavin, B. Gleeson

1 Introduction

This paper summarizes recent spatial decay results in linear homogeneous isotropic elasticity. They are derived by studying positive-definite cross-sectional measures of deformation/stress, using second-order differential inequality, or convexity, techniques.

Two contexts are considered:

(i) a hollow circular cylinder, in a state of axisymmetric (torsionless) stress, whose lateral boundaries are traction-free, the resultant stress on each cross-section being zero;
(ii) a semi-infinite rectangular strip in a state of plane stress, whose lateral edges are displacement free.

Section 2 discusses the former context, Sect. 3 the latter.

2 Estimates for an annular cylinder in an axisymmetric state of stress

Knowles and Horgan [1] established a spatial decay result (reflecting Saint-Venant's principle) for an energy-like functional in the context of a 'solid' circular cylinder in a state of axisymmetric stress. A central feature of this analysis was the use of an ingeniously defined 'scalar product' in order to facilitate the analysis. The estimates, discussed here, are based on a positive-definite cross-sectional measure of stress and second-order differential inequalities – as opposed to an energy-like functional together with a differential-integral inequality as used in [1] – and involve the use of a 'scalar product' analogous to that used in [1]. It should be noted that Horgan and Knowles [2] remarked that the methodology used by them in connection with the solid cylinder, could also be used for the analogous issue for the hollow cylinder.

It is relevant to mention that the methodology used in this paper is analogous to that used in [3], and that reviews of Saint-Venant's principle and related issues, together with copious references, are available in [2,4,5].

We first deal with notation and other preliminaries, including the representation of the stress field in terms of potential-like functions. We then define a positive-definite cross-sectional measure – based on a suitably defined 'scalar product' – and establish its convexity (with respect to the axial coordinate z). Next, an annular

cylinder $(0 < z < L)$ is considered subject to zero traction on its boundary except on the end $z = 0$ where the load is (necessarily) self-equilibrated, and the cross-sectional measure is proved to be a generalized convex function of z, using, inter alia, a conservation law (discussed in the Appendix). A spatial decay law for the cross-sectional measure appropriate to a semi-infinite cylinder $(L \to \infty)$ follows therefrom. The estimated decay constant depends on two eigenvalues which are discussed in the Appendix. The section concludes with a discussion of the estimated decay constant. A complete discussion of these issues appears in [6], together with further references.

We consider an annular circular cylinder of internal radius a and external radius b, consisting of homogeneous isotropic elastic material. We use cylindrical polar coordinates (r, θ, z) throughout. We are concerned with axisymmetric, torsionless, deformations of the cylinder, and employ the following notation: the non-vanishing components of the stress tensor $\boldsymbol{\tau}$ are denoted by τ_{rr}, $\tau_{\theta\theta}$, τ_{zz}, τ_{rz}. Moreover, all these are functions of r, z only. We assume that the displacement field $\mathbf{u}$ is three times continuously differentiable and that $\boldsymbol{\tau}$ is twice continuously differentiable in the relevant closed region. We also assume that the shear modulus μ and Poisson's ratio σ satisfy $\mu > 0$ and $-1 < \sigma < 1/2$, and that both are constant.

We are concerned with (torsionless) axisymmetric stress fields corresponding to free lateral boundaries and such that the resultant traction on each cross-section is zero (self-equilibration condition). These are expressed respectively as follows:

$$\tau_{rr} = \tau_{rz} = 0 \text{ on } r = a, b, \tag{1}$$

and

$$\int_a^b \tau_{zz} r dr = 0. \tag{2}$$

It may be noted that the latter is, of course, satisfied (for all z) provided that it holds on any particular cross-section (e.g., on $z = 0$).

In the early part of the section specific limits on z are not imposed, but later on we assume that the cylinder occupies

$$0 < z < L,$$

that (additionally) the end $z = L$ (constant) is traction-free, and therefore that (2) holds on $z = 0$ and hence on all cross-sections. The traction-free condition on $z = L$ is expressible as

$$\tau_{zz} = \tau_{zr} = 0 \text{ on } z = L. \tag{3}$$

Subsequently we will assume that $L \to \infty$, and that (3) continues to hold.

It is convenient to introduce the functions $\phi(r, z)$, $\psi(r, z)$ (e.g., Knowles and Horgan [1]), the former being analogous to the Airy stress function in plane elasticity. As pointed out by Knowles and Horgan [1], a virtual retracing of the argument given in Love [7] establishes that any stress field of the type envisaged, with the smoothness properties stated in the first paragraph of this section, exists if and only if there exist

functions $\phi(r,z)$, $\psi(r,z)$ which are four times continuously differentiable in the relevant closed region, such that

$$\tau_{rr} = \phi_{zz} + r^{-1}\phi_r - r^{-2}\psi_z, \tag{4}$$
$$\tau_{\theta\theta} = \sigma(\phi_{zz} + \phi_{rr}) - (1-\sigma) r^{-1}\phi_r + r^{-2}\psi_z, \tag{5}$$
$$\tau_{zz} = r^{-1}(r\phi_r)_r, \tag{6}$$
$$\tau_{rz} = -\phi_{rz}, \tag{7}$$

and which satisfy the differential equations

$$(1-\sigma)\left\{r^{-1}(r\phi_r)_r + \phi_{zz}\right\} = r^{-1}\psi_{rz}, \tag{8}$$
$$r\left(r^{-1}\psi_r\right)_r + \psi_{zz} = 0, \tag{9}$$

where suffixes attached to ϕ, ψ denote partial differentiation with respect to the appropriate variables, both here and subsequently. There is a certain arbitrariness inherent in the representation $(4)-(7)$; see [6]. The boundary conditions (1), (2) may be expressed in terms of ϕ, ψ as

$$\left.\begin{aligned} r^2\phi_{zz} + r\phi_r - \psi_z &= 0, \\ \phi_{rz} &= 0, \end{aligned}\right\} \text{ on } r = a, b. \tag{10}$$

Furthermore, one may express the self-equilibration condition (2) in the form

$$r\phi_r]_a^b = 0, \tag{11}$$

while the traction-free condition on $z = L$ – when relevant – may be expressed in the form

$$\left.\begin{aligned} r^{-1}(r\phi_r)_r &= 0, \\ \phi_{rz} &= 0, \end{aligned}\right\} \text{ on } z = L. \tag{12}$$

Throughout we employ a cross-sectional measure of stress, suggested both by [3] and the procedure adopted by Knowles and Horgan [1] wherein they define a scalar product appropriate to a solid cylinder. We define the cross-sectional measure of stress by

$$F(z) = (\phi_{zz}, \phi_{zz}) + (1-\sigma)\int_a^b r^{-1}\left\{(r\phi_r)_r\right\}^2 dr, \tag{13}$$

where the scalar product (u, v) is defined, for any two continuous functions $u(r,\cdot)$, $v(r,\cdot)$ defined in $a \leqslant r \leqslant b$, as follows:

$$(u,v) = (1-\sigma)\int_a^b ru(r,\cdot)\,v(r,\cdot)\,dr - r^2 u(r,\cdot)\,v(r,\cdot)\Big]_a^b. \tag{14}$$

[The cross-sectional measure may be expressed in terms of stress components using $(4)-(6)$]. In all cases where it arises, the class of functions upon which the scalar product is defined is constrained by either of the conditions

(a) $\quad u = 0$ on $r = a, b$,

$(b) \quad (u, 1) = 0.$
It can be shown that

$$(u, u) \geqslant 0 \text{ with equality if and only if } u \equiv 0 \tag{15}$$

in both these cases.

Since integration of (8), multiplied by r, with respect to r, together with (10_1) and (11), gives

$$(\phi_{zz}, 1) = 0, \tag{16}$$

the positive-definiteness of (ϕ_{zz}, ϕ_{zz}) is established.

Using the foregoing properties together with the arbitrariness inherent in the representation $(4) - (7)$, we may easily establish (see [6]) that $F(z)$, defined by (13), is positive-definite in the stress field, i.e., for each z, $F(z)$ is positive except when the stress field is identically zero. Thus $F(z)$ may be regarded as an acceptable global measure, in each cross-section, of the stress.

The foregoing notation and properties may be used (see [6]) to establish a convexity property of $F(z)$.

Theorem 1. *The cross-sectional measure of stress $F(z)$, defined by (13), in the context of an axisymmetric, torsionless stress field, corresponding to traction-free lateral boundaries and zero resultant traction on each cross-section, in the cylindrical region*

$$a < r < b, \, 0 \leqslant \theta < 2\pi,$$

is a convex function of the axial coordinate.

The proof of Theorem 1 is based upon the readily established equality

$$F''(z) = 2(\phi_{zzz}, \phi_{zzz}) + 4(\phi_{rzz}, \phi_{rzz}) + 2(1 - \sigma) \int_a^b r^{-1} \left\{ (r\phi_{rz})_r \right\}^2 dr. \tag{17}$$

Theorem 1 follows from this expression together with the non-negativity property of the scalar product.

Some implications of Theorem 1 are discussed in [6].

We now give a *generalised convexity* property, for a more restricted problem, and study its implications. We consider a cylinder with zero traction, not only on the lateral surfaces $r = a, b$, but also on the end $z = L$, all other conditions being as previously; naturally, the load applied on the end $z = 0$ is self-equilibrated. It is possible to replace the boundary conditions (10), (12) by simplified ones, on exploiting the arbitrariness inherent in ϕ, ψ:

$$r^2\phi_z - \psi = 0, \, \phi_r = 0, \text{ on } r = a, b, \tag{18}$$

and, on the unloaded end,

$$\phi = 0, \, \phi_z = 0, \text{ on } z = L. \tag{19}$$

Commencing from (17), and using Inequality 1 (see Appendix $(A1)$), etc., we obtain

$$F''(z) \geqslant 2(\phi_{zzz}, \phi_{zzz}) + 4\lambda_1(\phi_{zz}, \phi_{zz}) + 2(1-\sigma)\int_a^b r^{-1}\left\{(r\phi_{rz})_r\right\}^2 dr, \quad (20)$$

where λ_1 is an eigenvalue defined in Appendix $(A1)$.

Suitably combining this, (19), and the conservation law (see Appendix $(A4)$)

$$2(\phi_z, \phi_{zzz}) + (1-\sigma)\int_a^b r^{-1}\left\{(r\phi_r)_r\right\}^2 dr - (\phi_{zz}, \phi_{zz}) - 2(\phi_{rz}, \phi_{rz}) = E, \quad (21)$$

where E is a constant, we obtain

$$F''(z) - K\lambda_1 F(z) \geqslant 2(2-K)\lambda_1(\phi_{zz}, \phi_{zz}) + 2(1-\sigma)\int_a^b r^{-1}\left\{(r\phi_{rz})_r\right\}^2 dr - \tfrac{1}{2}K(K+4)\lambda_1(\phi_{rz}, \phi_{rz}). \quad (22)$$

On applying Inequality 2 (Appendix $(A2)$), etc., we obtain

$$F''(z) - K\lambda_1 F(z) \geqslant 2(2-K)\lambda_1(\phi_{zz}, \phi_{zz}) + \tfrac{1}{2}\left\{4\lambda_2 - K(K+4)\lambda_1\right\}(\phi_{rz}, \phi_{rz}). \quad (23)$$

We now choose the value of K so that the right-hand side of (23) is non-negative. In order to make the last term in (23) non-negative we choose

$$0 < K \leqslant 2\sqrt{\lambda_2/\lambda_1 + 1} - 2,$$

where the definitions of λ_1, λ_2 are given in the Appendix $(A1)$, $(A2)$, and we require

$$0 < K \leqslant 2$$

in order to secure the non-negativity of the first term on the right-hand side of (23). We thus have the following result.

Theorem 2. *The cross-sectional measure $F(z)$ (defined by (13)) of the axisymmetric stress field in an annular elastic cylinder $(0 < z < L)$ with null traction boundary conditions on its lateral surfaces $r = a, b$, and on its plane end $z = L$, satisfies the* (*generalised convexity*) *inequality*

$$F''(z) - k^2 F(z) \geqslant 0, \quad (24)$$

the (*positive*) *constant k being defined by*

$$k = \sqrt{K\lambda_1} \quad (25)$$

where K is given by

$$K = 2\min\left(1, \sqrt{\lambda_2/\lambda_1 + 1} - 1\right), \quad (26)$$

and the eigenvalues λ_1, λ_2 are defined in the Appendix.

One may deduce from (24) that, in the case of a semi-infinite cylinder ($L \to \infty$), one has the following result.

Theorem 3. *In the context of the cylinder in Theorem 2, whose length $L \to \infty$, we have*

$$F(z) \leqslant F(0) \exp[-kz] \tag{27}$$

where k is defined by (25), (26), *provided that*

$$\lim_{L\to\infty} F(L) \exp[-kL] = 0. \tag{28}$$

Remark 1. The bound (27) can be made fully explicit in either of the following two ways.
(a) Consider the case of normal loading on the end $z = 0$ (i.e., $\tau_{rz} = 0$ thereon), and assume $\phi_{zz} \to 0$ as $L \to \infty$. One may use the conservation law (A3) to obtain

$$F(0) = 2(1-\sigma) \int_a^b r \left\{\tau_{zz}(r,0)\right\}^2 dr.$$

(b) Suppose that τ_{zz} and the complementary displacement component u are both prescribed, as smooth functions, on the end $z = 0$. In these circumstances, one may express $F(0)$ in terms of τ_{zz} and u (the radial component of displacement).

Remark 2. With

$$b/a = 1 + \varepsilon,$$

an asymptotic analysis establishes that

$$\lambda_1/\lambda_2 \sim 1 \text{ as } \varepsilon \to 0;$$

it is found, accordingly, that the decay constant k is such that

$$ka \sim \pi\varepsilon^{-1} \left(2\sqrt{2} - 2\right)^{1/2} \text{ as } \varepsilon \to 0.$$

This is consistent with the decay constant of Knowles [8] for a plane elastic state, derived by a different methodology.

Remark 3. It is of interest to compare the estimated decay rate obtained here with that obtained by Stephen and Wang [9] who estimated the decay rate, in the case $\sigma = 0.25$, by means of an eigenfunction analysis and numerical techniques. A comparison is given in the accompanying table, in which the estimated decay rate ka (DE) is given

in the second column, and the constant ka obtained by Stephen and Wang (WS) is given in the third column.

b/a	ka**(DE)**	ka**(WS)**
1.05	52.1	166.5
1.1	26.1	84.2
1.5	5.3	16.8
2	2.7	8.4

3 Estimates for a semi-infinite rectangular strip in plane strain

In this section we consider an homogeneous isotropic linear elastic material in plane strain occupying the rectangular region $0 < x_2 < 1$, $0 < x_1 < \infty$, $((x_1, x_2)$ denoting rectangular cartesian coordinates). We suppose that the lateral boundaries $x_2 = 0, 1$ are displacement-free and that the deformation is generated by actions on the remaining ends. Specifically, we assume that the x_1, x_2 displacement components $u_1(x_1, x_2)$, $u_2(x_1, x_2) \in C^3$ satisfy

$$\left.\begin{array}{l} (\alpha+1)\, u_{1,11} + u_{1,22} + \alpha u_{2,12} = 0 \\ u_{2,11} + (\alpha+1)\, u_{2,22} + \alpha u_{1,12} = 0 \end{array}\right\} \tag{29}$$

subject to

$$u_1 = u_2 = 0 \text{ on } x_2 = 0, 1. \tag{30}$$

As usual, α is a constant such that

$$\alpha = (1 - 2\sigma)^{-1}, \tag{31}$$

where σ is Poisson's ratio, assumed to satisfy

$$-1 < \sigma < 1/2, \tag{32}$$

whence

$$1/3 < \alpha < \infty. \tag{33}$$

In previous work [10], a decay estimate was obtained for a cross-sectional measure, whence a pointwise estimate was obtained for the displacement component u_1. The main purpose of this section is to obtain an estimate for a different/modified cross-sectional measure, which yields a pointwise decay estimate for the complementary displacement component u_2. As pointed out in [2], such pointwise estimates are liable to be of considerable technical complexity.

Estimates of this, and related types, are obtained in [11]. The main result therein is now summarized. Defining the cross-sectional measure of deformation $K(x_1)$ as

$$K(x_1) = \int_0^1 \left[u_{1,2}^2 + u_{2,1}^2 + \alpha^{-1}\left(u_{1,1}^2 + u_{2,2}^2\right)\right] dx_2, \tag{34}$$

one finds

$$K''(x_1) - 2\pi^2 (\alpha+1)^{-1} K(x_1) = 0. \tag{35}$$

Under the asymptotic conditions

$$u_{\beta,\gamma} \to 0 \text{ as } x_1 \to \infty, \tag{36}$$

it follows from (35) that

$$K(x_1) \leqslant K(0) \exp\left[-\pi\sqrt{2}(\alpha+1)^{-1/2} x_1\right]. \tag{37}$$

It is possible to bound $K(0)$ above in terms of data, using a conservation law, provided that the displacement components are specified as smooth functions on the edge $x_1 = 0$:

$$K(0) \leqslant \int_0^1 \left[\left(\alpha + 1 + \alpha^{-1}\right) u_{2,2}^2 + 2u_{1,2}^2\right]_{x_1=0} dx_2. \tag{38}$$

A pointwise decay estimate for $u_2(x_1, x_2)$ – in terms of data – follows from (37), (38) and

$$\left|u_2(x_1, x_2)\right| \leqslant \sqrt{x_2(1-x_2)\alpha K(x_1)}, \tag{39}$$

essentially a consequence of Schwarz's inequality and (34).

Theorem 4. *In the context of a semi-infinite region, $0 < x_1 < \infty, 0 < x_2 < 1$, consisting of linear isotropic, homogeneous elastic material in plane strain. for which the boundary and asymptotic conditions (30), (36) hold, the displacement components u_1, u_2 being specified as smooth functions on the edge $x_1 = 0$, an explicit, pointwise, decay estimate for the displacement component u_2 is available from (37) − (39), if we assume that the elastic constant $\alpha > 0$.*

Appendix

The scalar product $(\ ,\ )$ used in this Appendix is that defined in (14).

A1. Inequality 1

For arbitrary $\chi(r) \in C^2(a, b)$ such that $(\chi, 1) = 0$,

$$(1-\sigma)\int_a^b r\chi_r^2 dr \geqslant \lambda_1(\chi, \chi), \tag{A1}$$

where λ_1 is the lowest (positive) eigenvalue of

$$r^2\chi_{rr} + r\chi_r + \lambda r^2\chi = 0$$

subject to

$$(1-\sigma)\,\chi_r + \lambda r\chi = 0 \text{ on } r = a, b.$$

The eigenvalue λ_1 is given by

$$\lambda_1 = s^2 a^{-2},$$

where s is the lowest positive root of

$$\begin{aligned}&[s\,\mathrm{J}_0(s) - (1-\sigma)\,\mathrm{J}_1(s)]\,[sv\,\mathrm{Y}_0(sv) - (1-\sigma)\,\mathrm{Y}_1(sv)]\\ &-[s\,\mathrm{Y}_0(s) - (1-\sigma)\,\mathrm{Y}_1(s)]\,[sv\,\mathrm{J}_0(sv) - (1-\sigma)\,\mathrm{J}_1(sv)] = 0,\end{aligned}$$

where $v = b/a$, and where J_n, Y_n denote Bessel functions of order n, of the first and second kind respectively.

A2. Inequality 2

For arbitrary $\chi(r) \in C^2(a, b)$ such that $\chi(a) = \chi(b) = 0$,

$$\int_a^b r^{-1}\chi_r^2 dr \geqslant \lambda_2 \int_a^b r^{-1}\chi^2 dr, \tag{A2}$$

where λ_2 is the lowest (positive) eigenvalue of

$$r^2\chi_{rr} + r\chi_r + \left(\lambda r^2 - 1\right)\chi = 0$$

subject to

$$\chi = 0 \text{ on } r = a, b.$$

The eigenvalue λ_2 is given by

$$\lambda_2 = t^2 a^{-2},$$

where t is the lowest positive root of

$$\mathrm{J}_1(t)\,\mathrm{Y}_1(tv) - \mathrm{Y}_1(t)\,\mathrm{J}_1(tv) = 0,$$

where the same notation is used as in the previous case.

A3. Conservation law

$$\begin{aligned}2(\phi_z, \phi_{zzz}) + (1-\sigma)\int_a^b r^{-1}\left(\left\{r\phi_r\right\}_r\right)^2 dr - (\phi_{zz}, \phi_{zz}) - 2(\phi_{rz}, \phi_{rz})\\ = E(\text{constant}).\end{aligned} \tag{A3}$$

The proof of this may be found in [6].

References

[1] Knowles, J.K., Horgan, C.O. (1969): On the exponential decay of stresses in circular elastic cylinders subject to axisymmetric self-equilibrated end loads. Internat. J. Solids Structures **5**, 33–50
[2] Horgan, C.O., Knowles, J.K. (1983): Recent developments concerning Saint-Venant's principle. Adv. Appl. Mech. **23**, 179–269
[3] Flavin, J.N., Knops, R.J. (1988): Some convexity considerations for a two-dimensional traction problem. Z. Angew. Math. Phys. **39**, 166–176
[4] Horgan, C.O. (1989): Recent developments concerning Saint-Venant's principle: an update. Appl. Mech. Rev. **42**, 295–303
[5] Horgan, C.O. (1996): Recent developments concerning Saint-Venant's principle: a second update. Appl. Mech. Rev. **49**, S101–S111
[6] Flavin, J.N., Gleeson, B. (2004): Decay and other estimates for an annular elastic cylinder in an axisymmetric state of stress. Math. Mech. Solids, to appear
[7] Love, A.E.H. (1944): A treatise on the mathematical theory of elasticity. 4th edition. Dover, New York
[8] Knowles, J.K. (1966): On Saint-Venant's principle in the two-dimensional linear theory of elasticity. Arch. Ration. Mech. Anal. **21**, 1–22
[9] Stephen, N.G., Wang, M.Z. (1992): Decay rates for the hollow circular cylinder. J. Appl. Mech. **59**, 747–753
[10] Flavin, J.N., Rionero, S. (1993): Decay and other estimates for an elastic cylinder. Quart. J. Mech. Appl. Math. **46**, 299–309
[11] Flavin, J.N., Gleeson, B. (2001): Pointwise and other decay estimates for an isotropic elastic strip. J. Elasticity **64**, 191–197

On nonlinear global stability of Jeffery-Hamel flows

M. Gentile, S. Rionero

Abstract. By using weighted energy methods, we prove a condition assuring nonlinear global stability for a large class of flows in a wedge.

1 Introduction

We consider an incompressible viscous fluid between two inclined impermeable plane walls with a line source, or sink, at the intersection of the walls. By introducing a system of cylindrical coordinates (x, r, θ), where x is the intersection of the walls, we denote by Ω the divergent channel defined by $\Omega = \left\{(x, r, \theta) : |\theta| \leq \theta_0\right\}$. Then the fluid motion is governed by the Navier-Stokes equations

$$\begin{aligned} &\boldsymbol{v}_t + \boldsymbol{v} \cdot \nabla \boldsymbol{v} = -\frac{1}{\rho}\nabla p + \nu \Delta \boldsymbol{v}, \\ &\nabla \cdot \boldsymbol{v} = 0, \end{aligned} \tag{1}$$

where ν is the kinematic viscosity, with the initial-boundary conditions

$$\boldsymbol{v}(\boldsymbol{x}, t_0) = \boldsymbol{v}_0(\boldsymbol{x}) \, \forall \boldsymbol{x} \in \Omega, \tag{2}$$

$$\boldsymbol{v}(x, r, \pm\theta_0, t) = 0 \, \forall t \geq 0. \tag{3}$$

It can be proved that Eqs. (1-3) admit the family of solutions

$$v_x = v_\theta = 0, \quad v_r = v(r, \theta) = \frac{F(\theta)}{r}, \tag{4}$$

$$\frac{p}{\rho} = \frac{1}{r^2}[2\nu F(\theta) - C_1] + \text{const}, \tag{5}$$

where $F(\theta)$ is the generic solution of the ordinary differential equation

$$2FF' + \nu F''' + 4\nu F' = 0, \tag{6}$$

and C_1 is a constant chosen to verify the boundary conditions $F(\pm\theta_0) = 0$. As is well-known, such motions are called Jeffery-Hamel flows. These motions were discovered, independently, by Jeffery (1915) and Hamel (1916). They were studied by many authors and completely classified in terms of elliptic functions [5,6,9]. Jeffery-Hamel flows are important not only from a mathematical point of view, but also in applications. For instance, they may be used to approximate locally the steady flow in a two-dimensional channel with walls of small curvature [5,6].

Recently, many authors have studied these motions. Much attention was devoted to stability with respect to linear or weakly nonlinear two-dimensional perturbations, and many results were established [1,3,7]. Nevertheless, it seems hard to determine stability conditions with respect to nonlinear three-dimensional perturbations. The aim of the present paper is to prove, by using weighted energy techniques, a nonlinear stability result for a general class of three-dimensional perturbations.

The scheme of the paper is as follows. In Sect. 2, the perturbation equations for a basic JH flow and a weighted energy equality are obtained. In Sect. 3, two embedding inequalities in a wedge are given. Then, in Sect. 4, we obtain a priori estimates. Lastly, in Sect. 5, a theorem assuring nonlinear global stability is proved.

2 The balance energy equation

Let $(\boldsymbol{u}, \pi)$ be a regular perturbation to the basic Jeffery-Hamel flow $(\boldsymbol{v}, p)$. From (1-3) it follows that

$$\begin{aligned} &\boldsymbol{u}_t + (\boldsymbol{v} + \boldsymbol{u}) \cdot \nabla \boldsymbol{u} + \boldsymbol{u} \cdot \nabla \boldsymbol{v} = -\nabla \pi + \nu \Delta \boldsymbol{u}, \\ &\nabla \cdot \boldsymbol{u} = 0, \end{aligned} \tag{7}$$

$$\boldsymbol{u}(\boldsymbol{x}, t_0) = \boldsymbol{u}_0(\boldsymbol{x}) \ \forall \boldsymbol{x} \in \Omega, \tag{8}$$

$$\boldsymbol{u}(x, r, \pm\theta_0, t) = 0 \ \forall t \geq 0. \tag{9}$$

We remark that a basic JH flow $(\mathbf{v}, p)$ belongs neither to $L^1(\Omega)$ nor to $L^2(\Omega)$. Therefore, it is natural to consider perturbations $(\mathbf{u}, \pi)$ having the same behaviour as the basic flow. In order to deal with the problem we introduce the one-parameter family of weight functions

$$\phi(x, r) = \exp\left[-\alpha(|x| + r)\right], \quad \alpha \in (0, 1). \tag{10}$$

This allows us to control the set $\wp$ of perturbations $(\mathbf{u}, \pi)$ such that:

1. $|\boldsymbol{u}| + |\pi| \leq C$ in Ω;
2. $\nabla \cdot \boldsymbol{u} = 0$;
3. $\boldsymbol{u}|_{\partial\Omega} = 0$;
4. $|\nabla \boldsymbol{u}| \leq C|x|^h r^k$ in Ω $(h \geq 0,\ k > 0)$;
5. $\pi \in L^2\left(\Omega \times [0, +\infty)\right)$.

Theorem 1. *Let $(\boldsymbol{u}, \pi) \in \mathcal{P}$. Then, for all $t \geq t_0$, we have*

$$\begin{aligned} \int_\Omega \phi u^2 d\Omega = \int_\Omega \phi u_0^2 d\Omega - 2\int_{t_0}^t \int_\Omega \phi \left[\nu|\nabla \boldsymbol{u}|^2 + \boldsymbol{u} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{u}\right] d\Omega d\tau \\ + \int_{t_0}^t \int_\Omega \nabla \phi \cdot \left[u^2(\boldsymbol{u} + \boldsymbol{v}) + 2\pi \boldsymbol{u} - 2\nu \nabla \boldsymbol{u} \cdot \boldsymbol{u}\right] d\Omega d\tau. \end{aligned} \tag{11}$$

Proof. From Eq. $(7)_1$ we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_\Omega \phi u^2 d\Omega = &-\int_\Omega \phi \nabla\pi\cdot\boldsymbol{u}d\Omega+\nu\int_\Omega \phi\boldsymbol{u}\cdot\Delta\boldsymbol{u}d\Omega\\ &-\int_\Omega \phi(\boldsymbol{v}+\boldsymbol{u})\cdot\nabla\boldsymbol{u}\cdot\boldsymbol{u}d\Omega-\int_\Omega \phi\boldsymbol{u}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{u}d\Omega\,.\end{aligned} \tag{12}$$

For any $\bar{x}>0$ and $\bar{r}>0$, we set

$$\begin{aligned}\bar{\Omega} &= \left\{(x,r,\theta) : |x|\le\bar{x},\ r\le\bar{r}\right\},\\ \bar{\Omega}^+ &= \left\{(x,r,\theta) : 0<x\le\bar{x},\ r\le\bar{r}\right\},\\ \bar{\Omega}^- &= \left\{(x,r,\theta) : -\bar{x}\le x<0,\ r\le\bar{r}\right\}.\end{aligned} \tag{13}$$

Now, by Gauss' theorem, we see that

$$\int_{\bar{\Omega}}\phi\,\nabla\pi\cdot\boldsymbol{u}d\Omega=\int_{\bar{\Omega}^+}\nabla\cdot\{\phi\pi\boldsymbol{u}\}\,d\Omega+\int_{\bar{\Omega}^-}\nabla\cdot\{\phi\pi\boldsymbol{u}\}\,d\Omega-\int_{\bar{\Omega}}\pi\nabla\phi\cdot\boldsymbol{u}d\Omega. \tag{14}$$

As $\bar{x}\to+\infty$ and $\bar{r}\to+\infty$, by the boundary conditions (9), we get

$$\int_\Omega \phi\nabla\pi\cdot\boldsymbol{u}d\Omega=-\int_\Omega \pi\nabla\phi\cdot\boldsymbol{u}d\Omega. \tag{15}$$

The following identities can be shown by using the same techniques:

$$\int_\Omega \phi\boldsymbol{u}\cdot\Delta\boldsymbol{u}d\Omega=-\int_\Omega \phi|\nabla\boldsymbol{u}|^2d\Omega-\int_\Omega \nabla\phi\cdot\nabla\boldsymbol{u}\cdot\boldsymbol{u}d\Omega, \tag{16}$$

$$\int_\Omega \phi(\boldsymbol{v}+\boldsymbol{u})\cdot\nabla\boldsymbol{u}\cdot\boldsymbol{u}d\Omega=-\frac{1}{2}\int_\Omega u^2\nabla\phi\cdot(\boldsymbol{v}+\boldsymbol{u})d\Omega\,. \tag{17}$$

From (12) and (15-17) we deduce that

$$\begin{aligned}\frac{d}{dt}\int_\Omega \phi u^2 d\Omega=&-2\int_\Omega \phi\left[\nu|\nabla\boldsymbol{u}|^2+\boldsymbol{u}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{u}\right]d\Omega\\ &+\int_\Omega \nabla\phi\cdot\left[u^2(\boldsymbol{u}+\boldsymbol{v})+2\pi\boldsymbol{u}-2\nu\nabla\boldsymbol{u}\cdot\boldsymbol{u}\right]d\Omega,\end{aligned} \tag{18}$$

and, finally, integrating in the interval $[t_0,t]$, we get (11).

3 Two embedding theorems

In the divergent channel a weighted Poincaré inequality can be obtained [4,8]:

$$\int_\Omega w\frac{\varphi^2}{r^2}d\Omega\le\gamma\int_\Omega w|\nabla\varphi|^2d\Omega \tag{19}$$

with different values for γ. We recall here inequalities obtained by Rionero for a wedge of angle $2\theta_0\in(0,2\pi]$ which, as far as we know, contain the smallest value of γ.

Theorem 2. *Let Ω be a wedge of angle $2\theta_0 \in (0, 2\pi]$ and let $w = w(x, r)$ be a nonnegative weight function. Then, for any φ such that $\sqrt{w}|\nabla\varphi| \in L^2(\Omega)$, the inequality (19) holds with*

$$\gamma = \frac{4\theta_0^2}{\pi^2}. \tag{20}$$

Proof. Because $\varphi(x, r, \pm\theta_0) = 0$, the usual Poincaré inequality [4] implies that

$$\int_{-\theta_0}^{\theta_0} \varphi^2 d\theta \le \frac{4\theta_0^2}{\pi^2} \int_{-\theta_0}^{\theta_0} \left(\frac{\partial\varphi}{\partial\theta}\right)^2 d\theta. \tag{21}$$

But, since

$$\left(\frac{\partial\varphi}{\partial\theta}\right)^2 \le r^2 \left[\left(\frac{\partial\varphi}{\partial x}\right)^2 + \left(\frac{\partial\varphi}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial\varphi}{\partial\theta}\right)^2\right] = r^2|\nabla\varphi|^2, \tag{22}$$

we have

$$\int_{-\theta_0}^{\theta_0} \frac{\varphi^2}{r^2} d\theta \le \frac{4\theta_0^2}{\pi^2} \int_{-\theta_0}^{\theta_0} |\nabla\varphi|^2 d\theta, \tag{23}$$

and hence, multiplying both sides by rw and integrating with respect to $x \in R$ and $r \in R^+$, as w does not depend on θ, we see that the inequality (19) holds with γ given by (20).

Remark 1. The special case of (19) with the weight function $\phi = e^{-\alpha(|x|+r)}$ will be of interest in what follows.

Theorem 3. *Let Ω be a wedge of angle $2\theta_0 \in (0, 2\pi]$, and φ such that $|\nabla\varphi| \in L^2(\Omega)$. Then the inequality*

$$\int_\Omega \frac{\varphi^2}{r^2} d\Omega \le \frac{4\theta_0^2}{\pi^2} \int_\Omega |\nabla\varphi|^2 d\Omega \tag{24}$$

holds.

Proof. On multiplying both sides of (23) by r, and integrating with respect to $x \in R$ and $r \in R^+$, the theorem immediately follows.

4 L^2 energy estimates

Definition 15. Introduce the Reynolds number

$$R_1 = \begin{cases} \dfrac{\theta_0 F^*}{\nu} & \text{if} \quad \max\limits_{[-\theta_0,\theta_0]} |F| \ge \max\limits_{[-\theta_0,\theta_0]} |F'|, \\ \dfrac{\theta_0^2 F^*}{\nu} & \text{if} \quad \max\limits_{[-\theta_0,\theta_0]} |F| < \max\limits_{[-\theta_0,\theta_0]} |F'|, \end{cases} \tag{25}$$

where $F^* = \max\limits_{[-\theta_0,\theta_0]} \{|F|, |F'|\}$.

Theorem 4. *Let* $(\boldsymbol{u}, \pi) \in \mathcal{P}$, $\int_\Omega u^2(\boldsymbol{x}, t_0)d\Omega < +\infty$, *and set*

$$\hat{R}_1 = \begin{cases} \theta_0 R_1 & \text{if} \quad \max\limits_{[-\theta_0,\theta_0]} |F| \geq \max\limits_{[-\theta_0,\theta_0]} |F'|, \\ R_1 & \text{if} \quad \max\limits_{[-\theta_0,\theta_0]} |F| < \max\limits_{[-\theta_0,\theta_0]} |F'|. \end{cases} \tag{26}$$

Then

$$\hat{R}_1 < \frac{\pi^2}{6} \tag{27}$$

implies that

$$\boldsymbol{u} \in L^2(\Omega), \quad \nabla \boldsymbol{u} \in L^2(\Omega \times [t_0, T]) \qquad \forall t \in [t_0, T] \tag{28}$$

according to

$$e^{-C(t-t_0)} \int_\Omega u^2 d\Omega + \frac{12\nu}{\pi^2} \left(\frac{\pi^2}{6} - \hat{R}_1 \right) \int_{t_0}^t \int_\Omega |\nabla \boldsymbol{u}|^2 d\Omega d\tau \leq \int_\Omega u_0^2 d\Omega, \tag{29}$$

where C is a positive constant independent of the time t.

Proof. The Cauchy-Schwarz inequalities and (19), where γ is given by (20), imply that:

$$\int_\Omega \pi \nabla \phi \cdot \boldsymbol{u} d\Omega \leq C \int_\Omega \phi u^2 d\Omega + \alpha^2 \int_\Omega \phi |\pi|^2 d\Omega, \tag{30}$$

$$\int_\Omega \nabla \phi \cdot \nabla \boldsymbol{u} \cdot \boldsymbol{u} d\Omega \leq C \int_\Omega \phi u^2 d\Omega + \alpha^2 \int_\Omega \phi |\nabla \boldsymbol{u}|^2 d\Omega, \tag{31}$$

$$\int_\Omega u^2 \nabla \phi \cdot (\boldsymbol{v} + \boldsymbol{u}) d\Omega \leq C \int_\Omega \phi u^2 d\Omega + \alpha^2 \int_\Omega \phi |\nabla \boldsymbol{u}|^2 d\Omega, \tag{32}$$

$$\int_\Omega \phi \boldsymbol{u} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{u} d\Omega \leq \frac{6\theta_0^2 F^*}{\pi^2} \int_\Omega \phi |\nabla \boldsymbol{u}|^2 d\Omega \,, \tag{33}$$

where C is a positive constant. The last inequality is obtained by taking into account the fact that

$$|\boldsymbol{u} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{u}| = \frac{1}{r^2} |F(\theta)\,(u_3^2 - u_2^2) + F'(\theta) u_2 u_3| \leq \frac{3}{2} F^* \frac{u^2}{r^2}. \tag{34}$$

From (11) and (30-33) it follows that

$$\int_\Omega \phi u^2 d\Omega \leq \int_\Omega \phi u_0^2 d\Omega - 2\nu \left(1 - \frac{6{\theta_0}^2 F^*}{\nu \pi^2} - \frac{\alpha^2}{\nu} \right) \int_{t_0}^t \int_\Omega \phi |\nabla \boldsymbol{u}|^2 d\Omega d\tau$$
$$+ C \int_{t_0}^t \int_\Omega \phi u^2 d\Omega d\tau + 2\alpha^2 \int_{t_0}^t \int_\Omega \phi \pi^2 d\Omega d\tau, \tag{35}$$

where C is a positive constant independent of α. Now, applying the Gronwall lemma in the interval $[t_0, t]$, we see that

$$e^{-C(t-t_0)} \int_\Omega \phi u^2 d\Omega + 2\nu \left(1 - \frac{6{\theta_0}^2 F^*}{\nu \pi^2} - \frac{\alpha^2}{\nu}\right) \int_{t_0}^t \int_\Omega \phi |\nabla \boldsymbol{u}|^2 d\Omega d\tau$$

$$\leq \int_\Omega \phi u_0^2 d\Omega + 4\alpha^2 \int_{t_0}^t \int_\Omega \phi \pi^2 d\Omega d\tau. \tag{36}$$

Therefore, as $\alpha \to 0$, in view of the monotone convergence theorem, we obtain (29).

Theorem 5. *Let $(\boldsymbol{u}, \pi) \in \mathcal{P}$ and let $t_0, \geq 0$, be such that $\int_\Omega u^2(\boldsymbol{x}, t_0) d\Omega < +\infty$. Then*

$$\hat{R}_1 < \frac{\pi^2}{6} \tag{37}$$

implies that, for all $t \geq t_0$,

$$\int_\Omega u^2(\boldsymbol{x}, t) d\Omega = \int_\Omega u^2(\boldsymbol{x}, t_0) d\Omega - 2 \int_{t_0}^t \int_\Omega \left[\nu |\nabla \boldsymbol{u}|^2 + \boldsymbol{u} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{u}\right] d\Omega d\tau. \tag{38}$$

Proof. By the Cauchy inequality and (24), we get:

$$\left| \int_\Omega \pi \nabla \phi \cdot \boldsymbol{u} d\Omega \right| \leq \frac{\sqrt{2}}{2} \alpha \left(\int_\Omega \pi^2 d\Omega + \int_\Omega u^2 d\Omega \right), \tag{39}$$

$$\left| \int_\Omega \nabla \phi \cdot \nabla \boldsymbol{u} \cdot \boldsymbol{u} d\Omega \right| \leq \frac{\sqrt{2}}{2} \alpha \left(\int_\Omega u^2 d\Omega + \int_\Omega |\nabla \boldsymbol{u}|^2 d\Omega \right), \tag{40}$$

$$\left| \int_\Omega u^2 \nabla \phi \cdot (\boldsymbol{v} + \boldsymbol{u}) d\Omega \right| \leq C\alpha \left(\int_\Omega u^2 d\Omega + \int_\Omega |\nabla \boldsymbol{u}|^2 d\Omega \right), \tag{41}$$

where C is a positive constant independent of α. In this way, by means of (29), we conclude that the left-hand sides of the previous inequalities go to zero as $\alpha \to 0$. Therefore, integrating (11) in the time interval $[t_0, t]$, as $\alpha \to 0$, by virtue of Lebesgue's theorem, we get (38).

5 The stability theorem

Now, we can prove the main stability theorem.

Theorem 6. *Let $(\boldsymbol{v}, p)$ be a Jeffery-Hamel basic flow in the wedge Ω. If $\hat{R}_1 < \pi^2/6$, then $(\boldsymbol{v}, p)$ is stable in the $L^2(\Omega)$-norm with respect to the perturbations $(\boldsymbol{u}, \pi) \in \wp$.*

Proof. From (38) and inequality (24) it follows that

$$\int_\Omega u^2(\boldsymbol{x}, t) d\Omega \leq \int_\Omega u^2(\boldsymbol{x}, t_0) d\Omega - \frac{12\theta_0^2 \nu}{\pi^2} \left(\frac{\pi^2}{6} - \hat{R}_1 \right) \int_{t_0}^t \int_\Omega |\nabla \boldsymbol{u}|^2 d\Omega d\tau, \tag{42}$$

which proves the result.

Aknowledgements

This work was carried out under the auspices of the GNFM of INDAM and MIUR (PRIN): “Nonlinear mathematical problems of wave propagation and stability in models of continuous media".

References

[1] McAlpine, A., Drazin, P.G. (1998): On the spatio-temporal development of small perturbations of Jeffery-Hamel flows. Fluid Dynam. Res. **22**, 123–138
[2] Batchelor, G.K. (1967): An introduction to fluid dynamics. Cambridge University Press, Cambridge
[3] Banks, W.H.H., Drazin, P.G., Zaturska, M.B. (1988): On perturbations of Jeffery-Hamel flow. J. Fluid Mech. **186**, 559–581
[4] Flavin, J.N., Rionero, S. (1996): Qualitative estimates for partial differential equations. CRC Press, Boca Raton, FL
[5] Fraenkel, L.E. (1962): Laminar flow in symmetrical channels with slightly curved walls. I. On the Jeffery-Hamel solutions for flow between plane walls. Proc. Roy. Soc. London Ser. A **267**, 119–138
[6] Fraenkel, L.E. (1963): Laminar flow in symmetrical channels with slightly curved walls. II. An asymptotic series for the stream function. Proc. Roy. Soc. London Ser. A **272**, 406–428
[7] Hamadiche, M., Scott, J., Jeandel, D. (1994): Temporal stability of Jeffery-Hamel flow. J. Fluid Mech. **268**, 71–88
[8] Maremonti, P., Russo, R. (1992): On asymptotic time decay of solutions to a parabolic equation in unbounded domains. Ricerche Mat. **41**, 311–326
[9] Rosenhead, L. (1940): The steady two-dimensional radial flow of viscous fluid between two inclined plane walls. Proc. Roy. Soc. London Ser. A **175**, 436–467

Energy penalty, energy barrier and hysteresis in martensitic transformations

Y. Huo, I. Müller

Abstract. Non-convex energy and interfacial energy have been considered as fundamental in modeling the martensitic-austenitic phase transformation with hysteresis. For a one-dimensional bar, an additional non-local energy penalizing the inhomogeneous deformation needs to be considered in order to obtain finely lamellated phase mixtures. We study the thermodynamic consequences of such energy penalization, in particular, the possible energy barriers that can lock the phase transition process so as to produce hysteresis. Under the assumption that the energy penalty for the interfaces and the penalty for inhomogeneity are both very small, we reduce the total energy functional into a function of the strain, the phase fraction and the number of interfaces. Minimization of the total energy determines the number of interfaces in terms of the phase fraction. The model also predicts that the martensitic transformation needs a large driving force for starting while it can proceed at a lower driving force. Also the phase transition nucleates with a small but finite amount of the new phase, i.e., with non-zero values of the phase fraction and of the number of interfaces. Possible mechanism for hysteresis of phase transition is discussed with the energy barriers in Gibbs free energy.

1 Introduction

Van der Waals first introduced a non-monotone state equation in his well-known work of 1873 (see [1]) on the gas-liquid phase transition. It is now widely recognized that a non-monotone state equation or, equivalently, a non-convex energy function should be taken as the starting point to model phase transitions. Once a phase transformation has occurred, a previously homogeneous body transforms to an inhomogeneous one composed of regions of different phases. On the two sides of the inter-phase boundary, the atoms or molecules are of the same type but with different arrangements. This leads to different inter-atomic interactions. To account for such differences, one way is to consider the inter-phase boundary as a singular surface with abrupt changes of the state variables and to endow the singular surface with a surface tension and a surface energy as was done by Maxwell in his work on capillary action [2]. Another approach is to assume that the state variables are continuous but change strongly in the neighborhood of the inter-phase boundary and the local free energy density depends not only on the state variables but also on their derivatives, as considered by, e.g., van der Waals [3], Cahn and Hilliard [4]. Thus, the total energy of a body under phase transformation is a sum of the bulk energy, with the energy density function being a non-convex function of the state variables, and the interfacial energy proportional to the area of the interphase boundaries for the first approach of singular surfaces, or as an integral of an energy density function of the derivatives of the state

variables for the second approach of continuous fields. The minimization of the total energy functional in three-dimensional space is a still an open problem [5], partly due to the fact that complicated patterns of the phase regions can form during phase transformations.

Finely lamellated microstructures have been observed for specimens under tension or compression in martensitic transformations and in twinning. It is suggestive to consider a one-dimensional approximation for such situations. In the singular surface approach, the interfacial energy is thus just proportional to the number of interfaces. However, the minimization of the total energy leads to only one interface. Therefore, a finely lamellated structure is not possible. Statistical arguments were used to derive a relation between the number of interfaces and the amount of the phase fraction [6]. In the continuous field approach, the interfacial energy density is a function (often assumed to be the square) of the derivative of the state variable, the strain gradient in the case of martensitic transformation. Also in this approach it was shown that the absolute minimizer of the total energy functional corresponds to a phase mixture with only one interface [7]. Thus, in a one-dimensional model, it is necessary to consider additional contributions to the total energy in order to obtain finely lamellated microstructures.

By adding the square of the displacement to the total energy density, finely oscillating minimizing sequences were shown to be the minimizers under zero displacement boundary condition [8]. Solutions of the corresponding Euler-Lagrange equation with symmetric displacement boundary conditions and with the number of interfaces equal to 0, 1, and 2 have been studied [9,10]. Bifurcation and stability were also investigated for the solutions of the above model [11] and a similar model that implies the square of the difference between the displacements of the bar in the phase mixture and in the homogeneous bar [12]. The results show that the additional energy contribution related to the displacement of the bar penalizes the inhomogeneity and leads to solutions with microstructures.

In this work, we follow the approach of singular interfaces and propose, in addition to the bulk energy and the interfacial energy, an energy penalizing the inhomogeneity which can be deduced from the square of the difference between the displacements of the bar in the phase mixture and in the homogeneous bar similar to the one used in [12]. The total energy becomes a function of the bulk strains, the number of interfaces and the phase fraction. The absolute and local minimizers of such an energy function can be studied without much mathematical difficulty. Minimization of the total energy determines the number of interfaces in terms of the phase fraction. According to the present model, the martensitic transformation needs a large driving force for starting while it proceeds at a lower driving force. This implies that the stress has jumps at the initiation of the phase transition: downward jumps in extension and upward jumps in compression. Moreover, the transition nucleates with a small but finite amount of the new phase.

2 Energy penalty for inhomogeneous deformations

Consider a nonlinearly elastic bar with a non-convex double-well energy density $f(u'(x))$. The total energy is

$$E = \int_0^1 [f(u'(x)) + \frac{\varepsilon^2}{2} u''(x)^2 + \frac{\gamma^2}{2}(u(x) - u_\circ(x))^2] dx, \tag{1}$$

with boundary conditions

$$u(0) = u^0, \quad u(1) = d + u^0, \tag{2}$$

where $u(x)$ is the displacement of the bar, $u_\circ(x) = u^0 + dx$ is the displacement of the homogeneous bar subject to the same boundary conditions (2), and ε and γ are two constants [10,12]. The corresponding Euler-Lagrange equation is

$$f''(u'(x))u''(x) - \varepsilon^2 u''''(x) - \gamma^2(u(x) - u_\circ(x)) = 0. \tag{3}$$

For the boundary conditions (2) on the displacement $u(x)$ and the natural boundary conditions $u''(0) = u''(1) = 0$, the solutions of (3) have jumps at points x_i ($i = 1,\ 2, ..., N$) if $\varepsilon = 0$ and $\gamma = 0$ holds and provided that d lies in the interval between the minima of the double-well energy density. Within each interval (x_i, x_{i+1}), $i = 0,\ 1,\ 2, ..., N$, with $x_0 = 0$ and $x_{N+1} = 1$, the strains are piecewise constant

$$u'(x) = \begin{cases} u_x^+ \\ u_x^- \end{cases}, \tag{4}$$

where $u_x^\pm$ are the strains in the two wells of the bulk energy density $f(u'(x))$, respectively, with the constraint

$$d = z u_x^+ + (1 - z) u_x^-, \tag{5}$$

where z is the length of all the intervals (x_i, x_{i+1}) with $u'(x) = u_x^+$, namely, the phase fraction of the "+" phase. The total energy of the bar for $\varepsilon = 0$ and $\gamma = 0$ has only the contribution from the first term in (1) and has the form

$$E_b = z f(u_x^+) + (1 - z) f(u_x^-). \tag{6}$$

Consider now very small values of the coefficients ε and γ. In that case the solutions of (3) with the previous boundary conditions should be very similar to the above solutions of piecewise constant strains except in the immediate neighborhoods of the points x_i, $i = 1,\ 2, ..., N$. Thus, the contribution of the bulk energy term in (1) to the total energy should approximately still be equal to E_b as defined by (6). The contribution of the second term in (1) is the interfacial energy and should be approximately proportional to the number N of jump points, namely, the number of interfaces. Thus

$$E_i = \tau_1 N, \tag{7}$$

where the proportionality coefficient τ_1, which represents the surface energy of the interfaces, is positive and depends on the constant ε, – and possibly on $(u_x^+ - u_x^-)^2$, the magnitude of the jump.

The last term in (1) penalizes the inhomogeneity of the above piecewise constant-strain solutions and forces the piecewise linear displacement function $u(x)$ to be close to the linear function $u_\circ(x)$. It was shown [13] that, for a given number N of jump points and given values of $u_x^\pm$, this energy of homogenisation has a minimal value when the jump points are at the positions

$$\begin{array}{l} x_i^{+-} = \frac{i-1+z}{N}, \text{ when the jump is from } u_x^+ \text{ to } u_x^-, \\ x_i^{-+} = \frac{i-z}{N}, \text{ when the jump is from } u_x^- \text{ to } u_x^+. \end{array} \tag{8}$$

The minimal value of this energy has the form

$$E_h = \frac{\tau_2}{2}\left(\frac{z(1-z)}{N}\right)^2, \tag{9}$$

where $\tau_2 = \frac{1}{3}(u_x^+ - u_x^-)^2\gamma^2$ is a positive coefficient. While the interfacial energy E_i defined by (7) is an increasing function of the number N of interfaces, the above energy of homogenisation decreases with N. So, it is this penalty for inhomogeneity which leads to a finely lamellated microstructure. The interfacial energy, however, prevents too fine a structure. Thus, these two energy terms together select the correct number of microstructures as explained in the next section.

The total energy is the sum of the three terms, $E = E_b + E_i + E_h$,

$$E = (1-z)f(u_x^+) + zf(u_x^-) + \tau_1 N + \frac{\tau_2}{2}\left(\frac{z(1-z)}{N}\right)^2. \tag{10}$$

3 Minimization of the total energy and condition of phase equilibrium

The minimization of the total energy $E(u_x^+, u_x^-, z, N)$ defined by (10) subject to the constraint (5) leads to the minimization of the function

$$\psi(u_x^+, u_x^-, z, N) = E - \lambda(zu_x^+ + (1-z)u_x^- - d), \tag{11}$$

with λ being the Lagrange multiplier which must be identified with the stress in the bar. We set the derivatives of the function ψ with respect to its variables equal to zero, and thus obtain the following equilibrium conditions:

$$\begin{aligned} \lambda &= f'(u_x^+) + \tau_1'(u_x^+ - u_x^-)\tfrac{N}{z} + \tfrac{1}{2}\tau_2'(u_x^+ - u_x^-)\left(\tfrac{1-z}{N}\right)^2 z \\ &= f'(u_x^-) - \tau_1'(u_x^+ - u_x^-)\tfrac{N}{1-z} - \tfrac{1}{2}\tau_2'(u_x^+ - u_x^-)\left(\tfrac{z}{N}\right)^2 (1-z), \end{aligned} \tag{12}$$

$$N = N_e(z) = \left(\frac{\tau_2}{\tau_1}\right)^{1/3}(z(1-z))^{2/3}, \tag{13}$$

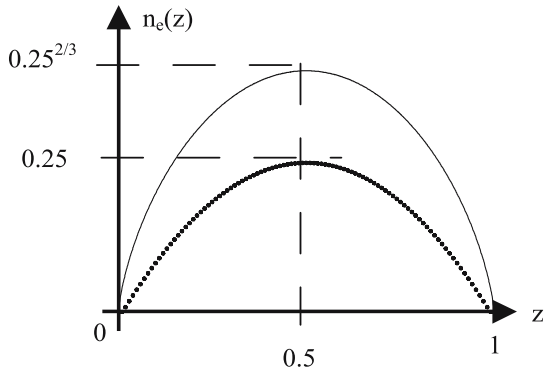


Fig. 1. The number of interfaces as a function of the phase fraction normalized by the maximum at z = 1/2. Solid line for $n_e(z) = (z(1-z))^{2/3}$ and dotted line for $n_e = z(1-z)$ of [6]

$$f(u_x^+) - f(u_x^-) + \tau_2 \frac{z(1-z)}{N^2}(1-2z) - \lambda(u_x^+ - u_x^-) = 0. \tag{14}$$

By (13), we obtain an explicit relation between the number N of interfaces and the phase fraction z that is similar to, but different from, the relation obtained by a statistical argument [8], as shown in Fig. 1. Such a relation comes from the interplay between the microstructure-preventing interfacial energy (7) and the microstructure-supporting homogenisation energy (9).

From Fig. 1, the above (N, z)-relation (13) shown by the solid line has infinite derivatives at z = 0, and 1, while the relation obtained by a statistical argument, viz., $N \propto z(1-z)$ shown by the dotted line, has finite derivatives. This fact has a strong influence on the nucleation of phases as is shown in the next section.

Equation (12) shows that, in general, the two energy terms, E_i and E_h, may also have contributions to the total stress. For simplicity, we neglect these contributions in (12) and assume that the two coefficients τ_1 and τ_2 in (10) are positive constants. Thus, we have

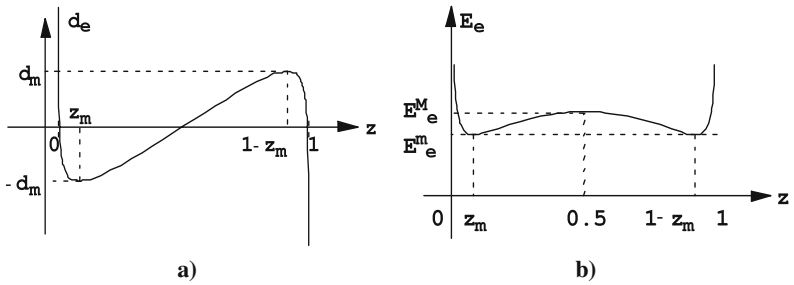
$$\lambda = f'(u_x^+) = f'(u_x^-). \tag{15}$$

In order to carry out all the calculations analytically, we accept the two-parabola bulk energy density function,

$$f(u') = \begin{cases} \frac{\alpha}{2}(u' + \Delta_d)^2, & \text{for } u' \le 0, \\ \frac{\alpha}{2}(u' - \Delta_d)^2, & \text{for } u' > 0, \end{cases} \tag{16}$$

where α is the elastic modulus and $\pm\Delta_d$ are the stress-free strains of the two stable states. Substituting (16) in (14) and considering the constraint (5), we obtain

$$\begin{aligned} &u_x^+ - u_x^- = 2\Delta_d, \quad d = u_x^- + z\Delta_d, \\ &\lambda = \lambda(d, z) = \alpha(d + (1-2z)\Delta_d). \end{aligned} \tag{17}$$



a) b)

Fig. 2. The strain and the energy of the phase equilibrium states as functions of the phase fraction. The parameters are chosen so that $A/\alpha\Delta_d^2 = 0.25$

This is the stress-strain relation of phase mixtures with phase fraction z. The phase equilibrium condition (14) is reduced to

$$\lambda = \lambda_e(z) = \frac{A}{3\Delta_d}\frac{1-2z}{(z(1-z))^{1/3}}, \tag{18}$$

where A is a positive constant and is related to the two coefficients $\tau_{1,2}$ through

$$A = \frac{3}{2}(\tau_1^2\tau_2)^{1/3}. \tag{19}$$

By use of (17) the total strain is related to the phase fraction in phase equilibrium through

$$d = d_e(z) = -(1-2z)\Delta_d + \frac{A}{3\alpha\Delta_d}\frac{1-2z}{(z(1-z))^{1/3}}. \tag{20}$$

As shown in Fig. 2a, $d_e(z)$ is non-monotone and has a minimum $-d_m$ in the interval $0 < z_m < 0.5$ and a maximum d_m at $1 - z_m$. As $d_e'(z_m) = 0$, we see that

$$\frac{z_m^{4/3}(1-z_m)^{4/3}}{1+2z_m} = \frac{A}{18\alpha\Delta_d^2} \quad \text{and} \quad d_m = -d_e(z_m). \tag{21}$$

Here z_m can be solved numerically or approximately for very small values of the constant A as

$$z_m \approx \left(\frac{A}{18\alpha\Delta_d^2}\right)^{3/4} \quad \text{and} \quad d_m \approx \Delta_d(1-8z_m) \quad \text{for} \quad A << \alpha\Delta_d^2. \tag{22}$$

The total energy for phase equilibrium states can be obtained by substituting (13) and (16-18) into (10),

$$E = E_e(z) = \frac{2A^2}{9\alpha\Delta_d^2}\frac{(1-2z)^2}{(z(1-z))^{2/3}} + A(z(1-z))^{2/3}. \tag{23}$$

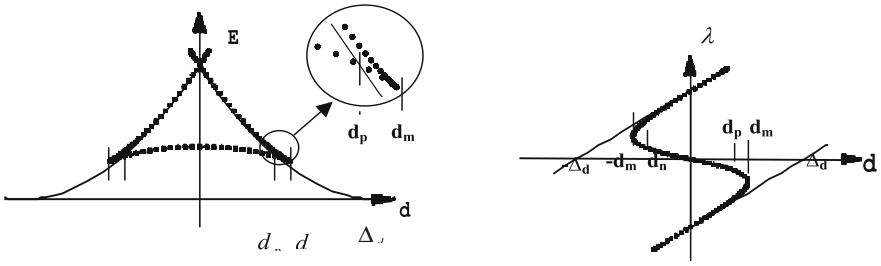


Fig. 3. The energy-strain and stress-strain relations for the pure phases (solid lines) and the phase equilibrium states (dotted lines). $A/\alpha\Delta_d^2 = 0.25$

As shown in Fig. 2b, it has two minima at z_m and $1 - z_m$ with $E_e^m = E_e(z_m) > 0$, and a maximum at z = 1/2 with $E_e^M = E_e(1/2) = A/2^{4/3}$.

In order to obtain the total energy as a function of the total strain, namely, $E_e(d)$, we need to invert the $(d, z)-$relation (18) for the three monotone branches: $(0, z_m)$, $(z_m, 1 - z_m)$ and $(1 - z_m, 1)$, respectively, and substitute them into (20). This can be done numerically and the resulting graph $E_e(d)$ is shown in Fig. 3a as dotted lines. There are two convex branches corresponding to $(0, z_m)$ and $(1 - z_m, 1)$, and one concave branch for $(z_m, 1 - z_m)$. The concave branch has lower values than the two convex branches and has two minima at $\pm d_m$ and a maximum at d = 0. In order that at least a part of $E_e(d)$ is below the energy of the pure phases, i.e., the two parabolas $f(d)$ shown as solid lines in Fig. 3a, the maximum of the concave branch of $E_e(d)$ should be below $f(d = 0)$. From (16) and (23) this requirement reads for our model that

$$E_e^M = E_e(z = 1/2) < \alpha\Delta_d^2/2 \quad \Rightarrow \quad A = \frac{3}{2}(\tau_1^2\tau_2)^{1/3} < 2^{1/3}\alpha\Delta_d^2. \tag{24}$$

Figure 3b shows the corresponding stress-strain relation which results from substituting the inverse of (20) into (18). It is interesting to observe that there is no point of intersection on the stress-strain diagram between phase equilibrium states (dotted line) and the two straight elastic curves (solid lines) which represent the states of the "-" and "+" phases. However, there are two intersection points on the energy-strain diagram between the pure phases and the phase equilibrium states.

The points of intersection, indicated by d_n and $d_p = -d_n$ in Fig. 3a, and the corresponding phase fractions: z_n and $z_p = 1 - z_n$, satisfy

$$d_n = d_e(z_n) \quad \text{and} \quad E_e(z_n) = f(d_n) = \frac{\alpha}{2}(d_n + \Delta_d)^2. \tag{25}$$

By (20) and (23), we have

$$\frac{z_n^{4/3}(1 - z_n)^{1/3}}{1 + z_n} = \frac{A}{6\alpha\Delta_d^2}. \tag{26}$$

As with (21), it can be solved numerically or approximately for very small values of the constant A as

$$z_n \approx \left(\frac{A}{6\alpha\Delta_d^2}\right)^{3/4} \quad \text{and} \quad d_n \approx -\Delta_d(1-4z_n) \quad \text{for} \quad A << \alpha\Delta_d^2. \tag{27}$$

4 Absolute minimizer and nucleation of phase transition

The absolute minimizer $E_m(d)$ of the total energy is the lowest graph in Fig. 3a and is shown in Fig. 4a,

$$E_m(d) = \begin{cases} E_e(d), & d_n < d < d_p, \\ f(d), & \text{otherwise.} \end{cases} \tag{28}$$

If a body begins to be extended from the "-" phase at the stress-free configuration $d = -\Delta_d$, it first follows the energy curve of $f(d)$ as far as d_n. Upon a further increase of d, it switches the energy branch to $E_e(d)$ at this point if the body chooses the lowest energy, i.e., it follows the absolute minimizer shown in Fig. 4a. Thus, at d_n, the body starts to transform from the "-" phase to a phase mixture with its phase fraction jumping from $z = 0$ to $z = z_n > 0$. At the same time, the number of interfaces jumps accordingly by (13) from $N = 0$ to $N = N_n = N_e(z_n)$. Moreover, since the slopes of $f(d)$ and $E_e(d)$ are not equal at d_n, the stress also has a jump as shown in Fig. 4b for the corresponding stress-strain relation. The stresses before and after the jump are

$$\begin{aligned} \lambda_0^n &= f'(d_n) = \alpha(d_n + \Delta_d), \\ \lambda_{z_n}^n &:= E_e'(d_n) = \lambda_e(d_n) = \alpha(d_n + \Delta_d - 2z_n\Delta_d) = \lambda_0^n - 2\alpha\Delta_d z_n. \end{aligned} \tag{29}$$

Because of symmetry, the above argument can be carried out similarly for compression tests starting from the "+" phase. The phase fraction, number of interfaces and the stress all jump at $d = d_p = -d_n$ from $z = 1$, $N = 0$ and $\lambda = -\lambda_0^n$ to $z = 1 - z_n$, $N = N_e(1 - z_n) = N_n$ and $\lambda = -\lambda_{z_n}^n$.

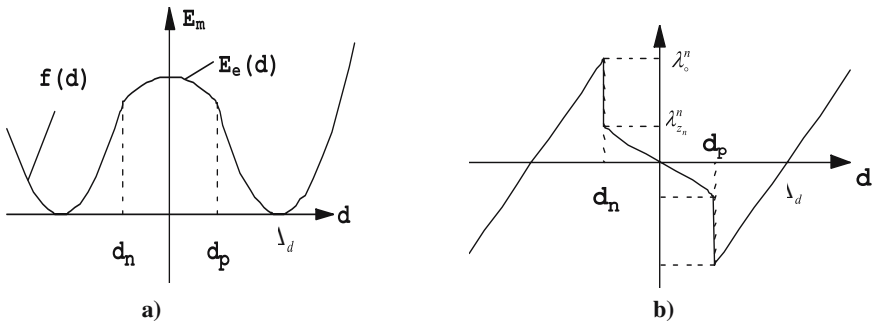


Fig. 4. The absolute minimizer of the energy and the corresponding stress-strain relation. $A/\alpha\Delta_d^2 = 0.25$

In both cases, z_n is the critical, i.e., minimal, mass of nucleation. This is very similar to the critical radius of a nucleus in gas-liquid phase transition in a volume-controlled process [14]. However, since the number of interfaces for this nucleus is larger than zero, the body is in the phase mixtures *with a microstructure*, once the phase transition has occurred at d_n. The stress decreases upon extension and it increases in compression. The stored elastic energy in the body decreases in both cases as can be calculated by (14), (17), (6) and (29),

$$\Delta E_f := E_b(z=0) - E_b(z=z_n) = \frac{\lambda_0^{n\,2}}{2\alpha} - \frac{\lambda_{z_n}^{n\,2}}{2\alpha} = A(z_n(1-z_n))^{2/3}. \tag{30}$$

Thus, a part of the stored elastic energy is changed into the interfacial energy and the energy penalizing inhomogeneity. So, in a sense, ΔE_f is the energy lost to form the microstructure of the phase mixture with the critical mass z_n and we may call it the *formation energy for nucleation*.

After the above initial nucleation at d_n or d_p, the body is in the phase equilibrium state. Upon further extension or compression, the body cannot follow the phase equilibrium line $\lambda_e(d)$, since the energy function is concave and the phase equilibrium states are unstable. Rather, hysteresis should occur and the body does not follow the absolute minimizer. To analyse this, local minimizers of the energy are considered in the next section.

5 Partial equilibrium states and hysteresis

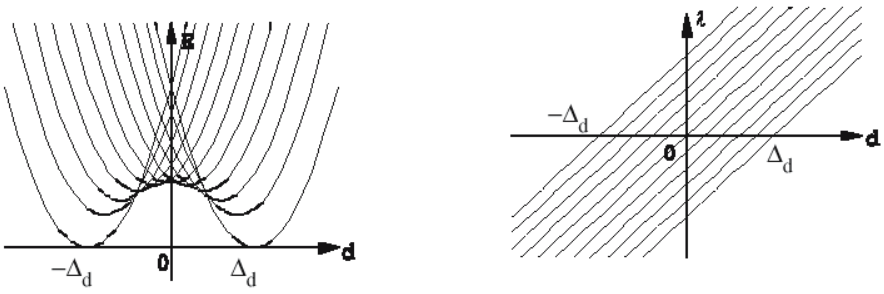
Four equilibrium conditions were obtained by (12-14) for phase equilibrium states. Equation (12) is the mechanical equilibrium condition, (13) represents the equilibrium condition for the microstructures and (14) is the phase equilibrium condition. Due to the different mechanisms for attaining those three equilibria, it seems reasonable to assume that some of them are attained much faster than the others. The mechanical equilibrium in a body is attained by elastic waves that have large speeds. The increase in the number of interfaces is because of the nucleation of the new phase, which is rather fast in diffusionless thermo-elastic martensitic transformations. If we assume that the mechanical equilibrium of (12) or (15) for the simplified model is attained first, we have the following total energy function for two-parabola potential by substituting (16), (17) into (10),

$$E(d,z,N) = \frac{\alpha}{2}[d + (1-2z)\Delta_d]^2 + \tau_1 N + \frac{\tau_2}{2}\left(\frac{z(1-z)}{N}\right)^{2/3}. \tag{31}$$

If we assume further that the equilibrium for the microstructures (13) is also satisfied, by substituting it into the above relation we obtain a simple form for the total energy as

$$E(d,z) = \frac{\alpha}{2}[d + (1-2z)\Delta_d]^2 + A(z(1-z))^{2/3}, \tag{32}$$

where the first term is the stored elastic energy and the second is the energy to form microstructures consisting of the interfacial energy and the energy penalizing



a) b)

Fig. 5. The energy-strain and the stress-strain relations of local minimizers for $z_i = 0, 0.1, 0.2, \ldots 0.9, 1$. $A/\alpha\Delta_d^2 = 0.25$

inhomogeneity. The corresponding stress-strain relation is still given by (17). Figure 5 shows the above energy function and the corresponding stress-strain relation for various values of the phase fraction z.

The stress-strain curves are all straight and parallel to each other. The energy-strain curves are all parabolas that intersect each other. In particular there are intersections with the energies of the pure phases. The above consideration of partial equilibrium is connected with the local minimizers that some authors prefer to use.

The strains $\pm d_i$ at the intersection points can be calculated by the equation

$$E(d_i, z) = E(d_i, z = 0) = f(d). \tag{33}$$

Note that $-d_i$ is the solution for the intersection with $z = 1$ because of symmetry. By (32), we obtain

$$d_i = d_i(z) = -(1 - z)\Delta_d + \frac{A}{2\alpha\Delta_d}\frac{(1 - z)^{2/3}}{z^{1/3}}. \tag{34}$$

As shown by the solid line in Fig. 6, there is a minimum of the above strain of intersection, which is exactly the intersection between $d_i(z)$ and the phase equilibrium condition $d_e(z)$ of (18) (dashed line in Fig. 6). Thus, the phase fraction z_n at the minimum defined by $d_i'(z_n) = 0$ satisfies (26) and $d_i(z_n) = d_e(z_n) = d_n$ as defined by (25).

It can be shown easily that

$$f(d) = E(d, z = 0) < E(d, z > 0) \quad \text{for} \quad d < d_n. \tag{35}$$

Therefore, the starting point of nucleation $d = d_n$ obtained in the last section by considering the absolute minimizer is the smallest strain at which there are other local minimizers with the same or lower energies than those of the pure "-" phase. And at $d = d_n$, the phase mixture with $z = z_n$ has the same energy as in the "-" phase and all the other phase mixtures have larger energies, i.e.,

$$f_n := f(d_n) = E(d_n, z = z_n) < E(d_n, z \neq 0,\ z_n). \tag{36}$$

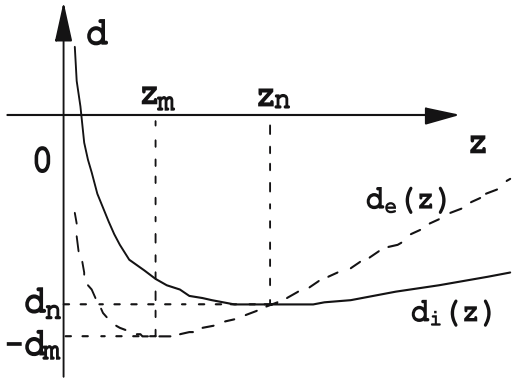


Fig. 6. The strain of the intersection $d_i(z)$ (solid lines) and the phase equilibrium strain $d_e(z)$ (dashed lines). $A/\alpha\Delta_d^2 = 0.25$

As shown in Fig. 7a, the phase mixture with $z = z_n$ has a lower energy than the "-" phase for $d > d_n$. Therefore, we may expect from considering the local minimizers that the body starts to transform its phase at $d = d_n$ and it transforms from $z = 0$ to $z = z_n$. The stress λ_0^n required for nucleation defined by (28) is essentially the stress needed to eliminate the energy barrier between $f(d)$ and $E(d, z_n)$ as made evident in Fig. 7b by considering $E - \lambda_0^n d$. Once the body has climbed over the energy barrier, it falls into the energy branch $E(d, z_n)$ and has a lower stress $\lambda_{z_n}^n = E_{,d}(d_n, z_n) < \lambda_0^n$ as defined by (29).

Thus, we may call the energy f_n at $d = d_n$ as defined by (36) the *driving force of nucleation*. The formation energy of nucleation ΔE_f defined by (30) is just the energy (32) of the local minimizer $z = z_n$ at its stress-free configuration. And the

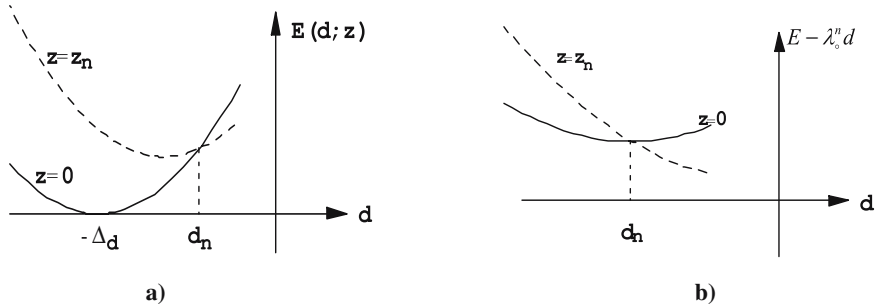


Fig. 7. The energy function of the local minimizer with the critical mass $z = z_n$ (dashed lines) and the "-" phase (solid lines). $A/\alpha\Delta_d^2 = 0.25$

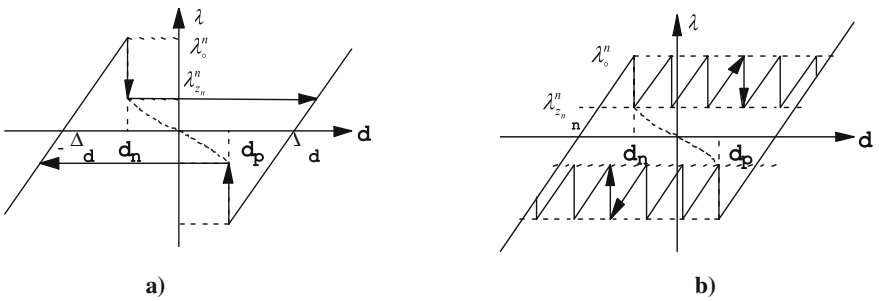


Fig. 8. Hysteresis in the stress-strain diagram for nucleation-then-growth and continuous-nucleation processes

difference between these two energies is

$$f_g := f_n - E_m = \frac{\lambda_{z_n}^{n\ 2}}{2\alpha}. \tag{37}$$

As discussed in the last section, the body under extension first transforms its phase by nucleation at $d = d_n$ and the stress drops from λ_0^n to $\lambda_{z_n}^n$. After the nucleation, the body is on the phase equilibrium branch but cannot proceed along the decreasing stress-strain branch of the phase equilibrium states, since they are unstable. Hysteresis should occur. Here we propose two possible mechanisms for hysteresis.

1. Nucleation-then-growth: The body proceeds in its phase transformation just by growth with the constant stress $\pm\lambda_{z_n}^n$. Hysteresis is observed as shown by the stress-strain diagram in Fig. 8a. The back transformation is shown by symmetry. The driving force of growth is f_g as defined by (37).
2. Continuous-nucleation: No growth but only nucleation with critical mass z_n and driving force f_n takes place. A zigzag stress-strain hysteresis is observed as shown in Fig. 8b.

The number of interfaces for process (1) first jumps from $N = 0$ to $N = N_n > 0$ by the initial nucleation and remains constant for a large part of the growth process. Then, it decreases to $N = 0$ at the end. However, the situation is entirely different for the continuous-nucleation process. The number of interfaces increases after each new nucleation as far as $z = 1/2$. Then, further nucleation of the new phase causes a decrease in the number of interfaces. The reality may lie somewhere between the above two extreme situations. Namely, both growth of existing new phases and new nucleation take place [15,16].

6 Load-controlled experiments and energy barrier

In the previous sections, we have always assumed that a hard device controls the end displacement of the bar. For a soft device, the stress at the two ends is controlled. The

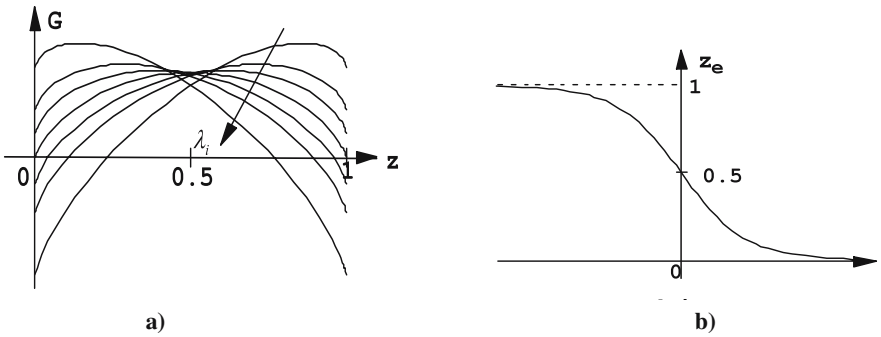


Fig. 9. The Gibbs free energy as a function of the phase fraction for various values of the stress and the phase fraction of their maximum as a function of the stress. $A/\alpha\Delta_d^2 = 0.25$

phase equilibrium conditions (13–15) are independent of the boundary conditions, so they remain unchanged. However, the proper potential for a body under load-controlled experiments is the Gibbs free energy $G = E - \lambda d$ with λ the stress of the bar. For the two-parabola bulk energy density (16), the equilibrium conditions are still (13), (17), (18). Under the assumption that the mechanical equilibrium (15) and the equilibrium for microstructures (13) are attained, we see from (32) and (15) that

$$G = G(z, \lambda) = E(d, z) - \lambda d = -\frac{\lambda}{2\alpha} + \lambda\Delta_d(1 - 2z) + A(z(1 - z))^{2/3}. \quad (38)$$

Figure 9a shows the Gibbs free energy above as a function of the phase fraction for various values of the stress. It is obvious that G is a concave function of z and has a maximum at z_e which satisfies the phase equilibrium condition (18) by $\lambda_e(z_e) = \lambda$ as shown in Fig. 9b. Thus, there is an energy barrier between the two pure phases, $z = 0$ and 1. Substituting (18) into (38), we find that the differences in G between the phase equilibrium state and the pure phases are

$$\begin{aligned} \Delta G_0(\lambda) &= G(z_e(\lambda), \lambda) - G(z = 0, \lambda) = \tfrac{A(1+z_e)}{3} \tfrac{z_e^{2/3}}{(1-z_e)^{1/3}} > 0, \\ \Delta G_1(\lambda) &= G(z_e(\lambda), \lambda) - G(z = 1, \lambda) = \tfrac{A(2-z_e)}{3} \tfrac{(1-z_e)^{2/3}}{z_e^{1/3}} > 0. \end{aligned} \quad (39)$$

From (18), we know that $0 < z_e < 1$ for any finite values of the stress. Thus, the above differences are always positive and the phase equilibrium states always have a larger Gibbs free energy than the two pure phases. Such an energy barrier in the Gibbs free energy can never be eliminated under finite values of the stress. This is very different from the displacement-controlled experiments as discussed in the previous section. Moreover, the derivatives of G with respect to z are infinite at z = 0 and 1. Therefore, any phase transition starting from z = 0 or 1 must start with a finite value of the phase fraction so that it will pass over the maximal point of G. Otherwise, there would be a very large driving force, -$G_{,z}$, to eliminate the nucleation of the new phase.

The above phenomenon predicted by the present model is very similar to the well-known theory of nucleation of a spherical droplet in gases in a pressure-controlled process; see [14] for a recent treatment. It is also known for the gas-liquid transition that there must be a fluctuation of the local pressure and the local density inside the gas in order that nucleation can take place in a pressure-controlled process. As a consequence, super-cooling and overheating may occur and lead to hysteresis. Similarly, we may expect for the present model that there is a fluctuation of the local stress and the local strain for a bar under load-controlled tests to start to transform. This is also consistent with the prediction that the stress has a jump at the initial nucleation of the phase transformation obtained by following the absolute minimizer of the energy function.

Acknowledgements

This work was supported by the Alexander von Humboldt Foundation of Germany and the National Natural Science Foundation of China (10372023).

References

[1] Rowlinson, J.S., van der Waals, J.D. (1988): On the continuity of gaseous and liquid states. In: Lebowitz, J.L. (ed.): Studies in statistical mechanics, Vol. 14. North-Holland, Amsterdam

[2] Maxwell, J.C. (1878): Capillary action. In: Encyclopaedia Britannica. 9th edition, 1878, Vol. 5. A. & C. Black, Edinburgh, pp. 56–71. Reprint: Niven, W.D. (ed.) (1965): The scientific papers of James Clerk Maxwell. Vol. 2. Dover, New York, pp. 541–596

[3] van der Waals, J.D. (1979): The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density. J. Statist. Phys. **20**, 197–244 [Original article published in 1893.]

[4] Cahn, J.W., Hilliard, J.E. (1958): Free energy of a nonuniform system. I. Interfacial free energy. J. Chemi. Phys. **28**, 258–267

[5] Kohn, R.V., Otto, F. (1997): Small surface energy, coarse-graining, and selection of microstructure. Phys. D **107**, 272–289

[6] Müller, I. (1989): On the size of the hysteresis in pseudoelasticity. Contin. Mech. Thermodyn. **1**, 125–142

[7] Carr, J., Gurtin, M.E., Slemrod, M. (1984): Structured phase transitions on a finite interval. Arch. Ration. Mech. Anal. **86**, 317–351

[8] Müller, S. (1993): Singular perturbations as a selection criterion for periodic minimizing sequences. Calc. Var. Partial Differential Equations **1**, 169–204

[9] Truskinovsky, L., Zanzotto G. (1995): Finite-scale microstructures and metastability in one-dimensional elasticity. Meccanica **30**, 577–589

[10] Truskinovsky, L., Zanzotto G. (1996): Ericksen's bar revisited: energy wiggles. J. Mech. Phys. Solids **44**, 1371–1408

[11] Vainchtein, A., Healey, T., Rosakis, P., Truskinovsky, L. (1998): The role of the spinodal region in one-dimensional martensitic transitions. Phys. D **115**, 29–48

[12] Vainchtein, A., Healey, T., Rosakis, P. (1999): Bifurcation and metastability in a new one-dimensional model for martensitic phase transitions. Comput. Methods Appl. Mech. Engrg. **170**, 407–421

[13] Huo, Y., Müller, I. (2003): Interfacial and inhomogeneity penalties in phase transitions. Contin. Mech. Thermodyn. **15**, 395–407
[14] Huo, Y., Müller, I. (2003): Nucleation of droplets in a binary mixture. Meccanica **38**, 493–504
[15] Sun, Q.P. private communication
[16] Müller, I., Seelecke, S. (2001): Thermodynamic aspects of shape memory alloys. Math. Comput. Modelling **34**, 1307–1355

On the applicability of generalized strain measures in large strain plasticity

M. Itskov

Abstract. In the present paper two thermodynamically consistent large strain plasticity models are examined and compared in finite simple shear. The first model (A) is based on the multiplicative decomposition of the deformation gradient, while the second one (B) on the additive decomposition of generalized strain measures. Both models are applied to a rigid-plastic material described by a von Mises-type yield criterion. Since both models include neither a hardening nor a softening law, a constant shear stress response, even for large amounts of shear, is expected. Indeed, model A exhibits true constant shear stress behavior independent of the elastic material law. This is not, however, the case for model B so that its applicability under finite shear deformations may be questioned.

1 Introduction

There are several different concepts enabling to consider large elasto-plastic strains in anisotropic materials. One is based on the multiplicative decomposition of the deformation gradient into an elastic and a plastic part. A thermodynamically consistent formulation of this concept naturally leads to a 9-dimensional flow rule and a yield criterion in terms of Mandel's stress tensor [1]. Since this tensor is generally non-symmetric, additional efforts are required to construct an anisotropic yield function and to formulate conditions of convexity of the yield surface resulting in the 9-dimensional stress space (see [2–4]). Another concept is based on the additive decomposition of the so-called generalized strain measures [5–7]. Thereby, the structure of the classical infinitesimal theory of plasticity is retained. A further remarkable feature of this concept is that the yield criterion is formulated in terms of the stress tensor work-conjugate to the underlying generalized strain (see, e.g., [8]). Since this stress tensor is a priori symmetric, a quadratic yield function preserves the form of the well-known Hill orthotropic criterion [9] and can easily be generalized to other material symmetries (see [10–12]). However, this concept has never been studied in the case of large plastic deformations accompanied by finite rotations which take place, for example, under simple shear.

Thus, the aim of the present paper is to examine and compare shear stress responses of the two above mentioned plasticity models in the case of finite simple shear. Both models are applied to an ideal-plastic material described by a von Mises-type yield criterion. Since both models include neither a hardening nor a softening law, a constant shear stress response even for large amounts of shear is expected. To avoid the influence of the elastic material law and the elastic strain energy, we consider small elastic but large plastic deformations (rigid-plastic material). This deformation

pattern takes place in many engineering problems, as, for example, metal forming processes, and is important for engineering practice.

Finally, a word of notation. Second-order tensors are denoted by bold face letters, e.g., $\mathbf{A}$, $\mathbf{B}$, ... Their scalar product, the quadratic norm and the deviator are defined by $\mathbf{A}:\mathbf{B}=\mathrm{tr}\left(\mathbf{A}\mathbf{B}^{\mathrm{T}}\right)$, $\|\mathbf{A}\|=\sqrt{\mathbf{A}:\mathbf{A}}$ and $\mathrm{dev}\mathbf{A}=\mathbf{A}-1/3\,\mathrm{tr}\left(\mathbf{A}\right)\mathbf{I}$ respectively, where $\mathbf{I}$ represents the second-order identity tensor. A linear mapping of one second-order tensor into another is described by $\mathbf{B}=\mathsf{C}:\mathbf{A}$, where C stands for a fourth-order tensor. Along with this "right" mapping we also define the "left" mapping, so that the relation $(\mathbf{A}:\mathsf{C}):\mathbf{X}=\mathbf{A}:(\mathsf{C}:\mathbf{X})$ holds for all second-order tensors $\mathbf{X}$ (see [13]). δ_{ij} denotes finally the Kronecker delta.

2 Thermodynamic and kinematic preliminaries

The derivation of evolution equations for both material models is based on the second law of thermodynamics and the principle of maximum plastic dissipation. In this section we begin with the second law of thermodynamics written in the Clausius-Planck form [14] as

$$\mathcal{D}=\boldsymbol{\tau}:\mathbf{L}-\dot{\psi}\geq 0, \tag{1}$$

where $\mathcal{D}$ denotes a dissipation defined as the difference between the stress power and the material time derivative of the free energy function ψ. Here $\boldsymbol{\tau}$ represents the Kirchhoff stress tensor work-conjugate to the velocity gradient $\mathbf{L}$ defined in terms of the deformation gradient $\mathbf{F}$ by

$$\mathbf{L}=\dot{\mathbf{F}}\mathbf{F}^{-1}. \tag{2}$$

The stress power in the dissipation inequality (1) can alternatively be written in terms of the so-called generalized strain measures. They represent isotropic tensor functions of the right Cauchy-Green tensor

$$\mathbf{C}=\mathbf{F}^{\mathrm{T}}\mathbf{F} \tag{3}$$

or the right stretch tensor

$$\mathbf{U}=\mathbf{C}^{1/2}. \tag{4}$$

The latter results from the polar decomposition of the deformation gradient

$$\mathbf{F}=\mathbf{R}\mathbf{U}, \tag{5}$$

where $\mathbf{R}=\mathbf{R}^{-\mathrm{T}}$ denotes a rotation tensor.

The generalized strains can be defined by means of the spectral decomposition of the stretch tensor

$$\mathbf{U}=\sum_{i}^{m}\lambda_i\mathbf{P}_i,\quad \mathbf{P}_i\mathbf{P}_j=\delta_{ij}\mathbf{P}_i,\quad i,j=1,...,m\leq 3, \tag{6}$$

in terms of the eigenprojections $\mathbf{P}_i$ and the corresponding pairwise distinct eigenvalues λ_i $(i = 1, ..., m \leq 3)$ [6,7] as

$$\mathbf{E} = \sum_{i}^{m} f(\lambda_i)\, \mathbf{P}_i, \tag{7}$$

where $f: \mathbb{R}^+ \to \mathbb{R}$ is a strictly increasing scalar function satisfying the conditions [7]

$$f(1) = 0, \quad f'(1) = 1. \tag{8}$$

Additionally, we require that the function $f(\lambda)$ is analytic everywhere except for the point $\lambda = 0$. For example, for the so-called Seth strains [5] the function f takes the form

$$f(\lambda) = \begin{cases} \dfrac{1}{r}(\lambda^r - 1) & \text{for } r \neq 0\,, \\ \ln\lambda & \text{for } r = 0\,. \end{cases} \tag{9}$$

In terms of the generalized strains (7) the dissipation inequality (1) can be rewritten as

$$\mathcal{D} = \mathbf{T} : \dot{\mathbf{E}} - \dot{\psi} \geq 0, \tag{10}$$

where $\mathbf{T}$ denotes a stress tensor work-conjugate to $\mathbf{E}$. Of special interest is a particular form of the dissipation inequality (10),

$$\mathcal{D} = \mathbf{S} : \frac{1}{2}\dot{\mathbf{C}} - \dot{\psi} \geq 0, \tag{11}$$

in terms of the second Piola-Kirchhoff stress tensor

$$\mathbf{S} = \mathbf{F}^{-1}\boldsymbol{\tau}\mathbf{F}^{-\mathrm{T}}. \tag{12}$$

Further derivation of the evolution equations depends on the assumption concerning the decomposition of strains into an elastic and a plastic part.

3 Multiplicative decomposition of the deformation gradient (model A)

In this section we assume the multiplicative decomposition of the deformation gradient into an elastic part $\mathbf{F}_{\mathrm{e}}$ and a plastic part $\mathbf{F}_{\mathrm{p}}$ [15],

$$\mathbf{F} = \mathbf{F}_{\mathrm{e}}\mathbf{F}_{\mathrm{p}}. \tag{13}$$

The strain energy function can further be represented by

$$\psi = \hat{\psi}(\mathbf{C}_{\mathrm{e}})\,, \tag{14}$$

where $\mathbf{C}_\mathrm{e} = \mathbf{F}_\mathrm{e}^\mathrm{T}\mathbf{F}_\mathrm{e}$ denotes the elastic right Cauchy-Green tensor. With the aid of the identity

$$\mathbf{C} = \mathbf{F}_\mathrm{p}^\mathrm{T}\mathbf{C}_\mathrm{e}\mathbf{F}_\mathrm{p} \tag{15}$$

resulting from (3) and (13), the material time derivative of $\mathbf{C}$ can be expressed as

$$\dot{\mathbf{C}} = \mathbf{F}_\mathrm{p}^\mathrm{T}\dot{\mathbf{C}}_\mathrm{e}\mathbf{F}_\mathrm{p} + \dot{\mathbf{F}}_\mathrm{p}^\mathrm{T}\mathbf{C}_\mathrm{e}\mathbf{F}_\mathrm{p} + \mathbf{F}_\mathrm{p}^\mathrm{T}\mathbf{C}_\mathrm{e}\dot{\mathbf{F}}_\mathrm{p}. \tag{16}$$

Thus, the dissipation inequality (11) takes the form

$$\mathscr{D} = \left(\mathbf{F}_\mathrm{p}\mathbf{S}\mathbf{F}_\mathrm{p}^\mathrm{T} - 2\frac{\partial \psi}{\partial \mathbf{C}_\mathrm{e}}\right) : \frac{1}{2}\dot{\mathbf{C}}_\mathrm{e} + \mathbf{S} : \left(\mathbf{F}_\mathrm{p}^\mathrm{T}\mathbf{C}_\mathrm{e}\dot{\mathbf{F}}_\mathrm{p}\right) \geq 0. \tag{17}$$

The first term in the expression of the dissipation (17) depends solely on the elastic strain rate, while the second term on the plastic strain rate. Since the elastic and plastic strain rates are independent of each other, the dissipation inequality (17) requires that

$$\mathbf{S} = 2\mathbf{F}_\mathrm{p}^{-1}\frac{\partial \psi}{\partial \mathbf{C}_\mathrm{e}}\mathbf{F}_\mathrm{p}^{-\mathrm{T}} = 2\frac{\partial \psi}{\partial \mathbf{C}}. \tag{18}$$

This leads to the so-called reduced dissipation inequality

$$\mathscr{D} = \boldsymbol{\Sigma} : \mathbf{L}_\mathrm{p} \geq 0, \tag{19}$$

where $\mathbf{L}_\mathrm{p} = \dot{\mathbf{F}}_\mathrm{p}\mathbf{F}_\mathrm{p}^{-1}$ denotes the plastic velocity gradient and

$$\boldsymbol{\Sigma} = \mathbf{F}_\mathrm{e}^\mathrm{T}\boldsymbol{\tau}\mathbf{F}_\mathrm{e}^{-\mathrm{T}} \tag{20}$$

is Mandel's stress tensor [1]. Among all admissible processes the real one maximizes the dissipation (19). This statement is based on the postulate of maximum plastic dissipation (see, e.g., [16]). According to the converse Kuhn-Tucker theorem (see, e.g., [17]) sufficient conditions for this maximum can be written as

$$\mathbf{L}_\mathrm{p} = \dot{\zeta}\frac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\Sigma}}, \quad \dot{\zeta} \geq 0, \quad \dot{\zeta}\boldsymbol{\Phi} = 0, \quad \boldsymbol{\Phi} \leq 0, \tag{21}$$

where $\boldsymbol{\Phi}$ represents a convex yield function and $\dot{\zeta}$ stands for a consistency parameter. In what follows we deal with an ideal-plastic isotropic material described by the von Mises yield function

$$\boldsymbol{\Phi} = \|\mathrm{dev}\boldsymbol{\tau}\| - \sqrt{\frac{2}{3}}\sigma_\mathrm{Y} = \sqrt{\mathrm{dev}\boldsymbol{\Sigma} : \mathrm{dev}\boldsymbol{\Sigma}^\mathrm{T}} - \sqrt{\frac{2}{3}}\sigma_\mathrm{Y}, \tag{22}$$

where σ_Y denotes the normal yield stress. Accordingly,

$$\mathbf{L}_\mathrm{p} = \dot{\zeta}\frac{\mathrm{dev}\boldsymbol{\Sigma}^\mathrm{T}}{\sqrt{\mathrm{dev}\boldsymbol{\Sigma} : \mathrm{dev}\boldsymbol{\Sigma}^\mathrm{T}}}. \tag{23}$$

Now, we specify the evolution equation (23) for small elastic (but large plastic) strains. In this case one may set $\mathbf{F}_{\mathrm{e}} = \mathbf{I}$, such that Mandel's stress tensor (20) becomes symmetric

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\mathrm{T}} = \boldsymbol{\tau}. \tag{24}$$

Thus, the reduced dissipation inequality (19) can be given as

$$\mathscr{D} = \boldsymbol{\Sigma} : \mathbf{D}_{\mathrm{p}} \geq 0 \tag{25}$$

in terms of the plastic rate-of-deformation tensor

$$\mathbf{D}_{\mathrm{p}} = \frac{1}{2}\left(\mathbf{L}_{\mathrm{p}} + \mathbf{L}_{\mathrm{p}}^{\mathrm{T}}\right). \tag{26}$$

In view of (22), (24) and (25) we can further write

$$\mathbf{D}_{\mathrm{p}} = \dot{\zeta}\frac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\Sigma}} = \dot{\zeta}\frac{\mathrm{dev}\boldsymbol{\tau}}{\|\mathrm{dev}\boldsymbol{\tau}\|}. \tag{27}$$

Taking the quadratic norm on the left- and right-hand sides of this equation shows that $\dot{\zeta} = \|\mathbf{D}_{\mathrm{p}}\|$. Thus, under consideration of the yield criterion $\boldsymbol{\Phi} = 0$ applied to (22), we obtain

$$\mathrm{dev}\boldsymbol{\tau} = \sqrt{\frac{2}{3}}\boldsymbol{\sigma}_{\mathrm{Y}}\frac{\mathbf{D}_{\mathrm{p}}}{\|\mathbf{D}_{\mathrm{p}}\|}. \tag{28}$$

This is a remarkable relation since it defines the deviator of the Kirchhoff stress tensor in terms of the plastic deformation rate *independent of the elastic material law*. In the case of simple shear we can further write:

$$\mathbf{F} = \begin{bmatrix} 1 & \gamma & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{e}_i \otimes \mathbf{e}_j, \quad \mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij,}$$

$$\mathbf{D}_{\mathrm{p}} = \mathbf{D} = \frac{1}{2}\left(\mathbf{L} + \mathbf{L}^{\mathrm{T}}\right) = \frac{1}{2}\begin{bmatrix} 0 & \dot{\gamma} & 0 \\ \dot{\gamma} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{e}_i \otimes \mathbf{e}_j, \tag{29}$$

where γ denotes the amount of shear. With consideration of (28) this leads to the classical constant shear stress response

$$\mathrm{dev}\boldsymbol{\tau} = \tau_{\mathrm{Y}}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{e}_i \otimes \mathbf{e}_j, \tag{30}$$

where $\tau_{\mathrm{Y}} = \sigma_{\mathrm{Y}} \Big/ \sqrt{3}$ denotes the shear yield stress.

4 Additive decomposition of the generalized strain measures (model B)

In this section we assume the additive decomposition of a generalized strain measure into an elastic part $\mathbf{E}_\mathrm{e}$ and a plastic part $\mathbf{E}_\mathrm{p}$,

$$\mathbf{E} = \mathbf{E}_\mathrm{e} + \mathbf{E}_\mathrm{p}. \tag{31}$$

Thus, the dissipation inequality (10) can be written as

$$\mathscr{D} = \left(\mathbf{T} - \frac{\partial \psi}{\partial \mathbf{E}_\mathrm{e}} \right) : \dot{\mathbf{E}}_\mathrm{e} + \mathbf{T} : \dot{\mathbf{E}}_\mathrm{p} \geq 0, \tag{32}$$

where $\psi = \check{\psi}(\mathbf{E}_\mathrm{e})$. By the same reasoning as in the previous section we obtain the constitutive relation

$$\mathbf{T} = \frac{\partial \check{\psi}(\mathbf{E}_\mathrm{e})}{\partial \mathbf{E}_\mathrm{e}} = \frac{\partial \check{\psi}(\mathbf{E} - \mathbf{E}_\mathrm{p})}{\partial \mathbf{E}} \tag{33}$$

and the reduced dissipation inequality

$$\mathscr{D} = \mathbf{T} : \dot{\mathbf{E}}_\mathrm{p} \geq 0, \tag{34}$$

which immediately leads to the evolution equation for the plastic strain rate as

$$\dot{\mathbf{E}}_\mathrm{p} = \dot{\zeta} \frac{\partial \Phi}{\partial \mathbf{T}}. \tag{35}$$

Equation (35) forces Φ to be a function of the stress tensor $\mathbf{T}$. Generally, a yield function formulated in terms of the stress work-conjugate to the elastic strain measure is a natural requirement of the thermodynamically based plasticity.

The Mises-type yield function (22) written in terms of the stress tensor $\mathbf{T}$ takes the form (see [10–12])

$$\Phi = \|\mathrm{dev}\mathbf{T}\| - \sqrt{\frac{2}{3}}\sigma_\mathrm{Y}. \tag{36}$$

As with (28) we thus obtain

$$\mathrm{dev}\mathbf{T} = \sqrt{\frac{2}{3}}\sigma_\mathrm{Y} \frac{\dot{\mathbf{E}}_\mathrm{p}}{\left\| \dot{\mathbf{E}}_\mathrm{p} \right\|}. \tag{37}$$

For small elastic (but large plastic) strains one may set $\mathbf{E}_\mathrm{e} = \mathbf{0}$. Thus, in the case of pairwise distinct principal stretches $\lambda_1 \neq \lambda_2 \neq \lambda_3 \neq \lambda_1$, the plastic strain rate can be given in view of (7) and (31) as

$$\dot{\mathbf{E}}_\mathrm{p} = \dot{\mathbf{E}} = \sum_i^3 f'(\lambda_i)\, \dot{\lambda}_i \mathbf{P}_i + \sum_i^3 f(\lambda_i)\, \dot{\mathbf{P}}_i. \tag{38}$$

Further, we specify our solution for simple shear loading. Thereby, the eigenvalues of the right Cauchy-Green tensor

$$\mathbf{C} = \begin{bmatrix} 1 & \gamma & 0 \\ \gamma & 1+\gamma^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{e}_i \otimes \mathbf{e}_j \tag{39}$$

take the form

$$\lambda_{1,2}^2 = 1 + \frac{\gamma^2 \pm \sqrt{4\gamma^2+\gamma^4}}{2}, \quad \lambda_3^2 = 1. \tag{40}$$

The relation (37) primarily requires that

$$\mathrm{tr}\dot{\mathbf{E}}_{\mathrm{p}} = 0. \tag{41}$$

Thus, in view of the identities $\mathrm{tr}\mathbf{P}_i = 1$ and $\mathrm{tr}\dot{\mathbf{P}}_i = 0$ $(i = 1, 2, 3)$ and with the aid of (38), we can write

$$\sum_i^3 f'(\lambda_i)\,\dot{\lambda}_i = 0. \tag{42}$$

By virtue of (40) this leads to the algebraic equation

$$f'(\lambda) - f'\left(\lambda^{-1}\right)\lambda^{-2} = 0 \quad \forall \lambda > 0, \tag{43}$$

where we set $\lambda_1 = \lambda$ and consequently $\lambda_2 = \lambda^{-1}$. With the aid of the Laurent series expansion

$$f'(\lambda) = \sum_{k=-\infty}^{\infty} a_k \lambda^k, \quad f'\left(\lambda^{-1}\right)\lambda^{-2} = \sum_{k=-\infty}^{\infty} a_k \lambda^{-k-2} \tag{44}$$

and in view of (8), the solutions of (43) can be given as

$$f_r(\lambda) = \begin{cases} \dfrac{1}{2r}\left(\lambda^r - \lambda^{-r}\right) & \text{for } r \neq 0\,, \\ \ln\lambda & \text{for } r = 0\,. \end{cases} \tag{45}$$

The functions f_r (45) yield the generalized strain measures

$$\mathbf{E}^{(r)} = \begin{cases} \dfrac{1}{2r}\left(\mathbf{U}^r - \mathbf{U}^{-r}\right) = \dfrac{1}{2r}\left(\mathbf{C}^{r/2} - \mathbf{C}^{-r/2}\right) & \text{for } r \neq 0\,, \\ \ln\mathbf{U} = \dfrac{1}{2}\ln\mathbf{C} & \text{for } r = 0\,, \end{cases} \tag{46}$$

among which only the logarithmic one $(r = 0)$ belongs to Seth's family (9). Henceforth, we deal only with the generalized strains (46) as able to provide the traceless deformation rate (41). For these strains (37) takes the form

$$\mathrm{dev}\mathbf{T}^{(r)} = \sqrt{\frac{2}{3}}\sigma_{\mathrm{Y}} \frac{\dot{\mathbf{E}}^{(r)}}{\left\|\dot{\mathbf{E}}^{(r)}\right\|}, \tag{47}$$

where $\mathbf{T}^{(r)}$ denotes the stress tensor work-conjugate to $\mathbf{E}^{(r)}$. Note that $\mathbf{T}^{(r)}$ itself has no physical meaning and should be transformed to the Cauchy stresses. In the case of the incompressible deformations which we deal with, the Cauchy stress tensor coincides with the Kirchhoff stress tensor given by

$$\boldsymbol{\tau} = \mathbf{F}\mathbf{S}\mathbf{F}^{\mathrm{T}} = 2\mathbf{F}\frac{\partial \psi}{\partial \mathbf{C}}\mathbf{F}^{\mathrm{T}} = \mathbf{F}\left[\frac{\partial \psi}{\partial \mathbf{E}^{(r)}} : 2\frac{\partial \mathbf{E}^{(r)}}{\partial \mathbf{C}}\right]\mathbf{F}^{\mathrm{T}} = \mathbf{F}\left[\mathbf{T}^{(r)} : \mathsf{P}_r\right]\mathbf{F}^{\mathrm{T}}, \quad (48)$$

where

$$\mathsf{P}_r = 2\frac{\partial \mathbf{E}^{(r)}}{\partial \mathbf{C}} \quad (49)$$

denotes a projection tensor of the fourth order. With the aid of the relation

$$\mathsf{P}_r : \mathbf{I} = 2\,\frac{d}{ds}\mathbf{E}^{(r)}\left(\mathbf{C} + s\mathbf{I}\right)\Big|_{s=0} = 2\frac{d}{ds}\sum_i^3 f_r\left(\sqrt{\lambda_i^2 + s}\right)\mathbf{P}_i\Big|_{s=0}$$

$$= \sum_i^3 f_r'\left(\lambda_i\right)\lambda_i^{-1}\mathbf{P}_i \quad (50)$$

and in view of (45) one gets the identities

$$\mathsf{P}_r : \mathbf{I} = \frac{1}{2}\left(\mathbf{C}^{r/2-1} + \mathbf{C}^{-r/2-1}\right), \quad \mathbf{F}\left[\mathsf{P}_r : \mathbf{I}\right]\mathbf{F}^{\mathrm{T}} = \frac{1}{2}\left(\mathbf{b}^{r/2} + \mathbf{b}^{-r/2}\right), \quad (51)$$

where $\mathbf{b} = \mathbf{F}\mathbf{F}^{\mathrm{T}}$ denotes the left Cauchy-Green tensor. Thus, $\boldsymbol{\tau}$ (48) takes the form

$$\boldsymbol{\tau} = \mathbf{F}\left[\mathsf{P}_r : \mathrm{dev}\mathbf{T}^{(r)}\right]\mathbf{F}^{\mathrm{T}} + \hat{\boldsymbol{\tau}} \quad (52)$$

with the abbreviation

$$\hat{\boldsymbol{\tau}} = \frac{1}{6}\mathrm{tr}\mathbf{T}^{(r)}\left(\mathbf{b}^{r/2} + \mathbf{b}^{-r/2}\right). \quad (53)$$

Since the tensors $\mathbf{b}^{r/2}$ and $\mathbf{b}^{-r/2}$ are coaxial we can represent $\hat{\boldsymbol{\tau}}$ (53) in the spectral form as

$$\hat{\boldsymbol{\tau}} = \frac{1}{6}\mathrm{tr}\mathbf{T}^{(r)}\left[\left(\lambda^r + \lambda^{-r}\right)\left(\mathbf{p}_1 + \mathbf{p}_2\right) + 2\mathbf{p}_3\right], \quad (54)$$

where λ is given by $(40)_1$ and $\mathbf{p}_i$ $(i = 1, 2, 3)$ denote the eigenprojections of $\mathbf{b}$. Thus, in the shear plane 1-2, $\hat{\boldsymbol{\tau}}$ has the double eigenvalue $\frac{1}{6}\mathrm{tr}\mathbf{T}^{(r)}\left(\lambda^r + \lambda^{-r}\right)$ and causes equibiaxial tension or compression. Hence, in this plane the stress tensor $\hat{\boldsymbol{\tau}}$ (53) is shear free and does not influence the shear stress response. Inserting (47) into (52) and taking (39) and (49) into account, we finally obtain

$$\boldsymbol{\tau} = \sqrt{\frac{2}{3}}\sigma_{\mathrm{Y}}\mathbf{F}\left[\frac{\mathsf{P}_r : \mathbf{A} : \mathsf{P}_r}{\|\mathsf{P}_r : \mathbf{A}\|}\right]\mathbf{F}^{\mathrm{T}} + \hat{\boldsymbol{\tau}}, \quad (55)$$

where

$$\mathbf{A} = \frac{1}{2\dot{\gamma}}\dot{\mathbf{C}} = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & \gamma & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{e}_i \otimes \mathbf{e}_j. \tag{56}$$

The projection operator P_r (49) appearing in (55) can be expressed by means of the closed-form basis-free solution for the derivative of an isotropic tensor function (see, e.g., [18–21]). Accordingly, for all second-order tensors $\mathbf{X}$,

$$\mathsf{P}_r\colon \mathbf{X} = 2\sum_{q,p=0}^{2} \eta_{qp}\mathbf{C}^q\mathbf{X}\mathbf{C}^p, \tag{57}$$

where the coefficients $\eta_{qp}\,(q,\, p = 0, 1, 2)$ are given as:

$$\eta_{00} = \sum_{i}^{3} \frac{\lambda_j^4\lambda_k^4 f_r'(\lambda_i)}{2\lambda_i D_i^2} - \sum_{i,j\neq i}^{3} \frac{\lambda_i^2\lambda_j^2\lambda_k^4\left[f_r(\lambda_i) - f_r(\lambda_j)\right]}{\left(\lambda_i^2 - \lambda_j^2\right)^3 D_k},$$

$$\eta_{01} = \eta_{10} = -\sum_{i}^{3} \frac{\left(\lambda_j^2 + \lambda_k^2\right)\lambda_j^2\lambda_k^2 f_r'(\lambda_i)}{2\lambda_i D_i^2} + \sum_{i,j\neq i}^{3} \frac{\left(\lambda_j^2 + \lambda_k^2\right)\lambda_i^2\lambda_k^2\left[f_r(\lambda_i) - f_r(\lambda_j)\right]}{\left(\lambda_i^2 - \lambda_j^2\right)^3 D_k},$$

$$\eta_{02} = \eta_{20} = \sum_{i}^{3} \frac{\lambda_j^2\lambda_k^2 f_r'(\lambda_i)}{2\lambda_i D_i^2} - \sum_{i,j\neq i}^{3} \frac{\lambda_i^2\lambda_k^2\left[f_r(\lambda_i) - f_r(\lambda_j)\right]}{\left(\lambda_i^2 - \lambda_j^2\right)^3 D_k},$$

$$\eta_{11} = \sum_{i}^{3} \frac{\left(\lambda_j^2 + \lambda_k^2\right)^2 f_r'(\lambda_i)}{2\lambda_i D_i^2} - \sum_{i,j\neq i}^{3} \frac{\left(\lambda_j^2 + \lambda_k^2\right)\left(\lambda_i^2 + \lambda_k^2\right)\left[f_r(\lambda_i) - f_r(\lambda_j)\right]}{\left(\lambda_i^2 - \lambda_j^2\right)^3 D_k},$$

$$\eta_{12} = \eta_{21} = -\sum_{i}^{3} \frac{\left(\lambda_j^2 + \lambda_k^2\right) f_r'(\lambda_i)}{2\lambda_i D_i^2} + \sum_{i,j\neq i}^{3} \frac{\left(\lambda_i^2 + \lambda_k^2\right)\left[f_r(\lambda_i) - f_r(\lambda_j)\right]}{\left(\lambda_i^2 - \lambda_j^2\right)^3 D_k},$$

$$\eta_{22} = \sum_{i}^{3} \frac{f_r'(\lambda_i)}{2\lambda_i D_i^2} - \sum_{i,j\neq i}^{3} \frac{f_r(\lambda_i) - f_r(\lambda_j)}{\left(\lambda_i^2 - \lambda_j^2\right)^3 D_k}, \quad i \neq j \neq k \neq i, \tag{58}$$

and

$$D_i = \left(\lambda_i^2 - \lambda_j^2\right)\left(\lambda_i^2 - \lambda_k^2\right), \quad i \neq j \neq k \neq i = 1, 2, 3. \tag{59}$$

Of particular interest is the shear stress as a function of the amount of shear. After algebraic manipulations with the equations (55-59) (performed with the aid of MAPLE) we obtain

$$\frac{\tau^{12}}{\tau_{\mathrm{Y}}} = \frac{2\sqrt{\left(4+\gamma^2\right)\Gamma^2 {f_r'}^2\left(\Gamma\right)+4f_r^2\left(\Gamma\right)}}{4+\gamma^2}, \tag{60}$$

where

$$\Gamma = \frac{\gamma}{2} + \frac{\sqrt{4+\gamma^2}}{2} \tag{61}$$

and the functions f_r are given by (45).

The formula (60) is illustrated graphically in Fig. 1. For all represented values of r a non-constant shear stress behavior is observable.

5 Discussion of results

We have examined and compared shear stress responses resulting from the multiplicative decomposition of the deformation gradient (model A) and the additive decomposition of the generalized strain measures (model B) in finite simple shear. Both models are applied to an ideal plastic material described by a von Mises-type yield criterion. Assuming small elastic but large plastic deformations (rigid-plastic material) we have obtained analytical solutions for both models. These solutions are valid *independent of the elastic material law* so that a particular elastic strain energy function or elastic material symmetry need not be specified. Since both models include neither a hardening nor a softening law, a constant shear stress response, even for large amounts of shear, is expected. Indeed, model A delivers true constant shear stress response. Examining model B, we have first seen that only one particular family of generalized strain measures (46) including the logarithmic one is able to provide the deviatoric deformation rate required by the pressure-independent yield criterion. However, even for these strain measures, model B exhibits a non-constant shear stress response (see Fig. 1). This restricts the applicability of this model to moderate plastic shears. Indeed, in the vicinity of the point $\gamma = 0$, the power series expansion of (60) takes the form

$$\frac{\tau^{12}}{\tau_{\mathrm{Y}}} = 1 + \frac{1}{4}r^2\gamma^2 + \left(\frac{1}{16}r^4 - \frac{3}{4}r^2 - 1\right)\gamma^4 + O\left(\gamma^6\right). \tag{62}$$

Thus, in the case of simple shear, the amount of shear is limited for the logarithmic strain $(r = 0)$ by $\gamma^4 \ll 1$ and for other generalized strain measures by $\gamma^2 \ll 1$.

It is seen that at moderate plastic shears the logarithmic strain is most appropriate for model B. Furthermore, in some loading cases, model B based on the logarithmic strain can be shown to coincide with model A. For an illustration we consider a special case in which the principal axes of the right Cauchy-Green tensor do not rotate during the deformation such that

$$\dot{\mathbf{P}}_i = \mathbf{0}, \quad i = 1, 2, 3. \tag{63}$$

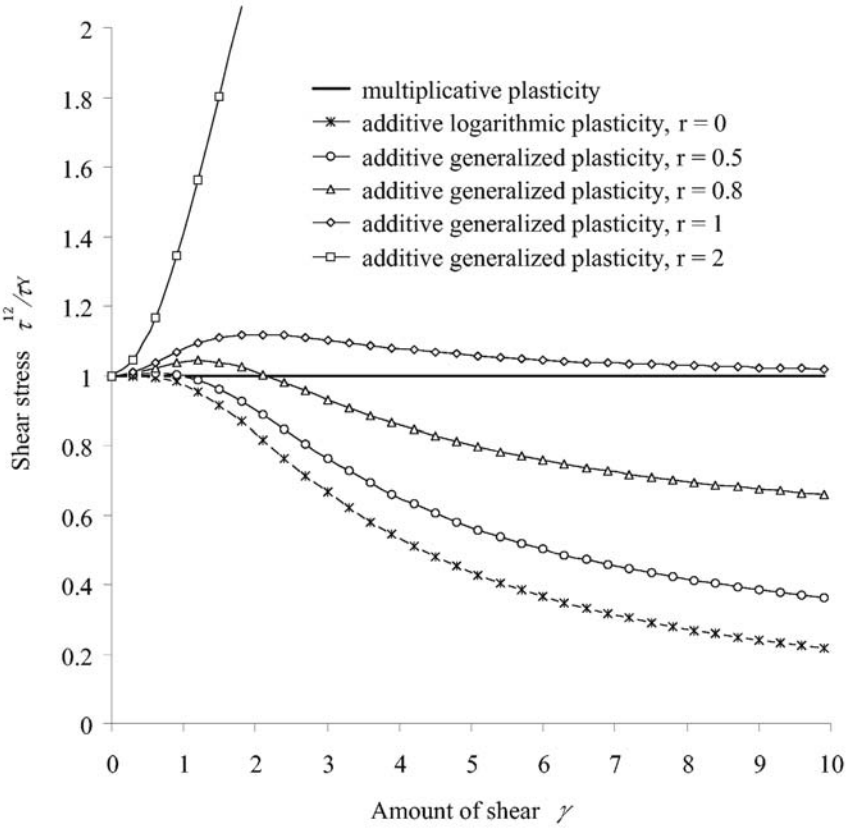


Fig. 1. Simple shear of an ideal plastic material: comparison between shear stress responses for the multiplicative decomposition of the deformation gradient and the additive decomposition of the generalized strain measures (46)

Such deformations take place, for example, under uniaxial loading. For the logarithmic strain rate the relation (38) yields in this case

$$\dot{\mathbf{E}}_{\mathrm{p}} = \dot{\mathbf{E}}^{(0)} = \sum_{i}^{3} \dot{\lambda}_i \lambda_i^{-1} \mathbf{P}_i. \tag{64}$$

With the aid of the expression (see, e.g., [13])

$$\mathbf{S} = \sum_{i}^{m} \frac{f'(\lambda_i)}{\lambda_i} \mathbf{P}_i \mathbf{T} \mathbf{P}_i + 2 \sum_{i,j \neq i}^{m} \frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i^2 - \lambda_j^2} \mathbf{P}_i \mathbf{T} \mathbf{P}_j \tag{65}$$

and in view of (5), (12), (47) and (64) we further obtain

$$\mathbf{T}^{(0)} = \mathbf{U}\mathbf{S}\mathbf{U} = \mathbf{R}^{\mathrm{T}} \boldsymbol{\tau} \mathbf{R}, \tag{66}$$

which implies the identity of the yield functions (22) and (36),

$$\boldsymbol{\Phi} = \left\| \mathrm{dev}\mathbf{T}^{(0)} \right\| - \sqrt{\frac{2}{3}}\boldsymbol{\sigma}_{\mathrm{Y}} \equiv \left\| \mathrm{dev}\boldsymbol{\tau} \right\| - \sqrt{\frac{2}{3}}\boldsymbol{\sigma}_{\mathrm{Y}}. \tag{67}$$

Thus, in the case of small elastic but large plastic strains, model B yields, by virtue of (47) and (66),

$$\mathrm{dev}\boldsymbol{\tau} = \mathbf{R}\left(\mathrm{dev}\mathbf{T}^{(0)}\right)\mathbf{R}^{\mathrm{T}} = \sqrt{\frac{2}{3}}\boldsymbol{\sigma}_{\mathrm{Y}}\frac{\mathbf{R}\dot{\mathbf{E}}^{(0)}\mathbf{R}^{\mathrm{T}}}{\left\|\dot{\mathbf{E}}^{(0)}\right\|}. \tag{68}$$

By virtue of the relation

$$\mathbf{D} = \frac{1}{2}\mathbf{R}\left(\dot{\mathbf{U}}\mathbf{U}^{-1} + \mathbf{U}^{-1}\dot{\mathbf{U}}\right)\mathbf{R}^{\mathrm{T}} = \mathbf{R}\dot{\mathbf{E}}^{(0)}\mathbf{R}^{\mathrm{T}} \tag{69}$$

the same result follows directly from (28) for model A as well. Thus, we observe that, under the additional condition (63), both plasticity models deliver the same stress response.

We conclude with remarks concerning the assumption of small elastic strains. This assumption is not restrictive as regards engineering practice. Indeed, in many engineering problems, as, for example, in metal forming processes, large plastic deformations are accompanied by small elastic strains. Further, since model B exhibits incorrect shear stress response already at small elastic strains, its applicability for large elastic and large plastic shears may likewise be questioned. A similar argument may also be applied in respect of the restriction to the isotropic yield function. Indeed, since model B fails already in the isotropic case, its application to anisotropic materials, at least at large plastic shears, is disputable.

References

[1] Mandel, J. (1972): Plasticité classique et viscoplasticité. (CISM course no. 97). Springer, Vienna

[2] Aravas, N. (1992): Finite elastoplastic transformations of transversely isotropic metals. Internat. J. Solids Structures **29**, 2137–2157

[3] Cleja-Ţigoiu, S. (2000): Nonlinear elasto-plastic deformations of transversely isotropic material and plastic spin. Internat. J. Engrg. Sci. **38**, 737–763

[4] Häusler, O., Schick, D., Tsakmakis, Ch. (2004): Description of plastic anisotropy effects at large deformations. II. The case of transverse isotropy. Internat. J. Plasticity **20**, 199–223

[5] Seth, B.R. (1964): Generalized strain measure with applications to physical problems. In: Reiner, M., Abir, D. (eds.): Second-order effects in elasticity, plasticity and fluid dynamics. Jerusalem Academic Press, Jerusalem, pp. 162–172

[6] Hill, R. (1968): On constitutive inequalities for simple materials. I. J. Mech. Phys. Solids **16**, 229–242

[7] Ogden, R.W. (1984): Nonlinear elastic deformations. Ellis Horwood, Chichester

[8] Papadopoulos, P., Lu, J. (1998): A general framework for the numerical solution of problems in finite elasto-plasticity. Comput. Methods Appl. Mech. Engrg. **159**, 1–18
[9] Hill, R. (1950): The mathematical theory of plasticity. Clarendon Press, Oxford
[10] Papadopoulos, P., Lu, J. (2001): On the formulation and numerical solution of problems in anisotropic finite plasticity. Comput. Methods Appl. Mech. Engrg. **190**, 4889–4910
[11] Miehe, C., Apel, N., Lambrecht, M. (2002): Anisotropic additive plasticity in the logarithmic strain space: modular kinematic formulation and implementation based on incremental minimization principles for standard materials. Comput. Methods Appl. Mech. Engrg. **191**, 5383–5425
[12] Schröder, J., Gruttmann, F., Löblein, J. (2002): A simple orthotropic finite elasto-plasticity model based on generalized stress-strain measures. Comput. Mech. **30**, 48–64
[13] Itskov, M. (2002): The derivative with respect to a tensor: some theoretical aspects and applications. ZAMM Z. Angew. Math. Mech. **82**, 535–544
[14] Truesdell, C., Noll, W. (1965): The nonlinear field theories of mechanics. In: Flügge, S. (ed.): Handbuch der Physik. Band III/3. Springer, Berlin, pp. 1–602
[15] Lee, E.H. (1969): Elastic-plastic deformation at finite strains. J. Appl. Mech. **36**, 1–6
[16] Lubliner, J. (1990): Plasticity theory. Macmillan, New York
[17] Brousse, P. (1988): Optimization in mechanics: problems and methods. North-Holland, Amsterdam
[18] Carlson, D.E., Hoger, A. (1986): The derivative of a tensor-valued function of a tensor. Quart. Appl. Math. **44**, 409–423
[19] Itskov, M., Aksel, N. (2002): A closed-form representation for the derivative of non-symmetric tensor power series. Internat. J. Solids Structures **39**, 5963–5978
[20] Itskov, M. (2003): Application of the Dunford-Taylor integral to isotropic tensor functions and their derivatives. R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **459**, 1449–1457
[21] Itskov, M. (2003): Computation of the exponential and other isotropic tensor functions and their derivatives. Comput. Meth. Appl. Mech. Engrg. **192**, 3985–3999

A nonlocal formulation of plasticity

F. Marotti de Sciarra, C. Sellitto

Abstract. This paper deals with a formulation of nonlocal plasticity with internal variables. The constitutive model complies with local internal variables which govern kinematic hardening and isotropic softening and with a nonlocal corrective internal variable. The constitutive problem is cast in the framework provided by convex analysis and the potential theory for monotone multivalued operators which provide suitable tools for performing a theoretical analysis of such nonlocal problems. Several variational formulations with different combinations of state variables are provided.

1 Introduction

Most materials usually adopted in engineering show a loss of positive definiteness of the tangent stiffness operator which yields to the localization of plastic deformations in narrow bands until cracks appear.

The deformation pattern in a body in which a localization phenomenon occurs suddenly evolves from relatively smooth into one in which shear bands of highly strained material appear whereas the remaining part of the body unloads.

The nonlocal theory introduces, in the constitutive model, state variables defined in an average form over a finite volume of the body and the material length parameter determines how the value of the variable at a certain point is weighted.

In this paper a nonlocal plasticity model with internal variables is addressed. We introduce local internal variables, which govern kinematic hardening and isotropic softening, and a nonlocal internal variable which is defined as the sum of a new internal variable and its spatial weighted average.

The nonlocal internal variable is added to the local variables, governing isotropic hardening, in the definition of the elastic domain.

Convex analysis and the potential theory for monotone multivalued operators provide suitable tools for performing a theoretical analysis of the nonlocal constitutive problem. The validity of the maximum dissipation theorem is assessed and constitutive variational formulations of the rate model are provided.

The structural problem is then formulated. The nonlocal variational formulation in the complete set of state variables is given and the methodology for deriving variational formulations, with different combinations of the state variables, is explicitly provided. In particular three variational formulations with different combinations of state variables are given.

These formulations show that two different approaches can be followed in order to perform a finite element approximation of the proposed nonlocal plastic model.

In the former approach the expression of the dissipation has to be approximated and in the latter formulation the indicator of the elastic domain can be expressed in terms of plastic multipliers and a predictor-corrector algorithm can be adopted. A discussion of approximation methods and of finite-step nonlocal plasticity deserves further analysis and will be the subject of a forthcoming paper.

2 Nonlocal plasticity

We analyze a nonlocal elastoplastic structural problem defined on a regular bounded domain Ω of a Euclidean space. The inelastic model is cast in the framework of internal variable theories of associated type and the *generalized standard material* (GSM), proposed by Halphen and Nguyen [3], is considered.

The dual spaces of strains ε and stresses σ will be labelled by $\mathcal{D}$ and $\mathcal{S}$ respectively. The internal variables account for the evolution of the hardening/softening phenomena; the kinematic internal variables are denoted by $\kappa \in \mathcal{Y}$, $\alpha_1 \in \mathcal{Y}_1$, $\alpha_2 \in \mathcal{Y}_2$ and the dual static internal variables are $X \in \mathcal{Y}'$, $\chi_1 \in \mathcal{Y}'_1$, $\chi_2 \in \mathcal{Y}'_2$. The symbol $(\!(\cdot,\cdot)\!)$ denotes the inner product in the dual spaces.

The *free energy* is provided by the saddle (convex-concave) differentiable functional $\Phi : \mathcal{D} \times \mathcal{Y} \times \mathcal{Y}_1 \times \mathcal{Y}_2 \mapsto \bar{\Re}$ and the convex elastic domain C is defined in the product space $\mathcal{S} \times \mathcal{Y}' \times \mathcal{Y}'_1 \times \mathcal{Y}'_2$.

The free energy is additively decomposed as the sum of a strictly convex potential $\Phi_e(e)$, representing the elastic energy, and a saddle functional $\Phi_{in}(\alpha_1, \alpha_2, \kappa)$, convex in (α_1, α_2) and concave in κ, which accounts for the inelastic phenomena. Such a decomposition, usually adopted in literature concerning local plasticity [6–8], corresponds to the mechanical assumption that the elastic behavior does not depend on the evolution of inelastic phenomena.

Nonlocal effects can then be modeled by giving the following expression to the free energy:

$$\Phi(e, \alpha_1, \alpha_2, \kappa) = \Phi_e(e) + \Phi_{in}(\alpha_1, \alpha_2, \kappa) = \Phi_e(e) + \Phi_L(\alpha_1, \kappa) + \Phi_{NL}(\alpha_2), \quad (1)$$

where the free energy component $\Phi_L(\alpha_1, \kappa)$ is convex in α_1, concave in κ and $\Phi_{NL}(\alpha_2) = \Phi_{NL}(\xi(\alpha_2))$ is the convex nonlocal part of the free energy.

In fact a nonlocal plastic behavior can be modeled by assuming that the functional Φ_{NL} at a point $\mathbf{x}$ of the body Ω depends on the entire field α_2. This task can be achieved by considering the nonlocal variable $\xi \in \mathcal{X}$ which has the parametric representation

$$\xi(\mathbf{x}) = (\mathbf{R}\alpha_2)(\mathbf{x}), \quad (2)$$

where $\mathbf{R} : \mathcal{Y}_2 \mapsto \mathcal{X}$ denotes a suitable linear regularization operator [10]. The kinematic internal variable ξ turns out to be nonlocal since its value at the point $\mathbf{x}$ of the body Ω depends on the entire field α_2.

A nonlocal field ξ can be obtained as a spatial weighted average of the variable α_2 in the form

$$\xi(\mathbf{x}) = (\mathbf{R}\alpha_2)(\mathbf{x}) = \frac{1}{V_r(\mathbf{x})} \int_\Omega \beta_{\mathbf{x}}(\mathbf{y}) \alpha_2(\mathbf{y}) d\mathbf{y}, \quad V_r(\mathbf{x}) = \int_\Omega \beta_{\mathbf{x}}(\mathbf{y}) d\mathbf{y}, \quad (3)$$

where $\beta_{\mathbf{x}}(\mathbf{y})$ is a spatial weighting function depending on a material parameter l called the internal length scale.

If a linear nonlocal softening behavior is assumed, the expression of Φ_{NL} is then given as

$$\begin{aligned}\Phi_{NL}(\alpha_2) \;=\; \tfrac{1}{2}(\!(\,\hat{h}\xi(\alpha_2),\xi(\alpha_2)\,)\!) \;&=\; \tfrac{1}{2}(\!(\,\hat{h}\mathbf{R}\alpha_2,\mathbf{R}\alpha_2\,)\!) \;= \\ &= \int_\Omega \hat{h}\left[\tfrac{1}{V_r(\mathbf{x})}\int_\Omega \beta_{\mathbf{x}}(\mathbf{y})\alpha_2(\mathbf{y})d\mathbf{y}\right]^2 d\mathbf{x},\end{aligned} \tag{4}$$

where $\hat{h} : \mathcal{Z} \mapsto \mathcal{Z}'$ is positive.

With the expression (2) of the nonlocal variable ξ, the constitutive relations can be obtained from the saddle free energy (1) as:

$$(\sigma,\chi_1,\chi_2,-X) \;=\; d\Phi(e,\alpha_1,\alpha_2,\kappa) \quad\Longleftrightarrow\quad \left\{\begin{array}{rcl}\sigma &=& d\Phi_e(e)\\ \chi_1 &=& d_{\alpha_1}\Phi_L(\alpha_1,\kappa)\\ \chi_2 &=& d_{\alpha_2}\Phi_{NL}(\xi(\alpha_2)) \;=\\ &=& \mathbf{R}'d\Phi_{NL}(\xi) \;=\; \mathbf{R}'\chi\\ -X &=& d_\kappa\Phi_L(\alpha_1,\kappa),\end{array}\right. \tag{5}$$

where $\chi = d\Phi_{NL}(\xi) \in \mathcal{Z}'$ and $\mathbf{R}' : \mathcal{Z}' \mapsto \mathcal{Y}'_2$ denotes the dual operator of $\mathbf{R}$. Assuming the expression (4) for the nonlocal convex part Φ_{NL} of the free energy, we have $\chi_2 = \mathbf{R}'\,\hat{h}\xi(\alpha_2) = (\mathbf{R}'\,\hat{h}\,\mathbf{R})\alpha_2 = h\alpha_2$.

Accordingly the static internal variable χ_2, which is dual of the (local) kinematic internal variable α_2, is defined as

$$\chi_2(\mathbf{x}) = (\mathbf{R}'\chi)(\mathbf{x}) \quad \forall\,\mathbf{x}\in\Omega, \tag{6}$$

and turns out to be a *nonlocal* variable since its pointwise value depends upon the entire field χ over the body Ω.

The constitutive model is completed by introducing the elastic domain C which is defined in the space of stresses and of static internal variables $\{\sigma,\chi_1,\chi_2,X\}$ as the level set of a convex *yield mode* $G : \mathcal{S}\times\mathcal{Y}'_1\times\mathcal{Y}'_2\times\mathcal{Y}' \mapsto \mathfrak{R}\cup\{+\infty\}$ in the form

$$C \;=\; \{(\sigma,\chi_1,\chi_2,X)\in\mathcal{S}\times\mathcal{Y}'_1\times\mathcal{Y}'_2\times\mathcal{Y}' \;:\; G(\sigma,\chi_1,\chi_2,X)\le 0\}\,, \tag{7}$$

provided that the minimum of G is negative.

The nonlocal elastoplastic constitutive model can be formulated in a more convenient way by defining the following generalized variables collecting together local and nonlocal variables:

$$\underset{\sim}{\boldsymbol{\varepsilon}} = \begin{bmatrix}\varepsilon\\0\\0\\0\end{bmatrix} = \begin{bmatrix}\boldsymbol{\varepsilon}\\0\end{bmatrix} \quad \underset{\sim}{\mathbf{e}} = \begin{bmatrix}e\\\alpha_1\\\alpha_2\\\kappa\end{bmatrix} = \begin{bmatrix}\mathbf{e}\\\kappa\end{bmatrix} \quad \underset{\sim}{\mathbf{p}} = \begin{bmatrix}p\\-\alpha_1\\-\alpha_2\\\kappa\end{bmatrix} = \begin{bmatrix}\mathbf{p}\\\kappa\end{bmatrix} \quad \underset{\sim}{\boldsymbol{\sigma}} = \begin{bmatrix}\sigma\\\chi_1\\\chi_2\\X\end{bmatrix} = \begin{bmatrix}\boldsymbol{\sigma}\\X\end{bmatrix}. \tag{8}$$

The vectors $\underset{\sim}{\boldsymbol{\varepsilon}}$, $\underset{\sim}{\mathbf{e}}$, $\underset{\sim}{\mathbf{p}}$ and $\underset{\sim}{\boldsymbol{\sigma}}$ represent the generalized vectors of total strain, elastic strain, plastic strain and stress. Accordingly two generalized spaces are introduced:

$$\underset{\sim}{\mathcal{D}} = \mathcal{D} \times \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Y} = \bar{\mathcal{D}} \times \mathcal{Y}, \qquad \underset{\sim}{\mathcal{S}} = \mathcal{S} \times \mathcal{Y}'_1 \times \mathcal{Y}'_2 \times \mathcal{Y}' = \bar{\mathcal{S}} \times \mathcal{Y}', \quad (9)$$

and the scalar product between generalized vectors is denoted by the symbol $\prec \cdot, \cdot \succ$ defined as

$$\prec \underset{\sim}{\boldsymbol{\sigma}}, \underset{\sim}{\mathbf{e}} \succ = (\!(\boldsymbol{\sigma}, \mathbf{e})\!) + (\!(X, \boldsymbol{\kappa})\!) = (\!(\sigma, e)\!) + (\!(\chi_1, \alpha_1)\!) + (\!(\chi_2, \alpha_2)\!) + (\!(X, \kappa)\!). \quad (10)$$

In the sequel, for simplicity, the term generalized is omitted since no confution can arise.

3 The elastic domain

In the applications the yield mode, defining the elastic domain C given by (7), is usually written in the form

$$G(\underset{\sim}{\boldsymbol{\sigma}}) = G(\boldsymbol{\sigma}, X) = G(\sigma, \chi_1, \chi_2, X) = g(\sigma, \chi_1) - \chi_2 - X - \sigma_o, \quad (11)$$

where g is a convex function and σ_o represents a constant scalar value which characterizes the initial yield limit.

The flow rule can be formulated in terms of the normal cone N_C to the elastic domain C as follows:

$$\dot{\underset{\sim}{\mathbf{p}}} \in N_C(\underset{\sim}{\boldsymbol{\sigma}}) = \partial \sqcup_C(\underset{\sim}{\boldsymbol{\sigma}}) \iff (\dot{p}, -\dot{\alpha}_1, -\dot{\alpha}_2, \dot{k}) \in N_C(\sigma, \chi_1, \chi_2, X), \quad (12)$$

and can be reformulated in three equivalent forms:

$$\dot{\underset{\sim}{\mathbf{p}}} \in N_C(\underset{\sim}{\boldsymbol{\sigma}}), \quad \underset{\sim}{\boldsymbol{\sigma}} \in \partial D(\dot{\underset{\sim}{\mathbf{p}}}), \quad \sqcup_C(\underset{\sim}{\boldsymbol{\sigma}}) + D(\dot{\underset{\sim}{\mathbf{p}}}) = \prec \underset{\sim}{\boldsymbol{\sigma}}, \dot{\underset{\sim}{\mathbf{p}}} \succ, \quad (13)$$

where $D : \underset{\sim}{\mathcal{D}} \mapsto \Re \cup \{+\infty\}$ is the support functional of the elastic domain C defined by

$$\begin{aligned} D(\dot{\underset{\sim}{\mathbf{p}}}) &= \sup\{\prec \underset{\sim}{\boldsymbol{\tau}}, \dot{\underset{\sim}{\mathbf{p}}} \succ \mid \underset{\sim}{\boldsymbol{\tau}} \in C\} = \\ &= \sup\{(\!(\bar{\sigma}, \dot{p})\!) - (\!(\bar{\chi}_1, \dot{\alpha}_1)\!) - (\!(\bar{\chi}_2, \dot{\alpha}_2)\!) + (\!(\bar{X}, \dot{k})\!) | (\bar{\sigma}, \bar{\chi}_1, \bar{\chi}_2, \bar{X}) \in C\}. \end{aligned} \quad (14)$$

The functional D has the physical meaning of dissipation associated with the plastic flow $\dot{\underset{\sim}{\mathbf{p}}}$.

Explicitly the relations (13) can be rewritten as

$$\begin{cases} (\dot{p}, -\dot{\alpha}_1, -\dot{\alpha}_2, \dot{k}) \in N_C(\sigma, \chi_1, \chi_2, X), \\ (\sigma, \chi_1, \chi_2, X) \in \partial D(\dot{p}, -\dot{\alpha}_1, -\dot{\alpha}_2, \dot{k}), \\ \sqcup_C(\sigma, \chi_1, \chi_2, X) + D(\dot{p}, -\dot{\alpha}_1, -\dot{\alpha}_2, \dot{k}) \;=\; (\!(\sigma, \dot{p})\!) - (\!(\chi_1, \dot{\alpha}_1)\!) + \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad -(\!(\chi_2, \dot{\alpha}_2)\!) + (\!(X, \dot{k})\!). \end{cases} \tag{15}$$

It is useful to express the flow rule $(13)_1$ in terms of the plastic multiplier λ. To this end we note that the indicator of the elastic domain C can be expressed in terms of the non-positive scalars $\mathfrak{R}^-$ as

$$\sqcup_C(\underset{\sim}{\boldsymbol{\sigma}}) \;=\; \sqcup_{\mathfrak{R}^-}[G(\underset{\sim}{\boldsymbol{\sigma}})]. \tag{16}$$

The assumption that G is continuous in C, a subdifferential rule given in [5], yields:

$$\partial \sqcup_C(\underset{\sim}{\boldsymbol{\sigma}}) \;=\; \partial(\sqcup_{\mathfrak{R}^-} \circ G)(\underset{\sim}{\boldsymbol{\sigma}}) \;=\; \partial \sqcup_{\mathfrak{R}^-}[G(\underset{\sim}{\boldsymbol{\sigma}})]\, dG(\underset{\sim}{\boldsymbol{\sigma}}) \qquad \forall\, (\underset{\sim}{\boldsymbol{\sigma}}) \in C. \tag{17}$$

Since the subdifferential of $\sqcup_{\mathfrak{R}^-}[G(\underset{\sim}{\boldsymbol{\sigma}})]$ coincides with the normal cone $N_{\mathfrak{R}^-}[G(\underset{\sim}{\boldsymbol{\sigma}})]$, the flow rule $(13)_1$ can be rewritten in the following equivalent forms:

$$\begin{cases} (\dot{\mathbf{p}}, \dot{k}) \in N_C(\boldsymbol{\sigma}, X), \\ (\dot{\mathbf{p}}, \dot{k}) \in N_{\mathfrak{R}^-}[G(\boldsymbol{\sigma}, X)]\, dG(\boldsymbol{\sigma}, X), \\ (\dot{\mathbf{p}}, \dot{k}) \;=\; \lambda\, dG(\boldsymbol{\sigma}, X) \qquad \text{with} \qquad \lambda \in N_{\mathfrak{R}^-}[G(\boldsymbol{\sigma}, X)], \\ (\dot{\mathbf{p}}, \dot{k}) \;=\; \lambda\, dG(\boldsymbol{\sigma}, X) \qquad \text{with} \qquad \lambda \geq 0,\, G(\boldsymbol{\sigma}, X) \leq 0,\, \lambda G(\boldsymbol{\sigma}, X) = 0. \end{cases} \tag{18}$$

From the expression (11) of the yield mode, the relation $(18)_4$ is explicitly given by

$$\begin{cases} \dot{p} = \lambda\, d_\sigma G(\sigma, \chi_1, \chi_2, X) \;=\; \lambda\, d_\sigma g(\sigma, \chi_1), \\ -\dot{\alpha}_1 = \lambda\, d_{\chi_1} G(\sigma, \chi_1, \chi_2, X) = \lambda\, d_{\chi_1} g(\sigma, \chi_1), \\ -\dot{\alpha}_2 = \lambda\, d_{\chi_2} G(\sigma, \chi_1, \chi_2, X) = -\lambda, \\ \dot{k} = \lambda\, d_X G(\sigma, \chi_1, \chi_2, X) \;=\; -\lambda \end{cases} \tag{19}$$

under the complementarity conditions

$$\lambda \geq 0, \quad g(\sigma, \chi_1) - \chi_2 - X - \sigma_o \leq 0, \quad \lambda[g(\sigma, \chi_1) - \chi_2 - X - \sigma_o] = 0. \tag{20}$$

As a result the rate of the kinematic internal variable α_2 coincides with the plastic multiplier and the kinematic internal variable κ is equal to the opposite of the plastic multiplier. Accordingly the parametric representation of $\dot{\xi}$ is given by $\dot{\xi} = \mathbf{R}\lambda$.

4 The constitutive model

The constitutive model of nonlocal plasticity can be formulated by considering the additivity of strains $\varepsilon = e + p$, the constitutive relations (5) and the flow rule (13) in the form:

$$\begin{cases} \boldsymbol{\varepsilon} = \mathbf{e} + \mathbf{p} & \text{additivity of strains,} \\ (\dot{\mathbf{p}}, \dot{k}) \in N_C(\boldsymbol{\sigma}, X) & \text{flow rule,} \\ (\boldsymbol{\sigma}, -X) = d\Phi(\mathbf{e}, \boldsymbol{\kappa}) & \text{elastic relation.} \end{cases} \tag{21}$$

For a linear elastic and hardening behavior of the type (4), the free energy is given by

$$\begin{aligned} \Phi(\mathbf{e}, \boldsymbol{\kappa}) &= \tfrac{1}{2} \prec \mathbf{H}(e, \alpha_1, \alpha_2, \boldsymbol{\kappa}), (e, \alpha_1, \alpha_2, \boldsymbol{\kappa}) \succ = \\ &= \tfrac{1}{2}(\!(\mathbf{E}e, e)\!) + \tfrac{1}{2}(\!(\mathbf{H}_1\alpha_1, \alpha_1)\!) + \tfrac{1}{2}(\!(\mathbb{H}_2\alpha_2, \alpha_2)\!) + \tfrac{1}{2}(\!(\mathbf{W}\boldsymbol{\kappa}, \boldsymbol{\kappa})\!)\,, \end{aligned} \tag{22}$$

where $\mathbf{H} = \text{diag}\,[\mathbf{E}, \mathbf{H}_1, \mathbb{H}_2, \mathbf{W}]$ denotes the matrix collecting the elastic and hardening/softening moduli. We note that, in the case of the linear nonlocal behavior previously introduced – see (4) – we have $\mathbb{H}_2 = \mathbf{R}'\hat{h}\mathbf{R}$.

In order to derive a variational formulation of the nonlocal elastoplastic model, it is compelling to consider alternative expressions of the free energy. To this end we introduce the conjugate [4] saddle functional $\Phi^* : \underset{\sim}{\mathcal{S}} \mapsto \bar{\mathfrak{R}}$, which represents the *complementary* free energy, defined by

$$\Phi^*(\boldsymbol{\sigma}, X) = \inf_{\boldsymbol{\kappa}} \sup_{\mathbf{e}} \{(\!(\boldsymbol{\sigma}, \mathbf{e})\!) + (\!(X, \boldsymbol{\kappa})\!) - \Phi(\mathbf{e}, \boldsymbol{\kappa})\} \tag{23}$$

and the convex functionals $\Xi : \mathcal{S} \times \mathcal{Y}'_1 \times \mathcal{Y}'_2 \times \mathcal{Y} \mapsto \bar{\mathfrak{R}}$ and $\Xi^* : \mathcal{D} \times \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Y}' \mapsto \bar{\mathfrak{R}}$, associated with the free energy Φ, defined as

$$\begin{aligned} \Xi(\boldsymbol{\sigma}, \boldsymbol{\kappa}) &= -\inf_X \{(\!(X, \boldsymbol{\kappa})\!) - \Phi^*(\boldsymbol{\sigma}, X)\} = \sup_{\mathbf{e}} \{(\!(\boldsymbol{\sigma}, \mathbf{e})\!) - \Phi(\mathbf{e}, \boldsymbol{\kappa})\}, \\ \Xi^*(\mathbf{e}, -X) &= -\inf_{\boldsymbol{\kappa}} \{(\!(X, \boldsymbol{\kappa})\!) - \Phi(\mathbf{e}, \boldsymbol{\kappa})\} = \sup_{\boldsymbol{\sigma}} \{(\!(\boldsymbol{\sigma}, \mathbf{e})\!) - \Phi^*(\boldsymbol{\sigma}, X)\}. \end{aligned} \tag{24}$$

The elastic relation $(21)_3$ can then be rewritten in the following equivalent forms:

$$\begin{cases} (\boldsymbol{\sigma}, -X) = d\Phi(\mathbf{e}, \boldsymbol{\kappa}), & (\mathbf{e}, \boldsymbol{\kappa}) = d\Phi^*(\boldsymbol{\sigma}, -X), \\ (\mathbf{e}, X) = d\Xi(\boldsymbol{\sigma}, \boldsymbol{\kappa}), & (\boldsymbol{\sigma}, \boldsymbol{\kappa}) = d\Xi^*(\mathbf{e}, X) \end{cases} \tag{25}$$

and in terms of Fenchel's equalities:

$$\begin{cases} -\Xi^*(\mathbf{e}, X) + \Phi(\mathbf{e}, \boldsymbol{\kappa}) = -(\!(X, \boldsymbol{\kappa})\!), \\ \Xi(\boldsymbol{\sigma}, \boldsymbol{\kappa}) + \Phi(\mathbf{e}, \boldsymbol{\kappa}) = (\!(\boldsymbol{\sigma}, \mathbf{e})\!), \\ -\Xi(\boldsymbol{\sigma}, \boldsymbol{\kappa}) + \Phi^*(\boldsymbol{\sigma}, -X) = -(\!(X, \boldsymbol{\kappa})\!), \\ \Xi^*(\mathbf{e}, X) + \Phi^*(\boldsymbol{\sigma}, -X) = (\!(\boldsymbol{\sigma}, \mathbf{e})\!), \\ \Xi^*(\mathbf{e}, X) + \Xi(\boldsymbol{\sigma}, \boldsymbol{\kappa}) = (\!(\boldsymbol{\sigma}, \mathbf{e})\!) + (\!(X, \boldsymbol{\kappa})\!), \\ \Phi(\mathbf{e}, k) + \Phi^*(\boldsymbol{\sigma}, -X) = (\!(\boldsymbol{\sigma}, \mathbf{e})\!) - (\!(X, \boldsymbol{\kappa})\!). \end{cases} \tag{26}$$

5 The structural problem for nonlocal plasticity

We now analyze an elastoplastic structural model having a nonlocal constitutive behavior. Displacements are assumed to belong to the Sobolev space $\mathcal{U} = H^m(\Omega)$ of fields which are square integrable in Ω together with their distributional derivatives up to order m [2]. Conforming displacement fields satisfy linear constraint conditions and belong to a closed linear subspace $\mathcal{L} \subset \mathcal{U}$.

The kinematic operator $\mathbf{B} \in \mathrm{Lin}\{\mathcal{U}, \mathcal{D}\}$ is a bounded linear operator from $\mathcal{U}$ to the Hilbert space of square integrable strain fields $\varepsilon \in \mathcal{D}$ [1].

With $\mathcal{F}$ denoting the subspace of external forces, which is the dual of $\mathcal{U}$, the continuous operator $\mathbf{B}' \in \mathrm{Lin}\{\mathcal{S}, \mathcal{F}\}$, the dual of $\mathbf{B}$, is the equilibrium operator. The symbol $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $\mathcal{U}$ and its dual $\mathcal{F}$.

Let $\ell = \{\mathbf{t}, \mathbf{b}\} \in \mathcal{F}$ be the load functional where $\mathbf{t}$ and $\mathbf{b}$ denote the tractions and the body forces. For simplicity, imposed strains and displacements are not considered.

The equilibrium equation between external forces f and stresses σ is

$$f = \mathbf{B}'\sigma, \qquad \sigma \in \mathcal{S},\ f \in \mathcal{F}, \tag{27}$$

and the compatibility condition is

$$\varepsilon = \mathbf{B}u, \qquad u \in \mathcal{U},\ \varepsilon \in \mathcal{D}. \tag{28}$$

The external relation between reactions and displacements is assumed to be given by

$$r \in \partial Y(u), \tag{29}$$

where $Y : \mathcal{U} \mapsto \mathfrak{R} \cup \{-\infty\}$ is a concave functional. Accordingly, the relation between external forces $f = \ell + r$ and displacements is expressed as

$$f \in \ell + \partial Y(u), \quad \text{or equivalently} \quad u \in \partial Y^*(f - \ell), \tag{30}$$

where the concave functional $Y^* : \mathcal{F} \mapsto \mathfrak{R} \cup \{-\infty\}$ represents the conjugate [4] of Y.

Different expressions can be given to the functional $Y(u)$ depending on the type of external constraints.

We now give the expression of Y in the case of external frictionless bilateral constraints with homogeneous boundary conditions. The orthogonal complement of the subspace $\mathcal{L}$ of conforming displacements is denoted by R and provides the subspace of the external constraint reactions. The functional Y and its dual Y^* are thus given in the form

$$Y(u) = \sqcap_{\mathcal{L}}(u) = \begin{cases} 0 & \text{if } u \in \mathcal{L} \\ -\infty & \text{otherwise,} \end{cases} \qquad Y^*(r) = \sqcap_{\mathcal{L}^\perp}(r). \tag{31}$$

Accordingly the relation $r \in \partial Y(u)$ is equivalent to the state $u \in \mathcal{L}$ and $r \in R = \mathcal{L}^\perp$.

We now derive the variational formulation for the nonlocal structural problem in order to provide the structural response of the body Ω to a given load ℓ starting from a known state.

With the pair of dual operators defined as

$$\overline{\mathbf{B}} = \begin{bmatrix} \mathbf{B} \\ 0 \\ 0 \end{bmatrix} : \mathcal{U} \mapsto \overline{\mathcal{D}}, \qquad \overline{\mathbf{B}'} = [\mathbf{B}', 0, 0] : \overline{\mathcal{S}} \mapsto \mathcal{F}, \tag{32}$$

the relations governing the nonlocal elastoplastic structural problem for a given load history $\ell(t)$ are:

$$\begin{cases} \overline{\mathbf{B}}'\boldsymbol{\sigma} = \ell + r & \text{equilibrium,} \\ \overline{\mathbf{B}}u = \mathbf{e} + \mathbf{p} & \text{compatibility,} \\ (\boldsymbol{\sigma}, -X) = d\Phi(\mathbf{e}, \boldsymbol{\kappa}) & \text{elastic relation,} \\ (\dot{\mathbf{p}}, \dot{k}) \in N_C(\boldsymbol{\sigma}, X) & \text{flow rule,} \\ u \in \partial Y^*(r) & \text{external relation.} \end{cases} \tag{33}$$

The evolutive analysis of a nonlocal elastoplastic constitutive problem can be performed by solving a sequence of problems in which the strain increment is applied and updating the state variables at the end of each increment [6,11].

Attention is focused on a single step of the procedure for which the strain increment is given. Accordingly we need to evaluate the finite increments of the unknown variables corresponding to the increment of strain when their values are assigned at the beginning of the step. Let $(\,\cdot\,)_o$ denote the known quantities $(\,\cdot\,)$ at the beginning of each step.

By adopting a fully implicit time integration scheme (Euler backward difference), the finite-step formulation of the elastoplastic constitutive model is achieved by enforcing the relations of the plastic flow rule at the end of the step in the form:

$$(\mathbf{p} - \mathbf{p}_o, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o) \in N_C(\boldsymbol{\sigma}, X) \quad \Longleftrightarrow \quad (\boldsymbol{\sigma}, X) \in \partial D(\mathbf{p} - \mathbf{p}_o, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o), \tag{34}$$

where the time derivatives $\dot{\mathbf{p}}$ and $\dot{k}$ are replaced by the relevant finite increment ratios and the time increment is neglected since N_C is a convex cone. In order to derive the variational formulation of the nonlocal elastoplastic finite-step structural problem it is convenient to reformulate the elastic relation in terms of the partial convex conjugate Ξ^* of Φ and the finite-step flow rule in terms of the dissipation D.

The nonlocal elastoplastic finite-step structural problem is then defined as:

$$\begin{cases} \overline{\mathbf{B}}'\boldsymbol{\sigma} = \ell + r & \text{equilibrium rate,} \\ \overline{\mathbf{B}}u = \mathbf{e} + \mathbf{p} & \text{compatibility rate,} \\ (\boldsymbol{\sigma}, \boldsymbol{\kappa}) = d\Xi^*(\mathbf{e}, X) & \text{elastic relation,} \\ (\boldsymbol{\sigma}, X) \in \partial D(\mathbf{p} - \mathbf{p}_o, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o) & \text{finite-step flow rule,} \\ u \in \partial Y^*(r) & \text{external relation.} \end{cases} \tag{35}$$

The structural problem can be recast in an operator form $\mathbf{0} \in \mathbf{S}(\mathbf{w}) + \mathbf{h}_o$ which is explicitly given as

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} 0 & \overline{\mathbf{B}}' & 0 & 0 & 0 & 0 & -I_{\overline{\mathscr{F}}} \\ \overline{\mathbf{B}} & 0 & -I_{\overline{\mathscr{D}}} & 0 & -I_{\overline{\mathscr{D}}} & 0 & 0 \\ 0 & -I_{\overline{\mathscr{F}}} & d\boldsymbol{\Xi}^* & 0 & 0 & 0 & 0 \\ 0 & 0 & & 0 & -I_{\mathscr{Y}} & 0 & \\ 0 & -I_{\overline{\mathscr{F}}} & 0 & 0 & \partial D & 0 & \\ 0 & 0 & 0 & -I_{\mathscr{Y}'} & & 0 & \\ -I_{\mathscr{U}} & 0 & 0 & 0 & 0 & 0 & \partial Y^* \end{bmatrix} \begin{bmatrix} u \\ \boldsymbol{\sigma} \\ \mathbf{e} \\ X \\ \mathbf{p} - \mathbf{p}_o \\ \boldsymbol{\kappa} - \boldsymbol{\kappa}_o \\ r \end{bmatrix} - \begin{bmatrix} \ell \\ \mathbf{p}_o \\ \mathbf{o} \\ \boldsymbol{\kappa}_o \\ \mathbf{o} \\ 0 \\ 0 \end{bmatrix}.$$

The conservativity of the structural operator is based on the conservativity of the constitutive operator and on the property that the subdifferential relations $\partial \sqcup_{\mathscr{N}}$ and ∂Y^* admit the potentials $\sqcup_{\mathscr{N}}$ and Y^* [9].

A direct integration can thus be performed:

$$\boldsymbol{\Omega}(\mathbf{w}) = \int_0^1 \prec \mathbf{S}(\mathbf{w}-\mathbf{w}_o), (\mathbf{w}-\mathbf{w}_o) \succ dt - \langle \ell, u \rangle - \langle \boldsymbol{\sigma}, \mathbf{p}_o \rangle - \langle X, \boldsymbol{\kappa}_o \rangle, \quad (36)$$

where $\mathbf{w} = (u, \boldsymbol{\sigma}, \mathbf{e}, X, \mathbf{p}, k, r)$ and $\mathbf{w}_o = (0, \mathbf{o}, \mathbf{o}, 0, -\mathbf{p}_o, -\boldsymbol{\kappa}_o, 0)$. The following potential in the complete set of state variables is thus obtained:

$$\begin{aligned} \boldsymbol{\Omega}(u, \boldsymbol{\sigma}, \mathbf{e}, X, \mathbf{p}, \boldsymbol{\kappa}, r) = {} & \boldsymbol{\Xi}^*(\mathbf{e}, X) + D(\mathbf{p} - \mathbf{p}_o, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o) + Y^*(r) + (\!(\boldsymbol{\sigma}, \bar{\mathbf{B}}u)\!) + \\ & -(\!(\boldsymbol{\sigma}, \mathbf{e} + \mathbf{p})\!) - (\!(X, \boldsymbol{\kappa})\!) - \langle \ell + r, u \rangle. \end{aligned} \quad (37)$$

The potential $\boldsymbol{\Omega}$ is linear in $(u, \boldsymbol{\sigma})$, convex in $(\mathbf{e}, X, \mathbf{p}, \boldsymbol{\kappa})$ and concave in r. The following proposition thus holds.

Proposition 4. *A set $(u, \boldsymbol{\sigma}, \mathbf{e}, X, \mathbf{p}, \boldsymbol{\kappa}, r)$ is a solution of the nonlocal elastoplastic finite-step structural problem if and only if it is a stationarity point for $\boldsymbol{\Omega}$.*

A family of potentials can be recovered from $\boldsymbol{\Omega}$ by enforcing the relations (33). Then a solution of the structural problem can be obtained as a stationary point of each of the potentials of the resulting family.

5.1 Variational principles

The external reactions can be eliminated from $\boldsymbol{\Omega}$ by enforcing the external relation $(35)_5$ in terms of Fenchel's equality,

$$Y(u) + Y^*(r) = \langle r, u \rangle, \quad (38)$$

to obtain

$$\begin{aligned} \boldsymbol{\Omega}_1(u, \boldsymbol{\sigma}, \mathbf{e}, X, \mathbf{p}, \boldsymbol{\kappa}) = {} & \boldsymbol{\Xi}^*(\mathbf{e}, X) + D(\mathbf{p} - \mathbf{p}_o, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o) - Y(u) + (\!(\boldsymbol{\sigma}, \bar{\mathbf{B}}u)\!) + \\ & -(\!(\boldsymbol{\sigma}, \mathbf{e} + \mathbf{p})\!) - (\!(X, \boldsymbol{\kappa})\!) - \langle \ell, u \rangle, \end{aligned} \quad (39)$$

which is convex in $(u, \mathbf{e}, X, \mathbf{p}, \boldsymbol{\kappa})$ and linear in $\boldsymbol{\sigma}$.

Then we have the following result.

Proposition 5. *The set* $(u, \boldsymbol{\sigma}, \mathbf{e}, X, \mathbf{p}, \boldsymbol{\kappa})$ *is a solution of the saddle problem*

$$\min_{u,\mathbf{e},X,\mathbf{p},\boldsymbol{\kappa}} \operatorname*{stat}_{\boldsymbol{\sigma}} \boldsymbol{\Omega}_1(u, \boldsymbol{\sigma}, \mathbf{e}, X, \mathbf{p}, \boldsymbol{\kappa})$$

if and only if it is a solution of the nonlocal elastoplastic structural model.

We now enforce in the expression (39) of $\boldsymbol{\Omega}_1$ the constitutive relation $(35)_3$, in terms of Fenchel's equality $(26)_1$, and the compatibility condition $(35)_2$ to get the potential

$$\boldsymbol{\Omega}_2(u, \mathbf{p}, \boldsymbol{\kappa}) \;=\; \Phi(\bar{\mathbf{B}}u - \mathbf{p}, \boldsymbol{\kappa}) + D(\mathbf{p} - \mathbf{p}_o, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o) - \Upsilon(u) - \langle \ell, u \rangle, \qquad (40)$$

which is convex in $(u, \mathbf{p})$ and locally subdifferentiable in $\boldsymbol{\kappa}$. We then have our next result.

Proposition 6. *The set* $(u, \mathbf{p}, \boldsymbol{\kappa})$ *is a solution of the optimization problem*

$$\min_{u,\mathbf{p}} \operatorname*{stat}_{\boldsymbol{\kappa}} \boldsymbol{\Omega}_2(u, \mathbf{p}, \boldsymbol{\kappa})$$

if and only if it is a solution of the nonlocal elastoplastic structural model (35).

Enforcing in (40) the flow rule

$$\sqcup_C(\boldsymbol{\sigma}, X) + D(\mathbf{p} - \mathbf{p}_o, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o) = (\!(\boldsymbol{\sigma}, \mathbf{p} - \mathbf{p}_o)\!) + (\!(X, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o)\!),$$

we see that

$$\begin{aligned} \boldsymbol{\Omega}_3(u, \mathbf{p}, \boldsymbol{\sigma}, X, \boldsymbol{\kappa}) \;=\; \Phi(\bar{\mathbf{B}}u - \mathbf{p}, \boldsymbol{\kappa}) - \sqcup_C(\boldsymbol{\sigma}, X) - \Upsilon(u) + \\ + (\!(\boldsymbol{\sigma}, \mathbf{p} - \mathbf{p}_o)\!) + (\!(X, \boldsymbol{\kappa} - \boldsymbol{\kappa}_o)\!) - \langle \ell, u \rangle, \end{aligned} \qquad (41)$$

which is convex in $(u, \mathbf{p})$ and concave in $(\boldsymbol{\sigma}, X, \boldsymbol{\kappa})$. Thus we obtain our last result.

Proposition 7. *The set* $(u, \mathbf{p}, \boldsymbol{\sigma}, X, \boldsymbol{\kappa})$ *is a solution of the saddle problem*

$$\min_{u,\mathbf{p}} \max_{\boldsymbol{\sigma},X,\boldsymbol{\kappa}} \boldsymbol{\Omega}_3(u, \mathbf{p}, \boldsymbol{\sigma}, X, \boldsymbol{\kappa})$$

if and only if it is a solution of the nonlocal elastoplastic structural model (35).

6 Conclusions

A nonlocal model of plasticity is presented and is cast in the framework of convex analysis and of the potential theory for monotone multivalued operators. As a consequence a theoretical analysis can be performed in analogy with local standard plasticity, and variational formulations for the nonlocal model are given. The proposed treatment of plasticity is rather general and can be applied to further different material behaviors which can be described within the theory of internal variables such as damage and rate-dependent plasticity.

Acknowledgements

The financial support of the Italian Ministry for Scientific and Technological Research is gratefully acknowledged.

References

[1] Romano, G. (2001): Theory of structural models. II. Structural models. (Italian). Doctoral Lectures. University of Napoli Federico II, Naples
[2] Brezis, H. (1983): Analyse fonctionnelle. Théorie et applications. Masson, Paris
[3] Halphen, B., Nguyen, Q.S. (1975): Sur les matériaux standards généralisés. J. Mécanique **14**, 39–63
[4] Hiriart-Urruty, J.-B., Lemaréchal, C. (1993): Convex analysis and minimization algorithms. I. Fundamentals. Springer, Berlin
[5] Romano, G. (1995): New results in subdifferential calculus with applications to convex optimization. Appl. Math. Optim. **32**, 213–234
[6] Reddy, B.D., Martin, J.B. (1991): Algorithms for the solution of internal variable problems in plasticity. Comput. Methods Appl. Mech. Engrg. **93**, 253–273
[7] Simo, J.C., Kennedy, J.J., Govindjee, S. (1988): Nonsmooth multisurface plasticity and viscoplasticity. Loading/unloading conditions and numerical algorithms. Internat. J. Numer. Methods Engrg. **26**, 2161–2185
[8] Lubliner, J. (1990): Plasticity theory. Macmillan, New York
[9] Romano, G., Rosati, L., Marotti de Sciarra, F., Bisegna, P. (1993): A potential theory for monotone multivalued operators. Quart. Appl. Math. **51**, 613–631
[10] Borino, G., Fuschi, P., Polizzotto, C. (1999): A thermodynamic approach to nonlocal plasticity and related variational principles. Trans. ASME J. Appl. Mech. **66**, 952–963
[11] Simo, J.C., Kennedy, J.G., Taylor, R.L. (1989): Complementary mixed finite element formulations for elastoplasticity. Comput. Methods Appl. Mech. Engrg. **74**, 177–206

Consistent order extended thermodynamics and its application to light scattering

I. Müller, D. Reitebuch

Abstract. The new theory of consistently ordered extended thermodynamics is described and compared to the earlier theory of extended thermodynamics with respect to the efficiency with which these theories describe light scattering spectra. It turns out that the new theory is more efficient, albeit only slightly.

1 Introduction

Extended thermodynamics is a field theory for gases, in particular rarefied gases. The fields are moments of the distribution function and the field equations are the equations of balance for the moments as they follow from the Boltzmann equation. The system of equations requires closure. Different closure procedures have been proposed and they distinguish different versions of extended thermodynamics.

There is extended thermodynamics proper, described in the monograph [1] and here abbreviated by ET, where closure is achieved by the exploitation of the entropy principle. And there is consistent order extended thermodynamics [2], called COET in the sequel, which makes use of combinations of moments as fields and those combinations – called G-moments – may be assigned an order of magnitude in a rational manner. Closure in this theory is an automatic consequence of the assignment of order.

We briefly motivate the choice of the G-moments and explain the criterion by which they are ordered. At this time COET has only been formulated for the BGK model for the Boltzmann collision term [3]. We write the full set of equations of first and second order in a one-dimensional setting and as appropriate for the BGK theory.

Light scattering in rarefied gases is a process that calls for the equations of ET and we recall from the published literature (cf. [4,5,1]) that very many moments have to be pressed into service in order to reach a satisfactory representation of light scattering spectra.

There is hope that COET, because of its more judicious selection of variables, gets away with a lesser number of G-moments and that expectation is confirmed. To be sure, however, we still need a considerable number of fields. The scattering spectra calculated by COET are compared with the predictions of the Yip & Nelkin exact solution of the BGK-Boltzmann equation (cf. [6]).

2 Field equations in consistent order extended thermodynamics

2.1 Equations of balance for moments

The field equations of the extended thermodynamics of gases are the equations of transfer for the moments of the phase density f, viz.,

$$\frac{\partial F_A}{\partial t} + \frac{\partial F_{iA}}{\partial x_i} = \frac{1}{\tau}\left(F_A^{\mathrm{E}} - F_A\right), \qquad \text{where} \qquad \begin{array}{l} F_A = m\int c_A f\, d\mathbf{c}, \\ F_{iA} = m\int c_i c_A f\, d\mathbf{c}. \end{array} \tag{2.1}$$

The index A here is a multi-index; we have

$$c_A = \begin{cases} 1 & A = 0 \\ c_{i_1} c_{i_2} \ldots c_{i_A} & A = 1, 2, \ldots \end{cases} . \tag{2.2}$$

It is often useful to replace the moments F_A by the internal moments $\hat{F}_A$ which correspond to the F_A's in the rest frame of the gas. The moments $\hat{F}_A$ and $\hat{F}_{iA}$ are also defined by $(2.1)_{2,3}$, except that the velocity c_i in the definitions must now be replaced by the relative velocity $C_i = c_i - v_i$. There is a one-to-one correspondence between F_A und $\hat{F}_A$ which reads[1]:

$$F_{i_1 i_2 \cdots i_A} = \hat{F}_{i_1 i_2 \ldots i_A} + \tbinom{A}{1} \hat{F}_{(i_1 i_2 \ldots i_{A-1}}\, v_{i_A)} + \tbinom{A}{2} \hat{F}_{(i_1 i_2 \ldots i_{A-2}}\, v_{i_{A-1}}\, v_{i_A)} + \\ + \ \ldots \ + \tbinom{A}{A-1} \hat{F}_{(i_1}\, v_{i_2} \ldots v_{i_A)} + \hat{F}\, v_{i_1}\, v_{i_2} \ldots v_{i_A}.$$

If we let A run from 0 to ∞, the system (2.1) represents an infinite system of balance equations. In gas dynamics and extended thermodynamics this system is cut off at a finite value N of A. Such a cut-off leaves us with a closure problem because the occurence of the last flux $\hat{F}_{i_N}$ prevents us from having a closed system for the moments $\hat{F}_A$ $(A = 1,\ 2,\ \ldots,\ N)$ automatically.

The question is where to close – i.e., at which N – and how to close? Generally the idea is to make N as big as possible and that idea seems to be a sound one. Indeed the treatment of light scattering in rarefied gases becomes more and more satisfactory when N increases. Even so, however, further moments – moments with $A > N$ – are by no means small; thus even for large values of A all even-ranked moments have non-vanishing equilibrium values.

In ET the last flux $\hat{F}_{i_N}$ is related to the moments $\hat{F}_A$ $(A = 1,\ 2,\ \ldots,\ N)$ by the exploitation of the entropy principle. This is the manner in which ET closes the system, e.g., see [1]. Recently objections have been raised to this closure agreement (cf. [7]) because the exploitation of the entropy principle may lead to a loss of hyperbolicity of the system in the immediate neighborhood of equilibrium.

[1] Round brackets indicate symmetrization and angular brackets denote traceless symmetric tensors.

2.2 *G*-moments

We have therefore proposed in [2] to choose new variables, namely, combinations of the moments $\hat{F}_A$ and these are introduced as moments of the orthonormal irreducible Hermite polynomials $\psi_{r,\langle i_1 i_2 \ldots i_l\rangle}$ of the velocity components C_i. Thus the new variables read

$$G_{r,\langle i_1 i_2 \ldots i_l\rangle} = m\sqrt{\Theta}^{2r+l} \int \psi_{r,\langle i_1 i_2 \ldots i_l\rangle}\, f\, d\mathrm{c}, \tag{2.3}$$

where m is the atomic mass and Θ stands for $\frac{k}{m}T$. In [2] we have motivated the choice of these moments at some length. The variable $G_{r,\langle i_1 i_2 \ldots i_l\rangle}$ represents a tensor of rank $2r+l$ in which r pairs of indices have been contracted.

There is a one-to-one correspondence between the moments $\mathbf{F}$ in (2.1) or the internal moments $\hat{\mathbf{F}}$, and the G-moments in (2.3) and for the first few low-ranked moments Table 1 exhibits that correspondence. For higher-ranked tensors the relationship may be taken from the formulae of [2]. All tensors $G_{r,\langle i_1 i_2 \ldots i_l\rangle}$ except G_0 vanish in equilibrium.

Since the $\mathbf{F}$'s and $\mathbf{G}$'s are related, the moment equations (2.1) dictate equations for the G-moments which, however, cannot be written in such a compact form as (2.1).

Table 1. Some Hermite polynomials and their moments

$l \setminus r$	0	1	2
0	$\psi_0 = 1$ $G_0 = \varrho$	$\psi_1 = -\frac{1}{\sqrt{6}}\frac{C^2-3\Theta}{\Theta}$ $G_1 = 0$	$\psi_2 = \frac{1}{2\sqrt{30}}\frac{C^4-10\Theta C^2+15\Theta^2}{\Theta^2}$ $G_2 = \frac{1}{2\sqrt{30}}\left(\hat{F}_{rrss} - 15\varrho\Theta^2\right)$
1	$\psi_{0,i} = \frac{C_i}{\sqrt{\Theta}}$ $G_{0,i} = 0$	$\psi_{1,i} = -\frac{1}{\sqrt{10}}\frac{(C^2-5\Theta)C_i}{\sqrt{\Theta}^3}$ $G_{1,i} = -\frac{1}{\sqrt{10}}\hat{F}_{irr}$	$\psi_{2,i} = \frac{1}{2\sqrt{70}}\frac{(C^4-14\Theta C^2+35\Theta^2)C_i}{\sqrt{\Theta}^5}$ $G_{2,i} = \frac{1}{2\sqrt{70}}\left(\hat{F}_{irrss} - 14\Theta\hat{F}_{irr}\right)$
2	$\psi_{0,\langle ij\rangle} = \frac{1}{\sqrt{2}}\frac{C_{\langle i}C_{j\rangle}}{\Theta}$ $G_{0,\langle ij\rangle} = \frac{1}{\sqrt{2}}\hat{F}_{\langle ij\rangle}$	$\psi_{1,\langle ij\rangle} = -\frac{1}{2\sqrt{7}}\frac{(C^2-7\Theta)C_{\langle i}C_{j\rangle}}{\Theta^2}$ $G_{1,\langle ij\rangle} = -\frac{1}{2\sqrt{7}}\left(\hat{F}_{\langle ij\rangle rr} - 7\Theta\hat{F}_{\langle ij\rangle}\right)$	$\psi_{2,\langle ij\rangle} = \frac{C^4-18\Theta C^2+63\Theta^2}{12\sqrt{7}\sqrt{\Theta}^5}C_{\langle i}C_{j\rangle}$ $G_{2,\langle ij\rangle} = \frac{\hat{F}_{\langle ij\rangle rrss}-18\Theta\hat{F}_{\langle ij\rangle rr}+63\Theta^2\hat{F}_{\langle ij\rangle}}{12\sqrt{7}}$
3	$\psi_{0,\langle ijk\rangle} = \frac{1}{\sqrt{6}}\frac{C_{\langle i}C_jC_{k\rangle}}{\sqrt{\Theta}^3}$ $G_{0,\langle ijk\rangle} = \frac{1}{\sqrt{6}}\hat{F}_{\langle ijk\rangle}$	$\psi_{1,\langle ijk\rangle} = -\frac{C^2-9\Theta}{6\sqrt{3}\sqrt{\Theta}^5}C_{\langle i}C_jC_{k\rangle}$ $G_{1,\langle ijk\rangle} = -\frac{\hat{F}_{\langle ijk\rangle rr}-9\Theta\hat{F}_{\langle ijk\rangle}}{6\sqrt{3}}$	$\psi_{2,\langle ijk\rangle} = \frac{C^4-22\Theta C^2+99\Theta^2}{12\sqrt{33}\sqrt{\Theta}^5}C_{\langle i}C_jC_{k\rangle}$ $G_{2,\langle ijk\rangle} = \frac{\hat{F}_{\langle ijk\rangle rrss}-22\Theta\hat{F}_{\langle ijk\rangle rr}+99\Theta^2\hat{F}_{\langle ijk\rangle}}{12\sqrt{33}}$
4	$\psi_{0,\langle ijkl\rangle} = \frac{1}{2\sqrt{6}}\frac{C_{\langle i}C_jC_kC_{l\rangle}}{\Theta^2}$ $G_{0,\langle ijkl\rangle} = \frac{1}{2\sqrt{6}}\hat{F}_{\langle ijkl\rangle}$	$\psi_{1,\langle ijkl\rangle} = -\frac{C^2-11\Theta}{4\sqrt{33}\Theta^2}C_{\langle i}C_jC_kC_{l\rangle}$ $G_{1,\langle ijkl\rangle} = -\frac{\hat{F}_{\langle ijkl\rangle rr}-11\Theta\hat{F}_{\langle ijkl\rangle}}{4\sqrt{33}}$	$\psi_{2,\langle ijkl\rangle} = \frac{C^4-26\Theta C^2+143\Theta^2}{8\sqrt{429}\sqrt{\Theta}^5}C_{\langle i}C_jC_kC_{l\rangle}$ $G_{2,\langle ijkl\rangle} = \frac{\hat{F}_{\langle ijkl\rangle rrss}-26\Theta\hat{F}_{\langle ijkl\rangle rr}+143\Theta^2\hat{F}_{\langle ijkl\rangle}}{8\sqrt{429}}$

2.3 Order of magnitude of *G*-moments

For essentially one-dimensional problems, i.e., problems with rotational symmetry about the x_1-axis, we may use the subset

$$\psi_{r,l} = \psi_{r,\underbrace{\langle 1\,1\,\ldots\,1\rangle}_{l\text{ times}}}$$

of Hermite polynomials and the corresponding subset

$$G_{r,l} = m\,\sqrt{\Theta}^{\,2r+l}\int \psi_{r,l} f\,d\mathbf{c} \tag{2.4}$$

of *G*-moments. In that case the conservation laws of mass, momentum and energy read, with $\mathbf{v} = (v,\ 0,\ 0)$:

$$\begin{aligned}
&\frac{\partial \rho}{\partial t} + v\,\frac{\partial \rho}{\partial x} + \rho\,\frac{\partial v}{\partial x} = 0,\\
&\rho\,\frac{\partial v}{\partial t} + \rho\,v\,\frac{\partial v}{\partial x} + \rho\,\frac{\partial \Theta}{\partial x} + \Theta\,\frac{\partial \rho}{\partial x} + \frac{2}{\sqrt{3}}\,\frac{\partial G_{0,2}}{\partial x} = 0,\\
&\rho\,\frac{\partial \Theta}{\partial t} + \rho\,v\,\frac{\partial \Theta}{\partial x} + \frac{2}{3}\,\rho\,\Theta\,\frac{\partial v}{\partial x} - \frac{\sqrt{10}}{3}\,\frac{\partial G_{1,1}}{\partial x} + \frac{2}{3}\,\sqrt{\frac{4}{3}}\,G_{0,2}\,\frac{\partial v}{\partial x} = 0.
\end{aligned} \tag{2.5}$$

All higher moment equations are of the generic form

$$G_{r,l} = \tilde{G}_{r,l}\left(\tau\,\frac{\partial \Theta}{\partial t},\ \tau\,\frac{\partial \Theta}{\partial x},\ \tau\,\frac{\partial v}{\partial t},\ \tau\,\frac{\partial v}{\partial x},\ G_{p,q},\ \tau\,\frac{\partial G_{r,s}}{\partial t},\ \tau\,\frac{\partial G_{r,s}}{\partial x}\right). \tag{2.6}$$

There are, of course, infinitely many of them and the system needs to be closed.

We use Eqs. (2.6) to calculate *n*th iterates $\overset{(n)}{G}_{r,l}$ from the $(n-1)$st iterates by virtue of the prescription

$$\overset{(n)}{G}_{r,l} = \tilde{G}_{r,l}\left(\tau\,\frac{\partial \Theta}{\partial t},\ \tau\,\frac{\partial \Theta}{\partial x},\ \tau\,\frac{\partial v}{\partial t},\ \tau\,\frac{\partial v}{\partial x},\ \overset{(n-1)}{G}_{p,q},\ \tau\,\frac{\partial \overset{(n-1)}{G}_{r,s}}{\partial t},\ \tau\,\frac{\partial \overset{(n-1)}{G}_{r,s}}{\partial x}\right) \tag{2.7}$$

with the initiation agreement $\overset{(0)}{G}_{r,l} = \rho\,\delta_{r0}\,\delta_{l0}$. Thus the iterates $\overset{(n)}{G}_{r,l}$ now contain expressions of the type

$$\left(\tau\,\frac{\partial v}{\partial t}\right)^n,\ \left(\tau^n\,\frac{\partial^n v}{\partial t^n}\right),\ \left(\tau\,\frac{\partial v}{\partial x}\right)^n,\ \left(\tau^n\,\frac{\partial^n v}{\partial x^n}\right)$$

and analogous ones for Θ instead of v. We regard these as of "order of magnitude *n*" in the rates of change and in the steepness of gradients of v and Θ.

In this manner the *G*-moments may be ordered as shown in Table 2. Inspection shows that – roughly speaking – the order of magnitude grows with the tensorial rank of the *G*-moments. Also we see that the so-called 14-moment theory – popular with mathematicians – in which the variables are ρ, v, Θ, $G_{0,2}$, $G_{1,1}$, $G_{2,0}$ has no standing within this ordering scheme. Indeed, if we adopt the second order quantity $G_{2,0}$ into the list of variables, there is no reason to leave out the other second order quantities, viz., $G_{0,3}$ through $G_{2,2}$.

Table 2. Orders of magnitude and tensorial rank $2r + l$ of the $\psi_{r,l}$-moments $G_{r,l}$.
The table holds for stationary and *instationary* heat conduction and one-dimensional motion.
Rows: Highest tensorial rank $2r + l$ of the $\psi_{r,l}$-moments $G_{r,l}$.
Columns: Order of magnitude

	0	1	2	3	4	5	6
0	ϱ	-	-	-	-	-	-
1	v	-	-	-	-	-	-
2	T	$G_{0,2}$	-	-	-	-	-
3	-	$G_{1,1}$	$G_{0,3}$	-	-	-	-
4	-	-	$G_{2,0}$ $G_{1,2}$ $G_{0,4}$	-	-	-	-
5	-	-	$G_{2,1}$ $G_{1,3}$	$G_{0,5}$	-	-	-
6	-	-	$G_{3,0}$ $G_{2,2}$	$G_{1,4}$ $G_{0,6}$	-	-	-
7	-	-	-	$G_{3,1}$ $G_{2,3}$ $G_{1,5}$	$G_{0,7}$	-	-
8	-	-	-	$G_{4,0}$ $G_{3,2}$ $G_{2,4}$	$G_{1,6}$ $G_{0,8}$	-	-
9	-	-	-	$G_{4,1}$ $G_{3,3}$	$G_{2,5}$ $G_{1,7}$	$G_{0,9}$	-
10	-	-	-	-	$G_{5,0}$ $G_{4,2}$ $G_{3,4}$	$G_{1,8}$ $G_{0,10}$	-
11	-	-	-	-	$G_{5,1}$ $G_{4,3}$ $G_{3,5}$	$G_{2,7}$ $G_{1,9}$	$G_{0,11}$
12	-	-	-	-	$G_{6,0}$ $G_{5,2}$ $G_{4,4}$	$G_{3,6}$ $G_{1,8}$	$G_{1,10}$ $G_{0,12}$

2.4 Field equations for the *G*-moments

We write the field equations (2.5), (2.6) of first order by omitting all terms of order 2 and obtain:

$$
\begin{aligned}
&\textit{first order}\\
&\textit{MB} \quad 0 = \ \tfrac{d\varrho}{dt} \ + \tfrac{d\varrho\, v}{dx},\\
&\textit{IB} \quad 0 = \ \varrho\, \tfrac{dv}{dt} \ + \tfrac{d\varrho\, \Theta}{dx} + \varrho\, v\, \tfrac{dv}{dx} + \tfrac{2}{\sqrt{3}}\, \tfrac{dG_{0,2}}{dx},\\
&\textit{EB} \quad 0 = \tfrac{3}{2}\, \varrho\, \tfrac{d\Theta}{dt} + \tfrac{3}{2}\, \varrho\, v\, \tfrac{d\Theta}{dx} + \varrho\, \Theta\, \tfrac{dv}{dx} - \sqrt{\tfrac{5}{2}}\, \tfrac{dG_{1,1}}{dx} + \tfrac{2}{\sqrt{3}}\, G_{0,2}\, \tfrac{dv}{dx},\\
&\qquad G_{0,2} = \qquad - \tfrac{2}{\sqrt{3}}\, \varrho\, \Theta\, \tau\, \tfrac{dv}{dx},\\
&\qquad G_{1,1} = \qquad \sqrt{\tfrac{5}{2}}\, \varrho\, \Theta\, \tau\, \tfrac{d\Theta}{dx}.
\end{aligned}
\tag{2.8}
$$

Similarly the field equations of second order read:

$$
\left.
\begin{aligned}
&\textit{second order}\\
&\textit{MB} \quad 0 = \ \tfrac{d\varrho}{dt} \ + \tfrac{\partial \varrho\, v}{\partial x},\\
&\textit{IB} \quad 0 = \ \varrho\, \tfrac{dv}{dt} \ + \tfrac{\partial \varrho\, \Theta}{\partial x} + \varrho\, v\, \tfrac{\partial v}{\partial x} + \tfrac{2}{\sqrt{3}}\, \tfrac{\partial G_{0,2}}{\partial x},\\
&\textit{EB} \quad 0 = \ \tfrac{3}{2}\, \varrho\, \tfrac{d\Theta}{dt} \ + \tfrac{3}{2}\, \varrho\, v\, \tfrac{\partial \Theta}{\partial x} + \varrho\, \Theta\, \tfrac{\partial v}{\partial x} - \sqrt{\tfrac{5}{2}}\, \tfrac{\partial G_{1,1}}{\partial x} + \tfrac{2}{\sqrt{3}}\, G_{0,2}\, \tfrac{\partial v}{\partial x},\\
&\qquad G_{0,2} = \underset{(2)}{-\tau}\, \tfrac{\partial G_{0,2}}{\partial t} \underset{(2)}{-} v\, \tau\, \tfrac{\partial G_{0,2}}{\partial x} \underset{(1)}{-} \tfrac{2}{\sqrt{3}}\, \varrho\, \Theta\, \tau\, \tfrac{\partial v}{\partial x} \underset{(2)}{-} \tfrac{7}{3}\, G_{0,2}\, \tau\, \tfrac{\partial v}{\partial x} \underset{(2)}{+} 2\sqrt{\tfrac{2}{15}}\, \tau\, \tfrac{\partial G_{1,1}}{\partial x},\\
&\qquad G_{1,1} = \underset{(2)}{-\tau}\, \tfrac{\partial G_{1,1}}{\partial t} \underset{(2)}{-} v\, \tau\, \tfrac{\partial G_{1,1}}{\partial x} \underset{(2)}{+} 2\sqrt{\tfrac{2}{15}}\, \Theta\, \tau\, \tfrac{\partial G_{0,2}}{\partial x}\\
&\qquad\qquad\qquad\qquad \underset{(1)}{+}\sqrt{\tfrac{5}{2}}\, \varrho\, \Theta\, \tau\, \tfrac{\partial \Theta}{\partial x} \underset{(2)}{+} 7\sqrt{\tfrac{2}{15}}\, G_{0,2}\, \tau\, \tfrac{\partial \Theta}{\partial x} \underset{(2)}{-} \tfrac{16}{5}\, G_{1,1}\, \tau\, \tfrac{\partial v}{\partial x},
\end{aligned}
\right\} \quad \text{(A)}
$$

$$
\left.
\begin{aligned}
G_{0,3} &= \underset{(2)}{-}\tfrac{3}{\sqrt{5}}\,\Theta\,G_{0,2}\,\tau\,\tfrac{\partial\Theta}{\partial x} \underset{(2)}{-}\tfrac{3}{\sqrt{5}}\,v\,G_{0,2}\,\tau\,\tfrac{\partial v}{\partial x} \underset{(2)}{+}\tfrac{2}{5}\sqrt{6}\,G_{1,1}\,\tau\,\tfrac{\partial v}{\partial x} \underset{(2)}{-}\tfrac{3}{\sqrt{5}}\,\Theta\,\tau\,\tfrac{\partial G_{0,2}}{\partial x},\\
G_{2,0} &= \underset{(2)}{+}\tfrac{7}{\sqrt{3}}\,G_{1,1}\,\tau\,\tfrac{\partial\Theta}{\partial x} \underset{(2)}{-}\tfrac{4}{3}\sqrt{\tfrac{2}{5}}\,\Theta\,G_{0,2}\,\tau\,\tfrac{\partial v}{\partial x} \underset{(2)}{+}\tfrac{2}{\sqrt{3}}\,v\,G_{1,1}\,\tau\,\tfrac{\partial v}{\partial x} \underset{(2)}{+}\tfrac{2\Theta}{\sqrt{3}}\,\tau\,\tfrac{\partial G_{1,1}}{\partial x},\\
G_{1,2} &= \underset{(2)}{+}\sqrt{\tfrac{7}{2}}\,G_{0,2}\,v\,\tau\,\tfrac{\partial\Theta}{\partial x} \underset{(2)}{-}4\sqrt{\tfrac{7}{15}}\,G_{1,1}\,\tau\,\tfrac{\partial\Theta}{\partial x} \underset{(2)}{+}\tfrac{11}{3}\sqrt{\tfrac{2}{7}}\,\Theta\,G_{0,2}\,\tau\,\tfrac{\partial v}{\partial x}\\
&\qquad \underset{(2)}{+}2\sqrt{\tfrac{7}{15}}\,v\,G_{1,1}\,\tau\,\tfrac{\partial v}{\partial x} \underset{(2)}{-}2\sqrt{\tfrac{7}{15}}\,\Theta\,\tau\,\tfrac{\partial G_{1,1}}{\partial x},\\
G_{0,4} &= \underset{(2)}{-}\tfrac{12}{\sqrt{35}}\,\Theta\,G_{0,2}\,\tau\,\tfrac{\partial v}{\partial x},\\
G_{2,1} &= \underset{(2)}{-}2\sqrt{\tfrac{14}{15}}\,\Theta\,G_{0,2}\,\tau\,\tfrac{\partial\Theta}{\partial x} \underset{(2)}{+}\sqrt{7}\,v\,G_{1,1}\,\tau\,\tfrac{\partial\Theta}{\partial x} \underset{(2)}{+}\tfrac{6}{5}\sqrt{7}\,\Theta\,G_{1,1}\,\tau\,\tfrac{\partial v}{\partial x},\\
G_{1,3} &= \underset{(2)}{+}\tfrac{9}{\sqrt{10}}\,\Theta\,G_{0,2}\,\tau\,\tfrac{\partial\Theta}{\partial x} \underset{(2)}{-}\tfrac{6}{5}\sqrt{3}\,\Theta\,G_{1,1}\,\tau\,\tfrac{\partial v}{\partial x},\\
G_{3,0} &= \underset{(2)}{-}\sqrt{14}\,\Theta\,G_{1,1}\,\tau\,\tfrac{\partial\Theta}{\partial x},\\
G_{2,2} &= \underset{(2)}{+}2\sqrt{\tfrac{21}{5}}\,\Theta\,G_{1,1}\,\tau\,\tfrac{\partial\Theta}{\partial x}.
\end{aligned}
\right\}\ \text{(B)}
\tag{2.9}
$$

The most important feature in these equations is that they are closed. This feature persists in higher orders as well and we conclude that: no specific closure agreement is needed. Or else: *closure is an automatic consequence of the assignment of order*.

Another feature, which is first seen in the equations of second order, is that that system – and all subsequent ones – decomposes into subsystem (A) which is itself closed and subsystem (B) which allows us to calculate the remaining G-moments of the order under consideration by differentiation after system (A) is solved. We observe that subsystem (A) of the equations (2.9) of second order is identical to the popular 13-moment equations of Grad [8], which represent the prototype of all extensions of thermodynamics.

The field equations of higher order than 2 are not listed here, because they are too long. However, subsystem (A) of third order is listed in [2]. Systems of still higher order are available on computer but cannot be printed on a few pages.

Such systems generally require boundary and initial values for the G-moments of higher order and that presents a problem. Indeed, there is no way to apply and control such values in practice. The problem of boundary values was recently considered in [9] in solving a fourth order system of stationary heat conduction in a gas at rest.

Here, however, we proceed with a phenomenon that does not require boundary and initial values, viz., light scattering in a rarefied gas.

3 Light scattering in ET

3.1 Experiment

In a light scattering experiment, a laser beam of frequency $\omega^{(i)}$ is scattered by density fluctuations of a gas. The experimental set-up is shown in Fig. 1. While most of the light passes the gas unscattered, a small part is deflected and most of this scattered light has the frequency $\omega^{(i)}$, the same frequency as the incident light. However, neighboring frequencies ω can also be detected in the scattering spectrum $S(\omega, p)$. That spectrum depends on the pressure p which is small for a strong degree of rarefaction.

Figure 1 also shows a typical schematic plot of $S(\omega, p)$ *measured* for a gas at large pressures. The graph is characterized by three peaks, the central one located at the incident frequency.

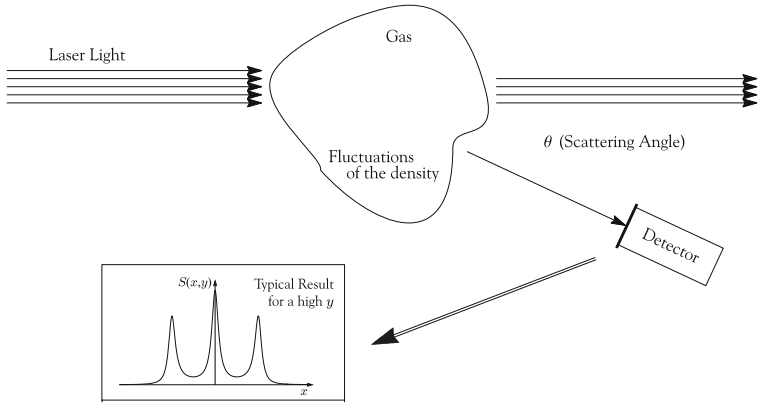


Fig. 1. Light scattering experiment

3.2 Calculation of the scattering spectrum

The Onsager hypothesis about the equivalence of the mean regression of a fluctuation and the solution of macroscopic field equations makes it possible to *calculate* a scattering spectrum from a thermodynamic field theory.

Thus, for instance, if we use the Navier-Stokes-Fourier theory, we obtain spectra of the type shown in Fig. 2 for different pressures. Here y is a dimensionless pressure and x is a dimensionless frequency. For the large pressure $y = 4$ the plot agrees well with experiments while for smaller values of y the Navier-Stokes-Fourier theory cannot describe the scattering spectrum properly.

This is where extended thermodynamics has found its most fruitful field of application. Indeed, small values of y correspond to an advanced degree of rarefaction

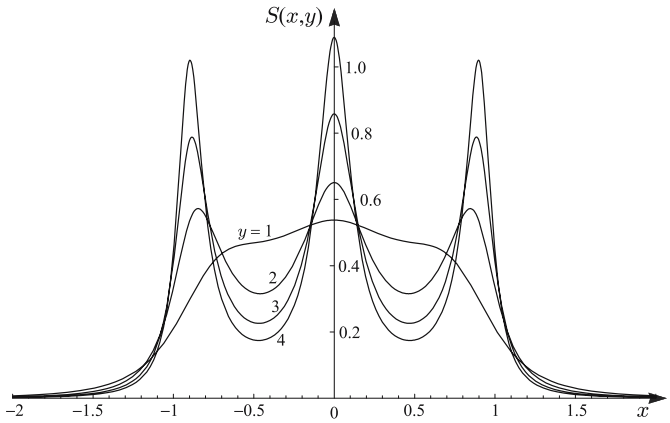


Fig. 2. Light scattering spectra, calculated with NSF theory

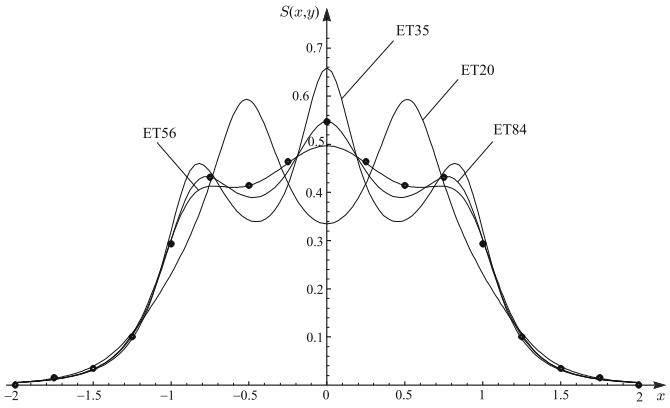


Fig. 3. Spectra measured and calculated with ET20, ET35, ET56, ET84

of the gas and that is the regime in which extended thermodynamics is expected to be useful. Weiss [4] calculated the scattering spectra for extended thermodynamics with $N = 3,\ 4,\ 5$ and 6 for $y = 1$ (cf. Fig. 3; see also [5,1]). He compared these spectra with measured data by Clark [10] and found that none of these theories is satisfactory.

This only means that N is not yet big enough. And indeed, if $N = 7$ is chosen, theory and experiment agree perfectly (cf. Fig. 4). What is more: there is a convergence, because when N is pushed up to $8,\ 9$, and 10 there is no longer significant improvement. This is quite satisfactory, because it shows that extended thermodynamics can reproduce the measured data. Moreover, the convergence indicates that we do not even need measured data in order to determine which value of N is appropriate

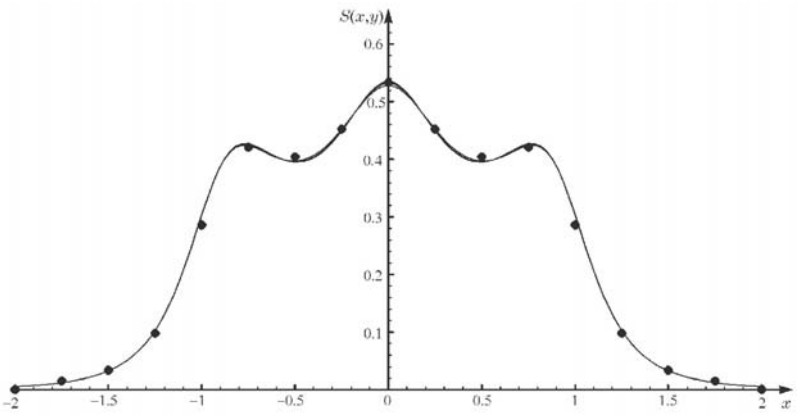


Fig. 4. Spectra measured and calculated with ET120, ET165, ET220, ET286

for a given degree of rarefaction: it is *the* value of N for which an increase does not significantly improve the calculated spectrum, e.g., $N = 7$ for $y = 1$.

However, there is also disappointment, because $N = 7$ means 120 moments. [The relation between the independent number n of moments and their highest tensorial rank N is given by

$$n = \frac{1}{6}(N+1)(N+2)(N+3)\ ] \tag{3.1}$$

Therefore for good agreement between theory and experiment we need many moments in ET, the theory based on the moments F_A (cf. (2.1)). It may be hoped that COET, the consistently ordered extended thermodynamics, succeeds with fewer G-moments. We proceed to investigate that proposition.

4 Consistent order extended thermodynamics and light scattering

4.1 The Yip & Nelkin solution as a reference

COET has so far been worked out only for the BGK model, which is known to be unsatisfactory if we require quantitatively correct results. Therefore it is not possible to compare the results of the theory with experimental data or with those of ET. The latter theory was applied to light scattering by use of the Maxwell interaction potential between atoms. Those are much more realistic than the BGK model.

It so happens, however, that Yip & Nelkin [6] have solved the Boltzmann equation for the BGK collision term exactly. We may calculate light scattering spectra from that solution and then use those as reference solution in order to judge the appropriateness of COET of any given order.

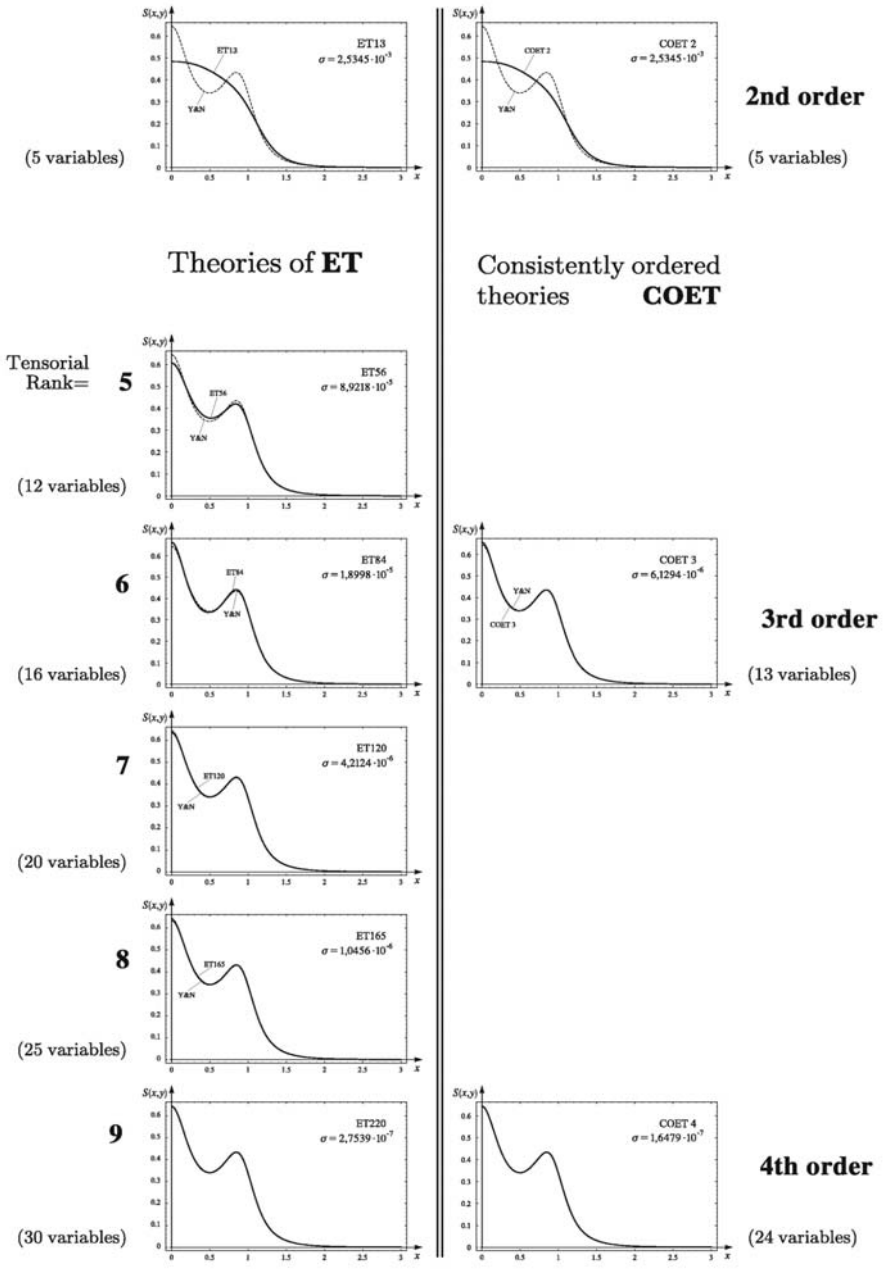


Fig. 5. Light scattering spectra for ET and COET for $y = 1$

4.2 Discussion of results

Figure 5 provides a three-way comparison of results. For the low dimensionless pressure $y = 1$ we have plotted:

- the spectra predicted by Yip & Nelkin;
- the spectra from ET, recalculated for the BGK-ansatz;
- the prediction of COET for second, third and fourth orders.

The two uppermost pictures are identical, because ET13 and COET in second order are governed by the same field equations as was mentioned in Sect. 2.4. Both disagree strongly with the Yip & Nelkin solution which is not surprising, since that solution is exact while both ET13 and second order COET provide poor descriptions for a rarefied gas with $y = 1$.

The main part of Fig. 5 on the left-hand side shows the emerging agreement between the exact Yip & Nelkin solution and ET for $N = 5$ through 9 corresponding to 56 through 220 independent moments. In order to have a quantitative measure for the agreement we have calculated a norm for the deviation, viz.,

$$\sigma = \frac{1}{100}\sum_{i=0}^{99}\left(S\left(x_i, y\right) - S_{\mathrm{Y\&N}}\left(x_i, y\right)\right)^2, \qquad \text{where}$$
$$x_i = \frac{3\,i}{99} \qquad (i = 0,\ 1,\ ...,\ 99). \tag{4.1}$$

The value of that norm is given in the frames of Fig. 5 and we observe that it becomes smaller rapidly for increasing N. For the fully satisfactory theory with $N = 9$ or 220 moments we have $\sigma = 2.75 \cdot 10^{-7}$.

The right-hand side of Fig. 5 exhibits the improving agreement of COET of increasing order with the Yip & Nelkin solution. In third order the agreement is already fairly good to the naked eye, but a glance at the norms shows that in fourth order it is really excellent, better than ET220.

What about the relative number of variables? In COET the system (A) in third order contains 13 variables (cf. (2.9)), while ET84 – which is comparable, actually bigger in norm – contains 16 one-dimensional equations for the 16 moments[2]

$$F,\ F_1,\ \hat{F}_{ll},\ \hat{F}_{\langle 11\rangle},\ \hat{F}_{ll1},\ \hat{F}_{\langle 111\rangle},\ \hat{F}_{llkk},\ \hat{F}_{ll\langle 11\rangle},\ \hat{F}_{\langle 1111\rangle}$$
$$\hat{F}_{llkk1},\ \hat{F}_{ll\langle 111\rangle},\ \hat{F}_{\langle 11111\rangle},\ \hat{F}_{llkkjj},\ \hat{F}_{llkk\langle 11\rangle},\ \hat{F}_{ll\langle 1111\rangle},\ \hat{F}_{\langle 111111\rangle}.$$

Therefore COET of third order is slightly more efficient than ET84 because the quotient of variables is 13:16. This increased efficiency becomes more pronounced for higher order, albeit slowly. Thus fourth order COET versus ET220 – which has a similar norm – has a quotient of 24:30. Therefore COET is preferable, since it needs less fields for the purpose of adequately describing the light scattering spectrum.

[2] It is not immediately obvious that the one-dimensional equations are appropriate for light scattering. But it can be shown that that is indeed the case: the equations that are coupled to the density field form exactly the same system as the one-dimensional equations.

References

[1] Müller, I, Ruggeri, T. (1998): Rational extended thermodynamics. (Springer Tracts in Natural Philosophy, vol. 37). Springer, New York

[2] Müller, I., Reitebuch, D., Weiss, W. (2003): Extended thermodynamics – consistent in order of magnitude. Contin. Mech. Thermodyn. **15**, 113–146

[3] Bhatnagar, P.L., Gross, E.P., Krook, M. (1954): A model for collision processes in gases I. Small amplitude processes in charged and neutral one-component systems. Phys. Rev. **94**, 511–525

[4] Weiss, W. (1990): Zur Hierarchie der Erweiterten Thermodynamik. Dissertation. Technische Universität Berlin, Berlin

[5] Weiss, W., Müller, I. (1995): Light scattering and extended thermodynamics. Contin. Mech. Thermodyn. **7**, 123–177

[6] Yip, S., Nelkin, M. (1964): Application of a kinetic model to time-dependent density correlations in fluids. Phys. Rev. (2) A **135**, 1241–1247

[7] Junk, M. (2004): Maximum entropy moment problems and extended Euler equations. In: Abdallah, N.B. et al. (eds.): Transport in transition regimes. Springer, New York, pp. 189–198

[8] Grad, H. (1949): On the kinetic theory of rarefied gases. Comm. Pure Appl. Math. **2**, 331–407

[9] Barbera, E., Müller, I., Reitebuch, D., Zhao, N.-R. (2004): Determination of boundary conditions in extended thermodynamics via fluctuation theory. Contin. Mech. Thermodyn., to appear. DOI: 10.1007/s00161-003-0165-x .

[10] Clark, N.A. (1975): Inelastic light scattering from density fluctuations in dilute gases. The kinetic-hydrodynamic transition in a monatomic gas. Phys. Rev. A **12**, 232–244

On instability sources in dynamical systems

S. Rionero

Abstract. A basic characteristic Lyapunov functional V is introduced for the dynamical systems generated by a pair of reaction-diffusion PDEs. with non-constant coefficients. The sign of the derivative of V, along the solutions, is linked through an immediate simple relation to the eigenvalues. This allows us to localize the *sources of instability*, i.e., the points at which the instability begins.

1 Introduction

We consider the dynamical systems generated by the binary system of PDEs.

$$\begin{cases} u_t = a_1(\mathbf{x}, R, C)\, u + a_2(\mathbf{x}, R, C)\, v + \gamma_1 \Delta u + f(u, v, \nabla u, \nabla v, \Delta u, \Delta v), \\ v_t = a_3(\mathbf{x}, R, C)\, u + a_4(\mathbf{x}, R, C)\, v + \gamma_2 \Delta v + g(u, v, \nabla u, \nabla v, \Delta u, \Delta v) \end{cases} \tag{1}$$

with f and g nonlinear and

$$\begin{cases} a_i : \ (\mathbf{x},\ R,\ C) \in \Omega \times (\boldsymbol{R}^+)^2 \to a_i(\mathbf{x},\ R,\ C) \in \boldsymbol{R}, \\ a_i \in C(\Omega \times (\boldsymbol{R}^+)^2), \qquad i \in [1, 2, 3, 4], \\ (u = v = 0) \Rightarrow f = g = 0, \\ u \ : \ (\mathbf{x}, t) \in \Omega \times \boldsymbol{R}^+ \to u(\mathbf{x}, t) \in \boldsymbol{R}, \\ v \ : \ (\mathbf{x}, t) \in \Omega \times \boldsymbol{R}^+ \to v(\mathbf{x}, t) \in \boldsymbol{R}, \end{cases} \tag{2}$$

where Ω is a bounded domain in $\boldsymbol{R}^3$ with smooth boundary $\partial\Omega$ and $R,\ C$ are positive dimensionless parameters characteristic of the phenomenon described by (1).

To (1) we append the Dirichlet boundary conditions

$$u = v = 0 \qquad \text{on } \partial\Omega \times \boldsymbol{R}^+ \tag{3}$$

or the Neumann boundary conditions, where $\mathbf{n}$ is the unit outward normal to $\partial\Omega$,

$$\frac{du}{d\mathbf{n}} = \frac{dv}{d\mathbf{n}} = 0 \qquad \text{on } \partial\Omega \times \boldsymbol{R}^+ \tag{4}$$

with the additional conditions

$$\int_\Omega u\, d\Omega = \int_\Omega v\, d\Omega = 0 \quad \forall\, t \in \mathbb{R}^+ \tag{5}$$

in case (4). Furthermore we denote the $L^2(\Omega)$-norm by $\|\cdot\|$ and by $H_0^1(\Omega),\ H_*^1(\Omega)$ the Sobolev functional spaces such that

$$\begin{cases} \varphi \in H_0^1(\Omega) \to \{\varphi^2 + (\nabla\varphi)^2 \in L^2(\Omega),\ \varphi = 0 \text{ on } \partial\Omega\}, \\ \varphi \in H_*^1(\Omega) \to \left\{\varphi^2 + (\nabla\varphi)^2 \in L^2(\Omega),\ \dfrac{d\varphi}{d\mathbf{n}} = 0 \text{ on } \partial\Omega,\ \bar\varphi = 0\right\} \end{cases}$$

with

$$\bar\varphi = \int_\Omega \varphi\, d\Omega\,;$$

we assume that the solutions belong to $H_0^1(\Omega)$ in case (3) and to $H_*^1(\Omega)$ in case (4–5) [9–12].

Our goal is to obtain conditions sufficient for the linear instability of the critical point $O \equiv (u \equiv v \equiv 0)$. To this end we associate to (1) its linear version

$$\begin{cases} u_t = b_1(\mathbf{x}, R, C)\, u + b_2(\mathbf{x}, R, C)\, v + f^*, \\ v_t = b_3(\mathbf{x}, R, C)\, u + b_4(\mathbf{x}, R, C)\, v + g^* \end{cases} \tag{6}$$

with

$$\begin{cases} b_1(\mathbf{x}, R, C) = a_1(\mathbf{x}, R, C) - \gamma_1\bar\alpha^2\,, & b_4(\mathbf{x}, R, C) = a_4(\mathbf{x}, R, C) - \gamma_2\bar\alpha^2, \\ b_2(\mathbf{x}, R, C) = a_2(\mathbf{x}, R, C)\,, & b_3(\mathbf{x}, R, C) = a_3(\mathbf{x}, R, C), \\ f^* = \gamma_1(\Delta u + \bar\alpha^2\, u)\,, & g^* = \gamma_2(\Delta v + \bar\alpha^2\, v), \end{cases} \tag{7}$$

where $\bar\alpha^2(\Omega)$ is the positive constant given by

$$\frac{1}{\bar\alpha^2} = \max\, \frac{\|\varphi\|^2}{\|\nabla\varphi\|^2}\,, \tag{8}$$

respectively, in $H_0^1(\Omega)$ and $H_*^1(\Omega)$ [1].

Our aim is to precisely characterize for (6) the critical instability values (R_C, C_C) for the parameters and the spatial location in Ω of the points $\mathbf{x}_0$ at which the instability with respect to $(\|u\|^2 + \|v\|^2)$ arises, at least in the class of the kinematically admissible perturbations $(u,\ v)$, i.e. those such that, according to their own reference spaces, $(u,\ v) \in (H_0^1)^2$ or $(u,\ v) \in (H_*^1)^2$. The points $\mathbf{x}_0$ will play then the role of *instability sources*.

The plan of the paper is as follows. In Sect. 2 we show that exists a *basic characteristic Lyapunov functional* such that the sign of its derivative along the solutions

of (6) strictly depends on the signs of the eigenvalues of (6) through the product $A\,I$, with

$$\begin{cases} A(\mathbf{x}, R, C) = b_1 b_4 - b_2 b_3, \\ \\ I(\mathbf{x}, R, C) = b_1 + b_4\,. \end{cases} \tag{9}$$

In Sect. 3, we obtain the following basic result.

Theorem 1. *Let* $(\bar{R},\ \bar{C}) \in (\boldsymbol{R}^+)^2$ *and suppose that there exists a point* $\mathbf{x}_0 \in \Omega$ *such that*

$$\begin{cases} A(\mathbf{x}_0, \bar{R}, \bar{C}) > 0, \\ \\ I(\mathbf{x}_0, \bar{R}, \bar{C}) > 0\,. \end{cases} \tag{10}$$

Then, for $(R = \bar{R},\ C = \bar{C})$*, the critical point* $O \equiv (u = v = 0)$ *is unstable.*

Further we determine a procedure for obtaining the critical values of the parameters R and C and for localizing in Ω the instability sources at least when one of the coefficients $a_i\ (i = 1, 2, 3, 4)$ depends on $\mathbf{x}$ (Sect. 3). Finally, in Sect. 4, we apply Theorem 1 to obtain an instability condition for doubly diffusive convection in a rotating porous medium, uniformly heated and salted from below. In this case the coefficients a_i are independent of $\mathbf{x}$. The dependence on $\mathbf{x}$ appears when the layer is not uniformly heated and/or salted from below [8].

2 The basic characteristic Lyapunov functional

We set

$$(\varphi_1, \varphi_2) = \int_\Omega \varphi_1\, \varphi_2\, d\Omega, \tag{11}$$

and introduce the basic characteristic Lyapunov functional

$$V = \frac{1}{2}\left[\int_\Omega A(u^2 + v^2)\, d\Omega + \|b_1 v - b_3 u\|^2 + \|b_2 v - b_4 u\|^2\right]. \tag{12}$$

Since

$$\frac{dV}{dt} = \begin{cases} [(A\,u, u_t) + (A\,v, v_t)] + ((b_1^2 + b_2^2)v, v_t) + \\ \\ ((b_3^2 + b_4^2)u, u_t) - [((b_1 b_3 + b_2 b_4)v, u_t) + ((b_1 b_3 + b_2 b_4)u, v_t)]\,, \end{cases} \tag{13}$$

and since along the solutions of (6) one easily obtains

$$\begin{cases} (u, u_t) = (b_1\,u,\ u) + (b_2\,u, v) + (u, f^*), \\ \\ (v, v_t) = (b_3\,u, v) + (b_4\,v,\ v) + (v, g^*), \\ \\ (v, u_t) = (b_1\,u, v) + (b_2\,v,\ v) + (v, f^*), \\ \\ (u, v_t) = (b_3\,u,\ u) + (b_4\,u, v) + (u, g^*), \end{cases} \tag{14}$$

by straightforward calculations we see that

$$\left.\frac{dV}{dt}\right|_{(6)} = \int_{\Omega} A\, I(u^2+v^2)\, d\Omega + \Psi \tag{15}$$

with

$$\begin{cases} \Psi = (\alpha_1 u - \alpha_3 v, f^*) + (\alpha_2 v - \alpha_3 u, g^*), \\ \alpha_1 = A + b_3^2 + b_4^2, \quad \alpha_2 = A + b_1^2 + b_2^2, \\ \alpha_3 = b_1 b_3 + b_2 b_4\,, \end{cases} \tag{16}$$

where $\left.\dfrac{dV}{dt}\right|_{(6)}$ denotes the derivative of V evaluated along the solutions of (6).

3 Instability sources: proof of Theorem 1

Let $\mathbf{x}_0 = (x_0, y_0, z_0)$ satisfy (10). In view of $(2)_2$ we may assume that exists a $d > 0$ such that the cube D

$$\begin{cases} x_0 - d \le x \le x_0 + d, \\ y_0 - d \le y \le y_0 + d, \\ z_0 - d \le z \le z_0 + d \end{cases} \tag{17}$$

is strictly contained in Ω and that, for all $\mathbf{x} \in D$,

$$\begin{cases} A(\mathbf{x}) \ge \dfrac{A(\mathbf{x}_0)}{2}, \\ I(\mathbf{x}) \ge \dfrac{I(\mathbf{x}_0)}{2}\,. \end{cases} \tag{18}$$

We now consider a kinematically admissible perturbation $(\bar{u}, \bar{v})$ such that $\bar{u}$ and $\bar{v}$ satisfy in Ω, at each instant, the system

$$\begin{cases} \Delta U = -\alpha^2 U & \text{in } D, \\ U = 0 & \text{on } \partial D, \end{cases} \tag{19}$$

and are defined by continuity in $\Omega - D$.[1] In view of (19), we obtain $f^* = g^* = 0$ and, along $(\bar{u}, \bar{v})$, we have

$$\begin{cases} V(\bar{u},\ \bar{v}) = \dfrac{1}{2}\left[\displaystyle\int_D A(\bar{u}^2 + \bar{v}^2)\, dD + \|b_1\bar{v} - b_3\bar{u}\|^2 + \|b_2\bar{v} - b_4\bar{u}\|^2\right], \\ \left.\dfrac{dV}{dt}\right|_{(6)} = \displaystyle\int_D AI(\bar{u}^2 + \bar{v}^2)\, dD \geq \frac{A(\mathbf{x}_0)I(\mathbf{x}_0)}{4}\int_D (\bar{u}^2 + \bar{v}^2)\, dD\,. \end{cases} \tag{20}$$

But one may assume that

$$\int_D (\bar{u}^2 + \bar{v}^2) dD \geq m = \text{positive constant} \tag{21}$$

and hence the instability immediately follows. In fact, with

$$\frac{k_2}{5} = \max\{A(\mathbf{x}_0), b_1^2(\mathbf{x}_0), b_2^2(\mathbf{x}_0), b_3^2(\mathbf{x}_0), b_4^4(\mathbf{x}_0)\}\,, \tag{22}$$

in view of (12) it follows that

$$V > \frac{k_2}{2}(\|u\|^2 + \|v\|^2)\,. \tag{23}$$

Setting

$$k_3 = \frac{2mA(\mathbf{x}_0)\, I(\mathbf{x}_0)}{k_2} \tag{24}$$

from (15) and (24) we conclude that, for all $t \geq 0$,

$$\left.\frac{dV}{dt}\right|_{(6)} > k_3\, V, \tag{25}$$

i.e., the instability when

$$V \geq V(0)e^{k_3 t}\,. \tag{26}$$

It remains to show that Theorem 1 allows to localize in Ω the instability sources (i.e., the points $\mathbf{x}_0 \in \Omega$ at which the instability begins) and the critical instability

[1] For instance,

$$\begin{cases} u = \dfrac{\alpha_1^2 d^2}{\pi^2} h(t) \sin\pi\dfrac{x - x_0}{d} \sin\pi\dfrac{y - y_0}{d} \sin\pi\dfrac{z - z_0}{d} \\ v = \dfrac{h_1(t)}{h(t)} u \end{cases}$$

with h_1 and h smooth functions defined on $\boldsymbol{R}^+$ and α_1^2 = constant.

values (R_C, C_C) for the parameters R and C. In fact let

$$\begin{cases} m_1 = \inf\limits_{\Omega\times(\mathbb{R}^+)^2} A(\mathbf{x}, R, C), \\ \\ m_2 = \inf\limits_{\Omega\times(\mathbb{R}^+)^2} I(\mathbf{x}, R, C) \end{cases} \tag{27}$$

under the conditions

$$\begin{cases} A(\mathbf{x}, R, C) > 0, \\ \\ I(\mathbf{x}, R, C) > 0\,. \end{cases} \qquad (\mathbf{x}, R, C) \in \Omega \times (\mathbb{R}^+)^2 \tag{28}$$

- In the case $m_1 \leq m_2$, if $\{(\mathbf{x}_i, R_i, C_i)\}$ $(i = 1, 2, ...)$ is the set of solutions of the equation

$$A(\mathbf{x}, R, C) = m_1, \qquad (\mathbf{x}, R, C) \in \Omega \times (\mathbb{R}^+)^2\,, \tag{29}$$

then the critical values R_C and C_C for R and C are

$$\begin{cases} R_C = \inf\ R_i, \\ \\ C_C = \inf\ C_i\,, \end{cases} \qquad i = 1, 2, ... \tag{30}$$

and the instability sources in Ω are the points $\mathbf{x}_0$ such that

$$A(\mathbf{x}_0, R_C, C_C) = m_1\,. \tag{31}$$

- In the case $m_1 \geq m_2$, one has to substitute in (29-31) m_2 for m_1 and I for A.

Remark 2. We emphasize the relevance of (30-31) in relation to the phenomena modeled by (1) with non-constant coefficients. For instance, if (1) models the growth of a disease in the human body, (30-31) allows us to determine the conditions for the onset of the disease and to localize where the disease will begin. Analogously if (1) models the perturbations to a financial equilibrium, (30-31) allows us to determine the condition for the loss of financial equilibrium and to localize where it will happen.

Remark 3. We observe that Theorem 1 also allows to determine instability conditions in the case of more than two reaction-diffusion equations. For instance, in the case

$$\begin{cases} u_t = a_1\, u + a_2\, v + a_3\, w + \gamma_1 \Delta u, \\ \\ v_t = a_4\, u + a_5\, v + a_6\, w + \gamma_2 \Delta v, \\ \\ w_t = a_7\, u + a_8\, v + a_9\, w + \gamma_3 \Delta w \end{cases} \tag{32}$$

with

$$\begin{cases} \gamma_i = \text{constant} > 0, \quad i = 1, 2, 3, \\ a_9 \neq \gamma_3 \bar{\alpha}^2 \end{cases}$$

under the boundary conditions

$$u = v = w = 0 \qquad \text{on} \quad \partial\Omega\,, \tag{33}$$

on considering the "perturbations" (u, v, w) with

$$w_t = 0, \qquad \Delta w = -\bar{\alpha}^2\, w$$

from $(32)_3$, we see that (32) reduces to the binary system $\left(\mu = \frac{1}{\gamma_3 \alpha^{-2} - a_9}\right)$

$$\begin{cases} u_t = (a_1 + \mu\, a_7)\, u + (a_2 + \mu\, a_8)\, v + \gamma_1 \Delta u, \\ v_t = (a_4 + \mu\, a_7)\, u + (a_5 + \mu\, a_8)\, v + \gamma_2 \Delta v\,. \end{cases} \tag{34}$$

to which one can apply Theorem 1.

4 Double diffusive convection in rotating porous media: instability conditions

The Darcy-Oberbeck-Boussinesq (DOB.) equations governing the motion of a binary porous fluid mixture bounded by two horizontal planes uniformly rotating around the vertical axis z are [2–4]:

$$\begin{cases} \nabla p = -\dfrac{\mu}{k}\, \mathbf{v} + \rho_f \mathbf{g} - 2\, \dfrac{\rho_0}{\epsilon} \underline{\omega} \times \mathbf{v}, \\ \nabla \cdot \mathbf{v} = 0, \\ A\ T_t + \mathbf{v} \cdot \nabla T = k_T \Delta T, \\ \epsilon C_t + \mathbf{v} \cdot \nabla C = k_C \Delta C, \end{cases} \tag{35}$$

where $\rho_f = \rho_0[1 - \gamma_T(T - T_0) + \gamma_C(C - C_0)]$, p_1 is the pressure field, $p = p_1 - \dfrac{1}{2}\rho_0[\underline{\omega} \times \mathbf{x}]^2$, $\underline{\omega} = \omega \mathbf{k}$ is the constant angular velocity, γ_C, γ_T are, respectively, the thermal and solute expansion coefficients, ϵ is the porosity of the medium, T_0 is a reference temperature, C_0 is a reference concentration, $\mathbf{v}$ is the seepage velocity field, C is the concentration field, μ is the viscosity, T is the temperature field, k_T, k_C are, respectively, the thermal and salt diffusivity, c is the specific heat of the solid, ρ_0 is the fluid density at reference temperature T_0, $A = \dfrac{(\rho_0\, c)_m}{(\rho_0\, c_p)_f}$, c_p is the specific

heat of fluid at constant pressure and the subscript, m and f refer the porous medium and the fluid, respectively. To (35) we append the boundary conditions

$$\begin{cases} T_L = T_0 + (T_1 - T_2)/2, \;\; C_L = C_0 + (C_1 - C_2)/2 \;\; \text{on } z = 0, \\ T_U = T_0 - (T_1 - T_2)/2, \;\; C_U = C_0 - (C_1 - C_2)/2 \;\; \text{on } z = d \end{cases} \tag{36}$$

with $T_1 > T_2$ and $C_1 > C_2$. We introduce the dimensionless quantities

$$\begin{cases} \mathbf{x} = d\,\mathbf{x}^*, & t = \dfrac{A\,d^2}{k_T} t^*, & \mathbf{v} = \dfrac{k_T}{d}\mathbf{v}^*, \\ P^* = \dfrac{k\,(p + \rho_0\, g\, z)}{\mu k_T}, & T^* = \dfrac{T - T_0}{T_1 - T_2}, & C^* = \dfrac{C - C_0}{C_1 - C_2}; \end{cases}$$

omitting the asterisks, the dimensionless equations are:

$$\begin{cases} \nabla P = -\mathbf{v} + (R\,T - \mathscr{C} C)\mathbf{k} + \mathscr{T}\,\mathbf{v} \times \mathbf{k}, \\ \nabla \cdot \mathbf{v} = 0, \\ T_t + \mathbf{v} \cdot \nabla T = \Delta T, \\ \varepsilon\, Le C_t + Le\,\mathbf{v} \cdot \nabla C = \Delta C, \end{cases} \tag{37}$$

where

$$\begin{cases} \nu = \mu/\rho_0 & \text{is the kinematic viscosity,} \\ \varepsilon = \epsilon/A & \text{is the normalized porosity,} \\ \mathscr{T} = 2k\omega/\epsilon\,\nu & \text{is the Taylor-Darcy number,} \\ Le = k_T/k_C & \text{is the Levis number,} \\ R = \dfrac{\gamma_T\, g\,(T_1 - T_2) d\, k}{\nu\, k_T} & \text{is the thermal Rayleigh number,} \\ \mathscr{C} = \dfrac{\gamma_C\, g\,(C_1 - C_2) d\, k}{\nu\, k_T} & \text{is the solutal Rayleigh number.} \end{cases}$$

To (37) we append the boundary data

$$\begin{cases} T_L = 1/2, & C_L = 1/2 & \text{on } z = 0, \\ T_U = -1/2, & C_U = -1/2 & \text{on } z = 1. \end{cases} \tag{38}$$

Thus (37-38) admit the steady state solution (motionless state)

$$\begin{cases} \mathbf{v}_s = 0, & \nabla p_s(z) = (-R + \mathscr{C}) \left(z - \dfrac{1}{2} \right) \mathbf{k}, \\ T(z) = -\left(z - \dfrac{1}{2} \right), & C(z) = -\left(z - \dfrac{1}{2} \right). \end{cases} \tag{39}$$

With $\mathbf{u} = (u,\ v,\ w)$, θ, Γ, π the dimensionless perturbations to the (seepage) velocity, temperature, concentration and pressure fields, respectively, the equations governing the perturbations $\mathbf{u} = (u,\ v,\ w)$, θ, $Le\,\Gamma$, π are:

$$\begin{cases} \nabla\pi = -\mathbf{u} + (R\,\theta - Le\,\mathscr{C}\Gamma)\mathbf{k} + \mathscr{T}\,\mathbf{u}\times\mathbf{k}, \\ \nabla\cdot\mathbf{u} = 0, \\ \theta_t + \mathbf{u}\cdot\nabla\theta = \Delta\theta, \\ \varepsilon\,Le\Gamma_t + Le\,\mathbf{u}\cdot\nabla\Gamma = w + \Delta\Gamma \end{cases} \tag{40}$$

with the boundary conditions

$$w = \theta = \Gamma = 0 \qquad \text{on } z = 0, 1\,. \tag{41}$$

In the sequel we assume that the perturbation fields are periodic functions of x and y of periods $2\pi/a_x$, $2\pi/a_y$, respectively and we denote the periodicity cell by $\Omega = [0, 2\pi/a_x]\times[0, 2\pi/a_y]\times[0, 1]$. Lastly, to ensure that the steady state (39) is unique, we assume that

$$\int_\Omega u\,d\Omega = \int_\Omega v\,d\Omega = 0\,.$$

By taking the third component of the double curl of $(40)_1$ and linearizing we obtain

$$\begin{cases} \Delta w + \mathscr{T}^2 w_{zz} = \Delta_1(R\theta - Le\,\mathscr{C}\Gamma), \\ \theta_t = w + \Delta\theta, \\ \varepsilon\,Le\,\Gamma_t = w + \Delta\Gamma, \end{cases} \tag{42}$$

where $\Delta_1 = \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$. We observe that the set $\mathscr{S}$ of the *kinematically admissible perturbations* is characterized by (41), $(42)_1$, the periodicity and regularity conditions.

It is easily verified that, setting

$$\begin{cases} \bar{w} = \alpha(R\bar{\theta} - Le\,\mathscr{C}\,\bar{\Gamma}), \\ \bar{\theta} = \hat{\theta}(x, y, t)\,\sin(\pi z), \\ \bar{\Gamma} = \hat{\Gamma}(x, y, t)\,\sin(\pi z) \end{cases} \tag{43}$$

with

$$\begin{cases} \alpha = \dfrac{a^2}{\xi + \mathscr{T}^2\pi^2}, \qquad \xi = a^2 + \pi^2, \\ a^2 = a_x^2 + a_y^2 \end{cases} \tag{44}$$

and $\hat{\theta}$, $\hat{\Gamma}$ satisfying the plan form equation

$$\Delta_1 \cdot = -a^2 \cdot , \tag{45}$$

it follows that $(\bar{w},\ \bar{\theta},\ \bar{\Gamma}) \in \mathcal{J}$. Along this, the perturbations $(42)_2 - (42)_3$ become

$$\begin{cases} \bar{\theta}_t = \alpha R\bar{\theta} - \alpha\, Le\, \mathcal{C}\, \bar{\Gamma} + \Delta\bar{\theta}, \\ \varepsilon\, Le\bar{\Gamma}_t = \alpha R\bar{\theta} - \alpha\, Le\, \mathcal{C}\, \bar{\Gamma} + \Delta\bar{\Gamma}\,. \end{cases} \tag{46}$$

As

$$\Delta\bar{\theta} = -\xi\bar{\theta}, \qquad \Delta\bar{\Gamma} = -\xi\bar{\Gamma},$$

it follows that the constant $\bar{\alpha}^2$ appearing in (8) is given by ξ, and (46) can be written (omitting the bar) as

$$\begin{cases} \theta_t = b_1\, \theta + b_2\, \Gamma, \\ \Gamma_t = b_3\, \theta + b_4\, \Gamma, \end{cases} \tag{47}$$

with

$$\begin{cases} b_1 = \alpha R - \xi, \quad b_2 = -\alpha\, Le\, \mathcal{C}, \\ b_3 = \dfrac{1}{\varepsilon\, Le}\alpha R, \quad b_4 = -\dfrac{\alpha}{\varepsilon}\mathcal{C} - \dfrac{1}{\varepsilon\, Le}\,\xi\,. \end{cases} \tag{48}$$

Therefore, introducing the functional (12) with the b_i given by (48), and observing that

$$\begin{cases} A = -\dfrac{\xi}{\varepsilon\, Le}(\alpha R - Le\, \alpha\, \mathcal{C} - \xi), \\ I = \alpha R - \dfrac{\alpha}{\varepsilon}\mathcal{C} - \left(1 + \dfrac{1}{\varepsilon\, Le}\right)\xi \end{cases} \tag{49}$$

are constant, we conclude from (15) that

$$\left.\frac{dV}{dt}\right|_{(9)} = I\, A(\|\theta\|^2 + \|\Gamma\|^2)\,. \tag{50}$$

We now consider the system

$$\begin{cases} A > 0, \\ I > 0, \end{cases} \tag{51}$$

i.e., in view of (49),

$$\begin{cases} \xi + Le\, \alpha\, \mathcal{C} - \alpha R > 0, \\ \alpha R - \dfrac{\alpha}{\varepsilon}\,\mathcal{C} - \left(1 + \dfrac{1}{\varepsilon\, Le}\right)\xi > 0\,. \end{cases} \tag{52}$$

But Eqs. (48) imply that

$$\left(Le - \frac{1}{\varepsilon}\right) \mathscr{C} > \frac{1}{\varepsilon\, Le}\, \frac{\xi}{\alpha} \geq \frac{1}{\varepsilon\, Le}\, R_B \tag{53}$$

with

$$R_B = \pi^2 \left(1 + \sqrt{1 + \mathcal{T}^2}\right) ; \tag{54}$$

hence, only if

$$\begin{cases} \varepsilon\, Le > 1, \\ \mathscr{C} \geq C^* = \dfrac{R_B}{Le(\varepsilon\, Le - 1)} , \end{cases} \tag{55}$$

can (49) hold.

Theorem 2. *Let*

$$\varepsilon\, Le > 1 , \qquad \mathscr{C} \geq C^* . \tag{56}$$

Then

$$R > R_C = \frac{\mathscr{C}}{\varepsilon} + \left(1 + \frac{1}{\varepsilon\, Le}\right) R_B \tag{57}$$

implies instability.

Proof. We have to show that, for any $k > 0$,

$$R = \frac{\mathscr{C}}{\varepsilon} + \left(1 + \frac{1}{\varepsilon\, Le}\right) (R_B + k) \tag{58}$$

implies instability. To this goal, we show that there exists suitable a^2 such that

$$\begin{cases} R - Le\, \mathscr{C} - \dfrac{\xi}{\alpha} < 0, \\ R - \dfrac{\mathscr{C}}{\varepsilon} - \left(1 + \dfrac{1}{\varepsilon\, Le}\right) \dfrac{\xi}{\alpha} > 0, \end{cases} \tag{59}$$

i.e.,

$$R - \frac{\mathscr{C}}{\varepsilon} + \left(\frac{1}{\varepsilon} - Le\right) \mathscr{C} < \frac{\xi}{\alpha} < \frac{\varepsilon\, Le}{1 + \varepsilon\, Le} \left(R - \frac{\mathscr{C}}{\varepsilon}\right) . \tag{60}$$

In view of (55) and (56), inequality (60) becomes

$$R_B - \frac{\mathscr{C} - C^*}{\varepsilon\, Le}(\varepsilon\, Le - 1) < \frac{\xi}{\alpha} < R_B + k\,. \tag{61}$$

But

$$\frac{\xi}{\alpha} = \frac{(a^2+\pi^2)[a^2+\pi^2(1+\mathcal{T}^2)]}{a^2} \tag{62}$$

takes its infimum value at $\bar{a}^2 = \pi^2\sqrt{1+\mathcal{T}^2}$:

$$\left(\frac{\xi}{\alpha}\right)_{a^2=\bar{a}^2} = R_B; \tag{63}$$

hence in $]\bar{a}^2, a_*^2[$ with

$$\bar{a}^2 < a_*^2: \quad \left(\frac{\xi}{\alpha}\right)_{a^2=a_*^2} = R_B + k \tag{64}$$

there exists suitable $a^2 > 0$ such that inequality (61) holds. Then for such a^2 we obtain

$$\left.\frac{dV}{dt}\right|_{(9)} = I\,A(\|\theta\|^2 + \|\Gamma\|^2) > 0 \tag{65}$$

which implies instability.

Remark 4. The problem arises of establishing, in case (56), whether R_C given by (57) represents the effective threshold of instability. That this is the case, is proved in [r2]

Remark 5. The stability-instability of a double diffusive convection in a porous medium (in different circumstances) was also considered in [5,6].

Acknowledgements

This work was carried out under the auspices of the GNFM of INDAM and MIUR (PRIN): "Nonlinear mathematical problems of wave propagation and stability in models of continuous media".

References

[1] Flavin, J.N., Rionero, S. (1996): Qualitative estimates for partial differential equations. CRC Press, Boca Raton, FL
[2] Nield, D.A., Bejan, A. (1992): Convection in porous media. Springer, New York
[3] Joseph, D.D. (1976): Stability of fluid motions. Vols. I, II. Springer, New York

[4] Straughan, B. (2004): The energy method, stability, and nonlinear convection. 2nd edition. (Applied Mathematical Sciences, vol. 91). Springer, New York
[5] Lombardo, S., Mulone, G., Straughan, B. (2001): Non-linear stability in the Bénard problem for a double-diffusive mixture in a porous medium. Math. Methods Appl. Sci. **24**, 1229–1246
[6] Mulone, G. (1994): On the nonlinear stability of a fluid layer of a mixture heated and salted from below. Contin. Mech. Thermodyn. **6**, 161–184
[7] Rionero, S. (2004): On the rigorous reduction of the L^2- stability of the solutions to a binary reaction - diffusion system of P.D.E. into the stability of the solutions to a binary system of O.D.E. Preprint.
[8] Capone, F., Rionero, S. (2004): On the instability of double diffusive convection in porous media under boundary data periodic in space. In: Romano, G., Rionero, S. (eds.): Recent trends in the applications of mathematics to mechanics. Springer, Berlin, pp. 1–8
[9] Ladyženskaja, O.A., Solonnikov, V.A. Uralćeva, N.N. (1968): Linear and quasilinear equations of parabolic type. (Translations of Mathematical Monographs, vol. 23). American Mathematical Society Providence, RI
[10] Friedman, A. (1964): Partial differential equations of parabolic type. Prentice-Hall, Englewood Cliffs, NJ
[11] Smoller, J. (1983): Shock waves and reaction – diffusion equations. (A "Series of Comprehensive Studies in Mathematics, **258**). Springer, New York
[12] Amann, H. (1986): Quasilinear parabolic systems under nonlinear boundary conditions. Arch. Rational Mech. Anal. **92**, 153–192
Amann, H. (1990): Dynamic theory of quasilinear parabolic equations. II. Reaction - diffusion systems. Differential Integral Equations **3**, 13–75
Amann, H. (1989): Dynamic theory of quasilinear parabolic systems. III. Global existence. Math. Z. **202**, 219–250

Tangent stiffness of elastic continua on manifolds

G. Romano, M. Diaco, C, Sellitto

Abstract. Non-linear models of beams, shells and polar continua are addressed from a general point of view with the aim of providing a clear motivation of the fact that the tangent stiffness of these structural models may be nonsymmetric. Classical and polar models of continua are investigated and a critical analysis of the commonly adopted strain measures is performed. It is emphasized that the kinematic space of a polar continuum is a non-linear differentiable manifold. Accordingly, by choosing a connection on the manifold, the Hessian operator of the elastic potential is defined as the second covariant derivative of the elastic potential. The Hessian operator can be expressed as the difference between the second directional derivative along the trial and test fields and the first directional derivative in the direction of the covariant derivative of the test field along the trial field. It follows that the evaluation of the Hessian operator requires the extension of the local virtual displacement to a vector field over the non-linear kinematic manifold. In any case the tensoriality of the Hessian operator ensures that the result is independent of the choice of the extension, and its symmetry depends on whether or not the assumed connection is torsionless. Conservative and nonconservative loads are considered and it is shown that, at equilibrium points, the tangent stiffness is independent of the chosen connection on the fiber manifold and symmetry holds for conservative loads.

1 Introduction

Polar models of beams and shells have been investigated by an ever increasing number of scholars since the pioneering contributions of J.C. Simo and co-workers who, in the years 1985-1989, the problem of providing a geometrically exact theory of polar beams and shells undergoing large deformations and a numerical implementation scheme for the related elastostatic and elastodynamic problems (see [4–6,8,9,12,13,15,17,18]).

By now, a number of papers have been devoted to the formulation of a suitable interpolation of the kinematic variables in finite element approximations of polar continua (see, e.g., [36,39,42,43,53]).

A list of recent contributions to the theoretical and computational analysis of polar beams and shells is provided in the references at the end of the paper.

Polar models of continua include one-dimensional polar beams (also called Timoshenko beams or shear deformable beams), two-dimensional polar shells (Reissner-Mindlin shells or shear deformable shells) and three-dimensional polar continua (Cosserat continua).

On special feature of polar models is that the evolution processes of the body take place in an ambient space which is no longer the usual three-dimensional euclidean space but instead a more general geometrical object, a non-linear manifold. This is

due to the fact that the polar structure of the continuum is represented by means of an additional set of kinematic variables which, at each point of the parent classical continuum, vary over a non-linear manifold, the fiber manifold. In polar beams the fiber manifold is the special orthogonal group of rotations which allows us to monitor the orientation of the cross sections of the beam, assumed to be rigid bodies, hinged to the beam axis, which can rotate indenpendently of the position of the beam axis. In polar shells the fiber manifold is the unit sphere, i.e., the locus to which the thickness-directors belong. Indeed the shell is described by a field of *needles* (or *rigid hairs*) attached at each point of the middle surface. The common length of the needles is equal to the constant thickness of the shell but they can be *combed* independently of the position of the middle surface. This model is referred to in the literature as a shell without drilling rotations since rotations of the needles around their axes are not taken into account. To accomodate for the interaction between shell and beam models assembled together to design a stiffened shell, another shell model has also been introduced, in which the polar structure is described by the rotations of a triad hinged at each point of the middle surface. This model is referred to in the literature as a shell with drilling rotations.

In Cosserat continua the fiber manifold is the special orthogonal group of rotations depicting the orientation of the rigid balls centered at each particle of the three-dimensional body which can rotate independently of the position of the parent particle.

The ambient spaces in which the evolution processes of these polar continua take place are trivial fiber bundles formed by the cartesian product of euclidean three-space and a non-linear fiber manifold.

The analysis of such polar models requires us to deal with non-linear geometrical objects and hence to rely on concepts and results of differential geometry. This aspect was underestimated in the initial investigations on polar beams, [4–6], and in the present authors' opinion has not yet been fully digested in spite of the contribution [14] provided by Simo to explain why the tangent stiffness of the polar beams evaluated in [5,6] was apparently nonsymmetric. Indeed the discussion given in [14] takes no account of the way in which the directional derivatives of the virtual displacement are defined, makes reference only to Riemannian connections and hence cannot explain why a nonsymmetric but tensorial tangent stiffness may occur. Further in [14] it is claimed that the right symmetric tangent stiffness can be obtained by simply taking the symmetric part of the nonsymmetric one, at least for conservative loadings. It can be shown [50] that this special property holds only for the polar beam model and that its validity is strictly connected to the special extension of virtual displacements considered in [14].

As we shall see, in general the expression of the tensorial tangent stiffness at nonequilibrium points depends on the choice of the connection over the fiber manifold which describes the polar behavior of the continuum. At an equilibrium point, however, the tangent stiffness is independent of the chosen connection and symmetry holds for conservative referential loads. At a nonequilibrium point a nonsymmetric but tensorial stiffness may occur if the torsion of the connection does not vanish and

the covariant derivative of the chosen extension of the virtual displacement vanishes identically [49,50].

The aim of the present paper is to provide an outline of a self-consistent treatment of non-linear equilibrium problems of an elastic continuum endowed with a polar structure. Special emphasis is put on the problem of the evaluation of the tangent stiffness of polar continua.

The basic notions of configuration maps and tangent (virtual) displacements are reformulated in an way suitable to deal with polar models.

The appropriate ambient space for polar continua is a non-linear manifold which has the geometric structure of a fiber bundle. In structural models of engineering interest this fiber bundle is simply the cartesian product of the physical space (three-dimensional euclidean space) and a non-linear manifold which characterizes the local structure of the polar continuum.

The space of configurations is a non-linear manifold of continuously differentiable mappings which map the base manifold of a reference placement into the actual placement in the ambient manifold. Virtual displacements are defined as tangent vectors to the manifold of admissible configurations.

A general discussion of finite strain measures is provided and the equilibrium condition of the polar continuum in a reference placement is formulated by invoking a consistency property of finite strain measures.

It is shown that the notion of a connection over the fiber manifold allows one to define, on the manifold of configuration maps, the covariant derivative of one-forms which have the physical meaning of force systems acting on the body. The covariant differentiation leads to the notion of absolute (or covariant) time derivative which, applied to the equilibrium condition, provides the incremental equilibrium condition governed by the tangent stiffness operator.

We emphasize the fact that the evaluation of the covariant derivative of one-forms requires that the virtual displacement tangent at a given configuration be extended to vector fields in a neighborhood of the configuration.

The roles played, in evaluating the tensorial tangent stiffness, by the connection assumed on the fiber manifold and by the chosen extension of the virtual displacements, are discussed in detail. It is shown that, at an equilibrium point, the tangent stiffness is independent of the assumed connection and its symmetry depends on whether the referential loads are conservative or not.

2 Differentiable manifolds

We provide here, for the sake of completeness and clarity, basic facts and definitions about differentiable manifolds (see, e.g., [7]).

- Let $\mathbb{M}$ be a set and E a Banach space. A *chart* $\{U, \varphi\}$ on $\mathbb{M}$ is a pair with $\varphi : U \mapsto E$ a bijection from the subset $U \subset \mathbb{M}$ onto an open set in E. A C^k-*atlas* $\mathcal{A}$ on $\mathbb{M}$ is a family of charts $\{\{U_i, \varphi_i\} \mid i \in I\}$ such that $\{\cup U_i \mid i \in I\}$ is a covering of $\mathbb{M}$ and that the overlap maps are C^k-diffeomorphisms.

- Two atlases are equivalent if their union is a C^k-atlas, and the union of all the atlases equivalent to a given one $\mathscr{A}$ is called the *differentiable structure* generated by $\mathscr{A}$.

- A *C^k-differentiable manifold* modeled on the Banach space E is a pair $\{\mathbb{M}, \mathscr{D}\}$, where $\mathscr{D}$ is an equivalence class of C^k-atlases on $\mathbb{M}$. The space E is called the model space.

- A subset $\mathbb{O}$ of a differentiable manifold $\mathbb{M}$ is said to be *open* if for each $\mathbf{x} \in \mathbb{O}$ there is a chart $\{U, \varphi\}$ such that $\mathbf{x} \in U$ and $U \subset \mathbb{O}$.

- A morphism between two differentiable manifolds $\mathbb{M}_1$ and $\mathbb{M}_2$ is a differentiable map $\phi : \mathbb{M}_1 \mapsto \mathbb{M}_2$.

- A C^k-diffeomorphism $\phi \in \mathrm{C}^k(\mathbb{M}_1 ; \mathbb{M}_2)$ is a morphism which is invertible and C^k, along with its inverse.

- The *tangent space* $\mathbb{T}_{\mathbb{M}}(\mathbf{x})$ at a point $\mathbf{x} \in \mathbb{M}$ is the linear space of *tangent vectors* $\{\mathbf{x}, \mathbf{v}\} : \mathrm{C}^r(\mathbf{x}, U) \mapsto \mathrm{C}^{r-1}(\mathbf{x}, U)$ where $\mathrm{C}^r(\mathbf{x}, U)$ is the germ of scalar functions which are r-times continuously differentiable in a neighborhood U of $\mathbf{x} \in \mathbb{M}$. Tangent vectors at a point are uniquely defined by requiring that they satisfy the formal properties of a *point derivation*:

$$\left.\begin{cases} (\mathbf{v}_1 + \mathbf{v}_2)(f) = \mathbf{v}_1(f) + \mathbf{v}_2(f), & \textit{additivity}, \\ \mathbf{v}(a\, f) = a\, \mathbf{v}(f), \quad a \in \mathscr{R}, & \textit{homogenity}, \end{cases}\right\} \quad \mathscr{R}\text{-linearity},$$

$$\mathbf{v}(f g) = \mathbf{v}(f)\, g + f(\mathbf{v}(g)), \qquad \text{Leibniz rule},$$

  where $f \in \mathrm{C}^r(\mathbf{x}, U)$. This point of view, which identifies the tangent vectors at a point with the directional derivatives of smooth scalar functions at that point, is the most convenient for obtaining the basic results of differential geometry.

- The *tangent bundle* $\mathbb{T}_{\mathbb{M}}$ of the manifold $\mathbb{M}$ is the disjoint union of the pairs $\{\mathbf{x}, \mathbb{T}_{\mathbb{M}}(\mathbf{x})\}$ with $\mathbf{x} \in \mathbb{M}$. An element $\{\mathbf{x}, \mathbf{v}\} \in \{\mathbf{x}, \mathbb{T}_{\mathbb{M}}(\mathbf{x})\}$ is said to be a tangent vector *applied at the base point* $\mathbf{x} \in \mathbb{M}$. We denote by $\boldsymbol{\tau}_{\mathbb{M}} : \mathbb{T}_{\mathbb{M}} \mapsto \mathbb{M}$ the projection on the base point: $\boldsymbol{\tau}_{\mathbb{M}}(\{\mathbf{x}, \mathbf{v}\}) = \mathbf{x}$.

- The *cotangent bundle* $\mathbb{T}^*_{\mathbb{S}}$ of the manifold $\mathbb{M}$ is the disjoint union of the pairs $\{\mathbf{x}, \mathbb{T}^*_{\mathbb{M}}(\mathbf{x})\}$, where $\mathbb{T}^*_{\mathbb{M}}(\mathbf{x})$ is the topological dual space of $\mathbb{T}_{\mathbb{M}}(\mathbf{x})$. The elements of the cotangent bundle are called *covectors*. We denote by $\mathbb{T}_{\mathbb{M}}(\mathscr{P}) \subseteq \mathbb{T}_{\mathbb{M}}$ the disjoint union of the pairs $\{\mathbf{x}, \mathbb{T}_{\mathbb{M}}(\mathbf{x})\}$ with $\mathbf{x} \in \mathscr{P} \subseteq \mathbb{M}$.

- A *finite-dimensional* differentiable manifold is a manifold modeled on a finite-dimensional normed linear space. All the tangent spaces of a finite-dimensional differentiable manifold are finite-dimensional linear spaces of the same dimension.

- A C^k-*fiber bundle* with typical fiber the C^k-manifold $\mathbb{F}$ and base the C^k-manifold $\mathbb{B}$ is a C^k-surjective map $\boldsymbol{\pi}_{\mathbb{S}} : \mathbb{S} \mapsto \mathbb{B}$ which is locally a cartesian product. This means that the C^k-manifold $\mathbb{B}$ has an open atlas $\{\{U_i, \varphi_i\} \mid i \in I\}$ such that for each $i \in I$ there is a C^k-diffeomorphism $\boldsymbol{\phi}_i : \boldsymbol{\pi}_{\mathbb{S}}^{-1}(U_i) \mapsto U_i \times \mathbb{F}$ such that $\boldsymbol{\tau}_i \circ \boldsymbol{\phi}_i = \boldsymbol{\pi}_{\mathbb{S}}$, where $\boldsymbol{\tau}_i : U_i \times \mathbb{F} \mapsto U_i$ is the canonical projection. If $\mathbb{S} = \mathbb{B} \times \mathbb{F}$ the fiber bundle is said to be *trivial*. If the fiber $\mathbb{F}$ is a vector space the bundle is said to be a *vector bundle*. The *tangent bundle* $\mathbb{T}_{\mathbb{M}}$ of a manifold $\mathbb{M}$ is a vector bundle whose fibers are the tangent spaces to $\mathbb{M}$.
- A *fiber bundle morphism* $\boldsymbol{\chi} : \mathbb{S} \mapsto \mathbb{S}'$ between two differentiable manifolds $\mathbb{S}, \mathbb{S}'$ is a morphism satisfying the *fiber preserving property*:

  $$\boldsymbol{\pi}_{\mathbb{S}}(\mathbf{x}) = \boldsymbol{\pi}_{\mathbb{S}}(\mathbf{y}) \quad \Longrightarrow \quad (\boldsymbol{\pi}_{\mathbb{S}'} \circ \boldsymbol{\chi})(\mathbf{x}) = (\boldsymbol{\pi}_{\mathbb{S}'} \circ \boldsymbol{\chi})(\mathbf{y}) \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{S}\,.$$

  A fiber bundle morphism induces a base morphism $\boldsymbol{\chi}_{\mathbb{B}} : \mathbb{B} \mapsto \mathbb{B}'$ according to the relation

  $$\boldsymbol{\chi}_{\mathbb{B}} \circ \boldsymbol{\pi}_{\mathbb{S}} = \boldsymbol{\pi}_{\mathbb{S}'} \circ \boldsymbol{\chi}\,.$$
- A *section* of the fiber bundle $\boldsymbol{\pi}_{\mathbb{S}} : \mathbb{S} \mapsto \mathbb{B}$ is a smooth map $\mathbf{s} : \mathbb{B} \mapsto \mathbb{S}$ such that

  $$(\boldsymbol{\pi}_{\mathbb{S}} \circ \mathbf{s})(\mathbf{x}) = \mathbf{x} \quad \forall\, \mathbf{x} \in \mathbb{B}\,.$$

  Vector fields $\hat{\mathbf{v}} : \mathbb{M} \mapsto \mathbb{T}_{\mathbb{M}}$ on a manifold $\mathbb{M}$ are sections of the tangent vector bundle $\boldsymbol{\tau}_{\mathbb{M}} : \mathbb{T}_{\mathbb{M}} \mapsto \mathbb{M}$; indeed they satisfy the property

  $$(\boldsymbol{\tau}_{\mathbb{M}} \circ \hat{\mathbf{v}})(\mathbf{x}) = \mathbf{x} \quad \forall\, \mathbf{x} \in \mathbb{M}\,.$$

  This means that the applied vector $\hat{\mathbf{v}}(\mathbf{x}) \in \mathbb{T}_{\mathbb{M}}$ has $\mathbf{x} \in \mathbb{M}$ as base point or, equivalently, that $\hat{\mathbf{v}}(\mathbf{x}) \in \mathbb{T}_{\mathbb{M}}(\mathbf{x})$.
- A *submanifold* $\mathbb{P} \subset \mathbb{M}$ is a subset of the manifold $\mathbb{M}$ such that, for each $\mathbf{x} \in \mathbb{P}$, there is a chart $\{U, \varphi\}$ in $\mathbb{M}$, with $\mathbf{x} \in U$, satisfying the *submanifold property*:

  $$\varphi : U \mapsto E = E_1 \times E_2\,, \quad \varphi(U \cap \mathbb{P}) = \varphi(U) \cap (E_1 \times \{0\})\,.$$

  Every open subset of the manifold $\mathbb{M}$ is a submanifold.

Let $\mathscr{A}, \mathscr{B}, \mathscr{C}$ be Banach spaces; we denote by $BL(\mathscr{A}, \mathscr{B}; \mathscr{C})$ the space of bounded maps taking values in $\mathscr{C}$ and separately linear in the arguments ranging in $\mathscr{A}$ and $\mathscr{B}$. In the sequel square brackets denote linear dependence on the enclosed arguments.

- A *Riemannian metric* on the manifold $\mathbb{S}$ is a field of twice covariant, symmetric and positive definite tensors $\mathbf{g}_{\mathbb{S}} : \mathbb{S} \mapsto BL(\mathbb{T}_{\mathbb{S}}, \mathbb{T}_{\mathbb{S}}; \mathscr{R})$.

Any tensor field, say, $\mathbf{T}_{\mathbb{S}} : \mathbb{S} \mapsto BL(\mathbb{T}_{\mathbb{S}}, \mathbb{T}_{\mathbb{S}}; \mathscr{R})$, *lives at points* in the sense that at each $\mathbf{x} \in \mathbb{S}$ there exists a tensor $\mathbf{T}_{\mathbf{x}} \in BL(\mathbb{T}_{\mathbb{S}}(\mathbf{x}), \mathbb{T}_{\mathbb{S}}(\mathbf{x}); \mathscr{R})$ such that

$$\mathbf{T}_{\mathbb{S}}(\mathbf{x})[\mathbf{X}, \mathbf{Y}] = \mathbf{T}_{\mathbf{x}}[\mathbf{X}(\mathbf{x}), \mathbf{Y}(\mathbf{x})] \quad \forall\, \mathbf{X}, \mathbf{Y} \in \mathbb{T}_{\mathbb{S}}\,.$$

A Riemannian metric is naturally induced in each submanifold $\mathbb{M} \subset \mathbb{S}$ of a Riemannian manifold $\{\mathbb{S}, \mathbf{g}_{\mathbb{S}}\}$ by the canonical injection of the tangent space $\mathbb{T}_{\mathbb{M}}(\mathbf{x})$ at any $\mathbf{x} \in \mathbb{M}$ into the tangent space $\mathbb{T}_{\mathbb{S}}(\mathbf{x})$ at the same point $\mathbf{x} \in \mathbb{S}$.

3 Polar continua

The description of a *polar continuum* in mechanics is based on the following concepts.

- The *ambient space* $\mathbb{S}$ is a finite-dimensional differentiable manifold without boundary in which the body undergoes evolution processes. The ambient space of a polar continuum is a fiber bundle with projection $\pi_{\mathbb{S}} : \mathbb{S} \mapsto \mathbb{E}$ and typical fiber $\mathbb{F}$. Then locally the manifold $\mathbb{S}$ can be diffeomorphically related to the cartesian product $\mathbb{E} \times \mathbb{F}$ of the *base manifold* $\mathbb{E}$ and the *fiber manifold* $\mathbb{F}$. Both are finite-dimensional differentiable manifolds without boundary. The fiber manifold $\mathbb{F}$ provides the geometric description of the local kinematics of the polar continuum. The base manifold $\mathbb{E}$ is called the *physical space* and its points are called *positions*.

- The *material body* $\mathcal{B}$ is a set of *particles* which, at each time $t \in I$, are located at points of a differentiable submanifold of the physical space $\mathbb{E}$.

- The *base configuration* map $\boldsymbol{\chi}_t : \mathcal{B} \mapsto \mathbb{E}$ is a bijection of the material body $\mathcal{B}$ onto the *base placement* $\mathbb{B}_t = \boldsymbol{\chi}_t(\mathcal{B}) \subseteq \mathbb{E}$ which is a submanifold of the physical space $\mathbb{E}$.

- The *polar structure* $\mathbf{s}_t : \mathbb{B}_t \mapsto \mathbb{S}$ is a map from the base placement at time t onto the placement $\mathbb{P}_t = \mathbf{s}_t(\mathbb{B}_t)$. The map $\mathbf{s}_t : \mathbb{B}_t \mapsto \mathbb{S}$ has the property

  $$(\boldsymbol{\pi}_{\mathbb{S}} \circ \mathbf{s}_t)(\mathbf{p}) = \mathbf{p} \qquad \forall\, \mathbf{p} \in \mathbb{B}_t \subset \mathbb{E},$$

  and is then a *section* of the fiber bundle $\mathbb{S}$ defined on the submanifold $\mathbb{B}_t \subset \mathbb{E}$.

- A *spatial configuration* of the polar body at time $t \in I$ is an injective map $\mathbf{u}_t : \mathcal{B} \mapsto \mathbb{S}$ which assigns a *placement* $\mathbb{P}_t := \mathbf{u}_t(\mathcal{B}) \subset \mathbb{S}$ to the material body $\mathcal{B}$ and is given by the composition of the base configuration map with the polar structure

  $$\mathbf{u}_t = \mathbf{s}_t \circ \boldsymbol{\chi}_t .$$

In nonpolar continua the section $\mathbf{s}_t : \mathbb{B}_t \mapsto \mathbb{S}$ reduces to the identity on $\mathbb{B}_t$.

Remark 1. An important property of the polar models of interest in structural mechanics is that the base manifold $\mathbb{E}$ and the fiber manifold $\mathbb{F}$ are both embedded in finite-dimensional affine spaces, respectively denoted by $\{E, \mathbf{g}_E\}$ and $\{F, \mathbf{g}_F\}$, which are endowed with the euclidean metrics $\mathbf{g}_E \in BL(\mathbb{T}_E, \mathbb{T}_E; \mathcal{R})$ and $\mathbf{g}_F \in BL(\mathbb{T}_F, \mathbb{T}_F; \mathcal{R})$.

The ambient space $\mathbb{S}$ is a then a Riemannian manifold with the metric $\mathbf{g}_{\mathbb{S}} \in BL(\mathbb{T}_{\mathbb{S}}, \mathbb{T}_{\mathbb{S}}; \mathcal{R})$ induced by the euclidean metrics in E and F via the inclusions $\mathbb{T}_{\mathbb{E}} \subseteq \mathbb{T}_E$ and $\mathbb{T}_{\mathbb{F}} \subseteq \mathbb{T}_F$.

Remark 2. In polar models of beams and shells and in Cosserat continua, the fiber bundle $\mathbb{S}$ is a trivial bundle, that is, a cartesian product $\mathbb{S} = \mathbb{E} \times \mathbb{F}$. The physical space $\mathbb{E}$ is the euclidean space $E(3)$.

The fiber manifold $\mathbb{F}$ is SO(3) (the special orthogonal group of rotations) for beams and Cosserat continua, and is S^2 (the unit sphere) for shells without drilling rotations and SO(3) for shells with drilling rotations.

Other examples of polar continua are provided by the mathematical models of *liquid crystals* (see, e.g., [3, p. 139]) which are modeled by assuming that $\mathbb{E} = E(3)$ and

- $\mathbb{F} = S^2$ for *cholesteric* liquid crystals (inextensible directed rod-like molecules), and
- $\mathbb{F} = P^2$ for *nematic* liquid crystals (inextensible undirected rod-like molecules), where P^2 is the real projective two-space obtained by identifying the antipodal points on S^2.

Let $\mathbf{u}_s : \mathcal{B} \mapsto \mathbb{S}$ and $\mathbf{u}_t : \mathcal{B} \mapsto \mathbb{S}$ be the reference and the current configuration of the body in the ambient space $\mathbb{S}$ and let $\boldsymbol{\chi}_s : \mathcal{B} \mapsto \mathbb{E}$ be the base map of the reference configuration.

- The *change of base configuration* from $\boldsymbol{\chi}_s$ to $\boldsymbol{\chi}_t$ is the diffeomorphism $\boldsymbol{\chi}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s ; \mathbb{B}_t)$ defined by

  $$\boldsymbol{\chi}_{t,s} \circ \boldsymbol{\chi}_s = \boldsymbol{\chi}_t \,,$$

  where the index k denotes a suitable integer.

- The *change of configuration* from $\mathbf{u}_s$ to $\mathbf{u}_t$ is the map $\mathbf{u}_{t,s} : \mathbf{u}_s(\mathcal{B}) \mapsto \mathbf{u}_t(\mathcal{B}) \subset \mathbb{S}$ defined by

  $$\mathbf{u}_{t,s} \circ \mathbf{u}_s = \mathbf{u}_t \,.$$

The composition rules are given by

$$\mathbf{u}_{\tau,t} \circ \mathbf{u}_{t,s} = \mathbf{u}_{\tau,s} \,, \qquad \boldsymbol{\chi}_{\tau,t} \circ \boldsymbol{\chi}_{t,s} = \boldsymbol{\chi}_{\tau,s} \,.$$

Since $\boldsymbol{\chi}_{s,s} \in \mathrm{C}^k(\mathbb{B}_s ; \mathbb{B}_s)$ and $\mathbf{u}_{s,s} : \mathbb{P}_s \mapsto \mathbb{P}_s$ are identity maps, the maps $\boldsymbol{\chi}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s ; \mathbb{B}_t)$ and $\mathbf{u}_{t,s} : \mathbf{u}_s(\mathcal{B}) \mapsto \mathbf{u}_t(\mathcal{B})$ are invertible and the inverses are given by

$$(\boldsymbol{\chi}_{t,s})^{-1} = \boldsymbol{\chi}_{s,t} \,, \qquad (\mathbf{u}_{t,s})^{-1} = \mathbf{u}_{s,t} \,.$$

Remark 3. The requirement of regularity of the configuration changes must be expressed in terms of maps between manifolds. Now, while the base placements $\mathbb{B}_s = \boldsymbol{\chi}_s(\mathcal{B})$ and $\mathbb{B}_t = \boldsymbol{\chi}_t(\mathcal{B})$ are manifolds, the placements $\mathbb{P}_s = \mathbf{u}_s(\mathcal{B})$ and $\mathbb{P}_t = \mathbf{u}_t(\mathcal{B})$ are not manifolds but instead images of sections of the fiber bundle $\mathbb{S}$ defined on submanifolds of the physical space. Accordingly we require that

$$\mathbf{u}_{t,s} \circ \mathbf{s}_s \in \mathrm{C}^k(\mathbb{B}_s ; \mathbb{S}) \,,$$

but simply write $\mathbf{u}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s ; \mathbb{S})$.

- The base configuration changes can be depicted as a two-parameter family of diffeomorphisms $\boldsymbol{\chi}_{t,s} : \mathcal{B} \mapsto \mathrm{C}^k(\mathbb{S}, \mathbb{S})$ which is called a *flow* of the material manifold $\mathcal{B}$ into the physical space $\mathbb{S}$. The flow $\boldsymbol{\chi}_{t,s}$ maps the position $\boldsymbol{\chi}_s(\mathbf{p})$ at time $s \in I$ of a particle $\mathbf{p} \in \mathcal{B}$ into its position $\boldsymbol{\chi}_t(\mathbf{p})$ at time $t \in I$ and, as seen above, satisfies the Chapman-Kolmogorov composition rule [7]

$$\boldsymbol{\chi}_{\tau,s} = \boldsymbol{\chi}_{\tau,t} \circ \boldsymbol{\chi}_{t,s}\,, \qquad \boldsymbol{\chi}_{t,t}(\mathbf{x}) = \mathbf{x}\,, \quad \forall\, \mathbf{x} \in \mathbb{B}_t\,.$$

- The *space of configuration changes* from $\mathbf{u}_s$ is the differentiable manifold $\mathbb{M} := \mathrm{C}^k(\mathbb{B}_s\,; \mathbb{S})$ modeled on the Banach space $\mathrm{C}^k(\mathbb{B}_s\,; \mathcal{R}^d)\,,\ d = \dim \mathbb{S}$.

When a reference configuration $\mathbf{u}_s$ is fixed, we often refer to a configuration change $\mathbf{u}_{t,s}$ simply as a configuration by identifying it with $\mathbf{u}_t = \mathbf{u}_{t,s} \circ \mathbf{u}_s$.

- The *push forward* of a vector field $\mathbf{v}_s \in \mathrm{C}^k(\mathbb{B}_s, \mathbb{T}_{\mathbb{S}})$ along the flow $\boldsymbol{\chi}_{t,s}$ is the vector field $\boldsymbol{\chi}_{t,s*}\mathbf{v}_s \in \mathrm{C}^k(\mathbb{B}_t, \mathbb{T}_{\mathbb{S}})$ locally defined by

$$((\boldsymbol{\chi}_{t,s*}\mathbf{v}_s)\, f)(\boldsymbol{\chi}_{t,s}\,\mathbf{p}) = (\mathbf{v}_s(f \circ \boldsymbol{\chi}_{t,s}))(\mathbf{p}) \quad \forall\, f \in \mathrm{C}^1(\boldsymbol{\chi}_{t,s}\,\mathbf{p}, U)\,, \mathbf{p} \in \mathbb{B}_s\,.$$

  The set $\mathrm{C}^1(\mathbf{x}, U)$ is the *germ* of continuously differentiable functions in the neighborhood U of $\mathbf{x} \in \mathbb{B}_t$. The push forward maps tangent vectors applied at points of a manifold into the corresponding deformed tangent vectors applied at the transformed points.

- The *pull back* $\boldsymbol{\chi}^*_{t,s} = \boldsymbol{\chi}^{-1}_{t,s\,*}$ is the push forward induced by the inverse diffeomorphism.

- The *push forward* of a tensor field $\mathbf{a}_s \in \mathrm{C}^k(\mathbb{S}, BL\,(\mathbb{T}_{\mathbb{S}}, \mathbb{T}_{\mathbb{S}}\,; \mathcal{R}))$ is the tensor field $\boldsymbol{\chi}_{t,s*}\mathbf{a}_s \in \mathrm{C}^k(\mathbb{S}, BL\,(\mathbb{T}_{\mathbb{S}}, \mathbb{T}_{\mathbb{S}}\,; \mathcal{R}))$ locally defined by the relation

$$(\boldsymbol{\chi}_{t,s*}\mathbf{a}_s)\,(\boldsymbol{\chi}_{t,s*}\mathbf{v}_s, \boldsymbol{\chi}_{t,s*}\mathbf{w}_s) := \boldsymbol{\chi}_{t,s*}\,(\mathbf{a}_s\,(\mathbf{v}_s, \mathbf{w}_s))$$

  for any $\mathbf{v}_s, \mathbf{w}_s \in \mathrm{C}^k(\mathbb{S}, \mathbb{T}_{\mathbb{S}})$.

- The Lie derivative of a tensor field $\mathbf{a}_t \in \mathrm{C}^k(\mathbb{S}, BL\,(\mathbb{T}_{\mathbb{S}}, \mathbb{T}_{\mathbb{S}}\,; \mathcal{R}))$ along a flow $\boldsymbol{\chi}_{\tau,t} : \mathcal{B} \mapsto \mathrm{C}^k(\mathbb{S}, \mathbb{S})$, evaluated at the configuration at time $t \in I$, is the time derivative of the tensor field pulled back to the configuration at time $t \in I$,

$$\mathscr{L}_{\mathbf{X}_t}\mathbf{a}_t := \frac{d}{d\tau}\bigg|_{\tau=t} \boldsymbol{\chi}_{t,\tau*}\mathbf{a}_\tau\,,$$

  where $\mathbf{X}_t$ is the velocity field $\mathbf{v}_t$ of the flow $\boldsymbol{\chi}_{\tau,t}$ at time $t \in I$,

$$\frac{d}{d\tau}\bigg|_{\tau=t} \boldsymbol{\chi}_{\tau,t} = \mathbf{X}_t\,.$$

3.1 Finite deformation measures

A *finite deformation measure* is a non-linear operator

$$\mathbf{A} \in \mathrm{C}^2(\mathbb{M}\,;\mathrm{C}^0(\mathbb{B}_s\,;D))\,,$$

that maps the configuration changes $\mathbf{u}_{t,s} \in \mathbb{M} = \mathrm{C}^k(\mathbb{B}_s\,;\mathbb{S})$ into the corresponding finite deformation fields $\mathbf{A}_{\mathbf{u}} = \mathbf{A}(\mathbf{u}_{t,s}) \in \mathrm{C}^0(\mathbb{B}_s\,;D)$. The space D is the finite-dimensional linear space of local strain values. Deformation measures are differential operators and hence the value $\mathbf{A}_{\mathbf{u}}(\mathbf{p})$ at a point $\mathbf{p} \in \mathbb{B}_s$ is independent of the values of the map $\mathbf{u}_{t,s}$ outside any given neighborhood of $\mathbf{p} \in \mathbb{B}_s$. This locality property is in fact characteristic of (linear or non-linear) differential operators (see [3, p. 189]).

The definition of the subset $\mathcal{R} \subset \mathbb{M}$ of *rigid configuration changes* is a cornerstone in the formulation of a continuous structural model. It is natural to assume that the identity map is a rigid configuration change.

The basic property satisfied by a deformation measure is that it vanishes if and only if the configuration change is rigid:

$$\mathbf{u}_{t,s} \in \mathcal{R} \quad \Longleftrightarrow \quad \mathbf{A}(\mathbf{u}_{t,s}) = 0 \in \mathrm{C}^0(\mathbb{B}_s\,;D)\,.$$

- Two deformation measures $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{C}^2(\mathbb{M}\,;\mathrm{C}^0(\mathbb{B}_s\,;D))$ are said to be *equivalent* if

  $$\mathbf{A}_1(\mathbf{u}_{t,s}) = 0 \quad \Longleftrightarrow \quad \mathbf{A}_2(\mathbf{u}_{t,s}) = 0\,.$$

  Let $D = D_1 \oplus D_2$ be a decomposition of the linear space D into the direct sum of two complementary subspaces and let the associated projectors be denoted by $\boldsymbol{\Pi}_1 \in BL\,(D\,;D_1)$, $\boldsymbol{\Pi}_2 \in BL\,(D\,;D_2)$.

- A deformation measure $\mathbf{A} \in \mathrm{C}^2(\mathbb{M}\,;\mathrm{C}^0(\mathbb{B}_s\,;D))$ is said to be *redundant* if there exists a nontrivial decomposition $D = D_1 \oplus D_2$ such that

  $$(\boldsymbol{\Pi}_1 \circ \mathbf{A})(\mathbf{u}_{t,s}) = 0 \quad \Longrightarrow \quad \mathbf{A}(\mathbf{u}_{t,s}) = 0\,.$$

  A *nonredundant* deformation measure is be said to be *minimal* in its equivalence class.

- In a referential description of kinematics it is also essential to require that the deformation measure satisfies the *consistency property*

  $$\mathbf{A}(\mathbf{u}_{\tau,s}) = \mathbf{A}(\mathbf{u}_{t,s}) + \mathbf{S}(\mathbf{A}(\mathbf{u}_{\tau,t}), \mathbf{u}_{t,s})\,,$$

  where $\mathbf{S}$ is a non-linear differentiable operator such that

  $$\mathbf{A}(\mathbf{u}_{\tau,t}) = 0 \quad \Longrightarrow \quad \mathbf{S}(\mathbf{A}(\mathbf{u}_{\tau,t}), \mathbf{u}_{t,s}) = 0 \quad \forall\, \mathbf{u}_{t,s} \in \mathbb{M}\,.$$

  The latter requirement ensures that the deformation measure is indifferent to superimposed rigid changes of configuration and hence also ensures the invariance of the deformation measure under a change of observer.

The relevance of the consistency property is clearly illustrated in Sect. 5.

Finite deformation fields are evaluated pointwise according to the following scheme. At any point $\mathbf{x} = (\mathbf{u} \circ \mathbf{s})(\mathbf{p})$, $\mathbf{p} \in \mathbb{B}$, we consider a local operator $\mathbf{A}_{\mathbf{x}} \in \mathrm{C}^2(\mathbb{M}; D)$ defined by

$$\mathbf{A}_{\mathbf{x}}(\mathbf{u}) := \mathbf{N}(\mathbf{D}\mathbf{u})_{\mathbf{x}},$$

where $\mathbf{D}$ is a linear differential operator of order k acting on the space variable $\mathbf{p} \in \mathbb{B}$ and $\mathbf{N}$ is a smooth local non-linear operator mapping the local values of the field $\mathbf{D}\mathbf{u}$ into the linear space D. The operator $\mathbf{A}$ is then defined pointwise by setting

$$\mathbf{A}_{\mathbf{u}}(\mathbf{x}) := \mathbf{A}_{\mathbf{x}}(\mathbf{u}) \quad \forall\, \mathbf{u} \in \mathbb{M}, \quad \forall\, \mathbf{x} \in (\mathbf{u} \circ \mathbf{s})(\mathbb{B}).$$

3.2 Virtual displacements

A *referential virtual displacement* at the configuration $\mathbf{u}_{t,s} \in \mathbb{M} = \mathrm{C}^k(\mathbb{B}_s; \mathbb{S})$ is a vector field tangent to $\mathbb{M}$ at $\mathbf{u}_{t,s} \in \mathbb{M}$, that is, a map $\mathbf{X} \in \mathrm{C}^k(\mathbb{M}; \mathbb{T}_{\mathbb{M}})$ such that

$$\mathbf{X}(\mathbf{u}_{t,s}) \in \mathbb{T}_{\mathbb{M}}(\mathbf{u}_{t,s}).$$

Virtual displacements are then vector fields which are defined on the space $\mathbb{M}$ of admissible configurations and take values in its tangent bundle $\mathbb{T}_{\mathbb{M}}$.

Since the reference placement $\mathbb{B}_s$ is fixed, virtual displacements can be represented as vector fields $\delta\mathbf{u}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s; \mathbb{T}_{\mathbb{S}})$ defined on the base reference placement $\mathbb{B}_s$ and taking values in the tangent bundle $\mathbb{T}_{\mathbb{S}}$ of the ambient space $\mathbb{S}$ ([3, p. 170]).

Accordingly the linear space of virtual displacements can be defined as

$$\mathbb{T}_{\mathbb{M}}(\mathbf{u}_{t,s}) = \{\, \delta\mathbf{u}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s; \mathbb{T}_{\mathbb{S}}) \mid \boldsymbol{\tau}_{\mathbb{S}} \circ \delta\mathbf{u}_{t,s} = \mathbf{u}_{t,s} \,\},$$

so that

$$\delta\mathbf{u}_{t,s}(\mathbf{p}) \in \mathbb{T}_{\mathbb{S}}(\mathbf{u}_{t,s}(\mathbf{p})) \quad \forall\, \mathbf{p} \in \mathbb{B}_s.$$

The fields $\delta\mathbf{u}_{t,s}$ are also called *referential virtual displacements*.

A *virtual displacement* at the configuration $\mathbf{u}_{t,s} \in \mathbb{M}$ is a field of vectors on $\mathbb{B}_t$ tangent to the ambient space $\mathbb{S}$, that is, a map $\mathbf{v}_t \in \mathrm{C}^k(\mathbb{B}_t; \mathbb{T}_{\mathbb{S}})$ such that

$$\mathbf{v}_t(\mathbf{x}) \in \mathbb{T}_{\mathbb{S}}(\mathbf{x}) \quad \forall\, \mathbf{x} \in \mathbb{P}_t = \mathbf{u}_{t,s}(\mathbb{P}_s).$$

Hence a virtual displacement $\mathbf{v}_t \in \mathrm{C}^k(\mathbb{B}_t; \mathbb{T}_{\mathbb{S}})$ and the corresponding referential virtual displacement $\delta\mathbf{u}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s; \mathbb{T}_{\mathbb{S}})$ are related by the composition rule

$$\delta\mathbf{u}_{t,s} = \mathbf{v}_t \circ \boldsymbol{\chi}_{t,s}.$$

4 Classical and polar models of continua

We present here basic models of classical and polar continua to illustrate the general topics analyzed in the previous sections.

4.1 Cauchy continua

A placement $\mathbb{B}$ of a Cauchy continuum is a regular region of the three-dimensional euclidean space E^3.

The tangent space at each point $\mathbf{p} \in \mathbb{B}$ is the three-dimensional linear space V^3 of translations in E^3 endowed with the usual inner product.

The tangent bundle $\mathbb{T}_{\mathbb{B}}$ is the disjoint union of copies of the linear translation space V^3 attached at each point of the affine space E^3.

The local structure of a Cauchy continuum reduces to that of its tangent bundle. Therefore Cauchy's model lacks a polar structure.

The ambient space $\mathbb{S}$ is the affine space E^3.

A placement at time $t \in I$ of the material body $\mathcal{B}$ is a diffeomorphism

$$\boldsymbol{\chi}_t \in \mathrm{C}^k(\mathcal{B}\,; E^3)\,.$$

A flow is given by a two-parameter family of diffeomorphisms $\boldsymbol{\chi}_{t,s} : \mathcal{B} \mapsto \mathrm{C}(E^3\,; E^3)$ defined by

$$\boldsymbol{\chi}_{t,s} \circ \boldsymbol{\chi}_s = \boldsymbol{\chi}_t\,.$$

Rigid configuration changes are isometric transformations in E^3 described by a translation vector and a rotation. The set of configuration changes from a given placement is then a six-dimensional manifold.

Green's finite deformation measure $\mathfrak{D}(\boldsymbol{\chi}_{t,s})$ associated with the flow $\boldsymbol{\chi}_{t,s}$ is the twice covariant tensor field defined (see, e.g., [40]) by

$$\mathfrak{D}(\boldsymbol{\chi}_{t,s})\,(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}(\boldsymbol{\chi}_{s,t\,*}\mathbf{g}_{\mathbb{S}} - \mathbf{g}_{\mathbb{S}})\,(\mathbf{X}, \mathbf{Y})\,,$$

where $\mathbf{X}, \mathbf{Y} \in \mathbb{T}_{\mathbb{S}}$ are tangent vector fields on $\boldsymbol{\chi}_s(\mathcal{B})$ and $\mathbf{g}_{\mathbb{S}}$ is the metric tensor of the euclidean space $\mathbb{S}$.

Green's strain measure satisfies the consistency property since

$$\boldsymbol{\chi}_{s,\tau\,*}\mathbf{g}_{\mathbb{S}} - \mathbf{g}_{\mathbb{S}} = (\boldsymbol{\chi}_{s,t\,*} \circ \boldsymbol{\chi}_{t,\tau\,*})\,\mathbf{g}_{\mathbb{S}} - \mathbf{g}_{\mathbb{S}} = \boldsymbol{\chi}_{s,t\,*}(\boldsymbol{\chi}_{t,\tau\,*}\mathbf{g}_{\mathbb{S}} - \mathbf{g}_{\mathbb{S}}) + (\boldsymbol{\chi}_{s,t\,*}\mathbf{g}_{\mathbb{S}} - \mathbf{g}_{\mathbb{S}})\,.$$

The tangent deformation at time $t \in I$ associated with Green's strain measure at the configuration $\boldsymbol{\chi}_t$ is given [40] by

$$\frac{1}{2}\,(\mathscr{L}_{\mathbf{v}}\,\mathbf{g}_{\mathbb{S}})_t(\mathbf{X}, \mathbf{Y}) = \frac{d}{ds}\bigg|_{s=t} (\boldsymbol{\chi}_{t,s\,*}\,\mathbf{g}_{\mathbb{S}})(\mathbf{X}, \mathbf{Y}) = \mathbf{g}_{\mathbb{S}}\,((\mathrm{sym}\,\partial\mathbf{v}_t)\mathbf{X}, \mathbf{Y})\,,$$

where $\partial\mathbf{v}_t$ is the spatial derivative of the velocity $\mathbf{v}_t$ of the flow $\boldsymbol{\chi}_{t,s}$ and $\mathbf{X}, \mathbf{Y} \in \mathbb{T}_{\mathbb{S}}$ are tangent vector fields on $\boldsymbol{\chi}_t(\mathcal{B})$.

4.2 Cables and membranes

A placement $\mathbb{B}$ of a cable is a one-dimensional manifold (a curve) embedded in the euclidean space $\mathbb{S} = E^3$. The tangent bundle $\mathbb{T}_{\mathcal{B}}$ is the disjoint union of the one-dimensional tangent spaces to $\mathbb{B}$.

A placement $\mathbb{B}$ of a membrane is a two-dimensional manifold embedded in the euclidean space $\mathbb{S} = E^3$ The tangent bundle $\mathbb{T}_{\mathcal{B}}$ is the disjoint union of the two-dimensional tangent spaces to $\mathbb{B}$.

The models of cables and membranes lack polar structures.

Rigid configuration changes are isometric transformations of the one- or two-dimensional manifolds and hence the set of configuration changes from a given placement is not a finite-dimensional manifold.

We next consider the metric tensor field on $\mathbb{B}$,

$$\mathbf{g}_{\mathbb{B}}(\mathbf{X}, \mathbf{Y}) := \mathbf{g}_{\mathbb{S}}(\boldsymbol{\Pi}^T\mathbf{X}, \boldsymbol{\Pi}^T\mathbf{Y}),$$

Green's deformation measure for the cable (or for the membrane) is given [40] by

$$\mathfrak{D}(\boldsymbol{\chi}_{t,s})(\mathbf{X}, \mathbf{Y}) = \tfrac{1}{2}\left[(\boldsymbol{\chi}_{s,t*}\mathbf{g}_{\mathbb{S}} - \mathbf{g}_{\mathbb{S}})(\boldsymbol{\Pi}^T\mathbf{X}, \boldsymbol{\Pi}^T\mathbf{Y})\right],$$

where $\mathbf{X}, \mathbf{Y} \in \mathbb{T}_{\mathbb{B}}$ are tangent vectors fields on $\boldsymbol{\chi}_s(\mathcal{B})$ and $\boldsymbol{\Pi} \in BL(\mathbb{T}_{\mathbb{S}}; \mathbb{T}_{\mathbb{B}})$ is the orthogonal projector from $\mathbb{T}_{\mathbb{S}}$ onto $\mathbb{T}_{\mathbb{B}}$. Its transpose $\boldsymbol{\Pi}^T \in BL(\mathbb{T}_{\mathbb{B}}; \mathbb{T}_{\mathbb{S}})$ is the canonical injection of $\mathbb{T}_{\mathbb{B}}$ into $\mathbb{T}_{\mathbb{S}}$.

The tangent deformation at time $t \in I$ associated with Green's strain measure is given [40] by

$$\begin{aligned}\tfrac{1}{2}(\mathscr{L}_{\mathbf{v}}\mathbf{g}_{\mathbb{B}})_t(\mathbf{X}, \mathbf{Y}) &:= \frac{d}{ds}\bigg|_{s=t} (\boldsymbol{\chi}_{t,s_*}\mathbf{g}_{\mathbb{S}})(\boldsymbol{\Pi}^T\mathbf{X}, \boldsymbol{\Pi}^T\mathbf{Y}) \\ &= \mathbf{g}_{\mathbb{B}}((\mathrm{sym}(\boldsymbol{\Pi}\,\partial\mathbf{v}_t\,\boldsymbol{\Pi}^T))\mathbf{X}, \mathbf{Y}),\end{aligned}$$

where $\mathbf{v}_t \in \mathrm{C}^k(\mathbb{B}, \mathbb{T}_{\mathbb{S}})$ is a virtual displacement and $\mathbf{X}, \mathbf{Y} \in \mathbb{T}_{\mathbb{S}}$ are tangent vector fields on $\mathbb{B} = \boldsymbol{\chi}_t(\mathcal{B})$.

4.3 Cosserat continua

In the Cosserat continuum the ambient space is the trivial fiber bundle defined by the projection $\boldsymbol{\pi}_{\mathbb{S}} : \mathbb{S} = E^3 \times \mathrm{SO}(3) \mapsto E^3$ onto the three-dimensional euclidean space E^3. The fiber manifold $\mathrm{SO}(3)$ is the compact three-dimensional special orthogonal group of rotations. The tangent bundle $\mathbb{T}_{\mathbb{S}}$ is the disjoint union of the linear spaces $V^3 \times (\mathrm{so}(3)\,\mathbf{Q})$ with $\mathbf{Q} \in \mathrm{SO}(3)$; here $\mathrm{so}(3) \subset BL(V^3; V^3)$ is the linear subspace of skew-symmetric mixed tensors and the linear space $\mathrm{so}(3)\,\mathbf{Q}$ is defined [40] by

$$\mathrm{so}(3)\,\mathbf{Q} = \{\mathbf{T} \in BL(V^3; V^3) \,:\, \mathbf{T} = \mathbf{W}\,\mathbf{Q}, \quad \mathbf{W} \in \mathrm{so}(3), \mathbf{Q} \in \mathrm{SO}(3)\}.$$

A base configuration at time $t \in I$ of a Cosserat continuum is an injective map $\boldsymbol{\chi}_t : \mathcal{B} \mapsto E^3$ whose image is a compact domain in E^3. A configuration at time $t \in I$ is an injective map $\mathbf{u}_t : \mathcal{B} \mapsto E^3 \times \mathrm{SO}(3)$ defined at each particle $\mathbf{p} \in \mathcal{B}$ by

$$\mathbf{u}_t(\mathbf{p}) = \{\boldsymbol{\chi}_t(\mathbf{p}), \mathbf{Q}_t(\mathbf{p})\} \in E^3 \times \mathrm{SO}(3),$$

where $\mathbf{Q}_t \in \mathcal{B} \mapsto \mathrm{SO}(3)$ is a rotation field with respect to a given reference triad. A flow is represented by a pair $\{\boldsymbol{\chi}_{t,s}, \mathbf{Q}_{t,s}\}$ with

$$\boldsymbol{\chi}_{t,s} \circ \boldsymbol{\chi}_s = \boldsymbol{\chi}_t, \qquad \mathbf{Q}_{t,s} \circ \mathbf{Q}_s = \mathbf{Q}_t.$$

A finite deformation measure for the Cosserat continuum is given [40] by

$$\mathfrak{D}(\boldsymbol{\chi}_{t,s}, \mathbf{Q}_{t,s}) = \{\mathbf{C}(\mathbf{Q}_{t,s}), \boldsymbol{\Delta}(\boldsymbol{\chi}_{t,s}, \mathbf{Q}_{t,s})\},$$

where

$$\begin{cases} \mathbf{C}(\mathbf{Q}_{t,s}) & := \boldsymbol{\Omega}_{t,s}, & \text{curvature change,} \\ \boldsymbol{\Delta}(\boldsymbol{\chi}_{t,s}, \mathbf{Q}_{t,s}) & := \mathbf{Q}_{t,s}^T\, \partial \boldsymbol{\chi}_{t,s} - \mathbf{I}_s, & \text{strain gap,} \end{cases}$$

with $\mathbf{I}_s \in \mathrm{L}(V^3; V^3)$ the identity at time $s \in I$ and

$$\boldsymbol{\Omega}_{t,s}[\mathbf{h}] := \mathrm{axial}\,(\mathbf{Q}_{t,s}^T\, \partial \mathbf{Q}_{t,s}[\mathbf{h}]) \quad \forall\, \mathbf{h} \in V^3.$$

Then $\boldsymbol{\Omega}_{t,s} \in \mathrm{L}(V^3; V^3)$ and $D = \mathrm{L}(V^3; V^3) \times \mathrm{L}(V^3; V^3)$.

4.4 Timoshenko beams

A placement of a Timoshenko beam is described by a regular curve in E^3 named the axis of the beam, and by a field of rotations $\mathbf{Q} \in \mathrm{SO}(3)$, attached at each point of the beam axis, which simulate the rigid body kinematics of the cross sections of the beam. The ambient space $\mathbb{S}$ is the trivial fiber bundle $\boldsymbol{\pi}_{\mathbb{S}} : E^3 \times \mathrm{SO}(3) \mapsto E^3$.

A base configuration at time $t \in I$ is an injective map $\mathbf{r}_t : \mathcal{B} \mapsto E^3$ whose image is a regular curve in E^3. A configuration at time $t \in I$ is an injective map $\mathbf{u}_t : \mathcal{B} \mapsto E^3 \times \mathrm{SO}(3)$ defined at each particle $\mathbf{p} \in \mathcal{B}$ by

$$\mathbf{u}_t(\mathbf{p}) = \{\mathbf{r}_t(\mathbf{p}), \mathbf{Q}_t(\mathbf{p})\} \in E^3 \times \mathrm{SO}(3),$$

where $\mathbf{Q}_t \in \mathcal{B} \mapsto \mathrm{SO}(3)$ is a rotation field with respect to a given reference triad. A flow is represented by a pair $\{\mathbf{r}_{t,s}, \mathbf{Q}_{t,s}\}$, where

$$\mathbf{r}_{t,s} \circ \mathbf{r}_s = \mathbf{r}_t, \qquad \mathbf{Q}_{t,s} \circ \mathbf{Q}_s = \mathbf{Q}_t.$$

A finite deformation measure is provided [5,40] by the pair

$$\mathfrak{D}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s}) = \{\mathbf{c}(\mathbf{Q}_{t,s}), \boldsymbol{\delta}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s})\},$$

where

$$\begin{cases} \mathbf{c}(\mathbf{Q}_{t,s}) := \text{axial}\,(\mathbf{Q}_{t,s}^T \mathbf{Q}_{t,s}') & \text{flexural-torsional curvature change,} \\ \boldsymbol{\delta}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s}) := \mathbf{Q}_{t,s}^T \mathbf{r}_{t,s}' - \mathbf{t}_s & \text{axial-shear sliding,} \end{cases}$$

with $\mathbf{t}_s \in V^3$ the unit tangent to the beam axis at time $s \in I$ and $\mathbf{c}(\mathbf{Q}_{t,s}) \in V^3$. Then $D = V^3 \times V^3$.

The prime $(\cdot)'$ denotes the derivative with respect to the curvilinear abscissa along the beam axis at the initial configuration of the time step, so that $\mathbf{Q}_{t,s}'$ is derived with respect to ξ_s and $\mathbf{Q}_{\tau,t}'$ is derived with respect to ξ_t.

4.5 Polar shells

Two basic models of polar shells have been proposed in the literature. In one the local polar structure is simulated by means of oriented rigid hairs attached at the points of the middle surface. No drilling rotations of the hairs (i.e., rotations around their axes) are considered. The other is the two-dimensional analogue of the three-dimensional Cosserat continuum and is referred to as the Cosserat shell model or shell with drilling rotations: a rigid trihedron is attached at the points of the middle surface and arbitrary rotations are allowed. Both models are illustrated briefly in the sequel.

Polar shells with drilling rotations. A placement of a Cosserat shell is described by a regular surface in E^3, the middle surface of the shell, and by a field of rotations $\mathbf{Q} \in \mathrm{SO}(3)$ defined at each point of the middle surface which simulate rigid body kinematics along the thickness of the shell. The ambient space $\mathbb{S}$ is the trivial fiber bundle $\boldsymbol{\pi}_{\mathbb{S}} : E^3 \times \mathrm{SO}(3) \mapsto E^3$. A base configuration at time $t \in I$ is an injective map $\boldsymbol{\chi}_t : \mathcal{B} \mapsto E^3$ whose image is a regular surface in E^3. A configuration at time $t \in I$ is an injective map $\mathbf{u}_t : \mathcal{B} \mapsto E^3 \times \mathrm{SO}(3)$ defined at each particle $\mathbf{p} \in \mathcal{B}$ by

$$\mathbf{u}_t(\mathbf{p}) = \{\boldsymbol{\chi}_t(\mathbf{p}), \mathbf{Q}_t(\mathbf{p})\} \in E^3 \times \mathrm{SO}(3),$$

where $\mathbf{Q}_t \in \mathcal{B} \mapsto \mathrm{SO}(3)$ is a rotation field with respect to a given reference triad. A finite deformation measure is provided [19,20] by the pair

$$\mathfrak{D}(\boldsymbol{\chi}_{t,s}, \mathbf{Q}_{t,s}) = \{\mathbf{C}(\mathbf{Q}_{t,s}), \boldsymbol{\Delta}(\boldsymbol{\chi}_{t,s}, \mathbf{Q}_{t,s})\},$$

where

$$\begin{cases} \mathbf{C}(\mathbf{Q}_{t,s}) := \boldsymbol{\Omega}_{t,s}, & \text{curvature change,} \\ \boldsymbol{\Delta}(\boldsymbol{\chi}_{t,s}, \mathbf{Q}_{t,s}) := \mathbf{Q}_{t,s}^T \partial \boldsymbol{\chi}_{t,s} - \mathbf{I}_s, & \text{strain gap,} \end{cases}$$

with $\mathbf{I}_s \in \mathrm{L}(V^3; V^3)$ the identity at time $s \in I$ and

$$\boldsymbol{\Omega}_{t,s}[\mathbf{h}] := \text{axial}\,(\mathbf{Q}_{t,s}^T \partial \mathbf{Q}_{t,s}[\mathbf{h}]) \quad \forall\, \mathbf{h} \in \mathbb{T}_{\mathbb{B}_t}(\boldsymbol{\chi}_t(\mathbf{p})).$$

Then $\boldsymbol{\Omega}_{t,s} \in \mathrm{L}(V^2; V^3)$ and $D = \mathrm{L}(V^2; V^3) \times \mathrm{L}(V^3; V^3)$.

Polar shells without drilling rotations. A placement of a polar shell without drilling rotations is described by a middle surface in E^3 and by a field of unit vectors attached at each of its points which simulate the kinematics of the shell in the transverse direction. The ambient space is the trivial fiber bundle $\boldsymbol{\pi}_{\mathbb{S}} : \mathbb{S} = E^3 \times S^2 \mapsto E^3$. The fiber manifold is the two-dimensional unit sphere S^2 in E^3. The finite deformation measure proposed and analyzed in [8,9] consists of the triplet

$$\mathbf{A}(\mathbf{u}_{t,s}) := \begin{vmatrix} \boldsymbol{\varepsilon}(\boldsymbol{\chi}_{t,s}) \\ \boldsymbol{\delta}(\mathbf{u}_{t,s}) \\ \boldsymbol{\chi}(\mathbf{u}_{t,s}) \end{vmatrix}$$

composed of:

$$\begin{aligned} \boldsymbol{\varepsilon}(\boldsymbol{\chi}_{t,s})(\mathbf{a},\mathbf{b}) &:= \mathbf{g}(\boldsymbol{\chi}_{t,s*}\mathbf{a},\, \boldsymbol{\chi}_{t,s*}\mathbf{b}) - \mathbf{g}(\mathbf{a},\, \mathbf{b})\,, && \text{membrane strain}\,, \\ \boldsymbol{\delta}(\mathbf{u}_{t,s})(\mathbf{a}) &:= \mathbf{g}(\mathbf{d}_t,\, \boldsymbol{\chi}_{t,s*}\mathbf{a}) - \mathbf{g}(\mathbf{d}_s,\, \mathbf{a})\,, && \text{shear sliding}\,, \\ \boldsymbol{\chi}(\mathbf{u}_{t,s})(\mathbf{a},\mathbf{b}) &:= \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{d}_t,\, \boldsymbol{\chi}_{t,s*}\mathbf{b}) - \mathbf{g}(\partial_{\mathbf{a}}\mathbf{d}_s,\, \mathbf{b})\,, && \text{curvature change}\,, \end{aligned}$$

where $\mathbf{a}\,, \mathbf{b} \in V^3 = \mathbb{T}_{E^3}$ and $\mathbf{g}$ is the metric tensor on E^3.

These measures vanish if and only if the shell undergoes a rigid body transformation, i.e., when the membrane deformation vanishes and the directors and tangent planes to the middle surface are rotated according to a constant rotation field. Indeed the vanishing of the membrane strain imposes that the middle surface transformation be isometric and the vanishing of the shear sliding implies that the directors be invariant when seen by observers co-rotating with the tangent planes.

From the vanishing of the flexural curvature, which is an extrinsic quantity, one infers that the second fundamental form of the surface does not change and hence that the surface must undergo a rigid body transformation with a rotation equal to the rotation of the directors.

This finite strain measure for shells is not redundant since the vanishing of any proper subset of the strain measures does not ensure that the transformation is rigid.

4.6 Consistency, redundancy and physical plausibility

It is interesting to underline the formal analogy existing between the deformation measures pertaining to Timoshenko beams, to polar shells with drilling rotations and to Cosserat continua. Such measures satisfy the consistency condition. Indeed for the Timoshenko beam,

$$\begin{aligned} \mathbf{Q}^T_{\tau,s}\mathbf{Q}'_{\tau,s} &= (\mathbf{Q}_{\tau,t}\mathbf{Q}_{t,s})^T(\mathbf{Q}_{\tau,t}\mathbf{Q}_{t,s})' = \mathbf{Q}^T_{t,s}\,\mathbf{Q}^T_{\tau,t}\,(\mathbf{Q}_{\tau,t}\mathbf{Q}_{t,s})' \\ &= \mathbf{Q}^T_{t,s}\,\mathbf{Q}^T_{\tau,t}\left(\mathbf{Q}'_{\tau,t}\,\frac{d\xi_t}{d\xi_s}\,\mathbf{Q}_{t,s} + \mathbf{Q}_{\tau,t}\mathbf{Q}'_{t,s}\right) \\ &= \mathbf{Q}^T_{t,s}\left(\mathbf{Q}^T_{\tau,t}\,\mathbf{Q}'_{\tau,t}\,\frac{d\xi_t}{d\xi_s}\right)\mathbf{Q}_{t,s} + \mathbf{Q}^T_{t,s}\mathbf{Q}'_{t,s}\,. \end{aligned}$$

Then the semisymmetric curvature tensor $\mathbf{C}(\mathbf{Q}_{t,s}) = \mathbf{Q}_{t,s}^T \mathbf{Q}'_{t,s}$ satisfies the relation

$$\mathbf{C}(\mathbf{Q}_{\tau,s}) = \mathbf{Q}_{t,s}^T \, \mathbf{C}(\mathbf{Q}_{\tau,t}) \mathbf{Q}_{t,s} \frac{d\xi_t}{d\xi_s} + \mathbf{C}(\mathbf{Q}_{t,s}) \, .$$

In the same way,

$$\begin{aligned} \mathbf{Q}_{\tau,s}^T \boldsymbol{\chi}'_{\tau,s} - \mathbf{r}'_s &= \mathbf{Q}_{t,s}^T \, \mathbf{Q}_{\tau,t}^T \, \mathbf{r}'_{\tau,s} - \mathbf{r}'_s \\ &= \mathbf{Q}_{t,s}^T \, (\mathbf{Q}_{\tau,t}^T \, \mathbf{r}'_{\tau,t} - \mathbf{r}'_t) \, \frac{d\xi_t}{d\xi_s} + (\mathbf{Q}_{t,s}^T \, \mathbf{r}'_{t,s} - \mathbf{r}'_s) \, . \end{aligned}$$

Then the axial-shear sliding satisfies the relation

$$\boldsymbol{\delta}(\mathbf{r}_{\tau,s}, \mathbf{Q}_{\tau,s}) = \mathbf{Q}_{t,s}^T \, \boldsymbol{\delta}(\mathbf{r}_{\tau,t}, \mathbf{Q}_{\tau,t}) \, \frac{d\xi_t}{d\xi_s} + \boldsymbol{\delta}(\mathbf{r}_{t,s}, \mathbf{Q}_{t,s}) \, ,$$

and the consistency property is proved.

A similar proof can be carried out for the deformation measure pertaining to polar shells and to Cosserat continua.

The formal analogy between the deformation measures of Timoshenko beams, of polar shells with drilling rotations and of Cosserat continua, has led some authors to consider the first two as special, respectively one- and two-dimensional, cases of the third [21].

!!! CE: refs?

In any case, despite the increasing popularity of Cosserat continua, and their application to modeling special phenomena in various field of structural mechanics (see, e.g.,), a simple analysis shows that the finite deformation measure of the three-dimensional Cosserat continuum is redundant (see Sect. 3.1). Indeed it can be proved [51] that the vanishing of the field of strain gaps implies the vanishing of the field of curvature changes:

$$\boldsymbol{\Delta}(\boldsymbol{\chi}_{t,s}, \mathbf{Q}_{t,s}) = 0 \quad \Longrightarrow \quad \mathbf{C}(\mathbf{Q}_{t,s}) = 0 \, .$$

The redundancy is due to the integrability conditions satisfied by the field $\partial \boldsymbol{\chi}_{t,s}$. A redundancy argument, which is more difficult to be prove, should then also apply to the two-dimensional model of polar shells with drilling rotations while the one-dimensional beam model is certainly nonredundant due to the absence of integrability conditions.

But for the three-dimensional Cosserat continua worse things are to come: if the redundant field of curvature changes is removed, in the attempt to obtain a nonredundant deformation measure, the three-dimensional Cosserat continuum collapses into a Cauchy continuum [51]. This shortcoming should lead to the conclusion that the three-dimensional Cosserat continuum is based on an ill-posed kinematical model.

On the other hand we observe that, despite their wide acceptance (see, e.g., [8,9,12,43]), the deformation measures reported in Sect. 4.5 and commonly adopted in the literature for polar shells without drilling rotations, lead to physically implausible results in the case of significant membrane strains. Indeed a simple computation

reveals an unrealistic behavior of an inflated polar spherical baloon since an increase of flexural curvature is measured when the radius increases. The effect is due to the amplification of the convected tangent vectors due to the deformation.

To eleminate this shortcoming we may redefine the deformation measures for polar shells without drilling rotations as:

$$\begin{aligned}
\boldsymbol{\varepsilon}(\boldsymbol{\chi}_{t,s})(\mathbf{a},\mathbf{b}) &:= \mathbf{g}(\boldsymbol{\chi}_{t,s*}\mathbf{a},\ \boldsymbol{\chi}_{t,s*}\mathbf{b}) - \mathbf{g}(\mathbf{a},\ \mathbf{b})\,, && \text{membrane strain}\,,\\
\boldsymbol{\delta}(\mathbf{u}_{t,s})(\mathbf{a}) &:= \mathbf{g}(\mathbf{d}_t,\ \boldsymbol{\chi}_{t,s*}\mathbf{a}) - \mathbf{g}(\mathbf{d}_s,\ \mathbf{a})\,, && \text{shear sliding}\,,\\
\boldsymbol{\chi}(\mathbf{u}_{t,s})(\mathbf{a},\mathbf{b}) &:= \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{d}_t,\ \mathbf{R}_{t,s}\mathbf{b}) - \mathbf{g}(\partial_{\mathbf{a}}\mathbf{d}_s,\ \mathbf{b})\,, && \text{curvature change}\,,
\end{aligned}$$

where $\mathbf{R}_{t,s}$ is the isometric transformation associated with the push forward $\boldsymbol{\chi}_{t,s*}$ according to the polar decomposition formula. The new expression for the curvature change correctly predicts no flexural curvature in the inflated polar spherical baloon when the radius is changed. A detailed discussion of these topics is provided in [57].

5 Equilibrium

The proof of the *virtual work principle*, which is the basic theoretical result in continuum mechanics, requires that virtual displacements be considered as vector fields belonging to a larger space. More precisely virtual displacements at any $\mathbf{u} \in \mathbb{M}$ are assumed to belong to the Sobolev space $H^k(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}}) \supset \mathrm{C}^k(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}})$ and the differential of the strain measure from $\mathbf{u} \in \mathbb{M}$ is assumed to be a bounded linear differential operator of Korn type [34,40],

$$\partial\mathbf{A}(\mathbf{i}_t) \in BL\left(H^k(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}})\,;\mathscr{L}^2(\mathbb{B}_t\,;D)\right),$$

where $\mathbf{i}_t = \mathbf{u}_{t,t}$ is the identity on $\mathbb{P}_t$.

Virtual displacements in the kernel of the tangent deformation operator $\partial\mathbf{A}(\mathbf{i}_t)$ are said to be *rigid* at $\mathbf{u} \in \mathbb{M}$.

Since virtual displacements belong to the Hilbert space $H^k(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}})$, the force systems $\mathbf{f}_t \in BL\left(H^k(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}})\,;\mathscr{R}\right)$ belong to the dual Hilbert space.

Equilibrium of a force system is expressed by the condition of orthogonality to any admissible rigid virtual displacement,

$$\langle \mathbf{f}_t, \mathbf{v}_t\rangle = 0 \qquad \forall\,\mathbf{v}_t \in \mathrm{Ker}\partial\mathbf{A}(\mathbf{i}_t)^{\perp} \subset H^k(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}})\,.$$

In the presence of kinematic constraints, admissible virtual displacements belong to a closed linear subspace $\mathscr{V}(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}}) \subseteq H^k(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}})$ and referential admissible virtual displacements to the closed linear subspace $\mathring{\mathscr{V}}(\mathbb{B}_s\,;\mathbb{T}_{\mathbb{S}}) \subseteq H^k(\mathbb{B}_s\,;\mathbb{T}_{\mathbb{S}})$.

- The *virtual work theorem* [40] ensures that, if a force system

  $$\mathbf{f}_t \in BL\left(H^k(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}})\,;\mathscr{R}\right)$$

  is in equilibrium, there exists a stress field $\boldsymbol{\sigma} \in \mathscr{L}^2(\mathbb{B}_t\,;S)$ satisfying the variational condition

  $$\int_{\mathbb{B}_t} \boldsymbol{\sigma}_{\mathbf{x}_t} \,:\, (\partial\mathbf{A}(\mathbf{i}_t)\cdot\mathbf{v})_{\mathbf{x}_t}\, d\mu_t = \langle \mathbf{f}_t, \mathbf{v}_t\rangle \qquad \forall\,\mathbf{v}_t \in \mathscr{V}(\mathbb{B}_t\,;\mathbb{T}_{\mathbb{S}})\,.$$

The local values $\boldsymbol{\sigma}_{\mathbf{x}_t}$ of the stress field belong to the finite-dimensional space S dual to D.

When transformed to the reference configuration, the virtual work condition reads

$$\int_{\mathbb{B}} \mathfrak{S}_{\mathbf{x}_s} \,:\, \big(\partial\mathbf{A}(\mathbf{u}_{t,s})\cdot\delta\mathbf{u}_{t,s}\big)_{\mathbf{x}_s} d\mu_s = \langle \mathbf{G}(\mathbf{u}_{t,s})\cdot\mathbf{f}_t, \delta\mathbf{u}_{t,s}\rangle$$
$$\forall\, \delta\mathbf{u}_{t,s} \in \mathcal{V}(\mathbb{B}_s;\mathbb{T}_{\mathbb{S}})\,,$$

where $\mathbf{G}(\mathbf{u}_{t,s})\cdot\mathbf{f}_t$ is the equivalent force in the reference configuration, defined by the identity

$$\langle \mathbf{G}(\mathbf{u}_{t,s})\cdot\mathbf{f}_t, \delta\mathbf{u}_{t,s}\rangle := \langle \mathbf{f}_t, \delta\mathbf{u}_{t,s}\circ\boldsymbol{\chi}_{t,s}^{-1}\rangle \qquad \forall\, \delta\mathbf{u}_{t,s} \in \mathcal{V}(\mathbb{B}_s;\mathbb{T}_{\mathbb{S}})\,,$$

and $\mathfrak{S} \in \mathscr{L}^2(\mathbb{B};S)$ is the referential stress measure conjugate to the finite deformation $\mathbf{A}(\mathbf{u}_{t,s}) \in \mathrm{C}^0(\mathbb{B}_s;D)$, locally defined as

$$\mathfrak{S}_{\mathbf{x}_s} = \mathbf{L}_{\mathbf{x}_s}(\mathbf{u}_{t,s})^{-T}\cdot\boldsymbol{\sigma}_{\mathbf{x}_t}\,,$$

where

$$\mathbf{L}_{\mathbf{x}_s}(\mathbf{u}_{t,s}) := \partial_1\mathbf{S}_{\mathbf{x}_s}(0,\mathbf{u}_{t,s}) \in BL\,(S;S)$$

is assumed to be invertible. The non-linear operator $\mathbf{S}$ was introduced in Sect. 3.1 in stating the consistency property and ∂_1 denotes the partial derivative with respect to the first argument.

The directional derivative $\partial\mathbf{A}(\mathbf{u}_{t,s})\cdot\delta\mathbf{u}_{t,s}$ is defined pointwise by considering a virtual trajectory through $\mathbf{u}_{t,s}$ with tangent $\delta\mathbf{u}_{t,s}$ and setting

$$(\partial\mathbf{A}(\mathbf{u}_{t,s})\cdot\delta\mathbf{u}_{t,s})_{\mathbf{x}_s} := \partial_{\delta\mathbf{u}_{t,s}}\mathbf{A}_{\mathbf{x}_s}(\mathbf{u}_{t,s}) = \frac{\partial}{\partial t}\,\mathbf{A}_{\mathbf{x}_s}(\mathbf{u}_{t,s})$$
$$= \partial\mathbf{N}((\mathbf{D}\mathbf{u}_{t,s})_{\mathbf{x}_s})\cdot\frac{\partial}{\partial t}(\mathbf{D}\mathbf{u}_{t,s})_{\mathbf{x}_s} = \partial\mathbf{N}((\mathbf{D}\mathbf{u}_{t,s})_{\mathbf{x}_s})\cdot(\mathbf{D}\delta\mathbf{u}_{t,s})_{\mathbf{x}_s}\,.$$

The dot denotes linear dependence on the subsequent term.

6 Elastic equilibrium

Green's *elastic energy* is a scalar function $\varphi_{\mathbf{x}_s} \in \mathrm{C}^2(D;\mathscr{R})$ that maps the local values of the finite elastic deformation $\mathfrak{D}_{\mathbf{x}_s} \in D$ into the corresponding elastic energy $\varphi_{\mathbf{x}_s}(\mathfrak{D}_{\mathbf{x}_s})$ per unit volume in the reference placement $\mathbb{P}_s$.

- The *elastic law* relates the *local deformation measure* $\mathfrak{D}_{\mathbf{x}_s} \in D$ to the conjugate *local stress state* $\mathfrak{S}_{\mathbf{x}_s} \in S$:

$$\mathfrak{S}_{\mathbf{x}_s} = \partial\varphi_{\mathbf{x}}(\mathfrak{D}_{\mathbf{x}_s})\,.$$

The reference placement $\mathbb{P}_s$ is assumed to be a *natural state* for the material. This means that $\mathfrak{S}_{\mathbf{x}_s} = (\partial\varphi_{\mathbf{x}_s})(\mathfrak{D}_{\mathbf{x}_s})$ vanishes if $\mathfrak{D}_{\mathbf{x}_s} = 0$. The *global elastic energy* $\varphi \in \mathrm{C}^2(\mathscr{L}^2(\mathbb{B}_s; D); \mathscr{R})$ of the body is the integral of the specific elastic energy over the base manifold,

$$\varphi(\mathfrak{D}) = \int_{\mathbb{B}} \varphi_{\mathbf{x}_s}(\mathfrak{D}_{\mathbf{x}_s})\, d\mu_s\,.$$

Hereafter the suffices t, s are dropped whenever not strictly necessary.

The *global elastic potential* $\phi \in \mathrm{C}^2(\mathrm{C}^k(\mathbb{B}_s, \mathbb{P}_t); \mathscr{R})$ provides the elastic energy associated with the configuration change $\mathbf{u} \in \mathrm{C}^k(\mathbb{B}_s, \mathbb{P}_t)$ and is given by

$$\phi(\mathbf{u}) := (\varphi \circ \mathbf{A})(\mathbf{u}) = \int_{\mathbb{B}} (\varphi_{\mathbf{x}} \circ \mathbf{A}_{\mathbf{x}})(\mathbf{u})\, d\mu\,.$$

Enforcing the constitutive law in terms of the elastic potential, the referential equilibrium of the body at time $t \in I$ is expressed by

$$\langle\, \partial\phi(\mathbf{u}), \delta\mathbf{u} \,\rangle \;=\; \langle\, \mathbf{G}(\mathbf{u}) \cdot \mathbf{f}, \delta\mathbf{u} \,\rangle \qquad\qquad \forall\, \delta\mathbf{u} \in \mathscr{V}(\mathbb{B}_s; \mathbb{T}_{\mathbb{S}})\,.$$

The bounded linear functionals $\mathbf{G}(\mathbf{u}) \cdot \mathbf{f} \in BL\,(\mathscr{V}(\mathbb{B}_s; \mathbb{T}_{\mathbb{S}}); \mathscr{R})$ and $\partial\phi(\mathbf{u}) \in BL\,(\mathscr{V}(\mathbb{B}_s; \mathbb{T}_{\mathbb{S}}); \mathscr{R})$ provide, respectively, the referential applied load and the referential elastic response of the body. In the sequel the terms *form*, *covector* and *bounded linear functional* should be considered as synonyms.

With $\mathbf{G}_{\mathbf{f}}(\mathbf{u}) := \mathbf{G}(\mathbf{u}) \cdot \mathbf{f}$, the equilibrium condition may be written equivalently by imposing the vanishing of the resultant force system on the body:

$$(\partial\phi - \mathbf{G}_{\mathbf{f}})(\mathbf{u}) \;=\; 0\,.$$

6.1 Incremental equilibrium

The incremental equilibrium is imposed by taking the total time derivative of the non-linear condition along the equilibrium path:

$$\frac{d}{dt}\Big[(\partial\phi - \mathbf{G}_{\mathbf{f}})(\mathbf{u})\Big] \;=\; 0\,.$$

Since both the configuration change $\mathbf{u}$ and the force map $\mathbf{f}$ depend on $t \in I$, the incremental equilibrium condition is given by

$$\partial_{\dot{\mathbf{u}}}(\partial\phi - \mathbf{G}_{\mathbf{f}})(\mathbf{u}) \;=\; \mathbf{G}(\mathbf{u}) \cdot\, \dot{\mathbf{f}}\,,$$

where as usual a superimposed dot denotes the time derivative.

The *total tangent stiffness* of the body is the directional derivative

$$\mathbf{K}(\mathbf{u}) := \partial(\partial\phi - \mathbf{G}_{\mathbf{f}})(\mathbf{u})\,,$$

and the incremental equilibrium is accordingly written as

$$\mathbf{K}(\mathbf{u}) \cdot \dot{\mathbf{u}} = \mathbf{G}(\mathbf{u}) \cdot \dot{\mathbf{f}}\,.$$

However, when dealing with polar continua, the directional derivative of the form-valued map $(\partial\phi - \mathbf{G_f})$ at a configuration $\mathbf{u} \in \mathbb{M}$ cannot be taken in the classical way since the ambient space $\mathbb{S}$ is a non-linear manifold and hence also the configuration space $\mathbb{M} = \mathrm{C}^k(\mathbb{B}_s\,;\mathbb{S})$ is a non-linear manifold of maps. Indeed in this case the evaluation of the directional derivative would require us to take the limit of differences between covectors defined on distinct tangent spaces and these differences would have no meaning until a further geometric structure is given to the space manifold. The issue is illustrated in the subsequent sections.

7 Affine connections and covariant differentiation

An *affine connection* on a differentiable manifold $\mathbb{M}$ is a map $\mathbf{X} \mapsto \nabla\mathbf{X}$ which associates to any vector field $\mathbf{X} : \mathbb{M} \mapsto \mathbb{T}_{\mathbb{M}}$ a tensor field

$$\nabla\mathbf{X} : \mathbb{M} \mapsto BL\,(\mathbb{T}_{\mathbb{M}}\,;\mathbb{T}_{\mathbb{M}})$$

of type $(1,1)$ such that, for any pair of tangent vectors $\mathbf{Y_u}, \mathbf{Z_u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$, the following characteristic properties of a derivation are met:

$i)$ $\quad \nabla f = \partial f\,;$

$ii)$ $\quad \nabla\mathbf{X}\,[\alpha\,\mathbf{Y_u} + \beta\,\mathbf{Z_u}] = \alpha\,\nabla\mathbf{X}\,[\mathbf{Y_u}] + \beta\,\nabla\mathbf{X}\,[\mathbf{Z_u}]\,;$

$iii)$ $\quad \begin{cases} \nabla(\mathbf{X}_1 + \mathbf{X}_2) = \nabla\mathbf{X}_1 + \nabla\mathbf{X}_2\,, \\ \nabla(f\,\mathbf{X})\,[\mathbf{Y_u}] = (\partial f\,[\mathbf{Y_u}])\,\mathbf{X} + f\,(\nabla\mathbf{X}\,[\mathbf{Y_u}])\,, \end{cases}$

where $\alpha, \beta \in \mathscr{R}$, $f \in \mathrm{C}^1(\mathbf{u}, U)$ where $U(\mathbf{u}) \subseteq \mathbb{M}$ is a neighborhood of $\mathbf{u} \in \mathbb{M}$ and ∂ denotes the directional differentiation.

- Property $i)$ asserts that directional and covariant derivative are the same for scalar fields.
- Property $ii)$ expresses the $(1,1)$ tensoriality of $\nabla\mathbf{X}$.
- Properties $iii_1, iii_2)$ are characteristic of a derivation.

The local value at $\mathbf{u} \in \mathbb{M}$ of the tangent vector field $\nabla_{\mathbf{Y_u}}\mathbf{X} : \mathbb{M} \mapsto \mathbb{T}_{\mathbb{M}}$ is the covariant derivative of the tangent vector field $\mathbf{X} : \mathbb{M} \mapsto \mathbb{T}_{\mathbb{M}}$ along the tangent vector $\mathbf{Y_u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$.

The covariant derivative of a tensor field $\mathbf{a} \in BL\,(\mathbb{T}_{\mathbb{M}}, \mathbb{T}_{\mathbb{M}}\,;\mathscr{R})$ is defined so that Leibniz rule holds:

$$(\nabla_{\mathbf{Z}}\,\mathbf{a})(\mathbf{X}, \mathbf{Y}) := \partial_{\mathbf{Z}}(\mathbf{a}(\mathbf{X}, \mathbf{Y})) - \mathbf{a}(\nabla_{\mathbf{Z}}\mathbf{X}, \mathbf{Y}) + \mathbf{a}(\mathbf{X}, \nabla_{\mathbf{Z}}\mathbf{Y})\,.$$

The definition is well-posed because the left-hand side does not depend on the extension of the vectors $\mathbf{X_u}, \mathbf{Y_u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$ to vector fields $\mathbf{X}, \mathbf{Y} : U(\mathbf{u}) \mapsto \mathbb{T}_{\mathbb{M}}$, even if each one of the summands in the right-hand side depends on such an extension.

This property ensures that the expression above defines a three-times covariant tensor field on $\mathbb{M}$ and can be easily assessed by applying the following tensoriality criterion [2,50].

Theorem 1. *A multilinear application*

$$A : \overbrace{\mathbb{T}_{\mathbb{M}} \times \ldots \times \mathbb{T}_{\mathbb{M}}}^{k \text{ times}} \mapsto \mathscr{R}\,,$$

which is linear on the space $\mathrm{C}^\infty(\mathbb{M})$ *in the sense that*

$$A(\mathbf{v}_1, \ldots, f\,\mathbf{v}_i, \ldots, \mathbf{v}_k) = f\,A(\mathbf{v}_1, \ldots, \mathbf{v}_k) \quad \forall\, i = 1, \ldots, k\,, \quad \forall\, f \in \mathrm{C}^\infty(\mathbb{M})\,,$$

can be pointwise represented by a unique tensor field T *on* $\mathbb{M}$*. In other words,* $A = A_T$*, where*

$$A_T(\mathbf{v}_1 \ldots \mathbf{v}_k)(\mathbf{p}) := T(\mathbf{p})(\mathbf{v}_1(\mathbf{p}), \ldots, \mathbf{v}_k(\mathbf{p})) \quad \forall\, \mathbf{p} \in \mathbb{M}\,,$$

is the multilinear application pointwise defined by the tensor field T *on* $\mathbb{M}$*.*

7.1 Parallel transport and connection

It is known from differential geometry (see, e.g., [3]) that the parallel transport $\mathscr{T}^{\mathbb{S}}_{\lambda,\xi} : \mathbb{T}_{\mathbb{S}}(\mathbf{c}(\xi)) \mapsto \mathbb{T}_{\mathbb{S}}(\mathbf{c}(\lambda))$ along a regular curve $\mathbf{c}$ in the ambient space manifold $\mathbb{S}$ is a solution of the ordinary differential equation

$$\nabla_{\dot{\mathbf{c}}(\lambda)}(\mathscr{T}^{\mathbb{S}}_{\lambda,\xi}\mathbf{v}_\xi) = 0 \quad \forall\, \lambda, \xi \in I\,.$$

By the uniqueness of the solution of an ODE we infer the validity of the composition rule

$$\mathscr{T}^{\mathbb{S}}_{\lambda,\mu} = \mathscr{T}^{\mathbb{S}}_{\lambda,\xi} \circ \mathscr{T}^{\mathbb{S}}_{\xi,\mu}\,.$$

Parallel transport induces a connection ∇ on the manifold according to the formula for covariant differentiation,

$$\nabla_{\dot{\mathbf{c}}(\lambda)}\mathbf{v}_\lambda := \frac{\partial}{\partial \xi}\bigg|_{\xi=\lambda} (\mathscr{T}^{\mathbb{S}}_{\lambda,\xi}\mathbf{v}_\xi)\,,$$

where $\mathbf{v}_\lambda := \mathbf{v}(\mathbf{c}(\lambda)) \in \mathbb{T}_{\mathbb{S}}(\mathbf{c}(\lambda))$ is a vector field tangent to $\mathbb{S}$. Note that the time derivative makes sense since

$$\mathscr{T}^{\mathbb{S}}_{\lambda,\xi}\mathbf{v}_\xi \in \mathbb{T}_{\mathbb{S}}(\mathbf{c}(\lambda)) \quad \forall\, \xi \in I\,.$$

It is easy to check that the field $\mathscr{T}^{\mathbb{S}}_{\lambda,\xi}\mathbf{v}_\xi$ is parallel-transported along $\mathbf{c}$ according to the connection since

$$\nabla_{\dot{\mathbf{c}}(\lambda)}(\mathscr{T}^{\mathbb{S}}_{\lambda,\xi}\mathbf{v}_\xi) = \frac{\partial}{\partial \mu}\bigg|_{\mu=\lambda} (\mathscr{T}^{\mathbb{S}}_{\lambda,\mu}\,\mathscr{T}^{\mathbb{S}}_{\mu,\xi}\mathbf{v}_\xi) = \frac{\partial}{\partial \mu}\bigg|_{\mu=\lambda} (\mathscr{T}^{\mathbb{S}}_{\lambda,\xi}\mathbf{v}_\xi) = 0\,.$$

A connection on the finite-dimensional space manifold $\mathbb{S}$, which is modeled on the linear space $\mathcal{R}^d$, induces a corresponding connection on the infinite-dimensional manifold $\mathbb{M} = \mathrm{C}^k(\mathbb{B}_s;\mathbb{S})$ of admissible configuration changes which is modeled on the Banach space $\mathrm{C}^k(\mathbb{B}_s;\mathcal{R}^d)$.

Indeed the notion of parallel transport $\mathcal{T}^{\mathbb{M}}_{\tau,t}$ of a vector feld $\delta\mathbf{u}_{t,s} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u}_{t,s})$ along curves $\{\,\mathbf{u}_{t,s}\,,\ t \in I\,\}$ on the manifold $\mathbb{M}$ is defined pointwise by setting

$$(\mathcal{T}^{\mathbb{M}}_{\tau,t}\delta\mathbf{u}_{t,s})(\mathbf{p}) := \mathcal{T}^{\mathbb{S}}_{\tau,t}\,(\delta\mathbf{u}_{t,s}(\mathbf{p})) \quad \forall\,\mathbf{p} \in \mathbb{B}_s\,.$$

Accordingly the covariant derivative on $\mathbb{M}$ is also defined pointwise by

$$(\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}_{t,s}}\,\delta\mathbf{u}_{t,s})(\mathbf{p}) := \nabla^{\mathbb{S}}_{\dot{\mathbf{u}}_{t,s}(\mathbf{p})}\,\delta\mathbf{u}_{t,s}(\mathbf{p}) \quad \forall\,\mathbf{p} \in \mathbb{B}_s\,,$$

and is related to the parallel transport by the relation

$$\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}_{t,s}}\,\delta\mathbf{u}_{t,s} = \frac{\partial}{\partial\tau}\bigg|_{\tau=t} (\mathcal{T}^{\mathbb{M}}_{t,\tau}\delta\mathbf{u}_{\tau,s})\,.$$

8 Tangent stiffness

Once a connection is defined on the manifold $\mathbb{M}$ of admissible configuration changes, the total tangent stiffness may be computed by performing covariant derivatives instead of directional derivatives to get the expression

$$\mathbf{K}(\mathbf{u}) := \nabla^{\mathbb{M}}(\partial\boldsymbol{\phi} - \mathbf{G}_{\mathbf{f}})(\mathbf{u}) = \nabla^{\mathbb{M}}\boldsymbol{\alpha}(\mathbf{u})\,,$$

where

$$\boldsymbol{\alpha} = \partial\boldsymbol{\phi} - \mathbf{G}_{\mathbf{f}}$$

is the equilibrium gap resulting from the difference between the covector fields representing the elastic response $\partial\boldsymbol{\phi} : \mathbb{M} \mapsto \mathbb{T}^*_{\mathbb{M}}$ and the referential load $\mathbf{G}_{\mathbf{f}} : \mathbb{M} \mapsto \mathbb{T}^*_{\mathbb{M}}$.

Accordingly the bounded linear functional $\boldsymbol{\alpha}(\mathbf{u}) \in \mathbb{T}^*_{\mathbb{M}}(\mathbf{u}) = BL\,(\mathbb{T}_{\mathbb{M}}(\mathbf{u})\,;\mathcal{R})$ provides the resultant referential force corresponding to the configuration change $\mathbf{u} \in \mathbb{M}$.

As shown in Sect. 7, the covariant derivative $\nabla_{\dot{\mathbf{u}}}\,\boldsymbol{\alpha}(\mathbf{u})$ is defined by means of a formal application of the Leibniz rule of calculus,

$$(\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}}\,\boldsymbol{\alpha}(\mathbf{u}))\,[\delta\mathbf{u}] := \partial_{\dot{\mathbf{u}}}\,(\boldsymbol{\alpha}(\mathbf{u})\,[\hat{\delta\mathbf{u}}]) - \boldsymbol{\alpha}(\mathbf{u})[\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}}\,\hat{\delta\mathbf{u}}]\,.$$

The vector field $\hat{\delta\mathbf{u}} \in \mathbb{T}_{\mathbb{M}}(U(\mathbf{u}))$ is an extension of the vector $\delta\mathbf{u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$ to a neighborhood $U(\mathbf{u}) \subseteq \mathbb{M}$ of $\mathbf{u} \in \mathbb{M}$. Recall that $\mathbf{u} \in \mathbb{M}$ is an admissible configuration and that $\delta\mathbf{u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$ is a virtual displacement from that configuration.

Although both derivatives in the right-hand side of the Leibniz formula depend on the assumed extension of the virtual displacement $\delta\mathbf{u}$, the left-hand side is independent of such an extension and hence is tensorial in $\delta\mathbf{u}$ by the tensoriality criterion provided in Theorem 1.

Hereafter the suffix $\mathbb{M}$ is dropped unless necessary.

The Hessian of the elastic potential $\phi = \varphi \circ \mathbf{A}$ provides the constitutive stiffness and is the twice covariant tensor field on the manifold $\mathbb{M}$ defined by

$$\nabla^2_{\dot{\mathbf{u}}\,\delta\mathbf{u}}\phi(\mathbf{u}) := (\nabla_{\dot{\mathbf{u}}}\partial\phi(\mathbf{u}))\,[\delta\mathbf{u}] = \partial_{\dot{\mathbf{u}}}\,\partial_{\hat{\delta\mathbf{u}}}\,\phi(\mathbf{u}) - \partial\phi(\mathbf{u})\,[\nabla_{\dot{\mathbf{u}}}\,\hat{\delta\mathbf{u}}]\,.$$

Applying the chain rule to $\phi(\mathbf{u}) = (\varphi \circ \mathbf{A})(\mathbf{u})$ and the Leibniz rule, the evaluation of the first term of the right-hand side yields

$$\begin{aligned}\partial_{\dot{\mathbf{u}}}(\partial\varphi(\mathbf{A}(\mathbf{u}))\cdot\partial\mathbf{A}(\mathbf{u})\cdot\hat{\delta\mathbf{u}}) = \partial^2\varphi(\mathbf{A}(\mathbf{u}))\cdot(\partial\mathbf{A}(\mathbf{u})\cdot\hat{\delta\mathbf{u}})\cdot(\partial\mathbf{A}(\mathbf{u})\cdot\dot{\mathbf{u}}) + \\ + \partial\varphi(\mathbf{A}(\mathbf{u}))\cdot(\partial_{\dot{\mathbf{u}}}\,\partial_{\hat{\delta\mathbf{u}}}\mathbf{A})(\mathbf{u}).\end{aligned}$$

The first term of the right-hand side is the *elastic tangent stiffness* which is a symmetric bilinear form in $\dot{\mathbf{u}}\,,\delta\mathbf{u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$. The symmetry of the second directional derivative of the functional $\varphi \in \mathrm{C}^2(\mathscr{L}^2(\mathbb{B}_s\,;D)\,;\mathscr{R})$ is a classical result since $\mathscr{L}^2(\mathbb{B}_s\,;D)$ is a linear space.

The remainder provides the *geometric tangent stiffness*, a bilinear form in $\dot{\mathbf{u}}$, $\delta\mathbf{u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$ given by

$$\partial\varphi(\mathbf{A}(\mathbf{u}))\cdot\Big[(\partial_{\dot{\mathbf{u}}}\partial_{\hat{\delta\mathbf{u}}} - \partial_{\nabla_{\dot{\mathbf{u}}}\hat{\delta\mathbf{u}}})\mathbf{A}\Big](\mathbf{u}) = \partial\varphi(\mathbf{A}(\mathbf{u}))\cdot\Big(\nabla^2_{\dot{\mathbf{u}}\,\delta\mathbf{u}}\,\mathbf{A}\Big)(\mathbf{u})\,.$$

We remark that the directional derivative of $\mathbf{A}$ at $\mathbf{u}$ is well defined since $\mathbf{A}(\mathbf{u})$ belongs to the linear space $\mathscr{L}^2(\mathbb{B}_s\,;D)$.

8.1 Torsion and symmetry

The *torsion* of the connection $\nabla^{\mathbb{S}}$ is the mixed tensor field $\boldsymbol{T}^{\mathbb{S}} \in \mathrm{L}(\mathbb{T}_{\mathbb{S}},\mathbb{T}_{\mathbb{S}}\,;\mathbb{T}_{\mathbb{S}})$, twice covariant and one time contravariant, defined by

$$\boldsymbol{T}^{\mathbb{S}}(\mathbf{v},\mathbf{w}) = \nabla^2_{\mathbf{v},\mathbf{w}} - \nabla^2_{\mathbf{w},\mathbf{v}} = [\mathbf{v},\mathbf{w}] - \nabla_{\mathbf{v}}\mathbf{w} + \nabla_{\mathbf{w}}\mathbf{v}\,.$$

The second equality follows from the formula for the second covariant derivative of a scalar field $f \in \mathrm{C}^2(\mathbb{S}\,;\mathscr{R})$:

$$\nabla^2_{\mathbf{v},\mathbf{w}}\,f = \partial_{\mathbf{v}}\partial_{\mathbf{w}}\,f - (\nabla_{\mathbf{v}}\mathbf{w})\,f\,,$$

where $\mathbf{h}\,f := \partial f\cdot\mathbf{h}$ denotes the directional derivative of $f \in \mathrm{C}^2(\mathbb{S}\,;\mathscr{R})$ along $\mathbf{h} \in \mathbb{T}_{\mathbb{S}}$. Hence,

$$\boldsymbol{T}^{\mathbb{S}}(\mathbf{v},\mathbf{w})\,f = (\nabla^2_{\mathbf{v},\mathbf{w}} - \nabla^2_{\mathbf{w},\mathbf{v}})\,f = (\partial_{\mathbf{v}}\partial_{\mathbf{w}} - \partial_{\mathbf{w}}\partial_{\mathbf{v}} - \nabla_{\mathbf{v}}\mathbf{w} + \nabla_{\mathbf{w}}\mathbf{v})\,f\,.$$

The formula then follows by recalling the definition of the Lie bracket,

$$[\mathbf{v},\mathbf{w}]\,f = (\partial_{\mathbf{v}}\partial_{\mathbf{w}} - \partial_{\mathbf{w}}\partial_{\mathbf{v}})\,f\,.$$

A well-known result of differential geometry states that the Lie bracket is equal to the Lie derivative, according to the formula

$$[\mathbf{X}, \mathbf{Y}]_s = (\mathscr{L}_{\mathbf{X}}\mathbf{Y})_s := \frac{d}{dt}\bigg|_{t=s} \boldsymbol{\chi}_{s,t*}\mathbf{Y}_t\,.$$

The *torsion* $\boldsymbol{T}^{\mathbb{M}} \in \mathrm{L}(\mathbb{T}_{\mathbb{M}}, \mathbb{T}_{\mathbb{M}}; \mathbb{T}_{\mathbb{M}})$ of the connection $\nabla^{\mathbb{M}}$ on the infinite-dimensional manifold $\mathbb{M} = \mathrm{C}^k(\mathbb{B}_s; \mathbb{S})$ is defined pointwise in terms of the parent torsion $\boldsymbol{T}^{\mathbb{S}}$ of $\nabla^{\mathbb{S}}$ by the identity

$$\Big(\boldsymbol{T}^{\mathbb{M}}(\mathbf{X}_{\mathbf{u}}, \mathbf{Y}_{\mathbf{u}})\Big)_{\mathbf{p}} = \boldsymbol{T}^{\mathbb{S}}((\mathbf{X}_{\mathbf{u}})_{\mathbf{p}}, (\mathbf{Y}_{\mathbf{u}})_{\mathbf{p}}) \in \mathbb{T}_{\mathbb{S}}(\mathbf{u}_{\mathbf{p}}) \qquad \forall\, \mathbf{p} \in \mathbb{B}_s\,.$$

Hence,

$$\boldsymbol{T}^{\mathbb{M}}(\mathbf{X}_{\mathbf{u}}, \mathbf{Y}_{\mathbf{u}}) = (\nabla^2_{\mathbf{X}_{\mathbf{u}},\mathbf{Y}_{\mathbf{u}}} - \nabla^2_{\mathbf{Y}_{\mathbf{u}},\mathbf{X}_{\mathbf{u}}}) = [\mathbf{X}_{\mathbf{u}}, \mathbf{Y}_{\mathbf{u}}] - \nabla_{\mathbf{X}_{\mathbf{u}}}\mathbf{Y}_{\mathbf{u}} + \nabla_{\mathbf{Y}_{\mathbf{u}}}\mathbf{X}_{\mathbf{u}}\,.$$

A vanishing torsion $\boldsymbol{T}^{\mathbb{S}}$ implies that the Hessian of any $f \in \mathrm{C}^2(\mathbb{S}; \mathscr{R})$ is symmetric. The finite dimensionality of D also ensures that the Hessian

$$(\nabla^2_{\mathbf{X}_{\mathbf{u}},\mathbf{Y}_{\mathbf{u}}}\mathbf{A}_{\mathbf{x}})(\mathbf{u}) \in D\,,$$

of the local deformation map $\mathbf{A}_{\mathbf{x}} \in \mathrm{C}^2(\mathbb{M}; D)$ is symmetric. If follows that the geometric tangent stiffness is also symmetric.

9 Conservative versus nonconservative loads

Suppose that the referential force system acting on the body is positional and conservative in the sense that there exists a scalar potential $F_{\mathbf{f}} \in \mathrm{C}^1(\mathbb{M}; \mathscr{R})$ linearly dependent on $\mathbf{f}$ such that

$$\mathbf{G}_{\mathbf{f}}(\mathbf{u}) = \mathbf{G}(\mathbf{u}) \cdot \mathbf{f} = -\partial F_{\mathbf{f}}(\mathbf{u})\,.$$

Then, in terms of the total potential $P = \phi + F_{\mathbf{f}}$, the sum of the elastic potential $\phi = \varphi \circ \mathbf{A}$ and the referential load potential $F_{\mathbf{f}}$, the condition of elastic equilibrium becomes

$$\partial P(\mathbf{u}) = \mathbf{o}\,.$$

A solution $\mathbf{u} \in \mathbb{M}$ is then a *critical point* of P. Accordingly, the incremental equilibrium condition can be expressed as

$$\nabla_{\dot{\mathbf{u}}}\, \partial P(\mathbf{u}) = -\partial F_{\dot{\mathbf{f}}}(\mathbf{u})\,,$$

and, in variational form, as

$$\nabla^2_{\dot{\mathbf{u}}\,\delta\mathbf{u}} P(\mathbf{u}) = \partial_{\dot{\mathbf{u}}}\, \partial_{\hat{\delta\mathbf{u}}} P(\mathbf{u}) - \partial P(\mathbf{u})\,[\nabla_{\dot{\mathbf{u}}}\, \hat{\delta\mathbf{u}}] = \partial_{\dot{\mathbf{u}}}\, \partial_{\hat{\delta\mathbf{u}}} P(\mathbf{u}) = -\langle \partial F_{\dot{\mathbf{f}}}(\mathbf{u}), \delta\mathbf{u}\rangle$$

$\forall\, \delta\mathbf{u} \in \mathbb{T}_{\mathbb{M}}(\mathbf{u})$ and $\forall\, \delta\hat{\mathbf{u}} \in \mathbb{T}_{\mathbb{M}}(U(\mathbf{u}))$, which is an extension of $\delta\mathbf{u}$ to a neighborhood $U(\mathbf{u}) \subseteq \mathbb{M}$. The second equality in the formula above holds since the derivative $\partial P(\mathbf{u})$ vanishes at equilibrium points $\mathbf{u} \in \mathbb{M}$.

From the previous results we see that the Hessian of the total potential at a critical point can be computed as the second directional derivative of the potential (the classical formula) by performing an arbitrary extension of the virtual displacement. Remarkably the result is tensorial and symmetric since it depends neither on the extension nor on the choosen connection. Since a torsionless connection can be considered, we infer that the Hessian has to be symmetric. It follows that the tangent stiffness $\mathbf{K}(\mathbf{u})$ at equilibrium points $\mathbf{u} \in \mathbb{M}$ is symmetric and defined by

$$\langle \mathbf{K}(\mathbf{u})\, \dot{\mathbf{u}}, \delta\mathbf{u} \rangle := \partial_{\dot{\mathbf{u}}}\, \partial_{\delta\hat{\mathbf{u}}}\, P(\mathbf{u})\,.$$

This observation was made in [5], but with some contradictions, and in a clearer but still incomplete form in [7]. Indeed the discussion given in [14] takes no account of the way in which the directional derivatives of the virtual displacement are defined and makes reference only to Riemannian connections.

Numerical evidence of the symmetry of the tangent stiffness at equilibrium points in the case of positional and conservative loads was provided in [5].

It is worth noting that the authors of [5] found a nonsymmetric but tensorial expression of the tangent stiffness for polar beams by adopting the expression above at nonequilibrium points. Indeed at noncritical points the Hessian should be evaluated by the tensorial formula

$$\langle \mathbf{K}(\mathbf{u})\, \dot{\mathbf{u}}, \delta\mathbf{u} \rangle := \nabla^2_{\dot{\mathbf{u}}\,\delta\mathbf{u}} P(\mathbf{u}) = \partial_{\dot{\mathbf{u}}}\, \partial_{\delta\hat{\mathbf{u}}}\, P(\mathbf{u}) - \partial P(\mathbf{u})\, [\nabla_{\dot{\mathbf{u}}}\, \delta\hat{\mathbf{u}}]\,,$$

which requires the definition of a connection and the choice of an extension of the virtual displacements.

The relevance of the role played by the torsion of the connection and by the extension chosen for the virtual displacement, in explaining why a nonsymmetric but tensorial stiffness may occur, was illuminated in [46] when the author was not yet aware of the paper [14].

More generally, if the referential load is nonconservative, the tangent stiffness has to be defined by the formula

$$\langle \mathbf{K}(\mathbf{u})\, \dot{\mathbf{u}}, \delta\mathbf{u} \rangle := (\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}}\, \boldsymbol{\alpha}(\mathbf{u}))\, [\delta\mathbf{u}] = \partial_{\dot{\mathbf{u}}}\, (\boldsymbol{\alpha}(\mathbf{u})\, [\delta\hat{\mathbf{u}}]) - \boldsymbol{\alpha}(\mathbf{u})[\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}}\, \delta\hat{\mathbf{u}}]\,,$$

where the resultant referential force, given by

$$\boldsymbol{\alpha} = \partial\boldsymbol{\phi} - \mathbf{G}_{\mathbf{f}} : \mathbb{M} \mapsto \mathbb{T}^*_{\mathbb{M}}\,,$$

vanishes at equilibrium points. In the general case the tangent stiffness is then tensorial but possibly nonsymmetric at equilibrium points as well. In any case, at these points the expression of the tangent stiffness is independent of the chosen connection and is given by the formula

$$\langle \mathbf{K}(\mathbf{u})\, \dot{\mathbf{u}}, \delta\mathbf{u} \rangle = (\nabla^{\mathbb{M}}_{\dot{\mathbf{u}}}\, \boldsymbol{\alpha}(\mathbf{u}))\, [\delta\mathbf{u}] = \partial_{\dot{\mathbf{u}}}\, (\boldsymbol{\alpha}(\mathbf{u})\, [\delta\hat{\mathbf{u}}])\,.$$

In [14] it was claimed that the correct symmetric stiffness for polar beams is obtained by taking the symmetric part of the nonsymmetric one. We remark that this statement is correct only for the special extension of the virtual dispacement chosen there. A comprehensive analysis of the evaluation of the tangent stiffness for polar beams can be found in [49,50].

10 Conclusions

On a non-linear manifold there is no preferential way of defining a connection among tangent spaces at different points.

The choice of a connection determines the covariant differentiation of vector fields belonging to the tangent bundle and of related covector and tensor fields. On the contrary, in the special case of an affine manifold, there is a standard connection, the euclidean connection.

If the non-linear manifold is embedded in an affine space endowed with a euclidean metric, there is a canonical way to define a Riemannian metric through the Levi–Civita connection. This connection is uniquely determined as the one that mimics some basic properties of euclidean geometry, namely, invariance of the local metric and symmetry of the second covariant derivative of scalar fields.

This connection is also the most natural to be considered due to the simple computation of the related covariant derivative in terms of the directional derivative in the ambient euclidean space.

In fact, in the polar models that we have considered, the fiber manifold is always embedded in a linear space with inner product and, according to the LeviCivita connection, the covariant derivative on the manifold is given by the orthogonal projection of the directional derivative in the parent linear space onto the tangent bundle.

Our analysis reveals that, in a general model of polar elastic continua, the tangent stiffness must be defined as the covariant derivative of the resultant referential force which is a covector field on the manifold of configuration changes.

At equilibrium points the resultant referential force vanishes and the tangent stiffness is indepenedet of the assumed connection on the fiber manifold but in general may fail to be symmetric.

The circumstance that at nonequilibrium points the expression of the tangent stiffness of polar continua and its symmetry property depend directly on the connection chosen on the fiber manifold, should not affect any *physical interpretation*. In fact it is known that, also in the euclidean space, nonconventional connections may be defined to obtain special geometric models capable, e.g., of providing mathematical models of continuous distributions of dislocations [1].

In the special case of conservative referential loads, the tangent stiffness is provided by the Hessian of the total potential and is then tensorial and symmetric at equilibrium points independently of the choice of the connection and of the extension of virtual displacements required for its evaluation.

Acknowledgements

The financial support of the Italian Ministry for University and Scientific Research (MIUR) is gratefully acknowledged.

References

[1] Bilby, B.A. (1960): Continuous distributions of dislocations. In: Sneddon, I.N., Hill, R. (eds.): Progress in solid mechanics. Vol. 1. North-Holland, Amsterdam, pp. 329–398
[2] Spivak, M. (1979): A comprehensive introduction to differential geometry. Vols. I-V. Publish or Perish, Wilmington, DE
[3] Marsden, J.E., Hughes, T.J.R. (1983): Mathematical foundations of elasticity. Prentice-Hall, Englewood Cliffs, NJ
[4] Simo, J.C. (1985): A finite strain beam formulation. The three-dimensional dynamic problem. I. Comput. Methods Appl. Mech. Engrg. **49**, 55–70
[5] Simo, J.C., Vu-Quoc, L. (1986): A three-dimensional finite-strain rod model. II. Computational aspects. Comput. Methods Appl. Mech. Engrg. **58**, 79–116
[6] Simo, J.C., Vu-Quoc, L. (1988): On the dynamics in space of rods undergoing large motions – a geometrically exact approach. Comput. Methods Appl. Mech. Engrg. **66**, 125–161
[7] Abraham, R., Marsden, J.E., Ratiu, T. (1988): Manifolds, tensor analysis, and applications. 2nd edition. Springer, New York
[8] Simo, J.C., Fox, D.D. (1989): On a stress resultant geometrically exact shell model. I. Formulation and optimal parametrization. Comput. Methods Appl. Mech. Engrg. **72**, 267–304
[9] Simo, J.C., Fox, D.D., Rifai, M.S. (1989): On a stress resultant geometrically exact shell model. II. The linear theory; computational aspects. Comput. Methods Appl. Mech. Engrg. **73**, 53–92
[10] Hughes, T.J.R., Brezzi, F. (1989): On drilling degrees of freedom. Comput. Methods Appl. Mech. Engrg. **72**, 105–121
[11] Arnold, V.I. (1989): Mathematical methods of classical mechanics. 2nd edition. Springer, New York
[12] Simo, J.C., Fox, D.D., Rifai, M.S. (1990): On a stress resultant geometrically exact shell model. III. Computational aspects of the nonlinear theory. Comput. Methods Appl. Mech. Engrg. **79**, 21–70
[13] Simo, J.C., Fox, D.D., Rifai, M.S. (1990): On a stress resultant geometrically exact shell model. IV. Variable thickness shells with through-the-thickness stretching. Comput. Methods Appl. Mech. Engrg. **81**, 91–126
[14] Simo, J.C. (1992): The (symmetric) Hessian for geometrically nonlinear models in solid mechanics: intrinsic definition and geometric interpretation. Comput. Methods Appl. Mech. Engrg. **96**, 189–200
[15] Simo, J.C., Kennedy, J.G. (1992): On a stress resultant geometrically exact shell model. V. Nonlinear plasticity: formulation and integration algorithms. Comput. Methods Appl. Mech. Engrg. **96**, 133–171
[16] Simo, J.C., Fox, D.D., Hughes, T.J.R. (1992): Formulations of finite elasticity with independent rotations. Comput. Methods Appl. Mech. Engrg. **95**, 277–288
[17] Fox, D.D., Simo, J.C. (1992): A drill rotation formulation for geometrically exact shells. Comput. Methods Appl. Mech. Engrg. **98**, 329–343

[18] Simo, J.C. (1993): On a stress resultant geometrically exact shell model. VII. Shell intersections with 5/6-DOF finite element formulations. Comput. Methods Appl. Mech. Engrg. **108**, 319–339

[19] Ibrahimbegović, A. (1994): Stress resultant geometrically nonlinear shell theory with drilling rotations. I. A consistent formulation. Comput. Methods Appl. Mech. Engrg. **118**, 265–284

[20] Ibrahimbegović, A., Frey, F. (1994): Stress resultant geometrically nonlinear shell theory with drilling rotations. II. Computational aspects. Comput. Methods. Appl. Mech. Engrg. **118**, 285–308

[21] Sansour, C., Bednarczyk, H. (1995): The Cosserat surface as a shell model, theory and finite-element formulation. Comput. Methods Appl. Mech. Engrg. **120**, 1–32

[22] Jelenić, G., Saje, M. (1995): A kinematically exact space finite strain beam model – finite element formulation by generalized virtual work principle. Comput. Methods Appl. Mech. Engrg. **120**, 131–161

[23] Ibrahimbegović, A. (1997): On the choice of finite rotation parameters. Comput. Methods Appl. Mech. Engrg. **149**, 49–71

[24] Petersen, P. (1998): Riemannian geometry. Springer, New York

[25] Li, M. (1998): The finite deformation of beam, plate and shell structures. II. The kinematic model and the Green-Lagrangian strains. Comput. Methods Appl. Mech. Engrg. **156**, 247–257

[26] Li, M. (1998): The finite deformation theory for beam, plate and shell. III. The three-dimensional beam theory and the FE formulation. Comput. Methods Appl. Mech. Engrg. **162**, 287–300

[27] Sansour, C. (1998): Large strain deformations of elastic shells constitutive modelling and finite element analysis. Comput. Methods Appl. Mech. Engrg. **161**, 1–18

[28] Betsch, P., Menzel, A., Stein, E. (1998): On the parametrization of finite rotations in computational mechanics. A classification of concepts with application to smooth shells. Comput. Methods Appl. Mech. Engrg. **155**, 273–305

[29] Bottasso, C.L., Borri, M. (1998): Integrating finite rotations. Comput. Methods Appl. Mech. Engrg. **164**, 307–331

[30] Crisfield, M.A, Jelenić, G. (1999): Objectivity of strain measures in the geometrically exact three-dimensional beam theory and its finite-element implementation. R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **455**, 1125–1147

[31] Ibrahimbegović, A., Al Mikdad, M. (2000): Quadratically convergent direct calculation of critical points for 3d structures undergoing finite rotations. Comput. Methods Appl. Mech. Engrg. **189**, 107–120

[32] Li, M., Zhan, F. (2000): The finite deformation theory for beam, plate and shell. IV. The Fe formulation of Mindlin plate and shell based on Green-Lagrangian strain. Comput. Methods Appl. Mech. Engrg. **182**, 187–203

[33] Li M., Zhan F. (2000): The finite deformation theory for beam, plate and shell. V. The shell element with drilling degree of freedom based on Biot strain. Comput. Methods Appl. Mech. Engrg. **189**, 743–759

[34] Romano, G. (2000): On the necessity of Korn's inequality. Symposium on Trends in Applications of Mathematics to Mechanics (STAMM 2000). Galway, Ireland, July 9-14, 2000

[35] Ibrahimbegović, A., Taylor, R.L. (2002): On the role of frame-invariance in structural mechanics models at finite rotations. Comput. Methods Appl. Mech. Engrg. **191**, 5159–5176

[36] Ibrahimbegović, A., Brank, B., Courtois, P. (2001): Stress resultant geometrically exact form of classical shell model and vector-like parameterization of constrained finite rotations. Internat. J. Numer. Methods Engrg. **52**, 1235–1252
[37] Vu-Quoc, L., Deng, H., Tan, X.G. (2001): Geometrically exact sandwich shells: the dynamic case. Comput. Methods Appl. Mech. Engrg. **190**, 2825–2873
[38] Ricci Maccarini, R., Saetta, A., Vitaliani, R. (2001): A non-linear finite element formulation for shells of arbitrary geometry. Comput. Methods Appl. Mech. Engrg. **190**, 4967–4986
[39] Mäkinen, J. (2001): Critical study of Newmark-scheme on manifold of finite rotations. Comput. Methods Appl. Mech. Engrg. **191**, 817–828
[40] Romano, G. (2002): Scienza delle costruzioni. Tomo I. Hevelius, Benevento
[41] Lee, W.J., Lee, B.C. (2001): An effective finite rotation formulation for geometrical non-linear shell structures. Comput. Mech. **27**, 360–368
[42] Bottasso, C.L., Borri, M., Trainelli, L. (2002): Geometric invariance. Comput. Mech. **29**, 163–169
[43] Romero, I., Armero, F. (2002): Numerical integration of the stiff dynamics of geometrically exact shells: an energy-dissipative momentum-conserving scheme. Internat. J. Numer. Methods Engrg. **54**, 1043–1086
[44] Romero, I., Armero, F. (2002): An objective finite element approximation of the kinematics of geometrically exact rods and its use in the formulation of an energy-momentum conserving scheme in dynamics. Internat. J. Numer. Methods Engrg. **54**, 1683–1716
[45] Betsch, P., Steinmann, P. (2002): Frame-indifferent beam finite elements based upon the geometrically exact beam theory. Internat. J. Numer. Methods Engrg. **54**, 1775–1788
[46] Romano, G. (2003): Scienza delle costruzioni. Tomo II. Hevelius, Benevento
[47] Kojic, M. (2002): An extension of 3-D procedure to large strain analysis of shells. Comput. Methods Appl. Mech. Engrg. **191**, 2447–2462
[48] Lee, Y., Park, K.C. (2002): Numerically generated tangent stiffness matrices for nonlinear structural analysis. Comput. Methods Appl. Mech. Engrg. **191**, 5833–5846
[49] Romano, G., Diaco, M., Romano, A. (2002): In: Tangent stiffness of Timoshenko beams undergoing large displacements. ISIMM Symposium. Maiori, Italia
[50] Romano, G., Diaco, M., Romano, A., Sellitto, C. (2003): When and why a nonsymmetric tangent stiffness may occur. 16th AIMETA Congress of Theoretical and Applied Mechanics. Ferrara, Italy, Sept. 9-12, 2003
[51] Romano, G., Romano, A., Sellitto, C. (2003): On the redundancy of 3D-Cosserat continuum. Preprint
[52] Kulikov, G.M., Plotnikova, S.V. (2003): Non-linear strain-displacement equations exactly representing large rigid-body motions. I. Timoshenko-Mindlin shell theory. Comput. Methods Appl. Mech. Engrg. **192**, 851–875
[53] Sansour, C., Wagner, W. (2003): Multiplicative updating of the rotation tensor in the finite element analysis of rods and shells – a path independent approach. Comput. Mech. **31**, 153–162
[54] Kapania, R.K., Li, J. (2003): On a geometrically exact curved/twisted beam theory under rigid cross-section assumption. Comput. Mech. **30**, 428–443
[55] Kapania, R.K., Li, J. (2003): A formulation and implementation of geometrically exact curved beam elements incorporating finite strains and finite rotations. Comput. Mech. **30**, 444–459
[56] Valente, R.A.F., Jorge, R.M.N., Cardoso, R.P.R., Cesarde Sa, J.M.A., Gracio, J.J.A. (2003): On the use of an enhanced transverse shear strain shell element for problems involving large rotations. Comput. Mech. **30**, 286–296

[57] Romano, G., Romano, A., Sellitto, C. (2003): On the physical plausibility of finite strains in polar shells. Preprint.

Basic issues in convex homogenization

G. Romano, A. Romano

Abstract. The basic results in homogenization theory are revisited in the abstract context of continuum mechanics in which the constitutive behaviour and the kinematic constraints are governed by pairs of conjugate convex potentials. The theory and the methods of this generalized elastic model are briefly recalled and applied to extend the classical linear theory of homogenization to the non-linear and possibly multivalued constitutive framework.

1 Prolegomena

The fundamentals of homogenization theory are here revisited with reference to an abstract structural model whose constitutive properties are characterized by monotone conservative multivalued laws governed by closed convex potentials.

The theory of such constitutive behavior, termed generalized elasticity, was developed by the first author and his co-workers in a number of papers (see [10,12]) and is illustrated in detail in [17].

The topic of nonlinear homogenization theory was investigated by Talbot, Willis and Toland in a series of papers [7–9]. Their approach was based on the theory of conjugate convex problems as developed in [5]. The present approach makes direct reference to an abstract structural problem and is carried out along the guidelines of the theory of generalized elasticity.

2 The continuum model

In continuum mechanics the fundamental theorems concerning the variational formulations of equilibrium and of tangent compatibility are founded on the property that the tangent kinematic operator has a closed range and a finite-dimensional kernel at every configuration in the admissible manifold.

The abstract framework is the following. Let V and D be the finite-dimensional linear spaces of local values of tangent (virtual) displacememts (also referred to as kinematic fields) and tangent strains respectively. Further, let S be the linear space of local values of stress fields, the dual space of D.

A continuous structural model, defined on a regular bounded connected domain Ω of an n-dimensional euclidean space E^n, is governed by a kinematic operator $\mathbf{B}$. This operator is the regular part of a distributional differential operator $\mathbb{B} : \mathcal{V}_\Omega \mapsto \mathbb{D}'_S$ of order m acting on Green-regular kinematic fields $\mathbf{u} \in \mathcal{V}_\Omega$ and ranging over the space of tangent strain distributions $\mathbb{B}\mathbf{u} \in \mathbb{D}'_S$ in Ω. Tangent strain distributions

are linear functionals, defined on the linear space $\mathbb{D}_S = \mathrm{C}_o^\infty(\Omega\,;\,S)$ of test stress fields which have compact support in Ω and which are continuous according to the uniform topology on compact subsets of Ω (see, e.g., [4,16]).

Piecewise Green-regular kinematic fields $\mathbf{u} \in \mathcal{V}_\Omega$ are square-integrable fields $\mathbf{u} \in H_{\mathrm{V}} = \mathcal{L}^2(\Omega\,;\,\mathrm{V})$ such that the corresponding distributional tangent strain fields $\mathbb{B}\mathbf{u} \in \mathbb{D}'_S$ are square-integrable on a finite subdivision $\mathcal{T}_{\mathbf{u}}(\Omega)$ of Ω (see [16,18,19]). The kinematic space $\mathcal{V}_\Omega$ is a pre-Hilbert space when endowed with the topology induced by the norm

$$\| \mathbf{u} \|^2_{\mathcal{V}_\Omega} = \| \mathbf{u} \|^2_{H_{\mathrm{V}}} + \| \mathbf{B}\mathbf{u} \|^2_{\mathcal{H}_D}\,,$$

where $\mathcal{H}_D = \mathcal{L}^2(\Omega\,;\,D)$ is the space of square-integrable tangent strain fields on Ω. The subdivision $\mathcal{T}_{\mathbf{u}}(\Omega)$ is said to be a support of regularity of the kinematic field $\mathbf{u} \in \mathcal{V}_\Omega$.

The kinematic constraints on a continuum are imposed by a sequence of two requirements. The first is a regularity requirement on the tangent displacements and is expresseed by considering a basic finite subdivision $\mathcal{T}(\Omega)$ of Ω and by requiring that the tangent displacements have $\mathcal{T}(\Omega)$ as a support of regularity. The closed linear subspace $\mathcal{V}(\mathcal{T}(\Omega)) \subset \mathcal{V}_\Omega$ of $\mathcal{T}(\Omega)$-regular tangent displacements is a Hilbert space for the topology of $\mathcal{V}_\Omega$.

The second requirement is that tangent displacements belong to a conformity subspace, a closed linear subspace $\mathcal{L} = \mathcal{L}(\mathcal{T}(\Omega)) \subset \mathcal{V}(\mathcal{T}(\Omega))$.

In boundary value problems the Hilbert space $\mathcal{L}$ is the kernel of a bounded linear operator which prescribes an additional linear constraint on the boundary values of the tangent displacements $\mathbf{u} \in \mathcal{V}(\mathcal{T}(\Omega))$.

The operator $\mathbf{B}_{\mathcal{L}} \in BL(\mathcal{L}\,;\,\mathcal{H}_D)$, which yields the regular tangent strain field $\mathbf{B}\mathbf{u} \in \mathcal{H}_D$ corresponding to a conforming tangent displacement $\mathbf{u} \in \mathcal{L}$, is linear and continuous.

The tangent kinematic operator $\mathbf{B} \in BL(\mathcal{V}_\Omega\,;\,\mathcal{H}_D)$ is assumed to be Korn-regular in the sense that, for any conformity subspace $\mathcal{L} \subset \mathcal{V}_\Omega$, the following conditions [13,14] are met:

$$\begin{cases} \dim\,\mathrm{Ker}\,\mathbf{B}_{\mathcal{L}} = \dim\,(\,\mathrm{Ker}\,\mathbf{B} \cap \mathcal{L}) < +\infty\,, \\ \| \mathbf{B}\mathbf{u} \|_{\mathcal{H}} \geq c_{\mathbf{B}}\, \| \mathbf{u} \|_{\mathcal{L}/\,\mathrm{Ker}\,\mathbf{B}_{\mathcal{L}}}\,, \quad \forall\,\mathbf{u} \in \mathcal{L} \quad \Longleftrightarrow \quad \mathrm{Im}\,\mathbf{B}_{\mathcal{L}} \text{ closed in } \mathcal{H}_D\,. \end{cases}$$

The requirement that these properties hold for any conformity subspace $\mathcal{L} \subset \mathcal{V}_\Omega$ is motivated by the requirement that, in engineering structural models, the equilibrium condition can be imposed by a finite number of scalar equations and that the existence results hold for any choice of linear kinematic constraints. The Korn-regularity of $\mathbf{B} \in BL(\mathcal{V}_\Omega\,;\,\mathcal{H}_D)$ is the basic tool for the proof of the theorem of virtual powers [16] which ensures the existence of a stress field $\boldsymbol{\sigma} \in \mathcal{H}_S = \mathcal{L}^2(\Omega\,;\,S)$ in equilibrium with an equilibrated system of active forces, i.e., bounded linear functionals $\mathbf{f} \in \mathcal{L}'$ such that $\langle\,\mathbf{f}\,,\,\mathbf{v}\,\rangle = 0$ for all $\mathbf{v} \in \mathrm{Ker}\,\mathbf{B} \cap \mathcal{L}$. It can be shown [14] that a necessary and sufficient condition for the Korn-regularity of $\mathbf{B} \in BL(\mathcal{V}_\Omega\,;\,\mathcal{H}_D)$ is the validity of an inequality of Korn's type,

$$\| \mathbf{B}\mathbf{u} \|_{\mathcal{H}_D} + \| \mathbf{u} \|_H \geq \alpha\, \| \mathbf{u} \|_m \quad \forall\,\mathbf{u} \in H^m(\Omega\,;\,\mathrm{V})\,,$$

where $H^m(\Omega\,;\mathrm{V})$ is the Sobolev space of tangent displacements which are square-integrable on Ω and which have distributional derivative up to the order m. The formal adjoint of $\mathbf{B} \in BL(\mathcal{V}_\Omega\,;\,\mathcal{H}_D)$ is the distributional differential operator $\mathbb{B}'_o : \mathcal{H}_S \mapsto \mathbb{D}'_{\mathrm{V}}$ of order m defined by the identity

$$\langle\, \mathbb{B}'_o\boldsymbol{\sigma}\,,\, \mathbf{v}\,\rangle := (\!(\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{v}\,)\!) \qquad \forall\, \mathbf{v} \in \mathbb{D}'_{\mathrm{V}}\,, \quad \forall\, \boldsymbol{\sigma} \in \mathcal{H}(\Omega)\,.$$

The space $\mathcal{S}_\Omega$ of piecewise Green-regular stress fields on Ω is then defined as the linear space of stress fields $\boldsymbol{\sigma} \in \mathcal{H}_S$ such that the corresponding body force distributions $\mathbb{B}'_o\boldsymbol{\sigma} \in \mathbb{D}'_{\mathrm{V}}$, are square-integrable on a finite subdivision $\mathcal{T}_{\boldsymbol{\sigma}}(\Omega)$ of Ω (see [16,18]). The space $\mathcal{S}_\Omega$ is a pre-Hilbert space when endowed with the induced norm

$$\|\,\boldsymbol{\sigma}\,\|^2_{\mathcal{S}_\Omega} = \|\,\boldsymbol{\sigma}\,\|^2_{\mathcal{H}_S} + \|\,\mathbf{B}'_o\boldsymbol{\sigma}\,\|^2_{H_F}\,,$$

where $\mathbf{B}'_o \in BL(\mathcal{S}_\Omega\,;\,\mathcal{H}_S)$ is the regular part of the distributional differential operator $\mathbb{B}'_o : \mathcal{H}_S \mapsto \mathbb{D}'_{\mathrm{V}}$. Any pair of Green-regular tangent displacement fields $\mathbf{v} \in \mathcal{V}_\Omega$ and Green-regular stress fields $\boldsymbol{\sigma} \in \mathcal{S}_\Omega$ satisfies Green's formula [15] for the operator $\mathbf{B} \in BL(\mathcal{V}_\Omega, \mathcal{H}_D)$,

$$(\!(\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{v}\,)\!) = (\,\mathbf{B}'_o\boldsymbol{\sigma}\,,\, \mathbf{v}\,) + \langle\!\langle\, \mathbf{N}\boldsymbol{\sigma}\,,\, \boldsymbol{\Gamma}\mathbf{v}\,\rangle\!\rangle \quad \forall\, \mathbf{v} \in \mathcal{V}_\Omega, \quad \forall\, \boldsymbol{\sigma} \in \mathcal{S}_\Omega,$$

where by definition

$$(\!(\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{v}\,)\!) := \int_\Omega \boldsymbol{\sigma} : \mathbf{B}\mathbf{v}\; \mathtt{d}\mu\,, \qquad (\,\mathbf{B}'_o\boldsymbol{\sigma}\,,\, \mathbf{v}\,) := \int_\Omega \mathbf{B}'_o\boldsymbol{\sigma}\cdot\mathbf{v}\; \mathtt{d}\mu\,,$$

and the duality pairing $\langle\!\langle\, \mathbf{N}\boldsymbol{\sigma}\,,\, \boldsymbol{\Gamma}\mathbf{v}\,\rangle\!\rangle$ is the extension by continuity of a sum of boundary integrals over $\partial\mathcal{T}_{\mathbf{v}\boldsymbol{\sigma}}(\Omega) = \cup\,\partial\Omega_e,\ e = 1, \ldots, n_{\mathrm{elements}}$,

$$\int_{\partial\mathcal{T}_{\mathbf{v}\boldsymbol{\sigma}}(\Omega)} \mathbf{N}\boldsymbol{\sigma}\cdot\boldsymbol{\Gamma}\mathbf{v}\; \mathtt{d}\,.$$

where $\mathcal{T}_{\mathbf{v}\boldsymbol{\sigma}}(\Omega) = \mathcal{T}_{\mathbf{v}}(\Omega) \vee \mathcal{T}_{\boldsymbol{\sigma}}(\Omega)$ is finer than $\mathcal{T}_{\mathbf{v}}(\Omega)$ and $\mathcal{T}_{\boldsymbol{\sigma}}(\Omega)$.

The trace $\boldsymbol{\Gamma}$ and the flux $\mathbf{N}$ are differential operators, with order ranging between 0 and $m-1$, associated to the operator $\mathbf{B}$ and defined by m subsequent applications of the rule of integration by parts.

2.1 Averaging operators

Let $\mathbf{M}_\Omega \in BL(\mathcal{H}_D\,;\,D)$ and $\textsc{med} \in BL(\mathcal{H}_D\,;\,D)$ be the surjective averaging operators defined by

$$\mathbf{M}_\Omega(\boldsymbol{\varepsilon}) := \int_\Omega \boldsymbol{\varepsilon}(\mathbf{x})\; \mathtt{d}\mu\,, \qquad \textsc{med}_\Omega = \frac{1}{\mathrm{vol}\,(\Omega)}\; \mathbf{M}_\Omega\,.$$

The dual operator $\mathbf{M}_\Omega^* \in BL(S\,;\,\mathcal{H}_S)$ of $\mathbf{M}_\Omega \in BL(\mathcal{H}_D\,;\,D)$ is defined by the identity

$$\langle\, \mathbf{M}_\Omega^*(\mathbf{T})\,,\, \boldsymbol{\varepsilon}\,\rangle = \langle\, \mathbf{T}\,,\, \mathbf{M}_\Omega(\boldsymbol{\varepsilon})\,\rangle \quad \forall\, \mathbf{T} \in S, \quad \forall \boldsymbol{\varepsilon} \in \mathcal{H}_D\,.$$

When applied to $\mathbf{T} \in S$ the operator $\mathbf{M}_\Omega^* \in BL(S\,;\,\mathcal{H}_S)$ provides the constant field in $\mathcal{H}_S = \mathcal{L}^2(\Omega\,;\,S)$ given by

$$(\mathbf{M}_\Omega^*(\mathbf{T}))(\mathbf{x}) = \mathbf{T} \quad \forall\, \mathbf{x} \in \Omega\,.$$

Note that the roles of the spaces D and S may be interchanged in the preceeding definitions. We remark that $\text{MED}_\Omega \in BL(\mathcal{H}_S\,;\,S)$ is a right inverse of $\mathbf{M}_\Omega^* \in BL(S\,;\,\mathcal{H}_S)$ since

$$(\text{MED}_\Omega \circ \mathbf{M}_\Omega^*)(\mathbf{T}) = \mathbf{T} \quad \forall\, \mathbf{T} \in S\,.$$

The surjectivity of $\mathbf{M}_\Omega \in BL(\mathcal{H}_D\,;\,D)$ yields

$$\operatorname{Im} \mathbf{M}_\Omega^* = (\operatorname{Ker} \mathbf{M}_\Omega)^\perp\,,$$

which implies that a square-integrable field, orthogonal to any square-integrable field with vanishing mean value, is constant. Trivially we also have that

$$\operatorname{Ker} \mathbf{M}_\Omega^* = (\operatorname{Im} \mathbf{M}_\Omega)^\perp = \{\,0\,\}\,.$$

To simplify the notation we also denote by the same symbols $\mathbf{M}_\Omega$ and $\mathbf{M}_\Omega^*$ the operators $\mathbf{M}_\Omega \in BL(\mathcal{L}^1(\Omega\,;\,\mathcal{R})\,;\,\mathcal{R})$, $\mathbf{M}_\Omega^* \in BL(\mathcal{R}\,;\,\mathcal{L}^\infty(\Omega\,;\,\mathcal{R}))$, where $\mathcal{L}^1(\Omega\,;\,\mathcal{R})$ is the space of real-valued integrable functions on Ω and $\mathcal{L}^\infty(\Omega\,;\,\mathcal{R})$ is the dual space of essentially bounded functions on Ω.

2.2 Conjugate convex potentials

A structural model is defined by considering a subdivision $\mathcal{T}(\Omega)$ of the domain Ω and the associated Hilbert space $\mathcal{V} = \mathcal{V}(\mathcal{T}(\Omega), \mathrm{V})$ of $\mathcal{T}(\Omega)$-regular displacements defined as those giving rise to distributional tangent strain fields which are square-integrable in each element of the subdivision. Force systems are the bounded linear functionals of the dual space $\mathcal{F} = BL(\mathcal{V}(\mathcal{T}(\Omega), \mathrm{V})\,;\,\mathcal{R})$.

The model is further characterized by a bounded linear tangent kinematic operator $\mathbf{B} \in BL(\mathcal{V}\,;\,\mathcal{H})$ which provides the tangent strain field corresponding to any $\mathcal{T}(\Omega)$-regular tangent displacement field. The operator $\mathbf{B} \in BL(\mathcal{V}\,;\,\mathcal{H})$ is assumed to satisfy an inequality of Korn type so that the kernel $\operatorname{Ker} \mathbf{B} \subset \mathcal{V}$ is finite-dimensional and, for any set of linear constraints defining a closed linear subspace $\mathcal{L} \subset \mathcal{V}$ of conforming displacements, the image $\mathbf{B}\mathcal{L}$ is closed in $\mathcal{H}$. The dual equilibrium operator $\mathbf{B}' \in BL(\mathcal{H}\,;\,\mathcal{F})$ is defined by the identity

$$\langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{v}\,\rangle = \langle\, \mathbf{B}'\boldsymbol{\sigma}\,,\, \mathbf{v}\,\rangle \qquad \forall\, \mathbf{v} \in \mathcal{V}\,, \quad \forall\, \boldsymbol{\sigma} \in \mathcal{H}_S = \mathcal{L}^2(\Omega\,;\,S)\,.$$

The constitutive properties of the elastic material are described, according to Green, by a field of local potentials $\varphi_{\mathbf{e}} : D \times \Omega \mapsto \overline{\mathcal{R}}$, where $\overline{\mathcal{R}} = \mathcal{R} \cup \{+\infty\}$ is the upper-extended real line [15]. We consider a generalized Green elasticity in which,

at each $\mathbf{x} \in \Omega$, the local potential is assumed to be proper, convex and everywhere subdifferentiable on its domain $\operatorname{dom} \varphi_{\mathbf{e}}(\cdot, \mathbf{x}) \subset D$. Convex analysis provides the mathematical tools to deal with such problems [3,5,6,12]. In this context a potential theory for monotone conservative multivalued operators was developed by the first author and his coworkers [10,17].

The convex global constitutive potential $\Phi_{\mathbf{e}} : \mathcal{H} \mapsto \overline{\mathcal{R}}$ is a function of the (small) strain fields $\boldsymbol{\varepsilon} \in \mathcal{H}$ and is defined by the integral

$$\Phi_{\mathbf{e}}(\boldsymbol{\varepsilon}) := \int_{\Omega} (\Phi_{\mathbf{e}}(\boldsymbol{\varepsilon}))(\mathbf{x}) \, \mathtt{d}\mu \,,$$

where the potential $\Phi_{\mathbf{e}} : \mathcal{H} \mapsto \mathcal{L}^2(\Omega\,; \overline{\mathcal{R}})$ is given by

$$(\Phi_{\mathbf{e}}(\boldsymbol{\varepsilon}))(\mathbf{x}) := \varphi_{\mathbf{e}}(\boldsymbol{\varepsilon}(\mathbf{x}), \mathbf{x}) \,,$$

and $\mathtt{d}\mu$ is the volume form on Ω.

We consider a general nondecreasing monotone and conservative stress-strain relation $\mathcal{G} \subset \mathcal{H}_S \times \mathcal{H}_D$. Monotonicity means that

$$\langle\, \boldsymbol{\sigma}_2 - \boldsymbol{\sigma}_1 \,,\, \boldsymbol{\varepsilon}_2 - \boldsymbol{\varepsilon}_1 \,\rangle \geq 0 \qquad \forall \{\boldsymbol{\sigma}_1 \,, \boldsymbol{\varepsilon}_1\} \in \mathcal{G}, \quad \forall \{\boldsymbol{\sigma}_2 \,, \boldsymbol{\varepsilon}_2\} \in \mathcal{G} \,,$$

and conservativity means that

$$\oint_{\Pi_{\boldsymbol{\varepsilon}}} \langle\, \mathcal{E}(\boldsymbol{\varepsilon}) \,,\, d\boldsymbol{\varepsilon} \,\rangle = 0 \quad \Longleftrightarrow \quad \oint_{\Pi_{\boldsymbol{\sigma}}} \langle\, \mathcal{E}^{-1}(\boldsymbol{\sigma}) \,,\, d\boldsymbol{\sigma} \,\rangle = 0 \,,$$

where $\Pi_{\boldsymbol{\varepsilon}} \subset \mathcal{H}_S$, $\Pi_{\boldsymbol{\sigma}} \subset \mathcal{H}_D$ are closed polylines and $\mathcal{E} : \mathcal{H}_D \mapsto \mathcal{H}_S$, $\mathcal{E}^{-1} : \mathcal{H}_S \mapsto \mathcal{H}_D$ are the left and right multivalued maps associated with the relation $\mathcal{G}$ and defined by

$$\mathcal{E}(\boldsymbol{\varepsilon}) := \{\, \boldsymbol{\sigma} \in \mathcal{L}^2(\Omega\,; S) \mid \{\boldsymbol{\sigma} \,, \boldsymbol{\varepsilon}\} \in \mathcal{G} \,\} \,,$$
$$\mathcal{E}^{-1}(\boldsymbol{\sigma}) := \{\, \boldsymbol{\varepsilon} \in \mathcal{L}^2(\Omega\,; D) \mid \{\boldsymbol{\sigma} \,, \boldsymbol{\varepsilon}\} \in \mathcal{G} \,\} \,.$$

The domains $\operatorname{dom} \mathcal{E} \in \mathcal{L}^2(\Omega\,; D)$, $\operatorname{dom} \mathcal{E}^{-1} \in \mathcal{L}^2(\Omega\,; S)$, the loci where the images $\mathcal{E}(\boldsymbol{\varepsilon})$ and $\mathcal{E}^{-1}(\boldsymbol{\sigma})$ are non-empty, are assumed to be convex sets.

It can be shown that the integrals along segments are independent of the special representative in the sets $\mathcal{E}(\boldsymbol{\varepsilon})$ and $\mathcal{E}^{-1}(\boldsymbol{\sigma})$ chosen to evaluate the integrands [10]. A multivalued monotone and conservative relation is governed by a pair of conjugate convex potentials $\Phi_{\mathbf{e}} : \mathcal{H}_D \mapsto \overline{\mathcal{R}}$ and $\Phi^*_{\mathbf{e}} : \mathcal{H}_S \mapsto \overline{\mathcal{R}}$ related by the involutive relation

$$\Phi^*_{\mathbf{e}}(\boldsymbol{\sigma}) := \sup_{\boldsymbol{\varepsilon} \in \mathcal{H}_D} \{\, \langle\, \boldsymbol{\sigma} \,,\, \boldsymbol{\varepsilon} \,\rangle - \Phi(\boldsymbol{\varepsilon}) \,\} \,,$$
$$\Phi_{\mathbf{e}}(\boldsymbol{\varepsilon}) := \sup_{\boldsymbol{\sigma} \in \mathcal{H}_S} \{\, \langle\, \boldsymbol{\sigma} \,,\, \boldsymbol{\varepsilon} \,\rangle - \Phi^*(\boldsymbol{\sigma}) \,\} \,.$$

The conjugate convex potentials $\Phi_{\mathbf{e}} : \mathcal{H}_D \mapsto \overline{\mathcal{R}}$ and $\Phi^*_{\mathbf{e}} : \mathcal{H}_S \mapsto \overline{\mathcal{R}}$ can be evaluated by direct integration of the multivalued maps along a segment or by the conjugacy relations above.

The effective domains $\mathrm{dom}\,\boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\varepsilon}) \subset \mathcal{H}_D$ and $\mathrm{dom}\,\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}) \subset \mathcal{H}_S$ are the convex sets where the potentials $\boldsymbol{\Phi}_{\mathbf{e}} : \mathcal{H}_D \mapsto \overline{\mathcal{R}}$ and $\boldsymbol{\Phi}^*_{\mathbf{e}} : \mathcal{H}_S \mapsto \overline{\mathcal{R}}$ assume finite values in $\mathcal{R}$. The convex potentials $\boldsymbol{\Phi}_{\mathbf{e}} : \mathcal{H}_D \mapsto \overline{\mathcal{R}}$ and $\boldsymbol{\Phi}^*_{\mathbf{e}} : \mathcal{H}_S \mapsto \overline{\mathcal{R}}$ are subdifferentiable in their domains. The subdifferentials are the convex sets defined by [3,6,11],

$$\partial\boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\varepsilon}) := \{\, \boldsymbol{\sigma} \in \mathcal{H}_S \mid \boldsymbol{\Phi}_{\mathbf{e}}(\overline{\boldsymbol{\varepsilon}}) - \boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\varepsilon}) \geq \langle\, \boldsymbol{\sigma}\,,\, \overline{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}\,\rangle \,\},$$
$$\partial\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}) := \{\, \boldsymbol{\varepsilon} \in \mathcal{H}_D \mid \boldsymbol{\Phi}^*_{\mathbf{e}}(\overline{\boldsymbol{\sigma}}) - \boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}) \geq \langle\, \overline{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\,,\, \boldsymbol{\varepsilon}\,\rangle \,\}.$$

The global generalized elastic law is expressed by the subdifferential maps

$$\boldsymbol{\sigma} \in \partial\boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\varepsilon}), \quad \boldsymbol{\varepsilon} \in \partial\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}).$$

By definition we have that

$$\boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\varepsilon}) + \boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}) \geq \langle\, \boldsymbol{\sigma}\,,\, \boldsymbol{\varepsilon}\,\rangle \qquad \forall\, \boldsymbol{\varepsilon} \in \mathcal{H}_D \quad \forall\, \boldsymbol{\sigma} \in \mathcal{H}_S,$$
$$\boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\varepsilon}) + \boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}) = \langle\, \boldsymbol{\sigma}\,,\, \boldsymbol{\varepsilon}\,\rangle \iff \boldsymbol{\sigma} \in \partial\boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\varepsilon}) \iff \boldsymbol{\varepsilon} \in \partial\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}).$$

Recalling that the elastic law is pointwise defined, we observe that the convex conjugate $\varphi^*_{\mathbf{e}} : S \times \Omega \mapsto \overline{\mathcal{R}}$ of the local potential $\varphi_{\mathbf{e}} : D \times \Omega \mapsto \overline{\mathcal{R}}$ is given by

$$\varphi^*_{\mathbf{e}}(\mathbf{T}, \mathbf{x}) := \sup_{\mathbf{D} \in D} \{\, \langle\, \mathbf{T}\,,\, \mathbf{D}\,\rangle - \varphi_{\mathbf{e}}(\mathbf{D}, \mathbf{x}) \,\}.$$

The global convex potential $\boldsymbol{\Phi}^*_{\mathbf{e}} : \mathcal{H}_S \mapsto \overline{\mathcal{R}}$, convex conjugate to $\boldsymbol{\Phi}_{\mathbf{e}} : \mathcal{H}_D \mapsto \overline{\mathcal{R}}$, can then be evaluated by each one of the following procedures [17]:

$$\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}) := \sup_{\boldsymbol{\eta} \in \mathcal{H}_D} \{\, \langle\, \boldsymbol{\sigma}\,,\, \boldsymbol{\eta}\,\rangle - \boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\eta}) \,\},$$

$$\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}) := \int_{\Omega} (\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}))(\mathbf{x})\, \mathtt{d}\mu.$$

In an analogous way, kinematic constraints are described by a conservative multivalued monotone nonincreasing relation $\mathcal{G} \subset \mathcal{F} \times \mathcal{V}$ and by the pair of conjugate proper superdifferentiable concave functional $J : \mathcal{V} \mapsto \underline{\mathcal{R}}$ and $J^* : \mathcal{F} \mapsto \underline{\mathcal{R}}$ where $\underline{\mathcal{R}} := \mathcal{R} \cup \{-\infty\}$ is the lower-extended real line [17].

We remark that kinematic constraint conditions are global in character and accordingly $J : \mathcal{V} \mapsto \underline{\mathcal{R}}$ and $J^* : \mathcal{F} \mapsto \underline{\mathcal{R}}$ are global functionals which may not be defined as integrals of local functionals.

The constraint map is nonincreasing since it provides the relation between the displacement fields of the constraint and the force systems that the constraint applies to the structure, that is, the opposite of the force systems of the structure on the constraint. This change in sign turns the monotone nondecreasing constitutive map into a nonincreasing one.

Multivaluedness of the constraint relations is the rule rather then the exception: the simplest linear frictionless bilateral kinematic constraint relation is described by

multivalued maps. If $\mathscr{L}$ is the subspace of conforming virtual displacements the constraint relation is

$$\mathscr{G} := \{\, \{\mathbf{r}\,,\mathbf{v}\} \in \mathscr{F} \times \mathscr{V} \mid \mathbf{v} \in \mathscr{L}\,, \quad \mathbf{r} \in \mathscr{L}^{\perp}\,\}\,.$$

Both the left and right maps are constant,

$$\mathscr{M}(\mathbf{v}) := \mathscr{L}^{\perp}\,, \quad \mathscr{M}^{-1}(\mathbf{r}) := \mathscr{L}\,.$$

In general reactive force systems are conjugate to the displacements with respect to the concave functional $J : \mathscr{V} \mapsto \underline{\mathscr{R}}$, namely,

$$\mathbf{r} \in \partial J(\mathbf{u}) \quad \Longleftrightarrow \quad J(\mathbf{v}) - J(\mathbf{u}) \le \langle\, \mathbf{r}\,,\, \mathbf{v}-\mathbf{u}\,\rangle \qquad \forall\, \mathbf{u} \in \mathscr{V}\,.$$

The inverse multivalued law is expressed by the condition

$$\mathbf{u} \in \partial J^*(\mathbf{r}) \quad \Longleftrightarrow \quad J^*(\overline{\mathbf{r}}) - J^*(\mathbf{r}) \le \langle\, \overline{\mathbf{r}}-\mathbf{r}\,,\, \mathbf{u}\,\rangle \qquad \forall\, \overline{\mathbf{r}} \in \mathscr{F}\,.$$

By definition,

$$\begin{aligned} &J(\mathbf{v}) + J^*(\mathbf{r}) \le \langle\, \mathbf{r}\,,\, \mathbf{v}\,\rangle \qquad \forall\, \mathbf{v} \in \mathscr{V} \quad \forall\, \mathbf{r} \in \mathscr{F}\,, \\ &J(\mathbf{u}) + J^*(\mathbf{r}) = \langle\, \mathbf{r}\,,\, \mathbf{u}\,\rangle \quad \Longleftrightarrow \quad \mathbf{r} \in \partial J(\mathbf{u}) \quad \Longleftrightarrow \quad \mathbf{u} \in \partial J^*(\mathbf{r})\,. \end{aligned}$$

The concave conjugate potentials $J : \mathscr{V} \mapsto \underline{\mathscr{R}}$ and $J^* : \mathscr{F} \mapsto \underline{\mathscr{R}}$ are related by

$$\begin{aligned} J^*(\mathbf{r}) &:= \inf_{\mathbf{u}\in\mathscr{V}} \{\, \langle\, \mathbf{r}\,,\, \mathbf{u}\,\rangle - J(\mathbf{u})\,\}\,, \\ J(\mathbf{u}) &:= \inf_{\mathbf{r}\in\mathscr{F}} \{\, \langle\, \mathbf{r}\,,\, \mathbf{u}\,\rangle - J^*(\mathbf{r})\,\}\,. \end{aligned}$$

2.3 Variational formulations

We consider a convex structural problem governed by a kinematic operator $\mathbf{B} \in BL(\mathscr{V}\,;\,\mathscr{H})$ under the constitutive law defined by a convex potential $\Phi : \mathscr{H} \mapsto \overline{\mathscr{R}}$ and the constraint condition defined by a concave potential $J : \mathscr{V} \mapsto \underline{\mathscr{R}}$, according to the rules

$$\begin{cases} \mathbf{B}\mathbf{u} = \boldsymbol{\varepsilon}\,, \\ \mathbf{B}'\boldsymbol{\sigma} = \mathbf{f}\,, \end{cases} \qquad \begin{cases} \boldsymbol{\sigma} \in \partial\Phi(\boldsymbol{\varepsilon})\,, \\ \mathbf{f} \in \partial J(\mathbf{u})\,, \end{cases}$$

which respectively impose the kinematic compatibility, the equilibrium, the global stress-strain law and the force-displacement law.

The stress-strain law is multivalued and monotone nondecreasing while the force-displacement law is multivalued and monotone nonincreasing.

Recalling the duality between the equilibrium operator $\mathbf{B}' \in BL(\mathscr{H}_S\,;\,\mathscr{F})$ and the kinematic operator $\mathbf{B} \in BL(\mathscr{V}\,;\,\mathscr{H}_D)$,

$$\langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{v}\,\rangle = \langle\, \mathbf{B}'\boldsymbol{\sigma}\,,\, \mathbf{v}\,\rangle \qquad \forall\, \mathbf{u} \in \mathscr{V}\,, \quad \forall\, \boldsymbol{\sigma} \in \mathscr{H}_S = \mathscr{L}^2(\Omega\,;\,S)\,,$$

we see that the equilibrium condition $\mathbf{B}'\boldsymbol{\sigma} = \mathbf{f}$ may be rewritten in variational terms by the virtual work principle

$$\langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{v}\,\rangle = \langle\, \mathbf{f}\,,\, \mathbf{v}\,\rangle \qquad \forall\, \mathbf{v} \in \mathcal{V}\,,$$

or, explicitly,

$$\int_{\Omega} \langle\, \boldsymbol{\sigma}(\mathbf{x})\,,\, (\mathbf{B}\mathbf{v})(\mathbf{x})\,\rangle\, \mathrm{d}\mu = \langle\, \mathbf{f}\,,\, \mathbf{v}\,\rangle \qquad \forall\, \mathbf{v} \in \mathcal{V}\,.$$

The convex structural problem defined above can be associated with a family of ten basic functionals whose stationary points are the solutions of the structural problem [17]. By introducing the product Hilbert spaces

$$\mathcal{H} = \mathcal{V} \times \mathcal{H}_S \times \mathcal{H}_D \times \mathcal{F}\,,$$
$$\mathcal{H}' = \mathcal{F} \times \mathcal{H}_D \times \mathcal{H}_S \times \mathcal{V}\,,$$

the operator $\mathbf{A} : \mathcal{H} \mapsto \mathcal{H}'$ governing the structural problem is given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{O} & \mathbf{B}' & \mathbf{O} & -\mathbf{I}_{\mathcal{F}} \\ \mathbf{B} & \mathbf{O} & -\mathbf{I}_{\mathcal{D}} & \mathbf{O} \\ \mathbf{O} & -\mathbf{I}_{\mathcal{S}} & \partial\boldsymbol{\Phi} & \mathbf{O} \\ -\mathbf{I}_{\mathcal{V}} & \mathbf{O} & \mathbf{O} & \partial J^* \end{bmatrix}.$$

The operator $\mathbf{A} : \mathcal{H} \mapsto \mathcal{H}'$ is clearly self-adjoint and, hence, by integrating along a ray in $\mathcal{H}$, we get the potential

$$\mathcal{L}(\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \mathbf{u}, \mathbf{f}) = \boldsymbol{\Phi}(\boldsymbol{\varepsilon}) + J^*(\mathbf{f}) + \langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{u}\,\rangle - \langle\, \boldsymbol{\sigma}\,,\, \boldsymbol{\varepsilon}\,\rangle - \langle\, \mathbf{f}\,,\, \mathbf{u}\,\rangle\,,$$

which is convex in $\boldsymbol{\varepsilon}$, concave in $\mathbf{f}$ and linear in $\mathbf{u}$ and $\boldsymbol{\sigma}$.

A solution $\{\,\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \mathbf{u}, f\,\}$ is then a minimum point with respect to $\boldsymbol{\varepsilon}$, a maximum point with respect to f and a stationary point with respect to $\mathbf{u}$ and $\boldsymbol{\sigma}$. A progressive elimination of state variables, based on the conjugacy relations, leads to a family of potentials according to the tree-shaped scheme:

$$\{\,\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \mathbf{u}, \mathbf{f}\,\}$$

$$\{\,\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \mathbf{u}\,\} \qquad\qquad \{\,\boldsymbol{\sigma}, \mathbf{u}, \mathbf{f}\,\}$$

$$\{\,\boldsymbol{\varepsilon}, \boldsymbol{\sigma}\,\} \qquad\qquad \{\,\boldsymbol{\sigma}, \mathbf{u}\,\} \qquad\qquad \{\,\mathbf{u}, \mathbf{f}\,\}$$

$$\{\,\boldsymbol{\varepsilon}\,\} \qquad\qquad \{\,\boldsymbol{\sigma}\,\} \qquad\qquad \{\,\mathbf{u}\,\} \qquad\qquad \{\,\mathbf{f}\,\}.$$

The family is composed of the ten basic functionals:

$L(\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \mathbf{u}, \mathbf{f}) = \boldsymbol{\Phi}(\boldsymbol{\varepsilon}) + J^*(\mathbf{f}) + \langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{u} - \boldsymbol{\varepsilon}\,\rangle - \langle\, \mathbf{f}\,,\, \mathbf{u}\,\rangle,$

$$
\begin{aligned}
H_1(\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \mathbf{u}) &= \Phi(\boldsymbol{\varepsilon}) - J(\mathbf{u}) + \langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{u} - \boldsymbol{\varepsilon}\,\rangle,\\
H_2(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{f}) &= -\Phi^*(\boldsymbol{\sigma}) + J^*(\mathbf{f}) + \langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{u}\,\rangle - \langle\, \mathbf{f}\,,\, \mathbf{u}\,\rangle,\\
R_1(\boldsymbol{\varepsilon}, \boldsymbol{\sigma}) &= \Phi(\boldsymbol{\varepsilon}) + J^*(\mathbf{B}'\boldsymbol{\sigma}) - \langle\, \boldsymbol{\sigma}\,,\, \boldsymbol{\varepsilon}\,\rangle,\\
R_2(\mathbf{u}, \boldsymbol{\sigma}) &= -\Phi^*(\boldsymbol{\sigma}) - J(\mathbf{u}) + \langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{u}\,\rangle,\\
R_3(\mathbf{u}, \mathbf{f}) &= \Phi(\mathbf{B}\mathbf{u}) + J^*(\mathbf{f}) - \langle\, \mathbf{f}\,,\, \mathbf{u}\,\rangle,\\
P(\boldsymbol{\varepsilon}) &= \Phi(\boldsymbol{\varepsilon}) - (J^* \circ \mathbf{B}')^*(\boldsymbol{\varepsilon}),\\
G(\boldsymbol{\sigma}) &= -\Phi^*(\boldsymbol{\sigma}) + J^*(\mathbf{B}'\boldsymbol{\sigma}),\\
F(\mathbf{u}) &= \Phi(\mathbf{B}\mathbf{u}) - J(\mathbf{u}),\\
\mathrm{Q}(\mathbf{f}) &= -(\Phi \circ \mathbf{B})^*(\mathbf{f}) + J^*(\mathbf{f}).
\end{aligned}
$$

All ten functionals of the family have the same value at a solution.

Assuming that the solution $\{\mathbf{u}\,,\boldsymbol{\sigma}\} \in \mathcal{V} \times \mathcal{H}_S$ of the structural problem is unique, we can determine as the minimum point of the extremum problem

$$F(\mathbf{u}) = \min_{\mathbf{v}\in\mathcal{V}} F(\mathbf{v}) = \min_{\mathbf{v}\in\mathcal{V}}\{\,\Phi(\mathbf{B}\mathbf{v}) - J(\mathbf{v})\,\}\,,$$

or as the maximum point of the extremum problem

$$G(\boldsymbol{\sigma}) = \max_{\mathbf{s}\in\mathcal{H}_S} G(\mathbf{s}) = \max_{\mathbf{s}\in\mathcal{H}_S}\{\,J^*(\mathbf{B}'\mathbf{s}) - \Phi^*(\mathbf{s})\,\}\,.$$

Moreover, at the solution,

$$\max_{\mathbf{s}\in\mathcal{H}_S} G(\mathbf{s}) = G(\boldsymbol{\sigma}) = F(\mathbf{u}) = \min_{\mathbf{v}\in\mathcal{V}} F(\mathbf{v})\,.$$

This relation provides a basis for bounding techniques which are applied in the sequel to the effective response of homogenized media.

3 Periodic homogenization

Let $\mathscr{C}$ be a periodicity cell (a parallelepiped in E^n) and $\mathbf{u}_\sharp \in \mathscr{L}^2(E^n\,;\mathrm{V})$ the $\mathscr{C}$-periodic extension of a vector field $\mathbf{u} \in H_\mathrm{V} = \mathscr{L}^2(\mathscr{C}\,;\mathrm{V})$, defined by

$$\mathbf{u}_\sharp(\mathbf{x} + k\,\mathbf{h}_i) \;:= \mathbf{u}(\mathbf{x}) \quad \forall\,\mathbf{x} \in \mathscr{C}\,,$$

for any integer k and each oriented side $\mathbf{h}_i$, $i = 1, ..., n$, of the periodicity cell.

We then consider a convex structural problem in the cell $\mathscr{C}$ with kinematic constraints which require that a conforming displacement field $\mathbf{u} \in \mathscr{L}_{\mathrm{PER}}(\mathscr{C})$, belonging to a conformity linear subspace $\mathscr{L}_{\mathrm{PER}}(\mathscr{C})$, is such that the corresponding $\mathscr{C}$-periodic extension $\mathbf{u}_\sharp \in \mathscr{L}^2(E^n, \mathrm{V})$ is Green regular, that is, such that

$$\int_\omega \|\,\mathbf{u}_\sharp(\mathbf{x})\,\|_V^2 + \|\,(\mathbf{B}\mathbf{u}_\sharp)(\mathbf{x})\,\|_D^2\;\mathrm{d}\mu < +\infty$$

for any compact subset ω in the euclidean space E^n.

That this condition is satisfied means that there are no jumps of the boundary traces of the $\mathscr{C}$-periodic extension displacement field across the interfaces of a regular mesh of repetitive periodicity cells. This condition is equivalent to requiring that the boundary traces of the displacement are equal on opposite faces of the cell. It follows that the mean value of the corresponding strain field vanishes, since

$$\text{MED}_{\mathscr{C}}\,(\mathbf{Bu}) = \text{sym}\int_{\partial\mathscr{C}} \boldsymbol{\Gamma}\mathbf{u}\otimes\mathbf{n}\,\mathrm{d}S = 0\,.$$

Homogenization can be carried out by solving the direct structural problem of the cell under the action of a constant strain field $\boldsymbol{\varepsilon} = \text{Im}\,\mathbf{M}^*_{\mathscr{C}} \subset \mathscr{H}(\mathscr{C})$ so that $\boldsymbol{\varepsilon}(\mathbf{x}) = \mathbf{D} \in D$ for almost all $\mathbf{x} \in \mathscr{C}$. Setting $\Omega = \mathscr{C}$ and $\mathscr{T}(\Omega) = \{\mathscr{C}\}$ we denote by $\mathscr{V}(\mathscr{C}\,;\mathrm{V}) \subset \mathscr{V}_{\mathscr{C}}$ the kinematic space of displacements fields which are Green-regular in $\mathscr{C}$.

Conforming displacements fields belong to the closed linear subspace $\mathscr{L}_{\text{PER}}(\mathscr{C}) \subset \mathscr{V}(\mathscr{C}\,;\mathrm{V})$. The problem is well-posed since strain fields corresponding to conforming displacements have null mean value and hence any constant strain field is effective as an imposed strain. The homogenized local constitutive law is the one that relates the mean value of the elastic stress field to the imposed constant strain field.

3.1 Orthogonal decomposition

A basic property of conforming displacements considered in periodic homogenization problems is that they have null mean value,

$$\mathscr{L}_{\text{PER}} \subset \text{Ker}\,\mathbf{M}_{\mathscr{C}} = (\,\text{Im}\,\mathbf{M}^*_{\mathscr{C}})^{\perp}\,,$$

where $\mathscr{L}_{\text{PER}}$ stands for $\mathscr{L}_{\text{PER}}(\mathscr{C})$. We then consider the closed linear subspace of displacement fields which can be expressed as the sum of a conforming field and of a constant-strain field,

$$\mathscr{L} := \{\,\mathbf{v} \in \mathscr{V}(\mathscr{C}\,;\mathrm{V}) \mid \mathbf{Bv} \in \mathbf{B}\mathscr{L}_{\text{PER}} \dotplus \text{Im}\,\mathbf{M}^*_{\mathscr{C}}\,\}\,.$$

Then the following relations hold:

$$\mathbf{B}\mathscr{L}_{\text{PER}} = \mathbf{B}\mathscr{L} \cap \text{Ker}\,\mathbf{M}_{\mathscr{C}}\,, \qquad (\mathbf{B}\mathscr{L}_{\text{PER}})^{\perp} = (\mathbf{B}\mathscr{L})^{\perp} \dotplus \text{Im}\,\mathbf{M}^*_{\mathscr{C}}\,,$$

$$\mathbf{B}\mathscr{L} = \mathbf{B}\mathscr{L}_{\text{PER}} \dotplus \text{Im}\,\mathbf{M}^*_{\mathscr{C}}\,, \qquad (\mathbf{B}\mathscr{L})^{\perp} = (\mathbf{B}\mathscr{L}_{\text{PER}})^{\perp} \cap \text{Ker}\,\mathbf{M}_{\mathscr{C}}\,.$$

By Korn's inequality the linear subspace $\mathbf{B}\mathscr{L}$ is closed in $\mathscr{H}_D$ and hence the following direct sum decomposition holds

$$\mathscr{H}_D = \mathbf{B}\mathscr{L} \dotplus (\mathbf{B}\mathscr{L})^{\perp}\,.$$

It follows that the Hilbert space $\mathscr{H}_D$ can be decomposed as the direct sum of three mutually orthogonal subspaces,

$$\begin{aligned}\mathscr{H}_D &= \text{Im}\,\mathbf{M}^*_{\mathscr{C}} \dotplus \mathbf{B}\mathscr{L}_{\text{PER}} \dotplus (\mathbf{B}\mathscr{L})^{\perp} \\ &= \text{Im}\,\mathbf{M}^*_{\mathscr{C}} \dotplus \mathbf{B}\mathscr{L}_{\text{PER}} \dotplus (\mathbf{B}\mathscr{L}_{\text{PER}})^{\perp} \cap \text{Ker}\,\mathbf{M}_{\mathscr{C}}\,.\end{aligned}$$

This direct sum decomposition in orthogonal complements plays a basic role in subsequent developments.

3.2 Conjugate potentials for the cell problem

The stress-strain law is assumed to be expressed by a generalized elastic law governed by two regular conjugate global convex potentials $\boldsymbol{\Phi}_{\mathbf{e}}(\boldsymbol{\varepsilon})$ and $\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma})$. The conjugate potentials governing the kinematic constraint for the cell problem are given by

$$\begin{aligned} J(\mathbf{u}) &:= \sqcap_{\mathscr{L}_{PER}}(\mathbf{u}-\mathbf{u}_D)\,, \\ J^*(\mathbf{f}) &:= \sqcap_{\mathscr{L}^{\perp}_{PER}}(\mathbf{f}) + \langle\, \mathbf{f}\,,\, \mathbf{u}_D \,\rangle\,, \end{aligned}$$

where $\mathbf{u}_D \in \mathscr{L}$ is a displacement field such that $(\mathbf{B}\mathbf{u}_D)(\mathbf{x}) = \mathbf{D}$ for all $\mathbf{x} \in \mathscr{C}$. Then $\mathbf{B}\mathbf{u}_D \in \operatorname{Im}\mathbf{M}^*_{\mathscr{C}}$. The symbol $\sqcap_{\mathscr{A}}$ denotes the concave indicator of the set $\mathscr{A}$, defined by

$$\sqcap_{\mathscr{A}}(\mathbf{x}) := \begin{cases} 0 & \mathbf{x} \in \mathscr{A}\,, \\ -\infty & \mathbf{x} \notin \mathscr{A}\,. \end{cases}$$

The functionals

$$\begin{aligned} F(\mathbf{u}) &= \boldsymbol{\Phi}(\mathbf{B}\mathbf{u}) - J(\mathbf{u})\,, && \mathbf{u} \in \mathscr{V}\,, \\ G(\boldsymbol{\sigma}) &= J^*(\mathbf{B}'\boldsymbol{\sigma}) - \boldsymbol{\Phi}^*(\boldsymbol{\sigma})\,, && \boldsymbol{\sigma} \in \mathscr{H}\,, \end{aligned}$$

take the explicit form

$$\begin{aligned} F(\mathbf{u}) &= \boldsymbol{\Phi}_{\mathbf{e}}(\mathbf{B}\mathbf{u})\,, && \mathbf{u} \in \mathbf{u}_D + \mathscr{L}_{\mathrm{PER}}\,, \\ G(\boldsymbol{\sigma}) &= \langle\, \boldsymbol{\sigma}\,,\, \mathbf{B}\mathbf{u}_D \,\rangle - \boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma})\,, && \boldsymbol{\sigma} \in (\mathbf{B}\mathscr{L}_{\mathrm{PER}})^{\perp}\,. \end{aligned}$$

Recalling the orthogonal decomposition $(\mathbf{B}\mathscr{L}_{\mathrm{PER}})^{\perp} = \operatorname{Im}\mathbf{M}^*_{\mathscr{C}} \dotplus (\mathbf{B}\mathscr{L})^{\perp}$, we can conveniently rewrite

$$\begin{aligned} F_{\mathbf{D}}(\mathbf{v}) &= \boldsymbol{\Phi}_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{D} + \mathbf{B}\mathbf{v})\,, && \mathbf{v} \in \mathscr{L}_{\mathrm{PER}}\,, \\ G_{\mathbf{D}}(\mathbf{s},\mathbf{T}) &= \langle\, \mathbf{M}^*_{\mathscr{C}}\mathbf{T}\,,\, \mathbf{M}^*_{\mathscr{C}}\mathbf{D} \,\rangle - \boldsymbol{\Phi}^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{T} + \mathbf{s})\,, && \mathbf{s} \in (\mathbf{B}\mathscr{L})^{\perp}\,, \quad \mathbf{T} \in S\,. \end{aligned}$$

3.3 Effective response

The global effective potential of the homogenized constitutive law is defined by

$$\begin{aligned} \boldsymbol{\Phi}_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{D}) &= \min\{\, F_{\mathbf{D}}(\mathbf{v}) \mid \mathbf{v} \in \mathscr{L}_{\mathrm{PER}} \,\} \\ &= \max\{\, G_{\mathbf{D}}(\mathbf{s},\mathbf{T}) \mid \mathbf{s} \in (\mathbf{B}\mathscr{L})^{\perp}, \mathbf{T} \in S \,\}\,, \end{aligned}$$

or, explicitly,

$$\begin{aligned} \boldsymbol{\Phi}_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{D}) &= \min\{\, \boldsymbol{\Phi}_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D} + \boldsymbol{\eta}) \mid \boldsymbol{\eta} \in \mathbf{B}\mathscr{L}_{\mathrm{PER}} \,\} \\ &= \max\{\, \langle\, \mathbf{M}^*_{\mathscr{C}}\mathbf{T}\,,\, \mathbf{M}^*_{\mathscr{C}}\mathbf{D} \,\rangle - \boldsymbol{\Phi}^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{T} + \mathbf{s}) \mid \mathbf{s} \in (\mathbf{B}\mathscr{L})^{\perp}, \mathbf{T} \in S \,\}\,. \end{aligned}$$

The global effective potential is convex, as it is the inf-convolution of the two convex functionals. Indeed,

$$\begin{aligned}\boldsymbol{\Phi}_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{D}) &= \min\{\,\boldsymbol{\Phi}_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}-\boldsymbol{\eta}) \mid \boldsymbol{\eta}\in\mathbf{B}\mathscr{L}_{\mathrm{PER}}\,\}\\ &= \min\{\,\boldsymbol{\Phi}_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}-\boldsymbol{\eta})+\sqcup_{\mathbf{B}\mathscr{L}_{PER}}(\boldsymbol{\eta})\,\}\\ &= (\boldsymbol{\Phi}_{\mathbf{e}}\,\square\sqcup_{\mathbf{B}\mathscr{L}_{PER}})(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D})\,.\end{aligned}$$

We recall that the epigraph of the inf-convolution of two convex functionals is the convex sum of the two convex epigraphs and that

$$\boldsymbol{\Phi}_{\mathbf{e}}\,\square\sqcup_{\mathbf{B}\mathscr{L}_{PER}} = (\boldsymbol{\Phi}^*_{\mathbf{e}}+\sqcup_{(\mathbf{B}\mathscr{L}_{PER})^{\perp}})^*\,.$$

The local potential of the homogenized constitutive law is then defined as

$$\boldsymbol{\varphi}_H(\mathbf{D}) = \frac{1}{\mathrm{vol}\,(\mathscr{C})}\;(\boldsymbol{\Phi}_H\circ\mathbf{M}^*_{\mathscr{C}})(\mathbf{D}) = \frac{1}{\mathrm{vol}\,(\mathscr{C})}\left[(\boldsymbol{\Phi}_{\mathbf{e}}\,\square\sqcup_{\mathbf{B}\mathscr{L}_{PER}})\circ\mathbf{M}^*_{\mathscr{C}}\right](\mathbf{D})\,.$$

Observing that $\boldsymbol{\Phi}_{\mathbf{e}} = \mathbf{M}_{\mathscr{C}}\,\boldsymbol{\varphi}_{\mathbf{e}}$ and that

$$\langle\,\mathbf{M}^*_{\mathscr{C}}\mathbf{T}\,,\;\mathbf{M}^*_{\mathscr{C}}\mathbf{D}\,\rangle = \mathrm{vol}\,(\mathscr{C})\,\langle\,\mathbf{T}\,,\;\mathbf{D}\,\rangle\,,$$

we get the following expression for the local homogenized potential:

$$\begin{aligned}\boldsymbol{\varphi}_H(\mathbf{D}) &= \min\{\,\textsc{med}_{\mathscr{C}}(\boldsymbol{\varphi}_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}-\boldsymbol{\eta}))\mid\boldsymbol{\eta}\in\mathbf{B}\mathscr{L}_{\mathrm{PER}}\,\}\\ &= \max\{\,\langle\,\mathbf{T}\,,\;\mathbf{D}\,\rangle-\textsc{med}_{\mathscr{C}}(\boldsymbol{\varphi}^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{T}+\mathbf{s}))\mid\mathbf{s}\in(\mathbf{B}\mathscr{L})^{\perp},\,\mathbf{T}\in S\,\}\\ &= \max_{\mathbf{T}\in S}\left\{\langle\,\mathbf{T}\,,\;\mathbf{D}\,\rangle-\frac{1}{\mathrm{vol}\,(\mathscr{C})}(\boldsymbol{\Phi}^*_{\mathbf{e}}\,\square\sqcup_{(\mathbf{B}\mathscr{L})^{\perp}})\circ\mathbf{M}^*_{\mathscr{C}}(\mathbf{T})\right\}.\end{aligned}$$

Hence, setting

$$\boldsymbol{\psi}_H(\mathbf{T}) = \left[\frac{1}{\mathrm{vol}\,(\mathscr{C})}(\boldsymbol{\Phi}^*_{\mathbf{e}}\,\square\sqcup_{(\mathbf{B}\mathscr{L})^{\perp}})\circ\mathbf{M}^*_{\mathscr{C}}\right](\mathbf{T})\,,$$

we get the conjugacy relation

$$\boldsymbol{\varphi}_H = (\boldsymbol{\psi}_H)^*\,.$$

By the properties of the inf-convolution we know that, setting

$$\begin{aligned}\boldsymbol{\Phi}_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{D}) &= \min\{\,\boldsymbol{\Phi}_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}-\boldsymbol{\eta})\mid\boldsymbol{\eta}\in\mathbf{B}\mathscr{L}_{\mathrm{PER}}\,\} = \boldsymbol{\Phi}_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}-\boldsymbol{\varepsilon}_{\mathbf{D}})\\ &= (\boldsymbol{\Phi}_{\mathbf{e}}\,\square\sqcup_{\mathbf{B}\mathscr{L}_{PER}})(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D})\,,\end{aligned}$$

with $\boldsymbol{\varepsilon}_{\mathbf{D}} = \mathbf{B}\mathbf{u}_{\mathbf{D}}$ and $\mathbf{u}_{\mathbf{D}}\in\mathscr{L}_{\mathrm{PER}}$, we have

$$\begin{cases}\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}-\boldsymbol{\varepsilon}_{\mathbf{D}}\in\partial\boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}_{\mathbf{D}})\,,\\ \boldsymbol{\varepsilon}_{\mathbf{D}}\in\partial\sqcup_{(\mathbf{B}\mathscr{L}_{PER})^{\perp}}(\boldsymbol{\sigma}_{\mathbf{D}})\,,\end{cases}$$

where

$$\boldsymbol{\sigma}_{\mathbf{D}}\in\partial(\boldsymbol{\Phi}_{\mathbf{e}}\,\square\sqcup_{\mathbf{B}\mathscr{L}_{PER}})(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}) = \partial\boldsymbol{\Phi}_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{D})\,,$$

is the stress solution of the direct problem [17].

By the chain rule of subdifferential calculus,

$$\partial(\boldsymbol{\Phi}_H \circ \mathbf{M}^*_{\mathscr{C}})(\mathbf{D}) = \mathbf{M}_{\mathscr{C}}\, \partial\boldsymbol{\Phi}_H(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D})\,,$$

and, from the definition of φ_H, we eventually get the relation

$$\text{MED}(\boldsymbol{\sigma}_{\mathbf{D}}) \in \partial\varphi_H(\mathbf{D})\,,$$

that justifies the homogenization role played by the potential φ_H.

3.4 Inverse effective response

An alternative procedure for carrying out the homogenization process, consists in solving the inverse structural problem of the cell under the action of a constant stress field $\boldsymbol{\sigma} = \operatorname{Im}\mathbf{M}^*_{\mathscr{C}} \subset \mathscr{H}_S(\mathscr{C}) = \mathscr{L}^2(\mathscr{C}\,;S)$ so that $\boldsymbol{\sigma}(\mathbf{x}) = \mathbf{T} \in S$ for almost all $\mathbf{x} \in \mathscr{C}$. Setting $\boldsymbol{\Omega} = \mathscr{C}$ and $\mathscr{T}(\boldsymbol{\Omega}) = \{\mathscr{C}\}$ we denote by $\mathscr{V}(\mathscr{C}\,;\mathrm{V})$ the kinematic space of displacements fields which are Green-regular in $\mathscr{C}$. Conforming displacement fields are assumed to belong to the subspace $\mathscr{L}_{\text{PER}}(\mathscr{C}) \subset \mathscr{V}(\mathscr{C},\mathrm{V})$. Self-equilibrated stresses then belong to the linear subspace $\mathscr{L}^{\perp}_{\text{PER}}(\mathscr{C})$. The problem is well-posed if every stress field is assumed to be the sum of the prescribed constant field and any self-equilibrated field with zero mean value. Indeed in this case any constant stress field is effective as an imposed stress.

According to the inverse homogenization procedure the homogenized local constitutive law is the one that relates the mean value of the strain field to the imposed constant stress field.

The conjugate pairs of convex potentials governing the monotone stress-strain and force-displacement relations are given as:

$$\begin{aligned}
\boldsymbol{\Phi}^*(\boldsymbol{\sigma}) &:= \boldsymbol{\Phi}^*_{\mathbf{e}}(\boldsymbol{\sigma}) + \sqcup_{\operatorname{Ker}\mathbf{M}_{\mathscr{C}}}(\boldsymbol{\sigma} - \mathbf{M}^*_{\mathscr{C}}\mathbf{T})\,,\\
\boldsymbol{\Phi}(\boldsymbol{\varepsilon}) &:= \Big(\boldsymbol{\Phi}_{\mathbf{e}} \,\square\, (\sqcup_{\operatorname{Im}\mathbf{M}^*_{\mathscr{C}}} + \langle\, \mathbf{M}^*_{\mathscr{C}}\mathbf{T}\,,\,\cdot\,\rangle\Big)(\boldsymbol{\varepsilon})\,,\\
J^*(\mathbf{f}) &:= \sqcap_{\mathscr{L}^{\perp}_{PER}}(\mathbf{f})\,,\\
J(\mathbf{u}) &:= \sqcap_{\mathscr{L}_{PER}}(\mathbf{u})\,.
\end{aligned}$$

Recalling that $(\mathbf{B}\mathscr{L})^{\perp} = (\mathbf{B}\mathscr{L}_{PER})^{\perp} \cap \operatorname{Ker}\mathbf{M}_{\mathscr{C}}$ and setting

$$\boldsymbol{\sigma} = \mathbf{M}^*_{\mathscr{C}}\mathbf{T} + \mathbf{s} \quad \text{with} \quad \mathbf{s} \in (\mathbf{B}\mathscr{L})^{\perp}\,,$$

we see that the functionals

$$\begin{aligned}
F(\mathbf{u}) &= \boldsymbol{\Phi}(\mathbf{B}\mathbf{u}) - J(\mathbf{u})\,, \qquad &\mathbf{u} \in \mathscr{V}\,,\\
G(\boldsymbol{\sigma}) &= J^*(\mathbf{B}'\boldsymbol{\sigma}) - \boldsymbol{\Phi}^*(\boldsymbol{\sigma})\,, \qquad &\boldsymbol{\sigma} \in \mathscr{H}\,,
\end{aligned}$$

take the explicit forms

$$\begin{aligned}
F_{\mathbf{T}}(\mathbf{v},\mathbf{D}) &= \inf_{\mathbf{D}\in D}\big\{\boldsymbol{\Phi}_{\mathbf{e}}(\mathbf{B}\mathbf{v} + \mathbf{M}^*_{\mathscr{C}}\mathbf{D}) - \langle\, \mathbf{M}^*_{\mathscr{C}}\mathbf{T}\,,\, \mathbf{M}^*_{\mathscr{C}}\mathbf{D}\,\rangle\big\} - \sqcap_{\mathscr{L}_{PER}}(\mathbf{v})\,,\\
G_{\mathbf{T}}(\mathbf{s}) &= -\Big(\boldsymbol{\Phi}^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{T} + \mathbf{s}) + \sqcup_{(\mathbf{B}\mathscr{L})^{\perp}}(\mathbf{s})\Big)\,.
\end{aligned}$$

The global effective potential of the homogenized medium is the convex functional $\Psi_H : \mathcal{H} \mapsto \overline{\mathcal{R}}$ defined by one of the equivalent relations

$$\begin{aligned}
-\Psi_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{T}) &:= \min_{\mathbf{v}\in\mathscr{L}_{PER}} F_{\mathbf{T}}(\mathbf{v}) = \min_{\mathbf{v}\in\mathscr{L}}\{\,\Phi_{\mathbf{e}}(\mathbf{B}\mathbf{v}) - \langle\, \mathbf{M}^*_{\mathscr{C}}\mathbf{T}\,,\,\mathbf{B}\mathbf{v}\,\rangle\,\}\,,\\
-\Psi_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{T}) &:= \max_{\mathbf{s}\in(\mathbf{B}\mathscr{L})^\perp} G_{\mathbf{T}}(\mathbf{s}) = \max_{\mathbf{s}\in(\mathbf{B}\mathscr{L})^\perp}\{-\Phi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{T}+\mathbf{s})\}\,,\\
&= -\min_{\mathbf{s}\in(\mathbf{B}\mathscr{L})^\perp}\{\,\Phi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{T}-\mathbf{s}) + \sqcup_{(\mathbf{B}\mathscr{L})^\perp}(\mathbf{s})\,\}\,,\\
&= -(\Phi^*_{\mathbf{e}}\,\square\,\sqcup_{(\mathbf{B}\mathscr{L})^\perp})(\mathbf{M}^*_{\mathscr{C}}\mathbf{T})\,.
\end{aligned}$$

The local potential of the homogenized constitutive law is then defined as

$$\psi_H(\mathbf{T}) := \frac{1}{\mathrm{vol}\,(\mathscr{C})}\;(\Psi_H \circ \mathbf{M}^*_{\mathscr{C}})(\mathbf{T})\,.$$

Recall that the corresponding convex potential for the direct problem is defined by the equivalent relations

$$\begin{aligned}
\Phi_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{D}) &:= \min_{\mathbf{v}\in\mathscr{L}_{PER}} F_{\mathbf{D}}(\mathbf{v}) \;=\; \min_{\mathbf{v}\in\mathscr{L}_{PER}} \Phi_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{D}+\mathbf{B}\mathbf{v})\,,\\
\Phi_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{D}) &:= \max_{\mathbf{s}\in(\mathbf{B}\mathscr{L}_{PER})^\perp} G_{\mathbf{D}}(\mathbf{s}) = \max_{\mathbf{s}\in(\mathbf{B}\mathscr{L}_{PER})^\perp}\{\,\langle\,\mathbf{s}\,,\,\mathbf{M}^*_{\mathscr{C}}\mathbf{D}\,\rangle - \Phi^*_{\mathbf{e}}(\mathbf{s})\,\}\,.
\end{aligned}$$

Hence, as $(\mathbf{B}\mathscr{L}_{\mathrm{PER}})^\perp = (\mathbf{B}\mathscr{L})^\perp \dotplus \mathrm{Im}\,\mathbf{M}^*_{\mathscr{C}}$, we have

$$\begin{aligned}
\Phi_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{D}) &= \max_{\mathbf{s}\in(\mathbf{B}\mathscr{L}_{PER})^\perp}\{\,\langle\,\mathbf{s}\,,\,\mathbf{M}^*_{\mathscr{C}}\mathbf{D}\,\rangle - \Phi^*_{\mathbf{e}}(\mathbf{s})\,\}\,,\\
&= \max_{\mathbf{T}\in S}\max_{\mathbf{s}\in(\mathbf{B}\mathscr{L})^\perp}\big\{\langle\,\mathbf{M}^*_{\mathscr{C}}\mathbf{T}\,,\,\mathbf{M}^*_{\mathscr{C}}\mathbf{D}\,\rangle - \Phi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\mathbf{T}+\mathbf{s})\big\}\\
&= \max_{\mathbf{T}\in S}\big\{\langle\,\mathbf{M}^*_{\mathscr{C}}\mathbf{T}\,,\,\mathbf{M}^*_{\mathscr{C}}\mathbf{D}\,\rangle - \Psi_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{T})\big\}\\
&= (\Psi_H)^*(\mathbf{M}^*_{\mathscr{C}}\mathbf{D})\,.
\end{aligned}$$

By the properties of the inf-convolution we also have

$$\begin{aligned}
\Psi_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{T}) &= \min\{\,\Phi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T}-\mathbf{s}) \mid \mathbf{s}\in(\mathbf{B}\mathscr{L})^\perp\,\} = \;\Phi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T}-\mathbf{s}_{\mathbf{T}})\\
&= (\Phi^*_{\mathbf{e}}\,\square\,\sqcup_{(\mathbf{B}\mathscr{L})^\perp})(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T})\,,
\end{aligned}$$

with $\mathbf{s}_{\mathbf{T}} \in (\mathbf{B}\mathscr{L})^\perp$ and

$$\begin{cases}
\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T} - \mathbf{s}_{\mathbf{T}} \in \partial\Phi_{\mathbf{e}}(\boldsymbol{\varepsilon}_{\mathbf{T}})\,,\\
\boldsymbol{\sigma}_{\mathbf{T}} \in \partial\,\sqcup_{\mathbf{B}\mathscr{L}}(\boldsymbol{\varepsilon}_{\mathbf{T}})\,,
\end{cases}$$

where

$$\boldsymbol{\varepsilon}_{\mathbf{T}} \in \partial(\Phi^*_{\mathbf{e}}\,\square\,\sqcup_{(\mathbf{B}\mathscr{L})^\perp})(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T}) = \partial\Psi_H(\mathbf{M}^*_{\mathscr{C}}\mathbf{T})\,,$$

is the strain solution of the inverse problem [17].

By the chain rule of subdifferential calculus we infer that

$$\partial(\boldsymbol{\Psi}_H \circ \mathbf{M}^*_{\mathscr{C}})(\mathbf{T}) = \mathbf{M}_{\mathscr{C}}\, \partial\boldsymbol{\Psi}_H(\mathbf{M}^*_{\mathscr{C}}\, \mathbf{T})\,,$$

and from the definition of ψ_H we eventually get the relation

$$\text{MED}(\boldsymbol{\varepsilon}_{\mathbf{T}}) \in \partial\psi_H(\mathbf{T})\,,$$

that justifies the homogenization role played by the potential ψ_H.

Remark 4. The conjugacy relation between the potentials of the direct and the inverse cell problems can also be obtained by applying the following conjugacy rules:

$$\begin{aligned}
&(\alpha\, f)^*(\mathbf{x}^*) = \alpha\, f^*(\frac{1}{\alpha}\,\mathbf{x}^*) \qquad \forall\, \alpha > 0\,,\\
&(f \circ \mathbf{L})^*(\mathbf{x}^*) = \inf\{\ f^*(\mathbf{y}^*) \mid \mathbf{L}'(\mathbf{y}^*) = \mathbf{x}^*\ \}\,,\\
&(f \,\square\, g)^*(\mathbf{x}^*) = \inf\{\ f^*(\mathbf{x}^*_1) + g^*(\mathbf{x}^*_2) \mid \mathbf{x}^*_1 + \mathbf{x}^*_2 = \mathbf{x}^*\ \}\,,
\end{aligned}$$

which hold under reasonable global regularity conditions of the involved potentials [3,5,6]. Less stringent local conditions were contributed in [12].

Indeed,

$$\begin{aligned}
(\varphi_H)^*(\mathbf{T}) &= \left[\frac{1}{\text{vol}\,(\mathscr{C})}(\boldsymbol{\Phi}_{\mathbf{e}} \,\square\, \sqcup_{\mathbf{B}\mathscr{L}_{PER}}) \circ \mathbf{M}^*_{\mathscr{C}}\right]^*(\mathbf{T})\\
&= \frac{1}{\text{vol}\,(\mathscr{C})}\left[(\boldsymbol{\Phi}_{\mathbf{e}} \,\square\, \sqcup_{\mathbf{B}\mathscr{L}_{PER}}) \circ \mathbf{M}^*_{\mathscr{C}}\right]^*(\text{vol}\,(\mathscr{C})\,\mathbf{T})\\
&= \frac{1}{\text{vol}\,(\mathscr{C})}\inf\left\{(\boldsymbol{\Phi}_{\mathbf{e}} \,\square\, \sqcup_{\mathbf{B}\mathscr{L}_{PER}})^*(\boldsymbol{\sigma}) \mid \mathbf{M}_{\mathscr{C}}(\boldsymbol{\sigma}) = \text{vol}\,(\mathscr{C})\,\mathbf{T}\right\}\\
&= \frac{1}{\text{vol}\,(\mathscr{C})}\inf\left\{(\boldsymbol{\Phi}^*_{\mathbf{e}} + \sqcup_{(\mathbf{B}\mathscr{L}_{PER})^\perp})(\boldsymbol{\sigma}) \mid \mathbf{M}_{\mathscr{C}}(\boldsymbol{\sigma}) = \text{vol}\,(\mathscr{C})\,\mathbf{T}\right\}\\
&= \frac{1}{\text{vol}\,(\mathscr{C})}\inf\left\{(\boldsymbol{\Phi}^*_{\mathbf{e}}(\mathbf{M}^*\mathbf{T} + \mathbf{s}) + \sqcup_{(\mathbf{B}\mathscr{L})^\perp}(\mathbf{s})\right\}\\
&= \left[\frac{1}{\text{vol}\,(\mathscr{C})}(\boldsymbol{\Phi}^*_{\mathbf{e}} \,\square\, \sqcup_{(\mathbf{B}\mathscr{L})^\perp}) \circ \mathbf{M}^*_{\mathscr{C}}\right](\mathbf{T}) = \psi_H(\mathbf{T})\,.
\end{aligned}$$

Note that, from the relation $(\mathbf{B}\mathscr{L}_{\text{PER}})^\perp = (\mathbf{B}\mathscr{L})^\perp \dotplus \text{Im}\,\mathbf{M}^*_{\mathscr{C}}$, we have argued that the conditions

$$\boldsymbol{\sigma} \in (\mathbf{B}\mathscr{L}_{PER})^\perp\,, \qquad \mathbf{M}_{\mathscr{C}}(\boldsymbol{\sigma}) = \text{vol}\,(\mathscr{C})\,\mathbf{T}\,,$$

are equivalent to the assumption that $\boldsymbol{\sigma} = \mathbf{M}^*\mathbf{T} + \mathbf{s}$ with $\mathbf{s} \in (\mathbf{B}\mathscr{L})^\perp$.

3.5 Bounds on the effective response

In computing the local potential of the homogenized constitutive law we can get a rough estimate by taking respectively $\boldsymbol{\eta} = 0$ and $\mathbf{s} = 0$ in the expressions to be

minimized and maximized as reported in Sect. 3.3. The upper and lower bounds so obtained are the generalized Voigt (upper) and Reuss (lower) bounds for the effective potential of the homogenized medium:

$$\max\{\langle\, \mathbf{T}\,,\, \mathbf{D}\,\rangle - \textsc{med}_{\mathscr{C}}(\varphi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T})) \mid \mathbf{T} \in S\,\} \le \varphi_H(\mathbf{D}) \le \textsc{med}_{\mathscr{C}}(\varphi_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}))\,.$$

To get the Voigt bound we consider a constant strain field $\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}$, evaluate the corresponding local potential φ at any point of the cell and take its mean value. In this way an *arithmetic mean* approximation is obtained.

In the linear elastic case the Voigt approximation amounts to taking the composition of the local elastic stiffnesses by a parallel scheme of elastic springs, and the effective elastic stiffness is given by the average of the local stiffnesses.

To get the Reuss bound we consider a constant stress field $\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T}$, evaluate the corresponding conjugate local potential φ^* at any point of the cell, take the mean value and evaluate the conjugate local potential. In this way a *harmonic mean* approximation is obtained.

In the linear elastic case the Reuss approximation amounts to taking the composition of the local elastic stiffnesses by a serial scheme of elastic springs, and the effective compliance is given by the average of the local compliances.

Better bounds can be found by computing approximate solutions of the cell problem either directly, in terms of conforming displacements with zero mean strain, to get upper bounds, or in the complementary way, in terms of self-stresses with zero mean value, to get lower bounds.

Another approach to the problem of bounding the effective properties of the homogenized medium is provided by polarization techniques which were first applied to elasticity problems by Hashin and Shtrikman in 1962 [1,2] and then extended and generalized to the non-linear setting by Talbot and Willis in 1985 [7] and by Willis and Toland-Willis in 1989 [8,9].

3.6 Uniform local bounds

We now assume that the field of local potentials $\varphi_{\mathbf{e}}$ is uniformly bounded from above and from below,

$$\varphi^- \le \varphi_{\mathbf{e}} \le \varphi^+\,,$$

where $\varphi^-, \varphi^+ : D \mapsto \overline{\mathscr{R}}$ are convex functions. From the Voigt-Reuss inequalities

$$\max\{\langle\, \mathbf{T}\,,\, \mathbf{D}\,\rangle - \textsc{med}_{\mathscr{C}}(\varphi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T})) \mid \mathbf{T} \in S\,\} \le \varphi_H(\mathbf{D}) \le \textsc{med}_{\mathscr{C}}(\varphi_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D}))\,,$$

where

$$(\varphi^+)^* \le \varphi^*_{\mathbf{e}} \le (\varphi^-)^*\,,$$

$$\textsc{med}_{\mathscr{C}}(\varphi_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{D})) \le \varphi^+(\mathbf{D})\,,$$

$$\textsc{med}_{\mathscr{C}}(\varphi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T})) \le (\varphi^-)^*(\mathbf{T})\,,$$

$$\begin{aligned}\varphi^-(\mathbf{D}) &= \max\{\langle\, \mathbf{T}\,,\, \mathbf{D}\,\rangle - (\varphi^-)^*(\mathbf{T}) \mid \mathbf{T} \in S\,\} \\ &\leq \max\{\langle\, \mathbf{T}\,,\, \mathbf{D}\,\rangle - \mathrm{MED}_{\mathscr{C}}(\varphi^*_{\mathbf{e}}(\mathbf{M}^*_{\mathscr{C}}\,\mathbf{T})) \mid \mathbf{T} \in S\,\}\,,\end{aligned}$$

we infer that the same bounds hold for the local potential of the homogenized constitutive law, that is,

$$\varphi^- \leq \varphi_{\mathbf{e}} \leq \varphi^+ \quad \Longrightarrow \quad \varphi^- \leq \varphi_H \leq \varphi^+\,.$$

3.7 Geometric constraints

We remark that the analysis carried out above relies only on the property that conforming displacements belonging to the subspace $\mathscr{L}_{\mathrm{PER}}$ have zero mean value, that is, that $\mathscr{L}_{\mathrm{PER}} \subset \mathrm{Ker}\,\mathbf{M}_{\mathscr{C}}$.

We could thus also choose the conforming subspace

$$\mathscr{L}_o(\mathscr{C}) := \{\,\mathbf{v} \in \mathscr{V}(\mathscr{C}) \mid \boldsymbol{\Gamma}\mathbf{v} = 0\,\} = \mathrm{Ker}\,\boldsymbol{\Gamma} \subset \mathrm{Ker}\,\mathbf{M}_{\mathscr{C}}\,,$$

instead of $\mathscr{L}_{\mathrm{PER}}(\mathscr{C})$. Since $\mathscr{L}_o(\mathscr{C}) \subset \mathscr{L}_{\mathrm{PER}}(\mathscr{C})$, denoting by φ^o_H and ψ^o_H the direct and inverse local effective potentials under the constraints defined by $\mathscr{L}_o(\mathscr{C})$, we get the inequalities

$$\varphi_H \leq \varphi^o_H\,, \qquad \psi_H \geq \psi^o_H\,.$$

References

[1] Hashin, Z., Shtrikman, S. (1962): On some variational principles in anisotropic and non-homogeneous elasticity. J. Mech. Phys. Solids **10**, 335–342

[2] Hashin, Z., Shtrikman, S. (1962): A variational approach to the theory of the elastic behaviour of polycrystals. J. Mech. Phys. Solids **10**, 343–352

[3] Rockafellar, R.T. (1970): Convex analysis. Princeton University Press, Princeton

[4] Yosida, K. (1974): Functional analysis. 4th edition. Springer, New York

[5] Ekeland, I., Temam, R. (1976): Convex analysis and variational problems. North-Holland, Amsterdam

[6] Ioffe, A.D., Tihomirov, V.M. (1974): Theory of extremal problems. (Russian). Nauka, Moscow. Translation (1979): North-Holland, Amsterdam

[7] Talbot, D.R.S., Willis, J.R. (1985): Variational principles for inhomogeneous nonlinear media. IMA J. Appl. Math. **35**, 39–54

[8] Willis, J.R. (1989): The structure of overall constitutive relations for a class of nonlinear composites. IMA J. Appl. Math. **43**, 231–242

[9] Toland, J.F., Willis, J.R. (1989): Duality for families of natural variational principles in nonlinear electrostatics. SIAM J. Math. Anal. **20**, 1283–1292

[10] Romano, G., Rosati, L., Marotti de Sciarra, F., Bisegna, P. (1993): A potential theory for monotone multivalued operators. Quart. Appl. Math. **51**, 613–631

[11] Hiriart-Urruty, J.B., Lemaréchal, C. (1993): Convex analysis and minimization algorithms. Vols. I, II. Springer, Berlin

[12] Romano, G. (1995): New results in subdifferential calculus with applications to convex optimization. Appl. Math. Optim. **32**, 213–234

[13] Romano, G. (2000): Structural mechanics. II. Continuous models. Libero, Napoli

[14] Romano, G. (2000): On the necessity of Korn's inequality. Symposium on Trends in Applications of Mathematics to Mechanics (STAMM 2000). Galway, Ireland, July 9–14, 2000
[15] Romano, G. (2001): Scienza delle costruzioni. Tomo Zero. Hevelius, Benevento
[16] Romano, G. (2002): Scienza delle costruzioni. Tomo I. Hevelius, Benevento
[17] Romano, G. (2003): Scienza delle costruzioni. Tomo II. Hevelius, Benevento
[18] Romano, G., Diaco, M., Sellitto, C. (2004): Tangent stiffness of elastic continua on manifolds. In: Romano, G., Rionero, S. (eds.): Recent trends in the applications of mathematics to mechanics. Springer, Berlin, pp. 155–184
[19] Romano, G., Diaco, M. (2004): A functional framework for applied continuum mechanics. In: Fergola, P., Capone, F. (eds.): New Trends in Mathematical Physics. World Scientific, Singapore, to appear

Tangent stiffness of polar shells undergoing large displacements

G. Romano, C. Sellitto

Abstract. The paper deals with the definition and evaluation of the tangent stiffness of hyperelastic polar shells without drilling rotations. The ambient space for such bodies is a non-linear differentiable manifold. As a consequence the incremental equilibrium must be expressed as the absolute time derivative of the non-linear equilibrium condition expressing the balance between the elastic response and the applied forces. In the absolute time derivative the classical directional derivative is replaced by the covariant derivative according to a fixed connection on the manifold. The evaluation of the tangent stiffness requires us to take the second covariant derivative of the finite deformation measure and this in turn requires an extension of the virtual displacement field in a neighborhood of the given configuration of the shell. It is explicitly shown that different choices of this extension lead to the same tangent stiffness, which is symmetric since the chosen connection is torsionless.

1 Introduction

The evaluation of the tangent stiffness of an hyperelastic body is of crucial importance when dealing with finite changes of configuration. The tangent stiffness provides the linear relationship between the rate of change of configuration and the corresponding rate of change of elastic response of the body in terms of forces. The analysis of small vibrations of a finitely deformed elastic body, the instability of equilibrium configurations and the prediction of the way in which an elastic body tends to move under a loading path, are all governed by the properties of the tangent stiffness.

There are many different ways of defining a deformation measure of the body and the choice of a special measure changes the way in which the modeling of the constitutive properties of the material is performed. The basic requirements with which a deformation measure has to conform, are that the measure must be independent of superimposed rigid changes of configuration and must be a local field in the sense that its value at a point must not be affected by a change of the placement map outside any neighborhood of that point.

The definition of a rigid change of configuration is a basic item that must be given in describing the kinematical properties of the body in its motion in the ambient space. In hyperelastic bodies Green's potential defines the local elastic properties of the material in terms of its deformation from a given natural state. The deformation field depends in turn on the map which defines the placement of the body with respect to a reference configuration in the ambient space.

Once a deformation measure has been chosen, the local elastic potential can be expressed as the composition of the local elastic energy and the deformation measure.

It is then a function of the configuration change from a reference configuration in which the material is assumed to be in a natural state. The global elastic potential is obtained by integrating the local elastic potential over the whole body in the reference configuration.

In finite deformation analysis all the state variables defined in the actual configuration are transformed into the corresponding ones in the reference configuration. Accordingly, in an evolution process, the equilibrium condition at the actual configuration is expressed by imposing the equality between the directional derivative of the global elastic potential along a conforming virtual (tangent) displacement and the corresponding virtual work of the referential forces. The derivative of the global elastic potential is the elastic response of the body to the change of configuration. Both the elastic response and the referential forces are bounded linear forms on the linear space of conforming virtual displacements. The condition of incremental equilibrium is then obtained by taking the time derivative of the equilibrium condition.

In classical structural analysis the time derivative of the elastic response is expressed by means of the chain rule, as the directional derivative of the elastic response along the velocity field of the body. When dealing with polar bodies this procedure must be revised to take into account the non-affine geometrical structure of the physical space. In such a situation the time derivative must be replaced by the absolute differentiation with respect to time, defined as the covariant derivative of the elastic response along the velocity field.

To grasp the motivation of this new approach one has to consider that, when the ambient space is a non-linear differentiable manifold, the tangent spaces of virtual displacements and their dual counterparts, the cotangent spaces of force systems, change from point to point. In general there is no way to perform a classical differentiation of a vector or of a covector field on a differentiable manifold since this would necessitate taking the difference of unrelated vectors belonging to different linear spaces.

In structural mechanics the non-linear differentiable manifold defining the ambient space is usually embedded into a larger affine space with a euclidean structure. In this case the covariant differentiation simply amounts to taking the component of the directional derivative on the subspace tangent to the manifold.

This definition of the covariant differentiation is equivalent to considering the Levi–Civita connection on the manifold associated with the Riemannian metric induced by the euclidean metric of the larger affine space.

One more essential point remains to be fixed. The directional derivative of a field of linear forms on a linear space satisfies the Leibniz rule of calculus: the directional derivative of a linear form at a vector field is equal to the difference between the directional derivative of its value at the vector field and its value corresponding to the directional derivative of the vector field.

By analogy the covariant differentiation of a linear form is defined by means of a formal application of the Leibniz rule: the value of the covariant derivative of a linear form at a vector field is equal to the difference between the covariant derivative of its value at the vector field and its value corresponding to the covariant derivative of the vector field. The definition is well-posed since, although both terms in the difference

depend on the values that the vector field takes in a neighborhood of the point, their difference is local and hence the covariant derivative of the linear form is tensorial.

From the discussion above it follows that the tangent stiffness must be properly defined as the covariant derivative of the elastic response. As the covariant derivative of a linear form, the tangent stiffness is then a two-times covariant tensor. The evaluation of the tangent stiffness of polar elastic bodies is then a remarkable example of the application of differential geometry, and specifically of calculus on manifolds, to issues of mechanics.

In previous treatments, in dealing with models of polar beams and shells, the geometric tangent stiffness was simply evaluated as the inner product of the referential stress and the second directional derivative of the deformation measure. It is apparent that such an evaluation requires the extension of the virtual displacement along which the first derivative is taken, to a vector field defined in a neighborhood of the given configuration.

In finite deformation analysis of polar shells without drilling rotations the ambient space is the trivial fiber bundle defined by the cartesian product of the euclidean space (the base manifold) and the unit sphere (the fiber). The corresponding tangent stiffness, computed by taking the second covariant derivative of the deformation measure, is local and symmetric when the space manifold is endowed with the Levi–Civita connection induced by the larger affine space.

Two different extensions of the virtual displacement are investigated and it is shown that the one yields a symmetric second directional derivative of the deformation measure while the other leads to a non-symmetric second directional derivative. It is further shown, by explicit calculation, that the corresponding second covariant derivative of the deformation measure is, however, symmetric in both cases, as required by the theory.

2 Polar shells

The general theory of polar models developed in [12] was applied in [13] to the analysis of the polar model of shear deformable beams undergoing finite configuration changes.

Here we investigate in detail a polar model of shear deformable shells in finite deformations which is referred to in the literature as the shell without drilling rotations [7].

Let E^3 be euclidean space and V^3 the associated linear space of translations.

The material shell $\mathcal{B}$ is a set of particles which, at each time $t \in I$, are located at points of a differentiable submanifold of the physical space $\mathbb{E} = E^3$.

The polar model of a shell without drilling rotations is a two-dimensional structural model characterized by a middle surface $\mathbb{B}$ and by vectors of prescribed length in V^3 attached at each of its points to simulate the constant thickness of the transversely undeformable shell. The corresponding versors, called directors, range over the unit sphere $\mathbf{S}^2$ which is a compact differentiable manifold without boundary embedded

in V^3,

$$\mathbf{S}^2 := (\mathbf{d} \in V^3 \,:\, \|\mathbf{d}\| = 1)\,.$$

The ambient space, in which the motion of the shell takes place, is then the differentiable manifold without boundary

$$\mathbb{S} = E^3 \times \mathbf{S}^2\,,$$

a trivial fiber bundle having the euclidean space E^3 as base manifold and the unit sphere $\mathbf{S}^2$ as typical fiber.

The base configuration map $\boldsymbol{\chi}_t : \mathcal{B} \mapsto \mathbb{E}$ of the shell at time $t \in I$ is a bijection of the material shell $\mathcal{B}$ onto the base placement $\mathbb{B}_t \subset \mathbb{E}$ which is the middle surface of the shell.

The polar structure $\mathbf{s}_t : \mathbb{B}_t \mapsto \mathbb{S}$ is a map from the middle surface at time t onto the placement $\mathbb{P}_t = \mathbf{s}_t(\mathbb{B}_t)$. The map $\mathbf{s}_t : \mathbb{B}_t \mapsto \mathbb{S}$, defined by

$$\mathbf{s}_t(\mathbf{p}_t) := \{\,\mathbf{p}_t\,,\mathbf{d}_t\,\} \in \mathbb{B}_t \times \mathbf{S}^2\,,$$

is a section of the fiber bundle $\mathbb{S}$ on the submanifold $\mathbb{B}_t \subset \mathbb{E}$.

A spatial configuration of the polar shell at time $t \in I$ is an injective map $\mathbf{u}_t : \mathcal{B} \mapsto \mathbb{S}$ which assigns a placement $\mathbb{P}_t := \mathbf{u}_t(\mathcal{B}) \subset \mathbb{S}$ to the material shell $\mathcal{B}$ and is given by the composition of the base configuration map with the polar structure

$$\mathbf{u}_t = \mathbf{s}_t \circ \boldsymbol{\chi}_t\,.$$

We consider the change of base configuration $\boldsymbol{\chi}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s\,;\mathbb{B}_t)$ from $\boldsymbol{\chi}_s$ to $\boldsymbol{\chi}_t$ defined by

$$\boldsymbol{\chi}_{t,s} \circ \boldsymbol{\chi}_s = \boldsymbol{\chi}_t\,.$$

The configuration change from $\mathbf{u}_s$ to $\mathbf{u}_t$ is the map $\mathbf{u}_{t,s} : \mathbf{u}_s(\mathcal{B}) \mapsto \mathbf{u}_t(\mathcal{B}) \subset \mathbb{S}$ defined by

$$\mathbf{u}_{t,s} \circ \mathbf{u}_s = \mathbf{u}_t\,.$$

To extract the base point and the director from a pair $\{\,\mathbf{p}_t, \mathbf{d}_t\,\}$ we introduce the cartesian projectors

$$\mathbf{P}_1\{\,\mathbf{p}_t, \mathbf{d}_t\,\} := \mathbf{p}_t\,, \qquad \mathbf{P}_2\{\,\mathbf{p}_t, \mathbf{d}_t\,\} := \mathbf{d}_t\,.$$

Accordingly we define the map $\hat{\mathbf{d}}_t : \mathbb{B}_t \mapsto S^2$, which provides the director associated to a base point on the middle surface,

$$\hat{\mathbf{d}}_t(\mathbf{p}_t) := (\mathbf{P}_2 \circ \mathbf{s}_t)(\mathbf{p}_t)\,, \qquad \mathbf{p}_t \in \mathbb{B}_t\,.$$

To simplify the notation we drop the $\hat{}$ and simply write $\mathbf{d}_t$ for $\hat{\mathbf{d}}_t$.

We consider the finite deformation measure for the polar shell model without drilling rotations that was proposed and analyzed in [7]. It consists of the triplet

$$\mathbf{A}(\mathbf{u}_{t,s}) := \begin{vmatrix} \boldsymbol{\varepsilon}(\boldsymbol{\chi}_{t,s}) \\ \boldsymbol{\delta}(\mathbf{u}_{t,s}) \\ \mathbf{C}(\mathbf{u}_{t,s}) \end{vmatrix}$$

composed of

$$\begin{aligned} \boldsymbol{\varepsilon}(\boldsymbol{\chi}_{t,s})(\mathbf{a},\mathbf{b}) &:= \mathbf{g}(\boldsymbol{\chi}_{t,s*}\mathbf{a},\, \boldsymbol{\chi}_{t,s*}\mathbf{b}) - \mathbf{g}(\mathbf{a},\,\mathbf{b})\,, && \text{membrane strain}\,, \\ \boldsymbol{\delta}(\mathbf{u}_{t,s})(\mathbf{a}) &:= \mathbf{g}(\mathbf{d}_t,\, \boldsymbol{\chi}_{t,s*}\mathbf{a}) - \mathbf{g}(\mathbf{d}_s,\,\mathbf{a})\,, && \text{shear sliding}\,, \\ \mathbf{C}(\mathbf{u}_{t,s})(\mathbf{a},\mathbf{b}) &:= \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\,\mathbf{d}_t,\, \boldsymbol{\chi}_{t,s*}\mathbf{b}) - \mathbf{g}(\partial_{\mathbf{a}}\mathbf{d}_s,\,\mathbf{b})\,, && \text{flexural curvature}\,, \end{aligned}$$

where $\mathbf{a}\,,\mathbf{b} \in \mathbb{T}_{\mathbb{B}_s}(\mathbf{p}_s)$. The push forward $\boldsymbol{\chi}_{t,s*} \in BL\,(\mathbb{T}_{\mathbb{B}_s}\,;\mathbb{T}_{\mathbb{B}_t})$ associated with the map $\boldsymbol{\chi}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s,\mathbb{B}_t)$ is defined (see [1–3]) by

$$\boldsymbol{\chi}_{t,s*}(\mathbf{p}_s,\mathbf{a}) := \{\boldsymbol{\chi}_{t,s}(\mathbf{p}_s)\,,\,\partial_{\mathbf{a}}\boldsymbol{\chi}_{t,s}(\mathbf{p}_s)\}\,.$$

The push forward maps a given tangent vector applied at a point of a manifold into the corresponding deformed tangent vector applied to the transformed point. The tangent space at $\{\mathbf{x}\,,\mathbf{d}\} \in \mathbb{S} = E^3 \times \mathbf{S}^2$ is the product manifold

$$\mathbb{T}_{\mathbb{S}}(\mathbf{x}\,,\mathbf{d}) = \mathbb{T}_{E^3}(\mathbf{x}) \times \mathbb{T}_{\mathbf{S}^2}(\mathbf{d}) = V^3 \times \mathbb{T}_{\mathbf{S}^2}(\mathbf{d})\ .$$

The virtual displacements $\delta\mathbf{u}_{t,s} \in H^k(\mathbb{B}_s\,;\mathbb{T}_{\mathbb{S}})$ are defined by

$$\delta\mathbf{u}_{t,s}(\mathbf{p}_s) = \{\mathbf{t}(\mathbf{u}_{t,s}(\mathbf{p}_s))\,,\mathbf{X}(\mathbf{u}_{t,s}(\mathbf{p}_s))\} \quad \text{with} \quad \begin{cases} \mathbf{t}\,(\mathbf{u}_{t,s}(\mathbf{p}_s)) & \in \mathbb{T}_{E^3}(\mathbf{p}_t)\,, \\ \mathbf{X}(\mathbf{u}_{t,s}(\mathbf{p}_s)) & \in \mathbb{T}_{\mathbf{S}^2}(\mathbf{d}_t) \end{cases}$$

for any $\mathbf{p}_s \in \mathbb{B}_s$ and $\{\,\mathbf{p}_t,\mathbf{d}_t\,\} = \mathbf{u}_{t,s}(\mathbf{p}_s)$, where $\mathbf{u}_{t,s}(\mathbf{p}_s)$ is an abbreviation for $(\mathbf{u}_{t,s} \circ \mathbf{s}_s)(\mathbf{p}_s)$.

Remark 1. We observe that, despite their wide acceptance (see, e.g., [4–6]) the deformation measures reported above in this section and commonly adopted in the literature for polar shells without drilling rotations, lead to physically implausible results in the case of significant membrane strains. Indeed a simple computation reveals an unrealistic behavior of an inflated polar spherical baloon since an increase of flexural curvature is measured when the radius increases. The effect is due to the amplification of the convected tangent vectors due to the deformation.

To eliminate this shortcoming we redefine the deformation measures for polar shells without drilling rotations as

$$\begin{aligned} \boldsymbol{\varepsilon}(\boldsymbol{\chi}_{t,s})(\mathbf{a},\mathbf{b}) &:= \mathbf{g}(\boldsymbol{\chi}_{t,s*}\mathbf{a},\, \boldsymbol{\chi}_{t,s*}\mathbf{b}) - \mathbf{g}(\mathbf{a},\,\mathbf{b})\,, && \text{membrane strain}\,, \\ \boldsymbol{\delta}(\mathbf{u}_{t,s})(\mathbf{a}) &:= \mathbf{g}(\mathbf{d}_t,\, \boldsymbol{\chi}_{t,s*}\mathbf{a}) - \mathbf{g}(\mathbf{d}_s,\,\mathbf{a})\,, && \text{shear sliding}\,, \\ \mathbf{C}(\mathbf{u}_{t,s})(\mathbf{a},\mathbf{b}) &:= \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\,\mathbf{d}_t,\, \mathbf{R}_{t,s}\mathbf{b}) - \mathbf{g}(\partial_{\mathbf{a}}\mathbf{d}_s,\,\mathbf{b})\,, && \text{curvature change}\,, \end{aligned}$$

where $\mathbf{R}_{t,s}$ is the isometric transformation associated with the push forward $\boldsymbol{\chi}_{t,s*}$ according to the polar decomposition formula $\boldsymbol{\chi}_{t,s*} = \mathbf{R}_{t,s}\,\mathbf{U}_{t,s}$ where $\mathbf{U}_{t,s}$ is the right Cauchy stretch tensor. The new expression for the curvature change correctly predicts no flexural curvature in the inflated polar spherical baloon when the radius is changed. Indeed in this problem the rotation $\mathbf{R}_{t,s}$ reduces to the identity and $\mathbf{d}_t \circ \boldsymbol{\chi}_{t,s} = \mathbf{d}_s$ so that

$$(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\,\mathbf{d}_t) \circ \boldsymbol{\chi}_{t,s} = \partial_{\mathbf{a}}\mathbf{d}_s\,.$$

The computation of the tangent stiffness for this new shell model is dealt with in a forthcoming paper.

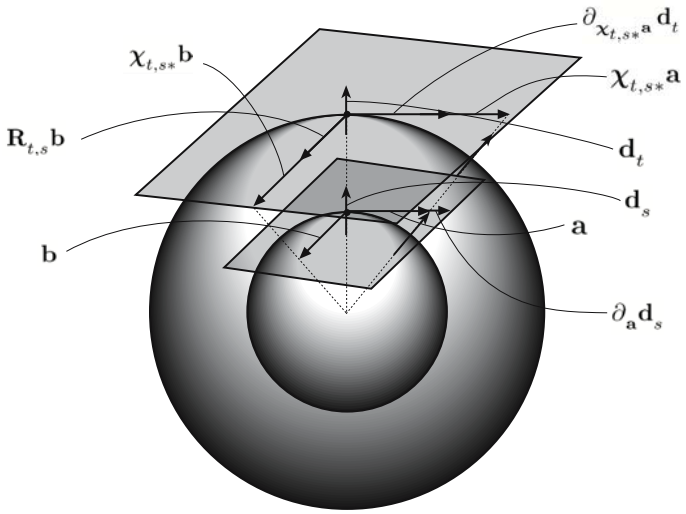


Fig. 1. Inflated polar spherical baloon

2.1 Tangent stiffness

Let $\mathbf{v}_{\mathbf{X}} := \{\mathbf{t}_{\mathbf{X}}\,, \mathbf{X}\}$ and $\mathbf{v}_{\mathbf{Y}} := \{\mathbf{t}_{\mathbf{Y}}\,, \mathbf{Y}\}$ be referential virtual displacements at the placement $\mathbb{P}_t$. For any $\mathbf{p}_s \in \mathbb{B}_s$ the position at time t is given by $\{\mathbf{x}_t, \mathbf{d}_t\} = \mathbf{u}_{t,s}(\mathbf{p}_s) \in \mathbb{P}_t$, and hence the referential virtual displacements are functions of the point $\mathbf{p}_s \in \mathbb{B}_s$ and of the configuration change $\mathbf{u}_{t,s} \in \mathrm{C}^k(\mathbb{B}_s, \mathbb{S})$. To simplify the notation we write $\mathbf{v}_{\mathbf{X}}$ or $\mathbf{v}_{\mathbf{X}}(\mathbf{u}_{t,s})$, dropping the explicit dependence on $\mathbf{p}_s \in \mathbb{B}_s$.

The constitutive tangent stiffness of the shell is evaluated by taking the directional derivative of the elastic potential along a virtual displacement and by subsequently taking the absolute time derivative of the directional derivative. As we have seen, by applying the Leibniz rule the tangent stiffness is decomposed as the sum of an elastic part and a geometric part.

The symmetric elastic tangent stiffness is the bilinear form in $\mathbf{v}_{\mathbf{X}}, \mathbf{v}_{\mathbf{Y}}$ given by the formula

$$\partial^2\varphi(\mathbf{A}(\mathbf{u}_{t,s})) \cdot (\partial\mathbf{A}(\mathbf{u}_{t,s}) \cdot \mathbf{v}_{\mathbf{Y}}) \cdot (\partial\mathbf{A}(\mathbf{u}_{t,s}) \cdot \mathbf{v}_{\mathbf{X}}),$$

where the virtual displacement $\mathbf{v}_{\mathbf{X}}$ is indeed the velocity vector along the equilibrium path, which is the unknown of the incremental elastic equilibrium problem.

The geometric tangent stiffness is the bilinear form in $\mathbf{v}_{\mathbf{X}}, \mathbf{v}_{\mathbf{Y}}$ given by

$$\partial\varphi(\mathbf{A}(\mathbf{u}_{t,s})) \cdot \left[\nabla^2_{\mathbf{v}_{\mathbf{X}}\mathbf{v}_{\mathbf{Y}}}(\mathbf{A}(\mathbf{u}_{t,s}))\right] = \partial\varphi(\mathbf{A}(\mathbf{u}_{t,s})) \cdot \left[(\partial_{\mathbf{v}_{\mathbf{X}}}\partial_{\hat{\mathbf{v}}_{\mathbf{Y}}} - \partial_{\nabla_{\mathbf{v}_{\mathbf{X}}}\hat{\mathbf{v}}_{\mathbf{Y}}})(\mathbf{A}(\mathbf{u}_{t,s}))\right].$$

To compute the second covariant derivative of the deformation measure it is tempting to choose a connection on the space manifold. Such a choice determines whether symmetry of the geometric tangent stiffness is ensured or not. Indeed a torsionless connection implies the symmetry of the second covariant derivative of the deformation measure and hence the symmetry of the geometric tangent stiffness. On the other hand, if the connection is not symmetric, the second covariant derivative can fail to be symmetric.

To provide a symmetric expression of the Hessian of the deformation measure, we assume that the manifold $\mathbb{S} = E^3 \times \mathbf{S}^2$ is endowed with the Riemannian metric $\mathbf{g} \in BL(\mathbb{T}_{\mathbb{S}}, \mathbb{T}_{\mathbb{S}}; \mathcal{R})$ induced by the usual metric in E^3. The Levi–Civita connection ∇ on $\{\mathbb{S}, \mathbf{g}\}$ is uniquely defined by the requirements that it is metric and torsionless:

i) $\partial_{\mathbf{c}}(\mathbf{g}(\mathbf{a}, \mathbf{b})) = \mathbf{g}(\nabla_{\mathbf{c}}\mathbf{a}, \mathbf{b}) + \mathbf{g}(\mathbf{a}, \nabla_{\mathbf{c}}\mathbf{b})$,

ii) $\boldsymbol{T}(\mathbf{a}, \mathbf{b}) := \nabla_{\mathbf{a}}\mathbf{b} - \nabla_{\mathbf{b}}\mathbf{a} - [\mathbf{a}, \mathbf{b}] = \mathbf{o}$,

where $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathrm{C}^1(\mathbb{S}; \mathbb{T}_{\mathbb{S}})$ are spatial vector fields.

The covariant derivative on $\mathbf{S}^2$ corresponding to this natural choice of the connection, can easily be computed as the projection of the directional derivative in E^3 on the tangent space to $\mathbf{S}^2$. Alternatively recourse can be made to the general formula due to Koszul [9]:

$$2\,\mathbf{g}(\nabla_{\mathbf{a}}\mathbf{b}, \mathbf{c}) = d_{\mathbf{a}}(\mathbf{g}(\mathbf{b}, \mathbf{c})) + d_{\mathbf{b}}(\mathbf{g}(\mathbf{c}, \mathbf{a})) - d_{\mathbf{c}}(\mathbf{g}(\mathbf{a}, \mathbf{b})) + \mathbf{g}([\mathbf{a}, \mathbf{b}], \mathbf{c}) + \\ -\mathbf{g}([\mathbf{b}, \mathbf{c}], \mathbf{a}) + \mathbf{g}([\mathbf{c}, \mathbf{a}], \mathbf{b}).$$

This more involved procedure, which requires the computation of the Lie brackets appearing in the last three terms, was adopted in [8].

The evaluation of both terms of the right-hand side in the expression of the second covariant derivative $\nabla^2_{\mathbf{v}_{\mathbf{X}}\mathbf{v}_{\mathbf{Y}}}(\mathbf{A}(\mathbf{u}_{t,s}))$ requires an extension $\hat{\mathbf{v}}_{\mathbf{Y}} := \{\hat{\mathbf{t}}_{\mathbf{Y}}, \hat{\mathbf{Y}}\}$ of the virtual displacement $\mathbf{v}_{\mathbf{Y}} := \{\mathbf{t}_{\mathbf{Y}}, \mathbf{Y}\}$ along virtual trajectories in the physical space. However, as we show, the second covariant derivative does not depend on how the extension is performed. Note that the extension of the vector $\mathbf{t}_{\mathbf{Y}}$ is trivial

and consists in assuming it to be constant in the affine euclidean space E^3. On the other hand, different extensions of the virtual displacement $\mathbf{Y}$ tangent to $\mathbf{S}^2$ at $\mathbf{d}_t$ change the second directional derivative while the second covariant derivative of the deformation measure is unchanged.

2.2 Extensions of the virtual displacements

We consider two extensions of the virtual displacement. The covariant derivative of the virtual displacement and the second directional derivative of the strain measure assume different expressions corresponding to the two extensions. In any case, as is to be expected from the general results, the same expression is obtained for the geometric tangent stiffness which is symmetric since the relevant connection is torsionless as it is induced by a Riemannian metric.

First extension. We first recall that, for any $\mathbf{p}_s \in \mathbb{B}_s$, we have $\{\mathbf{p}_t, \mathbf{d}_t\} = \mathbf{u}_{t,s}(\mathbf{p}_s) \in \mathbb{P}_t$. The tangent vectors $\mathbf{X}(\mathbf{u}_{t,s}), \mathbf{Y}(\mathbf{u}_{t,s}) \in \mathbb{T}_{\mathbf{S}^2}(\mathbf{d}_t)$ can be expressed as

$$\mathbf{X}(\mathbf{u}_{t,s}) = \mathbf{W}_{\mathbf{X}}\,\mathbf{P}_2\,\mathbf{u}_{t,s} = \mathbf{W}_{\mathbf{X}}\,\mathbf{d}_t = \boldsymbol{\omega}_{\mathbf{X}} \times \mathbf{d}_t\,,$$
$$\mathbf{Y}(\mathbf{u}_{t,s}) = \mathbf{W}_{\mathbf{Y}}\,\mathbf{P}_2\,\mathbf{u}_{t,s} = \mathbf{W}_{\mathbf{Y}}\,\mathbf{d}_t = \boldsymbol{\omega}_{\mathbf{Y}} \times \mathbf{d}_t\,,$$

where $\mathbf{W}_{\mathbf{X}}$ and $\mathbf{W}_{\mathbf{Y}}$ are semisymmetric tensors in V^3 characterized by axial vectors $\boldsymbol{\omega}_{\mathbf{X}}$ and $\boldsymbol{\omega}_{\mathbf{Y}}$ which are assumed to be orthogonal to $\mathbf{d}_t$.

We now consider a virtual trajectory $\mathbf{u}_{\tau,t} \in \mathrm{C}^k(\mathbb{B}_t, \mathbb{P}_\tau)$ starting at $\mathbb{P}_t$ and having velocity $\mathbf{v}_{\mathbf{X}}(\mathbf{u}_{t,s}) \in H^k(\mathbb{B}_s; \mathbb{T}_{\mathbb{S}})$ at time t. We may choose the following extension for the virtual displacement $\mathbf{v}_{\mathbf{Y}}(\mathbf{u}_{t,s}) = \{\mathbf{t}_{\mathbf{Y}}(\mathbf{u}_{t,s}), \mathbf{Y}(\mathbf{u}_{t,s})\}$:

$$\begin{cases} \hat{\mathbf{t}}_{\mathbf{Y}}(\mathbf{u}_{\tau,s}) := \mathbf{t}_{\mathbf{Y}}(\mathbf{u}_{t,s})\,, \\ \hat{\mathbf{Y}}(\mathbf{u}_{\tau,s}) \;\; := \mathbf{W}_{\mathbf{Y}}\,\mathbf{d}_\tau = \boldsymbol{\omega}_{\mathbf{Y}} \times \mathbf{d}_\tau\,, \end{cases}$$

where $\mathbf{u}_{\tau,s} = \mathbf{u}_{\tau,t} \circ \mathbf{u}_{t,s}$. Since the vector field $\hat{\mathbf{t}}_{\mathbf{Y}}(\mathbf{u}_{\tau,s})$ is taken constant in V^3 along the virtual trajectory, the evaluation of the covariant derivative of $\hat{\mathbf{v}}_{\mathbf{Y}} = \{\hat{\mathbf{t}}_{\mathbf{Y}}, \hat{\mathbf{Y}}\}$ at $\{\mathbf{x}_t, \mathbf{d}_t\}$ along $\mathbf{v}_{\mathbf{X}} = \{\mathbf{t}_{\mathbf{X}}, \mathbf{X}\}$ amounts to computing the covariant derivative of $\hat{\mathbf{Y}}(\mathbf{u}_{\tau,s})$ at $\mathbf{u}_{t,s}$ along $\mathbf{X}(\mathbf{u}_{t,s})$. To this end we observe that

$$\partial_{\mathbf{X}}\,\mathbf{d}_t = \frac{\partial}{\partial \tau}\bigg|_{\tau=t} \mathbf{P}_2 \circ \mathbf{u}_{\tau,s} = \mathbf{P}_2 \circ \mathbf{v}_{\mathbf{X}} = \mathbf{X}\,.$$

The directional derivative is then given by

$$\begin{aligned} (\partial_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s}) &= \mathbf{W}_{\mathbf{Y}}\mathbf{W}_{\mathbf{X}}\mathbf{d}_t = \boldsymbol{\omega}_{\mathbf{Y}} \times (\boldsymbol{\omega}_{\mathbf{X}} \times \mathbf{d}_t) = \\ &= \mathbf{g}(\boldsymbol{\omega}_{\mathbf{Y}}, \mathbf{d}_t)\,\boldsymbol{\omega}_{\mathbf{X}} - \mathbf{g}(\boldsymbol{\omega}_{\mathbf{Y}}, \boldsymbol{\omega}_{\mathbf{X}})\,\mathbf{d}_t = \\ &= -\mathbf{g}(\boldsymbol{\omega}_{\mathbf{Y}}, \boldsymbol{\omega}_{\mathbf{X}})\,\mathbf{d}_t = -\mathbf{g}(\mathbf{X}, \mathbf{Y})\,\mathbf{d}_t\,, \end{aligned}$$

since $\mathbf{g}(\boldsymbol{\omega}_{\mathbf{Y}}, \mathbf{d}_t) = 0$ by assumption.

With $\boldsymbol{\Pi}$ denoting the orthogonal projector in E^3 on the tangent space $\mathbb{T}_{\mathbf{S}^2}(\mathbf{d}_t)$ at the point $\mathbf{d}_t$, the formula of the covariant derivative yields

$$(\nabla_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s}) = \boldsymbol{\Pi}\,(\partial_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s}) = -\mathbf{g}(\boldsymbol{\omega}_{\mathbf{Y}}\,,\boldsymbol{\omega}_{\mathbf{X}})\,\boldsymbol{\Pi}\,\mathbf{d}_t = \mathbf{o}$$
$$\forall\,\mathbf{X}\in\mathbb{T}_{\mathbf{S}^2}(\mathbf{d}_t)\,,$$

since $\boldsymbol{\Pi}\,\mathbf{d}_t = \mathbf{o}$.

As a consequence $(\nabla\hat{\mathbf{v}}_{\mathbf{Y}})(\mathbf{u}_{t,s}) = \mathbf{o}$ and the second covariant derivative of the deformation measure at $\mathbf{u}_{t,s}$ coincides with the second directional derivative, that is,

$$\nabla^2_{\mathbf{v_X}\,\mathbf{v_Y}}(\mathbf{A}(\mathbf{u}_{t,s}))(\mathbf{d}_t) = \partial_{\mathbf{v_X}}\partial_{\hat{\mathbf{v}}_{\mathbf{Y}}}(\mathbf{A}(\mathbf{u}_{t,s}))(\mathbf{d}_t)\,.$$

We then compute the second directional derivative of the components of the strain measure. To this end we first observe that

$$\partial_{\mathbf{t_X}}\boldsymbol{\chi}_{t,s} = \mathbf{t}_{\mathbf{X}}\,,\qquad \boldsymbol{\chi}_{t,s*}\mathbf{a} = \partial_{\mathbf{a}}\boldsymbol{\chi}_{t,s}\,,$$
$$\partial_{\mathbf{t_X}}\partial_{\mathbf{a}}\boldsymbol{\chi}_{t,s} = \partial_{\mathbf{a}}\partial_{\mathbf{t_X}}\boldsymbol{\chi}_{t,s} = \partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}\,.$$

The second directional derivative of the membrane strain yields, for $\mathbf{a}\,,\mathbf{b}\in\mathbb{T}_{\mathbb{B}_s}$, the expression

$$\begin{aligned}\partial_{\mathbf{v_X}}\partial_{\mathbf{v_Y}}\Big[\boldsymbol{\varepsilon}(\boldsymbol{\chi}_{t,s})(\mathbf{a},\mathbf{b})\Big] &= \partial_{\mathbf{v_X}}\partial_{\mathbf{v_Y}}\Big[\mathbf{g}(\boldsymbol{\chi}_{t,s*}\mathbf{a},\,\boldsymbol{\chi}_{t,s*}\mathbf{b}) - \mathbf{g}(\mathbf{a},\,\mathbf{b})\Big]\\ &= \partial_{\mathbf{v_X}}\Big[\mathbf{g}(\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}\,,\boldsymbol{\chi}_{t,s*}\mathbf{b}) + \mathbf{g}(\boldsymbol{\chi}_{t,s*}\mathbf{a}\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}})\Big]\\ &= \mathbf{g}(\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{X}}) + \mathbf{g}(\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}})\,,\end{aligned}$$

which is clearly symmetric in $\mathbf{X},\mathbf{Y}$.

The second directional derivatives of the shear sliding yields, for $\mathbf{a}\in\mathbb{T}_{\mathbb{B}_s}$, the expression

$$\begin{aligned}\partial_{\mathbf{v_X}}\partial_{\mathbf{v_Y}}\Big[\boldsymbol{\delta}(\mathbf{u}_{t,s})(\mathbf{a})\Big] &= \partial_{\mathbf{v_X}}\partial_{\mathbf{v_Y}}\Big[\mathbf{g}(\mathbf{d}_t,\,\boldsymbol{\chi}_{t,s*}\mathbf{a}) - \mathbf{g}(\mathbf{d}_s,\,\mathbf{a})\Big]\\ &= \mathbf{g}(\partial_{\mathbf{X}}\hat{\mathbf{Y}}\,,\boldsymbol{\chi}_{t,s*}\mathbf{a}) + \mathbf{g}(\mathbf{Y}\,,\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}) + \mathbf{g}(\mathbf{X}\,,\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}})\\ &= -\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\mathbf{g}(\mathbf{d}_t\,,\boldsymbol{\chi}_{t,s*}\mathbf{a}) + \mathbf{g}(\mathbf{Y}\,,\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}) + \mathbf{g}(\mathbf{X}\,,\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}})\,.\end{aligned}$$

The second directional derivative of the flexural curvature is given by

$$
\begin{aligned}
&\partial_{\mathbf{v_X}}\partial_{\mathbf{v_Y}}\big[\mathbf{C}(\mathbf{u}_{t,s})(\mathbf{a},\mathbf{b})\big] = \partial_{\mathbf{v_X}}\partial_{\mathbf{v_Y}}\big[\mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{d}_t,\,\boldsymbol{\chi}_{t,s*}\mathbf{b}) - \mathbf{g}(\partial_{\mathbf{a}}\mathbf{d}_s,\,\mathbf{b})\big]\\
&= \partial_{\mathbf{v_X}}\Big[\mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{Y}\,,\boldsymbol{\chi}_{t,s*}\mathbf{b}) + \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{d}_t\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}}) + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}}\mathbf{d}_t\,,\boldsymbol{\chi}_{t,s*}\mathbf{b})\Big]\\
&= \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}(\partial_{\mathbf{X}}\hat{\mathbf{Y}})\,,\boldsymbol{\chi}_{t,s*}\mathbf{b}) + \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{Y}\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{X}}) + \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{X}\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}})\\
&\qquad + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}}\mathbf{Y}\,,\boldsymbol{\chi}_{t,s*}\mathbf{b}) + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}}\mathbf{d}_t\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}})\\
&\qquad + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}}\mathbf{X}\,,\boldsymbol{\chi}_{t,s*}\mathbf{b}) + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}}\mathbf{d}_t\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{X}})\,.\\
&= -\mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}(\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\mathbf{d}_t)\,,\boldsymbol{\chi}_{t,s*}\mathbf{b}) + \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{Y},\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{X}}) + \mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{X},\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}})\\
&\qquad + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}}\mathbf{Y}\,,\boldsymbol{\chi}_{t,s*}\mathbf{b}) + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}}\mathbf{d}_t\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}})\\
&\qquad + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}}\mathbf{X}\,,\boldsymbol{\chi}_{t,s*}\mathbf{b}) + \mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}}\mathbf{d}_t\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{X}})\,.
\end{aligned}
$$

From the expressions above it is apparent that the second directional derivatives of the shear sliding and of the flexural curvature are symmetric with respect to exchanging of $\mathbf{X}$ and $\mathbf{Y}$, as was to be expected. Indeed the second directional derivative coincides with the second covariant derivative for the adopted extension of the virtual displacements.

The same results are obtained by considering another, perhaps simpler, extension for the virtual displacement $\mathbf{Y} = \mathbf{Y}(\mathbf{u}_{t,s})$, defined as:

$$
\hat{\mathbf{Y}}(\mathbf{u}_{\tau,s}) := (\mathbf{I} - \mathbf{d}_\tau \otimes \mathbf{d}_\tau)\,\mathbf{Y}\,, \qquad \mathbf{d}_\tau = \mathbf{P}_2\mathbf{u}_{\tau,s}\,,
$$

so that

$$
\hat{\mathbf{Y}}(\mathbf{u}_{t,s}) = (\mathbf{I} - \mathbf{d}_t \otimes \mathbf{d}_t)\,\mathbf{Y} = \mathbf{Y}\,.
$$

The directional derivative of $\hat{\mathbf{Y}}$ along $\mathbf{X}$ at $\mathbf{u}_{t,s}$ is given by

$$
(\partial_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s}) = -(\mathbf{X}\otimes\mathbf{d}_t + \mathbf{d}_t\otimes\mathbf{X})\,\mathbf{Y} = -\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\mathbf{d}_t
$$

and the covariant derivative by

$$
(\nabla_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s}) = \boldsymbol{\Pi}\,(\partial_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s}) = -\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\boldsymbol{\Pi}\,\mathbf{d}_t = \mathbf{o}\,.
$$

Second extension. We now choose a different extension of the virtual displacement $\mathbf{v}_{\mathbf{Y}}(\mathbf{u}_{t,s})$ by setting

$$
\begin{cases}
\hat{\mathbf{t}}_{\mathbf{Y}}(\mathbf{u}_{\tau,s}) := \mathbf{t}_{\mathbf{Y}}(\mathbf{u}_{t,s})\,,\\
\hat{\mathbf{Y}}(\mathbf{u}_{\tau,s}) \;:= \big[1 - \mathbf{g}(\mathbf{d}_\tau\,,\mathbf{Y})\big]\,(\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{d}_\tau)\,, \qquad \mathbf{d}_\tau = \mathbf{P}_2\mathbf{u}_{\tau,s},
\end{cases}
$$

so that, since $\mathbf{g}(\mathbf{d}_t\,,\mathbf{Y})=0$ and $\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{d}_t=\mathbf{Y}$, we have

$$\hat{\mathbf{Y}}(\mathbf{u}_{t,s})=\left[1-\mathbf{g}(\mathbf{d}_t\,,\mathbf{Y})\right](\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{d}_t)=\mathbf{Y}\,.$$

The directional derivative of $\hat{\mathbf{Y}}$ along $\mathbf{X}$ at $\mathbf{d}_t$ is given by

$$\begin{aligned}(\partial_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s}) &= -\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,(\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{d}_t)+\left[1-\mathbf{g}(\mathbf{d}_t\,,\mathbf{Y})\right](\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{X})\\ &= -\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,(\boldsymbol{\omega}_{\mathbf{Y}}\times\mathbf{d}_t)+\boldsymbol{\omega}_{\mathbf{Y}}\times(\boldsymbol{\omega}_{\mathbf{X}}\times\mathbf{d}_t)\\ &= -\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\mathbf{Y}-\mathbf{g}(\boldsymbol{\omega}_{\mathbf{Y}}\,,\boldsymbol{\omega}_{\mathbf{X}})\,\mathbf{d}_t=-\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,(\mathbf{Y}+\mathbf{d}_t)\,,\end{aligned}$$

and the covariant derivative by

$$(\nabla_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s})=\boldsymbol{I\!I}\,(\partial_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s})=-\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\mathbf{Y}\,.$$

The second directional derivative of the shear sliding is now given by

$$-\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\mathbf{g}(\mathbf{Y}+\mathbf{d}_t\,,\boldsymbol{\chi}_{t,s*}\mathbf{a})+\mathbf{g}(\mathbf{Y}\,,\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}})+\mathbf{g}(\mathbf{X}\,,\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}})\,,$$

and that of the flexural curvature by

$$\begin{aligned}&-\mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\left(\mathbf{g}(\mathbf{X}\,,\mathbf{Y})(\mathbf{Y}+\mathbf{d}_t)\right),\boldsymbol{\chi}_{t,s*}\mathbf{b})+\mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{Y}\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{X}})\\ &+\mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}\mathbf{X}\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}})+\mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}}\mathbf{Y}\,,\boldsymbol{\chi}_{t,s*}\mathbf{b})+\mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{X}}}\mathbf{d}_t\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{Y}})\\ &+\mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}}\mathbf{X}\,,\boldsymbol{\chi}_{t,s*}\mathbf{b})+\mathbf{g}(\partial_{\partial_{\mathbf{a}}\mathbf{t}_{\mathbf{Y}}}\mathbf{d}_t\,,\partial_{\mathbf{b}}\mathbf{t}_{\mathbf{X}})\,.\end{aligned}$$

Both these expressions are non-symmetric due to the lack of symmetry of the first terms.

Symmetry is however recovered by taking into account the additional term appearing in the expression of the second covariant derivative of the deformation measure which does not vanish since $(\nabla_{\mathbf{X}}\hat{\mathbf{Y}})(\mathbf{u}_{t,s})=-\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\mathbf{Y}$.

In fact, for the shear sliding we have

$$\partial_{\nabla_{\mathbf{v}_{\mathbf{X}}}\hat{\mathbf{v}}_{\mathbf{Y}}}\left[\boldsymbol{\delta}(\mathbf{u}_{t,s})(\mathbf{a})\right]=-\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\mathbf{g}(\mathbf{Y}\,,\boldsymbol{\chi}_{t,s*}\mathbf{a})\,,$$

and for the flexural curvature

$$\partial_{\nabla_{\mathbf{v}_{\mathbf{X}}}\hat{\mathbf{v}}_{\mathbf{Y}}}\left[\mathbf{C}(\mathbf{u}_{t,s})(\mathbf{a},\mathbf{b})\right]=-\mathbf{g}(\partial_{\boldsymbol{\chi}_{t,s*}\mathbf{a}}(\mathbf{g}(\mathbf{X}\,,\mathbf{Y})\,\mathbf{Y})\,,\boldsymbol{\chi}_{t,s*}\mathbf{b})\,.$$

By subtracting the last two terms from the second directional derivatives we get the symmetric expressions of the second covariant derivatives of the shear sliding and of the flexural curvature, coinciding with the ones found with the first extension.

References

[1] Spivak, M. (1979): A comprehensive introduction to differential geometry. Vols. I-V. Publish or Perish, Wilmington, DE

[2] Marsden, J.E., Hughes, T.J.R. (1983): Mathematical foundations of elasticity. Prentice-Hall, Englewood Cliffs, NJ
[3] Abraham, R., Marsden, J.E., Ratiu, T. (1988): Manifolds, tensor analysis, and applications. 2nd edition. Springer, New York
[4] Simo, J.C., Fox, D.D.. (1989): On a stress resultant geometrically exact shell model. I. Formulation and optimal parametrization. Comput. Methods Appl. Mech. Engrg. **72**, 267-304
[5] Simo, J.C., Fox, D.D., Rifai, M.S. (1989): On a stress resultant geometrically exact shell model. II. The linear theory; computational aspects. Comput. Methods Appl. Mech. Engrg. **73**, 53–92
[6] Simo, J.C., Fox, D.D., Rifai, M.S. (1990): On a stress resultant geometrically exact shell model. III. Computational aspects of the nonlinear theory. Comput. Methods Appl. Mech. Engrg. **79**, 21-70
[7] Simo, J.C., Fox, D.D., Rifai, M.S. (1990): On a stress resultant geometrically exact shell model. IV. Variable thickness shells with through-the-thickness stretching. Comput. Methods Appl. Mech. Engrg. **81**, 91-126
[8] Simo, J.C. (1992): The (symmetric) Hessian for geometrically nonlinear models in solid mechanics: intrinsic definition and geometric interpretation. Comput. Methods Appl. Mech. Engrg. **96**, 189-200
[9] Petersen, P. (1998): Riemannian geometry. Springer, New York
[10] Romano, G. (2003): Scienza delle costruzioni. Tomo II. Hevelius, Benevento
[11] Romano, G., Diaco, M., Romano, A., Sellitto, C. (2003): When and why a nonsymmetric tangent stiffness may occur. AIMETA Congress of Theoretical and Applied Mechanics. Ferrara, Italy, Sept. 9-12, 2003
[12] Romano, G., Diaco, M., Sellitto, C. (2004): Tangent stiffness of elastic continua on manifolds. In: Romano, G., Rionero, S. (eds.): Recent trends in the applications of mathematics to mechanics. Springer, Berlin, pp. 155–184
[13] Diaco, M., Romano, A., Sellitto, C. (2004): Tangent stiffness of a Timoshenko beam undergoing large displacements. In: Romano, G., Rionero, S. (eds.): Recent trends in the applications of mathematics to mechanics. Springer, Berlin, pp. 49–66

Global existence of smooth solutions and stability of the constant state for dissipative hyperbolic systems with applications to extended thermodynamics

T. Ruggeri

Abstract. The entropy principle plays an important role in hyperbolic systems of balance laws: symmetrization, principal subsystems and nesting theories, equilibrium manifold. After a brief survey on these questions we present recent results concerning the local and global well-posedness of the Cauchy problem for smooth solutions with particular attention to the *genuine coupling* Kawashima condition. These results are applied to the case of extended thermodynamics and we prove that the K-condition is satisfied in the case of the 13-moment Grad theory with the consequence that there exist global smooth solutions for small initial data and the solutions converge to constant equilibrium states.

1 Introduction

Recently, non-equilibrium theories and in particular extended thermodynamics have generated a new interest in quasi-linear hyperbolic systems of balance laws with dissipation due to the presence of production terms (systems with relaxation). On this subject it is very important to find connections between properties of the full system and the associated subsystem obtained when certain parameters (relaxation coefficients) are equal to zero. The requirement that the system of balance laws satisfies an entropy principle with a convex entropy density gives strong restrictions. In fact, as is well-known, it was shown that eyery system of balance laws can be put into a very special hyperbolic symmetric system, given the introduction of the main field variables [1,2]. As was observed by Boillat and Ruggeri [3], the main field also allows us to recognize that non-equilibrium systems have the structure of nesting theories. In fact it is possible to define the principal subsystems so obtained by freezing those components of the main field which preserve the existence of a convex entropy law and for which the spectrum of the characteristic eigenvalues is contained in that of the full system (sub-characteristic conditions). A particular subsystem is the equilibrium subsystem. Here we give a brief summary on these results with a particular attention to the local and global well-posedness of the relative Cauchy problem for smooth solutions and to the stability of constant solutions. At the end we apply our results to the extended thermodynamics which governs the processes of rarefied gas.

2 Systems of balance laws, entropy and generators

We consider a general hyperbolic system of N balance laws:

$$\partial_\alpha \mathbf{F}^\alpha(\mathbf{u}) = \mathbf{F}(\mathbf{u}), \tag{1}$$

where the *densities* $\mathbf{F}^o$, the *fluxes* $\mathbf{F}^i$ and the *productions* $\mathbf{F}$ are $\mathcal{R}^N$-column vectors depending on the space variables x^i, $(i = 1, 2, 3)$ and the time $t = x^o$, $(\alpha = 0, 1, 2, 3;\ \partial_\alpha = \partial/\partial x^\alpha)$ through the field $\mathbf{u} \equiv \mathbf{u}(x^\alpha) \in \mathcal{R}^N$.

Now we assume, following Friedrichs and Lax [4], that the system (1) satisfies an entropy principle, i.e., there exists an entropy density $-h^o(\mathbf{u})$ and an entropy flux $-h^i(\mathbf{u})$, such that every solution of (1) also satisfies a new balance law (entropy law):

$$\partial_\alpha h^\alpha = \Sigma \leq 0 \tag{2}$$

with a non-negative entropy production $-\Sigma(\mathbf{u})$. The compatibility between (1) and (2) implies the existence of a *main field* $\mathbf{u}'$ such that [4,2]

$$\partial_\alpha h^\alpha - \Sigma \equiv \mathbf{u}' \cdot (\partial_\alpha \mathbf{F}^\alpha - \mathbf{F}) . \tag{3}$$

As a consequence of the above identity, we have

$$dh^\alpha = \mathbf{u}' \cdot d\mathbf{F}^\alpha, \qquad \Sigma = \mathbf{u}' \cdot \mathbf{F} \leq 0. \tag{4}$$

Boillat [1] (in a covariant formulation see Ruggeri and Strumia [2]) was able to introduce four potentials h'^α:

$$h'^\alpha = \mathbf{u}' \cdot \mathbf{F}^\alpha - h^\alpha, \tag{5}$$

such that from $(4)_1$

$$\mathbf{F}^\alpha = \frac{\partial h'^\alpha}{\partial \mathbf{u}'}. \tag{6}$$

It follows that, upon selecting the main field as the field variables, the original system (1) can be written with Hessian matrices in the symmetric form

$$\partial_\alpha \left(\frac{\partial h'^\alpha}{\partial \mathbf{u}'} \right) = \mathbf{F} \quad \Longleftrightarrow \quad \frac{\partial^2 h'^\alpha}{\partial \mathbf{u}' \partial \mathbf{u}'} \partial_\alpha \mathbf{u}' = \mathbf{F} \tag{7}$$

provided that h^0 is a convex function of $\mathbf{u} \equiv \mathbf{F}^0$ (or equivalently the Legendre transform h'^0 is a convex function of the dual field $\mathbf{u}'$). The Euler equations were already written in this form by Godunov [5].

3 Principal subsystems

We split the main field $\mathbf{u}' \in \mathcal{R}^N$ into two parts, $\mathbf{u}' \equiv (\mathbf{v}', \mathbf{w}')$, $\mathbf{v}' \in \mathcal{R}^M$, $\mathbf{w}' \in \mathcal{R}^{N-M}$ $(0 < M < N)$, so that the system (7) with $\mathbf{F} \equiv (\mathbf{f}, \mathbf{g})$ reads:

$$\partial_\alpha \left(\frac{\partial h'^\alpha(\mathbf{v}', \mathbf{w}')}{\partial \mathbf{v}'} \right) = \mathbf{f}(\mathbf{v}', \mathbf{w}'), \tag{8}$$

$$\partial_\alpha \left(\frac{\partial h'^\alpha(\mathbf{v}', \mathbf{w}')}{\partial \mathbf{w}'} \right) = \mathbf{g}(\mathbf{v}', \mathbf{w}'). \tag{9}$$

Given an assigned value $\mathbf{w}'_*(x^\alpha)$ of $\mathbf{w}'$ (in particular, a constant), we call the system [3]:

$$\partial_\alpha \left(\frac{\partial h'^\alpha(\mathbf{v}', \mathbf{w}'_*)}{\partial \mathbf{v}'} \right) = \mathbf{f}(\mathbf{v}', \mathbf{w}'_*) \tag{10}$$

a principal subsystem of 7. In other words a principal subsystem (there are $2^N - 2$ such subsystems) coincides with the first block of the system (8), (9), where $\mathbf{w}' = \mathbf{w}'_*$. Principal subsystems have two important properties: they also admit a convex sub-entropy law and the spectrum of characteristic velocities is contained in that of the full system (sub-characteristic conditions) [3].

4 Equilibrium subsystem

A particular case of (8), (9) is given when the first M equations are conservation laws, i.e., $\mathbf{f} \equiv 0$. In this case it is possible to define the equilibrium state as usual in thermodynamics.

Definition 16. An equilibrium state is a state for which the entropy production $-\Sigma|_E$ vanishes and hence attains its minimum value.

It is possible to prove the following theorem [3,6].

Theorem 1 (Equilibrium manifold). *In an equilibrium state, under the assumption of dissipative productions, i.e., if*

$$\mathbf{D} = \frac{1}{2} \left\{ \frac{\partial \mathbf{g}}{\partial \mathbf{w}'} + \left(\frac{\partial \mathbf{g}}{\partial \mathbf{w}'} \right)^T \right\} \Bigg|_E \quad \textit{is negative definite}, \tag{11}$$

the production vanishes and the main field components vanish except for the first M. Thus,

$$\mathbf{g}|_E = 0, \qquad \mathbf{w}'|_E = 0. \tag{12}$$

Therefore in the main field components the equilibrium manifold is linear, $\mathbf{w}' = 0$, and this confirms once again the importance of the main field.

There is another important characteristic property of the equilibrium state [7,8].

Theorem 2 (Maximum entropy). *At equilibrium the entropy density $-h$ is maximal, i.e.,*

$$h > h|_E \quad \forall\, \mathbf{u} \neq \mathbf{u}|_E, \quad \textit{where} \quad h|_E = h\left(\mathbf{v}, \mathbf{w}|_E(\mathbf{v})\right).$$

Hence we also find at this general level the well-known thermodynamical statement of maximum entropy in equilibrium.

In the present case, when we limit our attention to the case of one-dimensional space, the system (8), (9) assumes the form:

$$\begin{cases} \mathbf{v}_t + \left(k'_{\mathbf{v}'}\right)_x = 0 \\ \\ \mathbf{w}_t + \left(k'_{\mathbf{w}'}\right)_x = -\mathbf{G}\left(\mathbf{v}', \mathbf{w}'\right)\mathbf{w}' \end{cases} \tag{13}$$

where $\mathbf{v} = h'_{\mathbf{v}'}$, $\mathbf{w} = h'_{\mathbf{w}'}$ and $\mathbf{G}$ is a definite positive $(N-M)\times(N-M)$ matrix.

5 Qualitative analysis

In this section we discuss the importance of the entropy principle to the Cauchy problem.

5.1 Local well-posedness

In the general theory of hyperbolic conservation laws and hyperbolic-parabolic conservation laws, the existence of a strictly convex entropy function is a basic condition for well-posedness. In fact if the fluxes $\mathbf{F}^i$ and the production $\mathbf{F}$ are sufficiently smooth in a suitable convex open set $D \subseteq R^n$, it is well-known that system (1) has a unique local (in time) smooth solution for smooth initial data [4,9,10].

However, in the general case, and even for arbitrarily small and smooth initial data, there is no global continuation for these smooth solutions; continuations may develop singularities, shocks or blow-up in finite time (see, e.g., [11,12]).

On the other hand, in many physical examples, thanks to the interplay between the source term and the hyperbolicity, there exist global smooth solutions for a suitable set of initial data. This is the case for example of the isentropic Euler system with damping. Roughly speaking, for such a system the relaxation term induces a dissipative effect. This effect then competes with the hyperbolicity. If the dissipation is sufficiently strong to dominate the hyperbolicity, the system is *dissipative*, and we observe that the classical solution exists for all time and converges to a constant state. Otherwise, if the dissipation and the hyperbolicity are equally important, we expect that only part of the perturbation diffuses. In the latter case the system is called *of composite type* by Zeng [13].

5.2 The Kawashima condition

In general, there are several ways to identify whether a hyperbolic system with relaxation is dissipative or of composite type. One way is completely parallel to the

case of the hyperbolic-parabolic system, which was discussed first by Kawashima [9] and for this reason it is now called the *Kawashima condition* [14] or *genuine coupling* [8]:

in the equilibrium manifold no characteristic eigenvector is in the null space of $\nabla \mathbf{F}$.

5.3 Global existence and stability of constant state

For dissipative one-dimensional systems (13) satisfying the K-condition it is possible to prove the following global existence theorem due to Hanouzet and Natalini [14].

Theorem 3 (Global existence). *Assume that the system (13) is strictly dissipative and the K-condition is satisfied. Then there exists* $\delta > 0$ *such that, if* $\left\|\mathbf{u}'(x,0)\right\|_2 \leq \delta$, *there is a unique global smooth solution satisfying*

$$\mathbf{u}' \in C^0([0,\infty); H^2(R) \cap C^1([0,\infty); H^1(R)).$$

Moreover Ruggeri and Serre [8] have proved that the constant states are stable.

Theorem 4 (Stability of constant state). *Under natural hypotheses of strongly convex entropy, strict dissipativeness, genuine coupling and "zero mass" initial for the perturbation of the equilibrium variables, the constant solution stabilizes:*

$$\left\|\mathbf{u}(t)\right\|_2 = 0\left(t^{-1/2}\right).$$

The technique employed here via Liapunov function, may look rather classical, involving an "energy' (actually entropy) estimate, plus a compensation term as introduced by Kawashima for other purposes [9],

$$L_\varepsilon(\mathbf{u},\mathbf{p}) = h(\mathbf{u}) + \varepsilon\left\{\frac{1}{2}|\mathbf{p}|^2 - \frac{1}{2}\,\mathbf{p}^T\mathbf{A}\mathbf{v} - \mathbf{p}^T\mathbf{B}\mathbf{w}\right\},$$

where $\mathbf{p}$ is the potential,

$$\mathbf{p}_x = \mathbf{v}\ ;\quad \mathbf{p}_t = k'_{\mathbf{v}'},$$

$\mathbf{A}$ and $\mathbf{B}$ are suitable constants matrices and T denotes the transpose.

This method has the nice feature that it applies to weak entropy solutions. It is therefore valid in the presence of shock waves. Due to the finite propagation velocity of the support of a solution, it is natural to assume that the initial total mass of the conserved components of the unknown vanishes:

$$\int \mathbf{v}_0(x)\,dx = 0.$$

Under this condition, we find a $t^{-1/2}$ decay rate of the L^2-norm of the solution, although the decay can be no better than $t^{-1/4}$ in general, that is, when a non-zero mass is present at the initial time.

In [14] the authors report several examples of dissipative systems satisfying the K-condition: the p-system with damping, the Suliciu model for isothermal viscoelasticity, the Kerr-Debye model in nonlinear electromagnetism and the Jin-Xin relaxation model.

5.4 A counterexample of global existence without the K-condition

Zeng [13] considered a toy model for a vibrational non-equilibrium gas in Lagrangian variables, proving that, also if the system is of composite time, global existence holds. Therefore the K-condition is only a sufficient condition for the global existence of smooth solutions.

An intriguing open problem is whether there exists a weaker K-condition that is also necessary to ensure global solutions. And if there is such a condition, what is its physical meaning so as to consider it as a possible new principle of extended thermodynamics adding to the convexity of entropy [15]?

6 Extended thermodynamics

Kinetic theory describes the state of a rarefied gas through the phase density $f(\mathbf{x}, t, \mathbf{c})$, where $f(\mathbf{x}, t, \mathbf{c})d\mathbf{c}$ is the number density of atoms at point $\mathbf{x}$ and time t that have velocities between $\mathbf{c}$ and $\mathbf{c} + d\mathbf{c}$. The phase density obeys the Boltzmann equation

$$\frac{\partial f}{\partial t} + c^i \frac{\partial f}{\partial x^i} = Q, \tag{14}$$

where Q represents the collisional terms. Most macroscopic thermodynamic quantities are identified as moments of the phase density

$$F_{k_1 k_2 \cdots k_j} = \int f c_{k_1} c_{k_2} \cdots c_{k_j} dc, \tag{15}$$

and, due to the Boltzmann equation (14), the moments satisfy an infinity hierarchy of balance laws in which the flux in one equation becomes the density in the next:

$$\begin{aligned}
&\partial_t F + \partial_i F_i = 0 \\
&\qquad \swarrow \\
&\partial_t F_{k_1} + \partial_i F_{ik_1} = 0 \\
&\qquad \swarrow \\
&\partial_t F_{k_1 k_2} + \partial_i F_{ik_1 k_2} = P_{k_1 k_2} \\
&\qquad \swarrow \\
&\partial_t F_{k_1 k_2 k_3} + \partial_i F_{ik_1 k_2 k_3} = P_{k_1 k_2 k_3} \\
&\vdots \\
&\partial_t F_{k_1 k_2 \ldots k_n} + \partial_i F_{ik_1 k_2 \ldots k_n} = P_{k_1 k_2 \ldots k_n} \\
&\vdots
\end{aligned}$$

Taking into account that $P_{kk} = 0$, we recognize the first five equations as conservation laws which coincide with mass, momentum and energy conservation respectively, while the remaining ones are balance laws.

6.1 The closure of extended thermodynamics

When we cut the hierarchy at the density with tensor of rank n, we have the problem of closure because the last flux and the production terms are not in the list of densities. The idea of rational extended thermodynamics (Müller and Ruggeri [15]) was to view the truncated system as a phenomenological system of continuum mechanics and then to consider the new quantities as constitutive functions:

$$F_{k_1k_2\ldots k_nk_{n+1}} \equiv F_{k_1k_2\ldots k_nk_{n+1}}\left(F, F_{k_1}, F_{k_1k_2}, \ldots F_{k_1k_2\ldots k_n}\right)$$

$$P_{k_1k_2\ldots k_j} \equiv P_{k_1k_2\ldots k_j}\left(F, F_{k_1}, F_{k_1k_2}, \ldots F_{k_1k_2\ldots k_n}\right), \quad 2 \leq j \leq n.$$

In accordance with the continuum theory, the restrictions on the constitutive equations come only from *universal principles*, i.e., *the entropy principle*, *the objectivity Principle* and *causality and stability* (convexity of the entropy).

The restrictions are so strong (in particular the entropy principle) that, at least for processes not too far from the equilibrium, the system is completely closed and in the case of 13 moments the results are in perfect agreement with the kinetic closure procedure proposed by Grad [16].

6.2 Principal subsystems in ET

Now that we have stated that for any n we may use the closure of ET, the following question arises: what relation exists between two closure theories with different indices, a theory S_n and a theory S_m with $n > m$, say? Boillat and Ruggeri [3] have proved the following result.

Theorem 5 (Nesting theories). *For $n > m$, S_m is a principal subsystem of S_n obtained from S_n by setting $u'^{\alpha} = 0,\ \alpha = m+1, ..., n$, and neglecting the corresponding equations for $\alpha = m+1, ..., n$, i.e.,*

$$S_n : \begin{cases} \dfrac{\partial u^a(u'^b, u'^\beta)}{\partial t} + \dfrac{\partial F_i^a(u'^b, u'^\beta)}{\partial x_i} = \Pi^a(u'^b, u'^\beta), \\ \dfrac{\partial u^\alpha(u'^b, u'^\beta)}{\partial t} + \dfrac{\partial F_i^\alpha(u'^b, u'^\beta)}{\partial x_i} = \Pi^\alpha(u'^b, u'^\beta), \end{cases} \tag{16}$$
$$a = 0, \ldots, m, \ \ \alpha = m+1, \ldots, n.$$

$$S_m : \frac{\partial u^a(u'^b, 0)}{\partial t} + \frac{\partial F_i^a(u'^b, 0)}{\partial x_i} = \Pi^a(u'^b, 0). \tag{17}$$

In particular the Euler system becomes the equilibrium subsystem of any ET theory.

6.3 The one-dimensional case of the Grad 13-moment theory

The most simple case of extended thermodynamics is the 13-moment theory ET^{13} known as the Grad system [16]. In the one-dimensional case and with the usual

symbols the equations are:

$$\begin{cases} \dot{\rho} + \rho v_x = 0, \\ \rho \dot{v} + (p - \sigma)_x = 0, \\ \rho \dot{e} + q_x + (p - \sigma) v_x = 0, \\ \tau_\sigma \left[\dot{\sigma} - \frac{8}{15} q_x + \frac{7}{3} \sigma v_x \right] - \frac{4}{3} \mu v_x = -\sigma, \\ \tau_q \left[\dot{q} + \frac{16}{5} q v_x - \frac{7}{2} \left(\frac{p}{\rho} \right)_x \sigma - \frac{1}{\rho} (p + \sigma) \sigma_x + \frac{\sigma}{\rho} p_x \right] + \chi T_x = -q, \end{cases} \tag{18}$$

$$\tau_\sigma = \frac{\mu}{p}, \quad \tau_q = \frac{2}{5} \frac{\chi}{p^2} \rho \theta, \quad \frac{\tau_\sigma}{\tau_q} = \frac{2}{3}, \quad p = \mathcal{R} \rho T, \ \ \mathcal{R} = \frac{k}{m}.$$

The dot indicates the material derivative, $\cdot = \partial_t + v \partial_x$, while ρ, v, p, $e = 3p/(2\rho)$, q, σ are, respectively, mass density, velocity, pressure, internal energy, heat flux and shear stress and k, χ and μ denote the Boltzmann constant, heat conductivity and shear viscosity.

The first three equations are the usual conservation laws of mass, momentum and energy, while the remaining two are the new evolution balance laws corresponding to the non-equilibrium variables q and σ. The last two equations when the relaxation times are negligible reduce to the Navier-Stokes and Fourier equations respectively.

As the system (18) is compatible with a convex entropy principle and is a particular case of (13) we check if in the present case the K-condition holds.

With field $\mathbf{u} \equiv (\rho, v, p, \sigma, q)^T$ the associated eigenvectors $\mathbf{d}$ and eigenvalues λ in equilibrium are:

first and second sound:

$$\mathbf{d}_0^{(1,2,4,5)} \equiv \left(1, \frac{c}{\rho} W, \frac{5}{9} c^2 W^2, -\frac{4}{9} c^2 W^2, \frac{1}{6} c^3 W \left(-9 + 5W^2 \right) \right)^T$$

with

$$W = \frac{(v - \lambda)}{c} \quad \text{roots of} \quad 25W^4 - 78W^2 + 27 = 0,$$

$$c = \sqrt{\frac{5}{3} \mathcal{R} T} \ \text{ the sound velocity,}$$

and the *contact wave*:

$$\mathbf{d}_0^{(3)} \equiv \left(1, 0, \frac{5}{3} \frac{p}{\rho}, \frac{5}{3} \frac{p}{\rho}, 0 \right)^T, \qquad \lambda = v.$$

Taking into account that, in the present case,

$$\nabla\mathbf{F}_0 \equiv \begin{pmatrix} 0\,0\,0 & 0 & 0 \\ 0\,0\,0 & 0 & 0 \\ 0\,0\,0 & 0 & 0 \\ 0\,0\,0 & -\dfrac{1}{\tau_\sigma} & 0 \\ 0\,0\,0 & 0 & -\dfrac{1}{\tau_q} \end{pmatrix},$$

we may easily check that no eigenvector is in the null space of $\nabla\mathbf{F}_0$ and therefore the Kawashima condition is satisfied. For the previous theorems on the qualitative analysis we conclude that, if the initial data are sufficiently small, classical solutions of ET^{13} exist for all time and converge to a constant state of the equilibrium Euler manifold!

Acknowledgments

This paper was supported by MIUR (Progetto di Interesse Nazionale, *Problemi Matematici Non Lineari di Propagazione e Stabilità nei Modelli del Continuo,* Coordinator T. Ruggeri), by GNFM-INDAM, and by Istituto Nazionale di Fisica Nucleare (INFN).

References

[1] Boillat, G. (1974): Sur l'existence et la recherche d'équations de conservation supplémentaires pour les systèmes hyperboliques. C.R. Acad. Sci. Paris Sér. A **278**, 909–912
Boillat, G. (1996): Nonlinear hyperbolic fields and waves. In: Ruggeri, T. (ed.): Recent mathematical methods in nonlinear wave propagation. (Lecture Notes in Mathematics, vol. **1640**). Springer, Berlin, pp. 1–47

[2] Ruggeri, T., Strumia, A. (1981): Main field and convex covariant density for quasi-linear hyperbolic systems. Relativistic fluid dynamics. Ann. Inst. H. Poincaré Sect. A (N.S.) **34**, 65–84

[3] Boillat, G., Ruggeri, T. (1997): Hyperbolic principal subsystems: entropy convexity and subcharacteristic conditions. Arch. Rational Mech. Anal. **137**, 305–320

[4] Friedrichs, K.O., Lax, P.D. (1971): Systems of conservation equations with a convex extension. Proc. Nat. Acad. Sci. USA **68**, 1686–1688

[5] Godunov, S.K. (1961): An interesting class of quasi-linear systems (Russian). Dokl. Akad. Nauk SSSR **139**, 521–523; translated as: Sov. Math. Dokl. **2**, 947–949

[6] Boillat, G., Ruggeri, T. (1998): On the shock structure problem for hyperbolic system of balance laws and convex entropy. Contin. Mech. Thermodyn. **10**, 285–292

[7] Ruggeri, T. (2000): Maximum of entropy density in equilibrium and minimax principle for an hyperbolic system of balance laws. In: Albers, B. (ed.): Contributions to continuum theories. Anniversary volume for Krzysztoff Wilmanski. (WIAS-Report No. **18**). Weierstraß-Institut für Angewandte Analysis und Stochastik, Berlin, pp. 207–214

[8] Ruggeri, T., Serre, D. (2004): Stability of constant equilibrium state for dissipative balance laws system with a convex entropy. Quart. Appl. Math. **62**, 163–179
[9] Kawashima, S. (1987): Large-time behavior of solutions to hyperbolic-parabolic systems of conservation laws and applications. Proc. Roy. Soc. Edinburgh Sect. A **106**, 169–194
[10] Fischer, A.E., Marsden, J.E. (1972): The Einstein evolution equations as a first-order quasi-linear symmetric hyperbolic system. I. Comm. Math. Phys. **28**, 1–38
[11] Majda, A. (1984): Compressible fluid flow and systems of conservation laws in several space variables. Springer, New York
[12] Dafermos, C. (2000): Hyperbolic conservation laws in continuum physics. Springer, Berlin
[13] Zeng, Y. (1999): Gas dynamics in thermal nonequilibrium and general hyperbolic systems with relaxation. Arch. Ration. Mech. Anal. **150**, 225–279
[14] Hanouzet, B., Natalini, R. (2003): Global existence of smooth solutions for partially dissipative hyperbolic systems with a convex entropy. Arch. Rat. Mech. Anal. **169**, 89
[15] Müller, I., Ruggeri, T. (1998): Rational extended thermodynamics. 2nd edition. (Springer Tracts in Natural Philosophy, vol. **37**). Springer, New York
[16] Grad, H. (1949): On the kinetic theory of rarefied gases. Comm. Pure Appl. Math. **2**, 331–407

Central schemes for conservation laws with application to shallow water equations

G. Russo

Abstract. An overview is given of finite volume central schemes for the numerical solution of systems of conservation and balance laws. Well-balanced central schemes on staggered grids for the Saint-Venant model or river flow are considered. A scheme which is well-balanced for channels with variable cross section is introduced. Lastly, a scheme which preserves non-static equilibria is introduced, and some numerical results are presented.

1 Introduction

The numerical solution of hyperbolic systems of conservation laws has been a challenging and fascinating field of research for many decades.

The solution of conservation laws may develop jump discontinuities in finite time, and the uniqueness of the (weak) solution is guaranteed only by recurring to an additional selection rule, such as the entropy condition (see, e.g., [1] for a recent account of the theory of hyperbolic systems of conservation laws). The dissipation mechanism of a quasilinear hyperbolic system is concentrated at the shocks, and its effect can be described in terms of the balance laws and entropy condition.

The schemes more commonly used in this context are the so-called *shock capturing schemes*. At variance with *front tracking methods*, such schemes solve the field equations on a fixed grid, and the shocks are identified by the regions with large gradients. Among shock capturing schemes, the most commonly used are finite volume schemes, in which the basic unknowns represent the cell average of the unknown field. In finite difference schemes, the basic unknown represents the pointwise value of the field at the grid node.

The necessity of high accuracy and sharp resolution of the discontinuities encouraged the development of high order schemes for conservation laws.

Most modern high order shock capturing schemes are written in conservation form (in this way the conservation properties of the system are automatically satisfied), and are based on two main ingredients: the numerical flux function, and the non-oscillatory reconstruction. High order accuracy in the smooth regions, sharp resolution of discontinuities, and absence of spurious oscillations near shocks strongly depend on the characteristics of these two essential features (see, e.g., the books [2,3] or the lecture notes [4]).

Among finite volume methods, we distinguish between semidiscrete and fully discrete schemes. The first are obtained by integrating the conservation law in a spatial cell, by using a numerical flux function at the edge of the cell, and by providing a suitable reconstruction of the field at the two sides of each edge of the cell in terms of

the cell averages. In this way one obtains a set of ordinary differential equations that can be then solved by an ODE solver such as Runge-Kutta. High order semidiscrete schemes are described, e.g., by Shu in [4, Chap. 4].

Alternatively, fully discrete schemes are obtained by integrating the conservation law on a cell in space-time. The flux function appearing in the scheme is consistent, to the prescribed order of accuracy, with the time average of the flux at the edge of the cell in one time step. A second-order fully discrete method can be obtained, for example, by combining the second-order Lax-Wendroff method with a first-order method by suitable *flux limiter*, which prevents formation of spurious oscillations (see [2] for examples).

Another distinction can be made between upwind and central schemes. Roughly speaking, we say that a scheme is *upwind* if it makes extensive use of the characteristic information of the system, so that the scheme can take into account the direction of propagation of the signal, while a scheme is *central* if characteristic information is not used. The prototype of upwind schemes is first-order upwind, or its version for quasilinear systems, which is the first-order Godunov method, based on the solution of the Riemann problem at cell edges. The prototype of a central scheme is the first-order Lax-Friedrichs scheme, which requires neither Riemann solvers nor characteristic decomposition.

Generally speaking, upwind-based methods guarantee sharper resolution than central schemes for the same order of accuracy and grid spacing, but are usually more expensive, and more complicated to implement. For this reason, central schemes have attracted a good deal of attention in the last fifteen years. Following the original work of Nessyahu and Tadmor [5], where a second-order, shock capturing, finite volume central scheme on a staggered grid in space-time was introduced and analyzed, several extensions and generalizations have been made for central schemes, both fully discrete and semidiscrete (see [6] and its references for a review of central schemes).

The distinction between the upwind and the central worlds is not sharp, and in fact characteristic information can be used to improve the performance of central schemes. For example, by using different estimates of the negative and positive characteristic speeds, Kurganov et al. [7] improved the original semidiscrete central scheme [8]. The latter, in turn, is related to the finite volume schemes used by Shu, when a local Lax-Friedrics flux (also called a Rusanov flux) is used [4, Chap. 4]. Qiu and Shu [9] showed that, by using a reconstruction in characteristic variables for the computation of the staggered cell average in central schemes, one eliminates the spurious oscillations produced by high order central schemes for the integration of the Euler equation of gas dynamics near discontinuities.

Although semidiscrete schemes are attractive because of their flexibility and simplicity of use, fully discrete schemes sometimes provide better performance with the same grid spacing. For this reason, it is of interest to consider the use of fully discrete central schemes for systems with source term.

High order central schemes on staggered grids for conservation laws have been derived. See, e.g., [10,11] for recent results in this field, and [9] for a comparison between semidiscrete and fully discrete high order central schemes.

When a source is introduced in the system (i.e., when dealing with a quasilinear system of balance laws) then several new interesting problems arise in extending shock capturing schemes for conservation laws to this new case.

Straightforward extensions can be obtained by integrating the source term in space (for semidiscrete schemes) and in space and time (for fully discrete schemes) and using a suitable quadrature formula to compute the contribution of the source. Another general technique that is commonly used for systems with source is based on the fractional step (also called time-splitting) method. Both approaches, however, perform poorly in two cases, which require a more detailed "ad hoc" treatment.

One case concerns the problem of hyperbolic systems with stiff source. Here the source has to be treated by an implicit scheme, to avoid an excessive restriction on the time step due to the small characteristic time of the source term. The flux, on the other hand, is in general not stiff, and an explicit scheme is certainly more convenient because the nonlinearity of the space discretization (mainly due to the non-linear reconstruction) makes an implicit treatment excessively expensive.

Implicit-explicit (IMEX) time discretization is a natural choice for this kind of problem. See, e.g., [12] for a second-order IMEX scheme based on central discretization in space, and [13] for a review of recently developed IMEX schemes.

Another important problem consists in the integration of systems in which the source term is nearly balanced by flux gradients. In this case the solution is a small perturbation of a stationary one. For this problem it would be desirable to construct numerical schemes that maintain the stationary solutions at a discrete level.

Such schemes are often called *well-balanced*, after the paper by Greenberg and Leroux [14], and their development and analysis has interested many researchers in the recent years [15–20], although ideas based on characteristic decomposition were introduced earlier (see [21] for linear problems and [22] for the extension to quasilinear problems). See also [23] for a detailed explanation of various numerical methods for shallow water equations.

Most well-balanced schemes have been derived either for semidiscrete or for fully discrete schemes on a non-staggered grid. However, staggered central schemes are attractive since they have an automatic mechanism for controlling spurious oscillations. In many cases, they allow better resolution than the non-staggered counterpart. It is therefore attractive to explore the possibility of constructing well-balanced schemes on a staggered grid.

An example of a well-balanced central scheme on a staggered grid was presented at the HYP2000 conference, and is briefly described in [24]. A well-balanced second-order central scheme for the Saint-Venant equations that preserves static equilibria was derived.

Here we extend the result by presenting a well-balanced central scheme on a staggered grid that also works for channels of variable cross section, and a scheme that preserves non-static equilibria (in the case of subcritical flow).

The rest of the section is devoted to a brief review of hyperbolic systems of conservation laws and conservative schemes for their numerical approximation.

The next section is a review of shock capturing central schemes for balance laws. Section 3 presents central schemes that preserve static solutions, with application

to the Saint-Venant model of shallow water. Section 4 is techincal, and describes in detail the derivation of a central scheme for the Saint-Venant equations that preserves stationary, non-static equilibria. Applications of the schemes are given in Sects. 3 and 4.

1.1 Hyperbolic systems

We consider a hyperbolic system of balance laws. It takes the form

$$\frac{\partial u}{\partial t} + \frac{Df(x,u)}{Dx} = R(x,u), \tag{1}$$

where $u \in R^m$; $f, R : \mathbb{R}^m \to \mathbb{R}^m$, $A = \partial f/\partial u$ has real eigenvalues and basis of eigenvectors. Here $Df/Dx = \partial f/\partial x + Au_x$.

Such a system may develop discontinuities in finite time (shocks) and therefore one has to abandon the hope of finding regular solutions (strong solutions), and one looks for weak solutions.

An admissible discontinuity that propagates in the media has to satisfy the so-called jump conditions, which can be derived directly from the balance law, written in the original integral form. Such conditions, also called *Rankine-Hugoniot conditions* can be written in the form

$$-V_\Sigma [\![u]\!] + [\![f]\!] = 0, \tag{2}$$

where V_Σ denotes the speed of the moving discontinuity Σ, and, for any function $h(x,t)$, $[\![h]\!] \equiv (h^+ - h^-)$ denotes the jump of the quantity across the discontinuity Σ.

A function u that satisfies Eq. (1) in the regions of regularity and conditions (2) at discontinuities is a weak solution of the balance equation. However, uniqueness is not guaranteed for such a solution. In order to restore uniqueness of the solution, one has to resort to additional selection rules. The entropy condition, for example, is often used to select the unique solution of a conservation law. For a review of the modern theory of hyperbolic systems of conservation laws see, e.g., the book by Dafermos [1].

Conservation form, jump conditions, and entropy conditions are used as guidelines in the development of modern shock capturing schemes, in order to guarantee that the numerical solution of the scheme converges to the unique entropic solution when the grid is refined.

1.2 Numerical schemes

A numerical scheme for balance laws has to admit the possibility of capturing discontinuous solutions, giving correct shock speed. For a system of conservation laws (i.e., in the case of zero source term) it is desirable that a numerical scheme maintains the conservation properties of the exact solution of the system. For both purposes it is essential that the scheme is written in conservation form.

The so-called *shock capturing finite volume schemes* are usually derived by integrating the system of balance laws on a suitable region of space-time.

We divide space into equally spaced cells $I_j \equiv [x_{j-1/2}, x_{j+1/2}]$ of size Δx, centered at x_j, $j \in \mathbb{Z}$.

Semidiscrete schemes are obtained by integrating the conservation law (1) in each cell in space, and approximating the flux function at the border of the cell by a suitable *numerical flux function* that depends on the values of the field across the edge of the cell. These values, in turn, are obtained by a suitable reconstruction of the function from the cell averages. In this way, one obtains a system of ordinary differential equations for the cell averages, of the form

$$\frac{d\bar{u}_j}{dt} = -\frac{F_{j+1/2} - F_{j-1/2}}{\Delta x} + \bar{R}_j,$$

where $\bar{u}_j$ is an approximation of the cell average

$$\bar{u}_j \approx \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t)\, dx,$$

and

$$F_{j+1/2} = F(u^-_{j+1/2}, u^+_{j+1/2})$$

is the numerical flux function, $u^{\pm}_{j+1/2}$ are the reconstructed values of the field across the edge $x_{j+1/2}$.

For an account of modern semidiscrete high order schemes for conservation laws, see, e.g., the chapter by Shu in [4].

Fully discrete schemes are obtained by integrating the conservation law on a suitable cell in space-time. Integrating the equation on a cell $I_j \times [t^n, t^{n+1}]$ one typically obtains a scheme of the form

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}(F_{j+1/2} - F_{j-1/2}) + \Delta t\, \bar{R}_j^{n+1/2},$$

where $F_{j+1/2}$ is the so-called numerical flux function, which is consistent, to the prescribed order of accuracy, with the time average of the flux f at the edge $x_{j+1/2}$ of the cell, and $\bar{R}_j^{n+1/2}$ is an approximation of the space-time cell average of the source.

The piecewise constant solution $\tilde{U}^n(x) = \sum_j \chi_j(x) U_j^n$, where $\chi_j(x)$ is the characteristic function of the interval I_j, satisfies discrete jump conditions, and, if it converges to a function $U(x,t)$ as $\Delta x \to 0$, then $U(x,t)$ is a weak solution of (1) (Lax-Wendroff theorem, see [25]). A nice description of a fully discrete scheme of this form is presented, e.g., in [2].

2 Staggered central schemes for balance laws

A different family of schemes is obtained by integrating the conservation laws on a staggered grid in space-time (see Fig. 1).

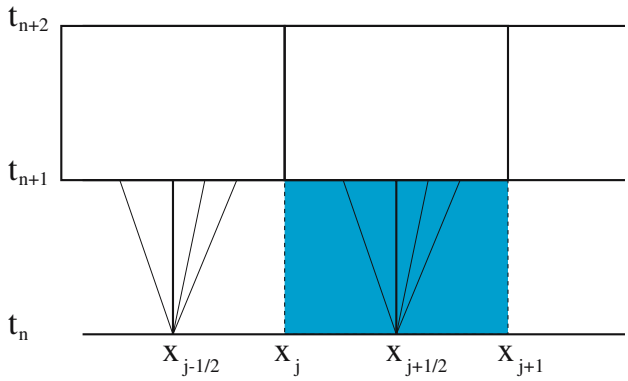


Fig. 1. Staggered grid in space-time and central scheme

After integration one obtains

$$\Delta x\, \bar{u}^{n+1}_{j+1/2} = \int_{x_j}^{x_{j+1}} u(x,t^n)\,dx - \int_{t^n}^{t^{n+1}} \left(f(u_{j+1}(t)) - f(u_j(t))\right) dt + \int_{x_j}^{x_{j+1}} dx \int_{t^n}^{t^{n+1}} dt\, R(x,u(x,t)),$$

with $u_j \equiv u(x_j,t)$. The above formula is exact for piecewise smooth solutions. To convert the formula in a numerical scheme one must:

(i) reconstruct $u(x,t^n)$ from $\bar{u}^n_j$ and use it to compute $\bar{u}^n_{j+1/2}$;

(ii) approximate integrals in time by a quadrature formula;

(iii) compute an approximation of $u_j(t)$ on the quadrature nodes;

(iv) approximate the integral of the source on the cell in space-time by a quadrature formula.

The celebrated second-order Nessyahu-Tadmor (NT) scheme [5] is obtained (for $R = 0$) by:

(i) approximating $u(x,t^n)$ by a piecewise linear function;

(ii) integrating the flux by the midpoint rule;

(iii) using the first-order Taylor expansion for the computation of $u(x_j, t^n + \Delta t/2)$.

Its generalization to a system of balance laws (with no explicit dependence of flux and source on x) can be written as a simple two-line predictor-corrector scheme:

$$\begin{aligned} u_j^{n+1/2} &= u_j^n - \tfrac{\lambda}{2} f'_j + \tfrac{\Delta t}{2} R(u_j^{n+\beta}) && \textit{predictor} \\ u_{j+1/2}^{n+1} &= \tfrac{1}{2}(u_j^n + u_{j+1}^n) + \tfrac{1}{8}(u'_j - u'_{j+1}) - \lambda\big(f(u_{j+1}^{n+1/2}) \\ &\quad - f(u_j^{n+1/2})\big) + \tfrac{\Delta t}{2}(R(u_j^{n+\beta}) + R(u_{j+1}^{n+\beta}) && \textit{corrector}, \end{aligned} \tag{3}$$

where $\lambda = \Delta t/\Delta x$ denotes the mesh ratio, and $u'_j/\Delta x$, $f'_j/\Delta x$ are first-order approximations of space derivatives.

The NT scheme can be made (discretely) entropic and total variation diminishing (TVD). These properties depend on the reconstruction of the derivatives. In order to avoid spurious oscillations, suitable *slope limiters* for u' and f' are required. The simplest choice is given by the so-called *minmod* limiter, which is defined as

$$\mathrm{MinMod}(a, b) = \begin{cases} \mathrm{sign}(a)\min(|a|, |b|) & \text{if } ab > 0 \\ 0 & \text{if } ab \le 0 \end{cases}$$

Therefore u'_j can be computed as

$$u'_j = \mathrm{MinMod}(u_{j+1} - u_j, u_j - u_{j-1}),$$

and f'_j can be computed either by using the minmod function or by $f'_j = A(u_j)u'_j$.

Better slope limiters (e.g., Harten's UNO limiter) can be used. For an account of different slope limiters see, e.g., [2] or [5].

Note that the contribution of the source term can be completely explicit ($\beta = 0$) or implicit ($\beta = 1/2$). Both cases result in a second-order scheme in space and time. The time restriction due to the flux term, in absence of source, is the Courant-Friedrichs-Lewy (CFL) condition, which for the NT scheme reads

$$\lambda C_{\max} \le \frac{1}{2}, \tag{4}$$

where $C_{\max}$ is the maximum spectral radius of the Jacobian matrix A on the computational domain.

High order central schemes for conservation laws ($R \equiv 0$) are obtained by using high order non-oscillatory reconstruction, such as WENO, and higher order time evolution, such as Runge-Kutta schemes with natural continuous extension (see [10]) or central Runge-Kutta [11].

The time step restriction due to the source term depends on the use of explicit ($\beta = 0$) or implicit ($\beta = 1/2$) predictor, and on the stiffness of the source, i.e., on its relaxation time. If the restriction introduced by explicit treatment of the source is more severe than the CFL condition (4), then it is preferable to use an implicit discretization of the predictor step. Note that the above scheme with $\beta = 1/2$ is a simple example of implicit-explicit (IMEX) time discretization, obtained by coupling an explicit RK2 scheme (modified Euler scheme) with an A-stable second-order scheme (midpoint

method). In the case of very stiff relaxation terms, the above scheme is not suitable, and an L-stable scheme is needed for a proper treatment of the source. An example of an IMEX central scheme for hyperbolic systems with stiff sources is presented in [12]. Other examples of IMEX schemes applied to relaxation systems are given in [26] and its references.

We remark here that finite volume schemes are not suitable for high order approximation of hyperbolic systems with stiff source, because the averaging of the source couples all the cells, making implicit schemes expensive. Finite difference schemes are more natural in this case, because the pointwise value of the function rather than its cell average is used as basic unknown, and therefore the cells are decoupled (at the level of the source term). See [4, Chap. 4], for an illustration of high order finite difference schemes in conservation form.

3 A well-balanced scheme that preserves static equilibria

In this section we develop a well-balanced central scheme for a system of balance laws, with particular application to the Saint-Venant equations of shallow water.

Consider a problem in which the solution of the system

$$\frac{\partial u}{\partial t} + \frac{Df(x,u)}{Dx} = R(x,u)$$

is a small deviation from the stationary solution $\tilde{u}(x)$, for which

$$\frac{\partial f(\tilde{u})}{\partial x} = R(\tilde{u}). \tag{5}$$

In this case, fractional step schemes or a scheme of the form (3) perform poorly, because they do not preserve the equilibria (5), even at a discrete level.

We consider here the specific case of the Saint Venant model of shallow water equations. We start with the one-dimensional equations in a channel of constant cross section. The equations can be written in the form:

$$\frac{\partial h}{\partial t} + \frac{\partial q}{\partial x} = 0, \tag{6}$$

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q^2}{h} + \frac{1}{2}gh^2\right) = -ghB_x, \tag{7}$$

where $h(x,t)$ denotes the water depth, $q(x,t)$ the water flux, $B(x)$ the bottom profile, and g the constant gravity acceleration; see Fig. 2.

Several well-balanced schemes have been developed in the literature; they satisfy different requirements. We mention here the works by Greenberg and Leroux [14], Gosse [15], LeVeque [16], Perthame et al. [20], Jin [18], Kurganov and Levy [19], Bouchut et al. [27], Gallouët et al. [17], just to mention a few names.

The usual requirement for a well-balanced scheme is the preservation of static equilibria, which means a stationary solution of Eq. (5), for which the fluid does not flow, i.e., $q = 0$.

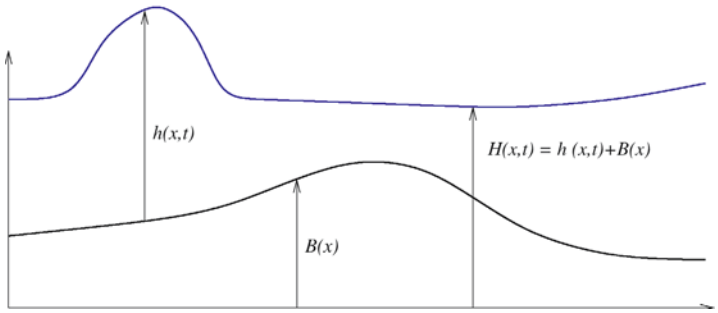


Fig. 2. Water height and bottom profile of the Saint-Venant model of shallow water

In addition, other common requirements are:

(i) preservation of all equilibria;
(ii) preservation of non-negativity of h;
(iii) capability of treating dry zones (i.e., zones for which $h = 0$);
(iv) numerical entropy condition.

A recent scheme derived by Bouchut and collaborators [27] is able to fulfill all these requirements.

We consider, as a test, the initial condition

$$q(x,0) = q_0, \quad h(x,0) = \begin{cases} 1.01 \text{ if } |x - 0.2| < 0.05 \\ 1 \quad \text{otherwise} \end{cases}, \tag{8}$$

with $q_0 = 0$, and we let the bottom profile be given by

$$B(x) = \begin{cases} 1 + \cos(10x - 5) \text{ if } |x - 0.5| < 0.1 \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases}. \tag{9}$$

If we use scheme (3) (with $\beta = 0$) then we obtain the solution shown in Fig. 3, where the total height $H = h + B$ is reported at initial time (dashed line) and final time $t = 0.7$. In all calculations we set the constant g of gravity to 1. The number of grid points used in the calculation is $N = 200$. The dashed line represents the initial condition.

Flat boundary conditions were used here and for all calculations presented in the paper.

The spurious effect present in the center of the computational domain is due to the fact that the scheme is not able to preserve stationary solutions of Eqs. (6), (7).

How can we construct well-balanced schemes that preserve equilibria? Several approaches have been considered in the literature.

The paper by Kurganov and Levy [19], for example, provides an example of a semidiscrete central scheme that preserves static equilibria. The scheme they present, however, cannot be generalized straightforwardly in the case of staggered grids, because the NT scheme is not able to preserve solutions of the equation

$$\frac{\partial u}{\partial t} = 0,$$

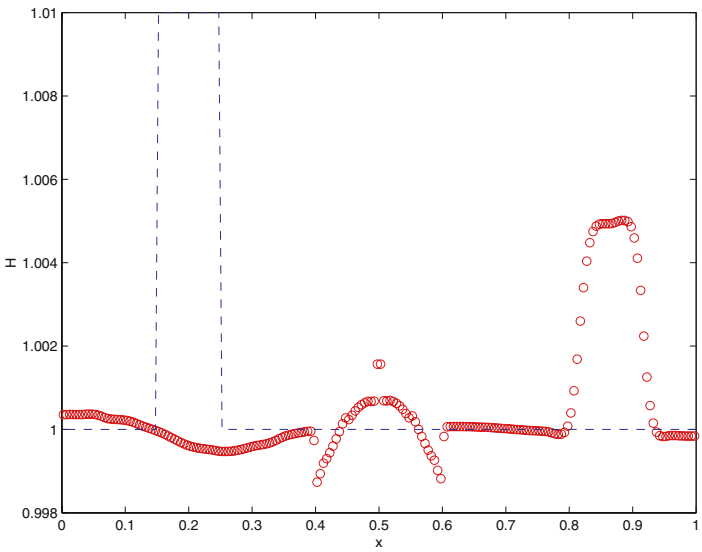


Fig. 3. Numerical solution of system (6), (7), with the use of scheme (3), with $\beta = 0.$ 200 grid points

unless $u \equiv$ constant.

With this taken into account, a well-balanced central scheme that preserves static solutions was derived and presented at the hyperbolic conference in Magdeburg [24].

The guidelines in the development of such a scheme are:

(1) reformulate the problem using $H = h + B$ as conserved variable;
(2) compute a predictor by a non-conservative form using $f' = \delta f + A(u)u'$, where $\delta f \approx \Delta x\, \partial f/\partial x$;
(3) in the corrector use a suitable approximation of functions and space derivatives.

The last requirements are obtained, for example, by setting (at even time steps):

$$B_j = \frac{1}{2}\left(B(x_j + \Delta x/2) + B(x_j - \Delta x/2)\right),$$
$$\frac{B'_j}{\Delta x} = \frac{B(x_j + \Delta x/2) - B(x_j - \Delta x/2)}{\Delta x}.$$

The scheme applied to the SV equations takes the form:
predictor:

$$\begin{aligned} H_j^{n+1/2} &= H_j^n - \frac{\lambda}{2} q'_j, \\ q_j^{n+1/2} &= q_j^n - \frac{\lambda}{2}\left(2v_j^n q'_j - (v_j^n)^2 (H'_j - B'_j) + g H'_j (H_j - B_j)\right); \end{aligned} \tag{10}$$

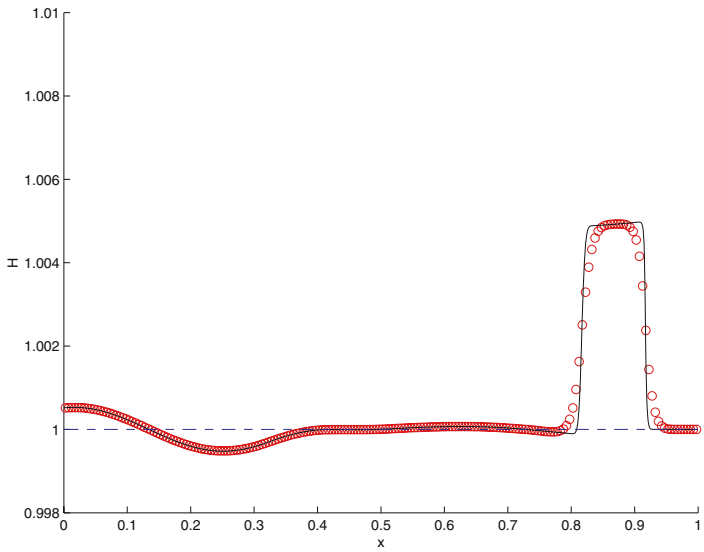


Fig. 4. Numerical solution of system (6), (7), with the well-balanced central scheme (10), (11). The thin line represents the reference solution and is obtained by the well-balanced scheme with $N = 1600$ gridpoints

corrector:

$$\begin{aligned} H^{n+1/2}_{j+1/2} &= H^n_{j+1/2} - \lambda(q^{n+1/2}_{j+1} - q^{n+1/2}_j), \\ q^{n+1/2}_{j+1/2} &= q^n_{j+1/2} - \lambda(\psi^{n+1/2}_{j+1} - \psi^{n+1/2}_j) \\ &\quad - g\frac{\lambda}{2}(H^{n+1/2}_j B'_j + H^{n+1/2}_{j+1} B'_{j+1}), \end{aligned} \tag{11}$$

where the staggered cell averages $H^n_{j+1/2}$ and $q^n_{j+1/2}$ are computed as in scheme (3), $v \equiv q/(H-B)$, $\psi^{n+1/2}_j \equiv \psi(H^{n+1/2}_j, q^{n+1/2}_j)$ and

$$\psi(H, q) \equiv \frac{q^2}{H-B} + \frac{1}{2}H(H-B)\ .$$

It is easy to check that $H =$ constant, $q = 0$ is a solution for this scheme.

The scheme is applied to problem (6)-(8), and the numerical results are shown in Fig. 4. Note that no spurious profile appears where $B \neq 0$.

Variable cross section. The Saint-Venant equations for a channel of variable cross section have the form:

$$\begin{aligned} &\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0, \\ &\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{Q^2}{A}\right) + gA\frac{\partial H}{\partial x} = 0, \end{aligned}$$

where $A(x,t)$ is the cross section of the part of the channel occupied by the water, and Q is the flux. The above approach can be extended to the case of a channel with rectangular cross section, $A(H,x) = hW(x)$. A well-balanced scheme is obtained as follows. Let $A = W(x)h(x,t) = W(x)(H(x,t) - B(x))$. Reformulate the problem using H and $q = Q/W$ as unknown conservative variables. This choice will not alter the jump conditions. Then the system becomes

$$\begin{aligned} &\frac{\partial H}{\partial t} + \frac{\partial q}{\partial x} = -q\frac{W_x(x)}{W(x)}, \\ &\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q^2}{H - B(x)} + \frac{g}{2}H(H - 2B(x))\right) \\ &\qquad = -gHB_x(x) - \frac{q^2}{H - B(x)}\frac{W_x(x)}{W(x)}. \end{aligned} \tag{12}$$

The scheme used before preserves static solutions of this system, because the two additional terms vanish as $q = 0$.

As an application of the scheme we consider two test cases. The initial conditions are given by Eq. (8), and the bottom profile by Eq. (9).

The results corresponding to these two cases are reported in Fig. 5. The first corresponds to the choice

$$W(x) = 1 - B(x),$$

while, in the second, it is

$$W(x) = \frac{1}{1 - B(x)}.$$

Notice that in the second case the area of the cross section of the channel corresponding to the static solution $H = 1$ is basically constant, and this reduces the amplitude of the reflected wave, while in the first case the cross-sectional area at the center of the channel becomes narrower than that in the case of constant cross section, resulting in a larger reflected wave.

4 A well-balanced scheme for subcritical flows

The approach described above cannot be directly applied to the case in which the stationary solution is not static. Even if the scheme preserves the equilibrium

$$\frac{Df}{Dx} = R$$

at a discrete level, the Nessyahu-Tadmor scheme does not preserve a solution of the trivial equation $\partial u/\partial t = 0$ unless u is also constant in space.

One possibility to obtain a scheme that preserves stationary solutions is to perform a change of variables in such a way that in the new variables the equilibrium is represented by constants.

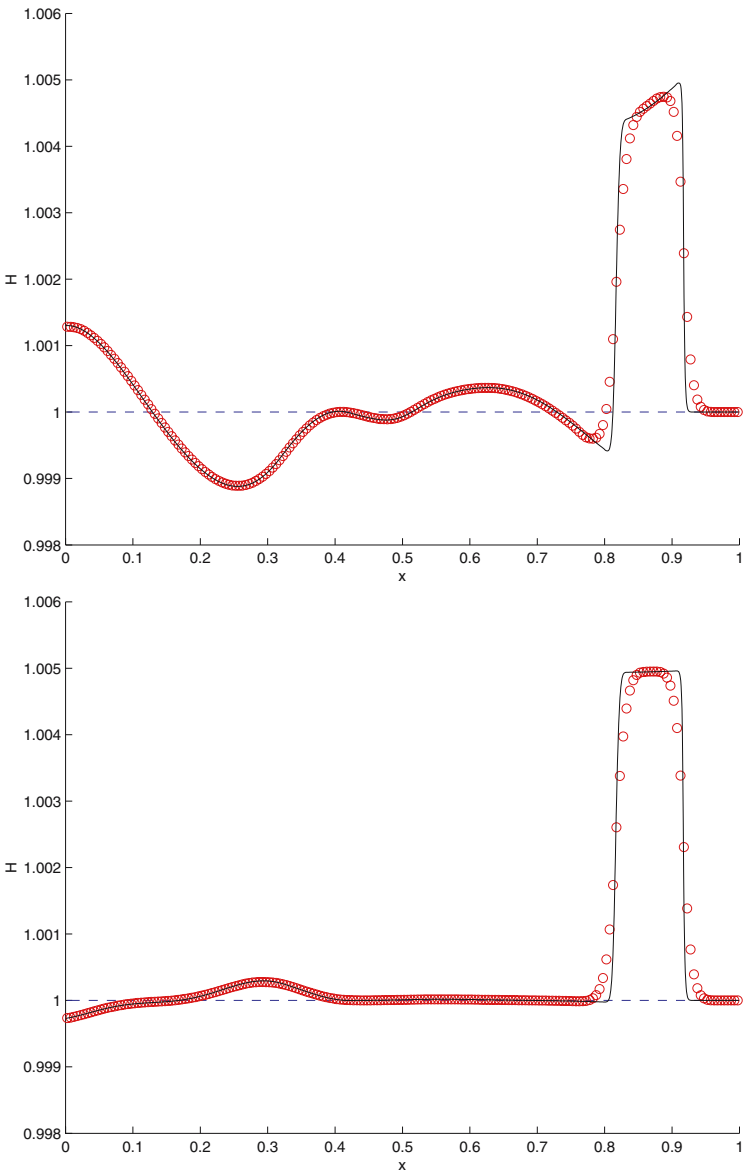


Fig. 5. Numerical solution of Saint-Venant equations with variable cross section, at time $t = 0.7$, 200 grid points. The thin dotted line represents the reference solution obtained by 1600 points. Upper: $W(x) = 1 - B(x)$; lower: $W(x) = (1 - B(x))^{-1}$

Scalar equation. We consider the scalar case first. We denote a stationary solution by $\tilde{u}(x)$, i.e.,

$$\frac{Df(x,\tilde{u})}{Dx} = R(x,\tilde{u}). \tag{13}$$

Then we look for a solution of the form

$$u(x,t) = \tilde{u}(x)v(x,t).$$

The equation for u becomes

$$\tilde{u}(x)\frac{\partial v}{\partial t} + \frac{Df(x,\tilde{u}v)}{Dx} = R(x,\tilde{u}v).$$

Integrating on a staggered cell in space-time one has

$$\begin{aligned}\Delta x\,\bar{u}_{j+1/2}^{n+1} &= \int_{x_j}^{x_{j+1}} \tilde{u}(x)v^{n+1}(x)\,dx = \int_{x_j}^{x_{j+1}} \tilde{u}(x)v(x,t^n)\,dx \\ &\quad - \int_{t^n}^{t^{n+1}} (f(x_{j+1},\tilde{u}_{j+1}v_{j+1}(t)) - f(x_j,\tilde{u}_j v_j(t)))\,dt \\ &\quad + \int_{x_j}^{x_{j+1}} dx \int_{t^n}^{t^{n+1}} dt\, R(x,\tilde{u}(x)v(x,t)).\end{aligned} \tag{14}$$

A second-order discretization of the conservation equations is obtained as follows. We define the quarter cell values of the equilibrium solution (which coincide with their average to second order) by

$$\tilde{u}_{j\pm1/4} \equiv \tilde{u}(x_j \pm \Delta x/4) = \tilde{u}(x_{j\pm1/4}).$$

The cell average at time t^{n+1} (left-hand side of Eq. (14)) is discretized as

$$\frac{1}{\Delta x}\int_{x_j}^{x_{j+1}} \tilde{u}(x)v^{n+1}(x)\,dx \approx \frac{1}{2}(\tilde{u}_{j+1/4} + \tilde{u}_{j+3/4})\bar{v}_{j+1/2}^{n+1}.$$

The first term on the right-hand side (staggered cell average) is discretized as follows. Assume $v^n(x)$ is approximated by a piecewise linear function

$$v^n(x) \approx \sum_j \chi_j(x)L_j(x),$$

where χ_j is the characteristic function of the jth interval, and

$$L_j(x) = v_j^n + v_j'(x - x_j)/\Delta x.$$

Here $v_j'/\Delta x$ denotes a first-order approximation of the space derivative of $v^n(x)$, and v_j^n denotes an approximation of the pointwise value of $v^n(x)$ (which agrees with its cell average to second order).

In each cell one has

$$\bar{u}_j^n \approx \frac{1}{2}(\tilde{u}_{j-1/4} + \tilde{u}_{j+1/4})v_j^n.$$

The staggered cell average at time n is computed as

$$\int_{x_j}^{x_{j+1}} \tilde{u}(x)v^n(x)\,dx = I_j^R + I_{j+1}^L$$
$$= \int_{x_j}^{x_{j+1/2}} \tilde{u}(x)L_j(x)\,dx + \int_{x_{j+1/2}}^{x_{j+1}} \tilde{u}(x)L_{j+1}(x)\,dx.$$

To second-order accuracy, the integrals are evaluated as

$$I_j^R = \Delta x\,\tilde{u}_{j+1/4}\left(\frac{1}{2}v_j^n + \frac{1}{8}v_j'\right),$$
$$I_j^L = \Delta x\,\tilde{u}_{j-1/4}\left(\frac{1}{2}v_j^n - \frac{1}{8}v_j'\right),$$

and therefore the staggered cell average is

$$\frac{1}{\Delta x}\int_{x_j}^{x_{j+1}} \tilde{u}(x)v^n(x)\,dx \approx \frac{1}{2}(\tilde{u}_{j+1/4}v_j^n + \tilde{u}_{j+3/4}v_{j+1}^n)$$
$$+ \frac{1}{8}(\tilde{u}_{j+1/4}v_j' - \tilde{u}_{j+3/4}v_{j+1}').$$

Remark. A better approximation of the staggered cell value can be obtained by using

$$\int_0^{h/2} \tilde{u}(x_j + \xi)\xi\,d\xi = \frac{1}{8}\tilde{u}_{x_j+h/3}h^2 + O(h^4);$$

however, this requires the storage of an additional value of the stationary solution, and it does not improve the overall order of accuracy.

The contribution of flux and source is computed by a predictor-corrector type scheme

$$\frac{1}{\Delta x}\int_0^{\Delta t} f(x_j, u_j(t^n + \tau)\,d\tau \approx f(x_j, \tilde{u}(x_j)v_j^{n+1/2}).$$

Predictor step. This can be computed by a non-conservative scheme

$$u_j^{n+1/2} = u_j^n - \frac{\lambda}{2}\left(\delta f_j + \frac{\partial f}{\partial u}u_j'\right) + \frac{\Delta t}{2}R(x_j, u_j^n),$$

where $\delta f_j/\Delta x$ denotes a first-order approximation of the space derivative of f and

$$u_j' = \tilde{u}_j'v_j^n + u_j v_j'.$$

Here

$$u'_j = \Delta x \left.\frac{du}{dx}\right|_{(x_j,u_j)}, \quad v'_j = \Delta x \left.\frac{\partial v}{\partial x}\right|_{(x_j,u_j)} + O(\Delta x^2), \quad \tilde{u}'_j = \Delta x \frac{d\tilde{u}}{dx}.$$

Once the stationary equation (13) is solved for $\tilde{u}$, the quantity $\tilde{u}'$ is obtained from the differentiation of Eq. (13) as

$$A\tilde{u}' = \Delta x \left(R(x,\tilde{u}) - \frac{\partial f}{\partial x} \right).$$

Corrector. The contribution of the source term is obtained by a suitable quadrature formula

$$\frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} \int_{t^n}^{t^{n+1}} R(x,u)\, dx\, dt \approx \Delta t\, \tilde{R}(\tilde{u}_j v_j^{n+1/2}, \tilde{u}_{j+1} v_{j+1}^{n+1/2}).$$

The formula has to be consistent with the well-balanced property of the scheme, i.e.,

$$\lambda(f(x_j,\tilde{u}_j) - f(x_{j+1},\tilde{u}_{j+1})) + \Delta t\, \tilde{R}(\tilde{u}_j,\tilde{u}_{j+1}) = 0;$$

in fact, this relation can be used to define the function $\tilde{R}$ in the numerical scheme.

The extension to a system of equations is obtained by repeating the above steps component by component.

We apply this technique to the shallow water equations. We consider 1D shallow water equations in the form:

$$\frac{\partial H}{\partial t} + \frac{\partial q}{\partial x} = 0,$$
$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{q^2}{H - B(z)} + \frac{1}{2} g H (H - 2B(x)) \right) = -gHB_x.$$

The stationary solution is $q = q_0$ and $H = \tilde{H}(x)$, obtained by solving the equation

$$\frac{\partial}{\partial x}\left(\frac{q_0^2}{\tilde{H} - B(z)} + \frac{1}{2} g \tilde{H} (\tilde{H} - 2B(x)) \right) = -g\tilde{H}B_x.$$

Integrating the equation one has

$$\frac{q_0^2}{2(\tilde{H} - B)^2} + g\tilde{H} = gH_0 + \frac{q_0^2}{2(H_0 - B_0)^2}.$$

Therefore $\tilde{H}$ is obtained as solution of a cubic equation.

If, in the whole channel, the stationary flow is subcritical, i.e., if $|u| < \sqrt{gh}$, then the cubic equation has only one real solution. This is the case we consider.

The quantities $\tilde{H}_j$, $\tilde{H}_{j+1/2}$, $\tilde{H}_{j+1/4}$, $\tilde{H}_{j-1/4}$ are precomputed and stored at the beginning of the calculation.

Particular care has to be used in the approximation of the derivatives. Here we distinguish between predictor and corrector steps.

Predictor. We use the non-conservative form of the equation. In particular

$$q_j^{n+1} = q_j^n - \frac{\lambda}{2} K,$$

with

$$K = 2\frac{q}{h} q_j' - \left(\frac{q^2}{h^2} - gh \right) (H' - B') + gHB',$$

$$H' = v'\tilde{H} + v\tilde{H}'.$$

Here $h = H - B$, $B' = B_x \Delta x$, $\tilde{H}' = \tilde{H}_x \Delta x$, and $\tilde{H}_x$ is computed *exactly* by differentiating the equation for $\tilde{H}$.

Corrector. The source has to be computed with a well-balanced formula. For example, for the flux one has

$$\begin{aligned} q_{j+1/2}^{n+1} = {} & q_{j+1/2}^n - \lambda (f_{j+1}^{n+1/2} - f_j^{n+1/2}) \\ & + (\tilde{H}_j v_j^{n+1/2} + \tilde{H}_{j+1} v_{j+1}^{n+1/2}) S_{j+1/2}, \end{aligned}$$

with

$$S_{j+1/2} = \frac{f(\tilde{H}(x_j), q_0) - f(\tilde{H}(x_{j+1}), q_0)}{\tilde{H}_j + \tilde{H}_{j+1}}.$$

Remark. In the new scheme the derivative of the bottom is computed *exactly*, and not by finite difference, as in the earlier well-balanced scheme (10), (11).

4.1 Numerical results

We consider here a case in which the solution is a small perturbation of a stationary non-static equilibrium. The initial condition is given by Eq. (8), but with $q_0 = 0.17$. The numerical solution for the total height at two different times is reported in Fig. 6, where only 50 grid points are used. The main difference between the earlier well-balanced scheme and the new one which preserves non-static equilibria is noticeable near the middle of the channel, where the bottom is higher, and the water profile becomes lower. To enhance the effect, in Fig. 7 we report the quantity

$$I(x, t) = \frac{q^2}{2(H - B)^2} + gH,$$

which is invariant at equilibrium. Notice the spurious effect near the center of the channel in the numerical result obtained using scheme (10), (11). We remark, however, that these effects are rather small, and quickly disappear as the grid is refined. Fig. 8, for example, represents the same quantity with 200 grid points, and the effect is barely noticeable.

In all cases, the reference solution is obtained using the earlier well-balanced scheme with 1600 points.

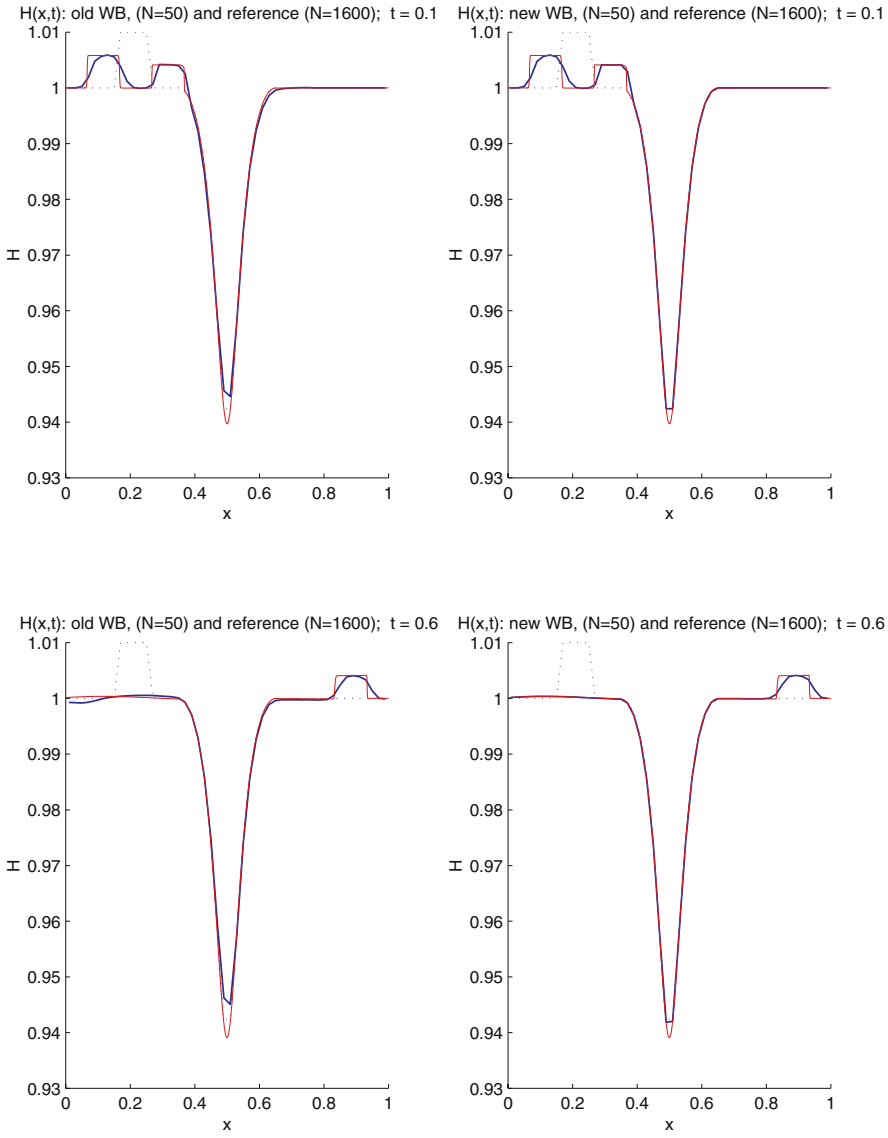


Fig. 6. Small perturbation of stationary, non-static, equilibria. $H(x,t)$ at different times: $t = 0.1$ (top) and $t = 0.6$ (bottom). Well-balanced central scheme that preserves static equilibria (left), and scheme that preserves non-static equilibria (right). Number of grid points $N = 50$

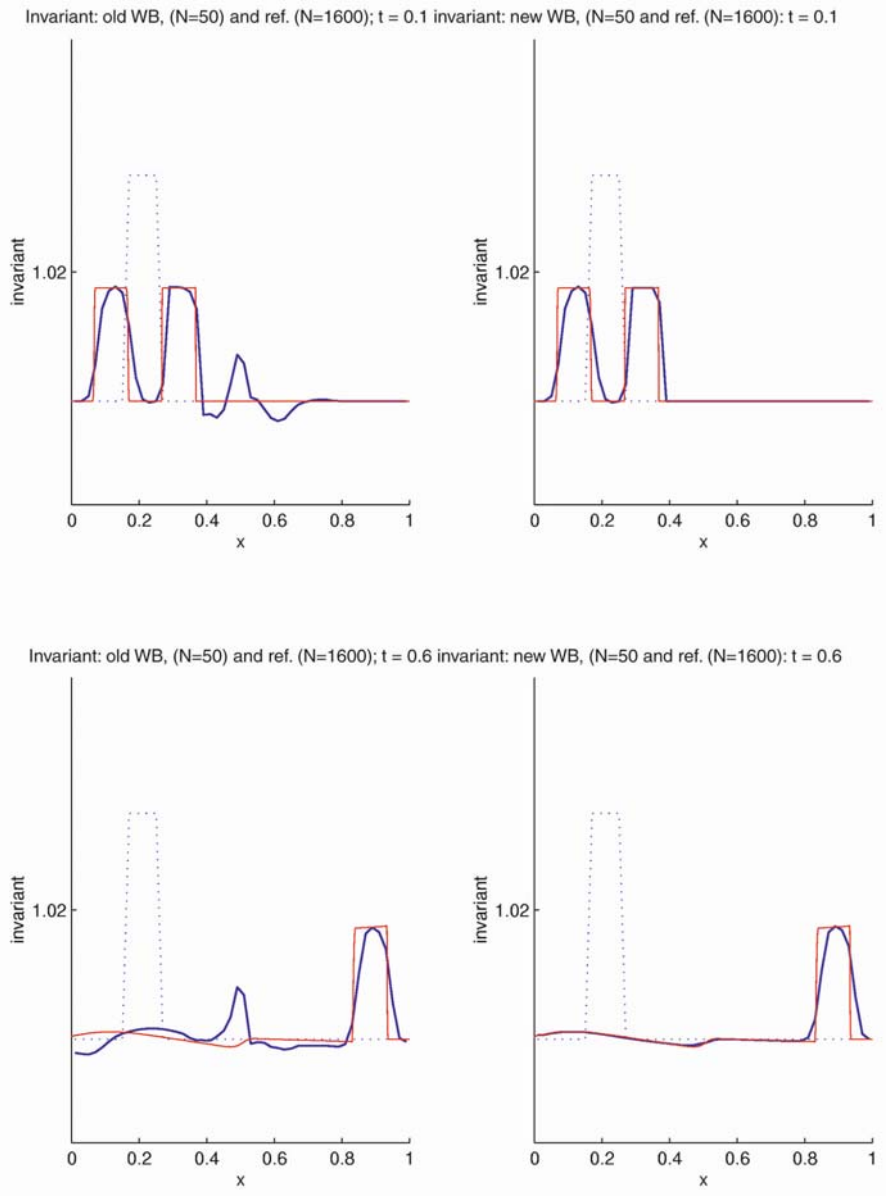


Fig. 7. Plot of the equilibrium invariant at different times. Time $t = 0.1$ (top) and $t = 0.6$ (bottom). Earlier well-balanced scheme (left) and new well-balanced scheme (right). Number of grid points $N = 50$

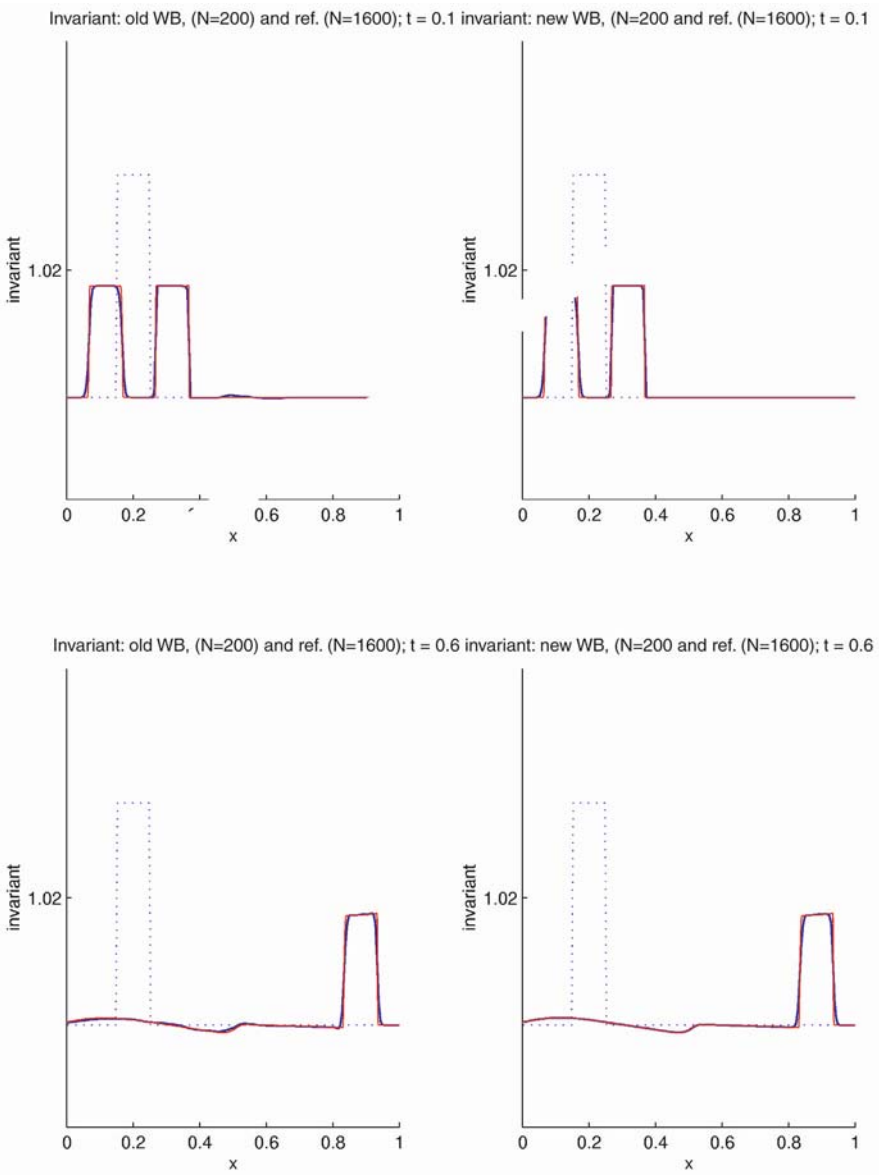


Fig. 8. Same as in the previous figure, but with $N = 200$ grid points

5 Conclusions

Staggered central schemes can be used for:

- systems with stiff source (implicit-explicit schemes);
- quasi-stationary flows (well-balanced schemes);
- accurate solutions (high order central schemes).

They may be more effective than non-staggered central schemes in some cases (higher resolution with the same number of grid points).

High order schemes can be constructed for problems with stiff source (high order finite difference discretization + IMEX time discretization).

The construction of well-balanced central schemes on staggered grids that preserve static equilibria is possible, but requires the solution of the stationary problem.

Future work in this topic may include the construction of schemes that do not require this information. Second-order schemes that preserve only static equilibria perform well even when the unperturbed solution is non-static. The performance improves with the increase of the accuracy of the scheme. An interesting problem is the construction of a third-order well-balanced scheme (on staggered or non-staggered grids).

References

[1] Dafermos, C.M. (2000): Hyperbolic conservation laws in continuum physics. (Grundlehren der Mathematischen Wissenschaften, vol. 325). Springer, Berlin
[2] LeVeque, R.J. (1992): Numerical methods for conservation laws. 2nd ed. (Lectures in Mathematics ETH Zurich). Birkhäuser, Basel
[3] Godlewski, E., Raviart, P.-A. (1996): Numerical approximation of hyperbolic systems of conservation laws. (Applied Mathematical Sciences, vol. 118). Springer, New York
[4] Cockburn, B., Johnson, C., Shu, C.-W., Tadmor, E. (1998): Advanced numerical approximation of nonlinear hyperbolic equations. (Lecture Notes in Mathematics, vol. 1697). Springer, Berlin
[5] Nessyahu, H., Tadmor, E. (1990): Nonoscillatory central differencing for hyperbolic conservation laws. J. Comput. Phys. **87**, 408–463
[6] Russo, G. (2002): Central schemes and systems of balance laws. In: Meister, A., Struckmeier J. (eds.): Hyperbolic partial differential equations. Vieweg, Braunschweig, pp. 59–114
[7] Kurganov, A., Noelle, S., Petrova, G. (2001): Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton-Jacobi equations. SIAM J. Sci. Comput. **23**, 707–740
[8] Kurganov, A., Tadmor, E. (2000): New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations. J. Comput. Phys. **160**, 241–282
[9] Qiu, J., Shu, C.-W. (2002): On the construction, comparison, and local characteristic decomposition for high-order central WENO schemes. J. Comput. Phys. **183**, 187–209
[10] Levy, D., Puppo, G., Russo, G. (1999): Central WENO schemes for hyperbolic systems of conservation laws. M2AN Math. Model. Numer. Anal. **33**, 547–571
[11] Pareschi, L., Puppo, G., Russo, G. (2004): Central Runge-Kutta schemes for conservation laws. SIAM J. Sci. Comput., accepted

[12] Liotta, S.F., Romano, V., Russo, G. (2000): Central schemes for balance laws of relaxation type. SIAM J. Numer. Anal. **38**, 1337–1356
[13] Kennedy, C.A., Carpenter, M.H. (2003): Additive Runge-Kutta schemes for convection-diffusion-reaction equations. Appl. Numer. Math. **44**, 139–181
[14] Greenberg, J. M., Leroux, A.Y. (1996): A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. SIAM J. Numer. Anal. **33**, 1–16
[15] Gosse, L. (2001): A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms. Math. Models Methods Appl. Sci. **11**, 339–365
[16] LeVeque, R.J. (1998): Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. J. Comput. Phys. **146**, 346–365
[17] Gallouët, T., Hérard, J.-M., Seguin, N. (2003): Some approximate Godunov schemes to compute shallow-water equations with topography. Comput. & Fluids **32**, 479–513
[18] Jin, S. (2001): A steady-state capturing method for hyperbolic systems with geometrical source terms. M2AN Math. Model. Numer. Anal. **35**, 631–645
[19] Kurganov, A., Levy, D. (2002): Central-upwind schemes for the Saint-Venant system. M2AN Math. Model. Numer. Anal. **36**, 397–425
[20] Botchorishvili, R., Perthame, B., Vasseur, A. (2003): Equilibrium schemes for scalar conservation laws with stiff sources. Math. Comp. **72**, 131–157
[21] Roe, P.L. (1987): Upwind differencing schemes for hyperbolic conservation laws with source terms. In: Carasso, C. et al. (eds.): Nonlinear hyperbolic problems. (Lecture Notes in Math., vol. 1270). Springer, Berlin, pp. 41–51
[22] Bermudez, A., Vazquez, M.E. (1994): Upwind methods for hyperbolic conservations laws with source terms. Comput. & Fluids **23**, 1049–1071
[23] Toro, E.F. (2001): Shock-capturing methods for free-surface shallow flows. Wiley, Chichester
[24] Russo, G. (2001): Central schemes for balance laws. In: Freistühler, H., Warnecke, G. (eds.): Hyperbolic problems: theory, numerics, applications. Vol. II. (Internat. Ser. Numer. Math., vol. 141). Birkhäuser, Basel, pp. 821–829
[25] Richtmyer, R.D., Morton, K.W. (1967): Difference methods for initial-value problems. 2nd edition. Wiley, New York
[26] Pareschi, L., Russo, G. (2004): Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. J. Sci. Comput., accepted
[27] Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., Perthame, B. (2004): A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. Sci. Comp., to appear

Regularized 13 moment equations for rarefied gas flows

H. Struchtrup, M. Torrilhon

Abstract. A new closure for Grad's 13 moment equations is presented that adds terms of super-Burnett order to the balances of the pressure deviator and heat flux vector. The resulting system of equations contains the Burnett and super-Burnett equations when expanded in a series in the Knudsen number. However, other than the Burnett and super-Burnett equations, the new set of equations is linearly stable for *all* wavelengths and frequencies. The dispersion relation and damping for the new equations agree better with experimental data than those for the Navier-Stokes-Fourier equations, or the original 13 moments system. The new equations allow the description of Knudsen boundary layers, and yield smooth shock structures for all Mach numbers in good agreement with experiments and DSMC simulations.

1 Introduction

Processes in rarefied gases are well described by the Boltzmann equation [1,2], a non-linear integro-differential equation that describes the evolution of the particle distribution function $f(\mathbf{x}, t, \mathbf{c})$ in phase space. Here $\mathbf{x}$ and t are the space and time variables, respectively, and $\mathbf{c}$ denotes the microscopic velocities of particles. The distribution function is defined so that $f(\mathbf{x}, t, \mathbf{c})\, d\mathbf{c}d\mathbf{x}$ gives the number of gas particles in the phase space cell $d\mathbf{c}d\mathbf{x}$. Thus f is a function of seven independent variables, and the numerical solution of the Boltzmann equation, either directly [3] or via the direct simulation Monte Carlo (DSMC) method [4], is very time expensive. In particular that is the case at low Mach numbers in the transition regime. Since this regime is important for the simulation of microscale flows, e.g., in MEMS, there is a strong desire for accurate models which allow the calculation of processes in rarefied gases at lower computational cost.

Macroscopic models can be derived from the Boltzmann equations, in particular for smaller values of the Knudsen number Kn, defined as the ratio between the mean free path of the molecules and the relevant macroscopic length scale. In this paper we present a new model, the regularized 13 moment equations, or R13 equations [5,6], which agrees with the Boltzmann equation up to third order in the Knudsen number.

Before we discuss the new equations in detail, we give an overview of methods for deriving macroscopic equations from the Boltzmann equation. Then we introduce the R13 equations and discuss their main features.

2 Macroscopic models for rarefied gas flows

2.1 Chapman-Enskog expansion

The best known approach to deriving macroscopic transport equations from the Boltzmann equation is the Chapman-Enskog method [1,2,7] where the distribution function is expanded in powers of the Knudsen number, $f = f^{(0)} + \mathrm{Kn} f^{(1)} + \mathrm{Kn}^2 f^{(2)} + \cdots$. The expansion parameters $f^{(\alpha)}$ are determined successively by plugging this expression into the Boltzmann equation, and equating terms with the same factors in powers of the Knudsen number; see, e.g., [1,2,7] for details.

To zeroth order the expansion yields the Euler equations, the first order correction results in the equations of Navier-Stokes and Fourier, the second order expansion yields the Burnett equations [2,7], and the third order expansion yields the so-called super-Burnett equations [8,9].

The equations of Navier-Stokes and Fourier cease to be accurate for Knudsen numbers above 0.05, and one is lead to think that the Burnett and super-Burnett equations are valid for larger Knudsen numbers. Unfortunately, however, the higher order equations become linearly unstable for processes involving small wavelengths, or high frequencies [8], and they lead to unphysical oscillations in steady state processes [10], and thus cannot be used in numerical simulations .

There is no clear argument why the Chapman-Enskog expansion leads to unstable equations. It seems, however, that a first order Chapman-Enskog expansion leads generally to stable equations, while higher order expansions generally yield unstable equations, although exceptions apply; e.g., see [11,12].

In recent years, several authors presented modifications of the Burnett equations that contain additional terms of super-Burnett order (but not the actual super-Burnett terms) to stabilize the equations to produce the "augmented Burnett equations" [13,14], or derived regularizations of hyperbolic equations that reproduce the Burnett equations when expanded in the Knudsen number [15,16]. These models are only partially successful: the augmented Burnett equations still are unstable in space [6], and both approaches lack a rational derivation from the Boltzmann equation [6].

2.2 Grad's moment method

In the method of moments of Grad [17,18], the Boltzmann equation is replaced by a set of moment equations, first order partial differential equations for the moments of the distribution function. Which and how many moments are needed depends on the particular process, but experience shows that the number of moments must be increased with increasing Knudsen number [19–23].

For the closure of the equations, the phase density is approximated by an expansion in Hermite polynomials about the equilibrium distribution (the local Maxwellian), where the coefficients are related to the moments.

Only a few moments have an intuitive physical meaning, i.e., density ϱ, momentum density ϱv_i, energy density ϱe, heat flux q_i and pressure tensor p_{ij}. This set of 13 moments forms the basis of Grad's well-known 13 moment equations [17] which

are commonly discussed in textbooks. However, the 13 moment set does not allow the computation of boundary layers [24,25,20] and, since the equations are symmetric hyperbolic, leads to shock structures with discontinuities (sub-shocks) for Mach numbers above 1.65 [19,26]. With increasing number of moments, one can compute Knudsen boundary layers [27,20,28] and smooth shock structures up to higher Mach numbers [26,22]. As becomes evident from the cited literature, for some problems, in particular for large Mach or Knudsen numbers, one has to face hundreds of moment equations.

2.3 Reinecke-Kremer-Grad method

In most of the available literature, both methods - moment method and Chapman-Enskog expansion – are treated as being completely unrelated. However, using a method akin to the Maxwellian iteration of Truesdell and Ikenberry [29,30], Reinecke and Kremer extracted the Burnett equations from Grad-type moment systems [31,32].

Which set of moments one has to use for this purpose depends on the model for the collisions of particles. For Maxwell molecules it is sufficient to consider Grad's classical set of 13 moments.

In [25] it was shown that this iteration method is equivalent to the Chapman-Enskog expansion of the moment equations. In the original Chapman-Enskog method one first expands, and then integrates, the resulting distribution function to compute its moments. In the Reinecke-Kremer-Grad method, the order of integration and expansion is exchanged.

2.4 Regularization of Grad's 13 moment equations

The original derivation of the regularized 13 moment equations uses a different combination of the methods of Grad and Chapman-Enskog. The basic idea is to assume different time scales for the 13 basic variables of the theory on one side, and all higher moments on the other. Under that assumption, one can perform a Chapman-Enskog expansion around a non-equilibrium state which is defined through the 13 variables.

This idea was also presented by Karlin et al. [33] for the linearized Boltzmann equation. Based on the above idea they compute an approximation to the distribution function, which they then use to derive a set of 13 linear equations for the 13 moments. These equations correspond to Grad's 13 moment equations for linear processes plus additional terms.

Our derivation of the R13 equations in [5] exchanges the order of expansion and integration. The derivation of the equations is based on the non-linear moment equations for 26 moments, instead of the linearized Boltzmann equation, so that we obtain a set of *non-linear* equations. Also, the use of moment equations allows for a much faster derivation of the equations, and yields explicit numerical expressions for coefficients that were not specified in [33]. The Karlin et al. equations follow from our equations by linearization.

A closer inspection of the regularized equations shows that the terms added to the original Grad equations are of super-Burnett order. The additional terms, which are obtained from the moment equations for higher moments, place the new equations in between the super-Burnett and Grad's 13 moment equations in as much as the new equations keep the desirable features of both, while discarding the unwelcome features.

In particular, the R13 equations

- contain the Burnett and super-Burnett equations as can be seen by means of a Chapman-Enskog expansion in the Knudsen number,
- are linearly stable for all wavelengths, and/or frequencies,
- show phase speeds and damping coefficients that match experiments better than those for the Navier-Stokes-Fourier equations, or the original 13 moments system,
- exhibit Knudsen boundary layers,
- lead to smooth shock structures for all Mach numbers.

The most important of these features are discussed in the sequel.

Hyperbolic partial differential equations imply finite wave speeds and discontinuities that make them difficult to handle with standard analysis. Regularization is a method for adding parabolic terms which change the character of the equations so that no discontinuities occur, but a narrow smooth transition zone [34,35]. We decided to adopt the notion of regularization for the new equations since the additional terms indeed are smoothing out the discontinuities (sub-shocks) that occur in Grad's 13 moment system for Mach numbers above 1.65. It is important to note, though, that the shocks in Grad's moment equations (at $\mathrm{Ma} = 1.65$ for 13 moments, at higher Mach numbers for extended moment sets; see [26]) are artefacts of the method, and thus unphysical. The parameter that controls our regularization is the mean time of free flight, which is a physical parameter. In other words, the regularization of Grad's 13 moment system removes artificial discontinuities, and replaces them by a shock structure which is based in physics.

2.5 Order of magnitude / order of accuracy approach

The weak point in the derivation of the R13 equations as outlined above is the assumption of different time scales for the basic 13 moments, and higher moments. While this assumption leads to a set of equations with the desired behavior, it is difficult to justify, since the characteristic times of all moments are of the same order.

Only recently, an alternative approach to the problem was presented by Struchtrup in [36]. This approach is an extension of an idea developed by Müller et al. in [37].

Müller et al. [37] considered the infinite system of coupled moment equations of the BGK equation [38]. From these they determined the order of magnitude of moments in terms of orders in powers of the Knudsen number, and declared that a theory of order λ needs to consider all terms in all moment equations up to the

order $\mathbb{O}\left(\varepsilon^{\lambda}\right)$. At first, these are the moment equations for all moments of order $\beta \leq \lambda$ under omission of higher order terms. However, these equations split into two independent subsystems, and only a smaller number of equations (and variables) remain as equations of importance [37].

The extension of this idea in [36] does not ask for the order of terms in all moment equations, but of the order of magnitude of their influence in the conservations laws, i.e., on the heat flux and stress tensor. This is quite different. For example, in order to compute the heat flux with third order accuracy, as is necessary in a third order theory, other moments are needed only with second order accuracy, while others can be ignored completely. Müller et al. [37], on the other hand, requires higher order accuracy for these moments, and a larger number of moments. The new method was applied to the special cases of Maxwell molecules and the BGK model in [36], and it was shown that it yields the Euler equations at zeroth order, the Navier-Stokes-Fourier equations at second order, Grad's 13 moment equations (with omission of a non-linear term) at second order, and the regularized 13 moment equations at third order.

The lone scaling parameter in this method is the Knudsen number, and the assumption of different time scales is not needed for the derivation of the R13 equations. Thus, one can consider this derivation of the R13 equations to be better founded than the original derivation in [5].

3 Regularized 13 moment equations

The regularized 13 moment equations for monatomic gases were derived in [5,36], and here we just present the results. The R13 equations are a set of field equations for the 13 variables $\rho_A = \left\{\varrho, \varrho v_i, \varrho\varepsilon = \frac{3}{2}\varrho RT, \sigma_{ij}, q_i\right\}$, where ϱ is the mass density, v_i is the gas velocity, ε is the specific internal energy, T is the temperature, R is the specific gas constant, σ_{ij} is the trace-free part of the pressure tensor, and q_i is the heat flux vector. The field equations for these variables are the conservation laws for mass, momentum and energy,

$$\begin{aligned}
\frac{\partial \rho}{\partial t} + \frac{\partial \varrho v_k}{\partial x_k} &= 0 \,, \\
\frac{\partial \rho v_i}{\partial t} + \frac{\partial}{\partial x_k}\left(\rho v_i v_k + p\delta_{ik} + \sigma_{ik}\right) &= 0 \,, \\
\frac{\partial}{\partial t}\left(\rho\varepsilon + \frac{1}{2}\rho v_i^2\right) + \frac{\partial}{\partial x_k}\left(\rho\varepsilon v_k + \frac{1}{2}\rho v_i^2 v_k + p v_k + \sigma_{ik} v_i + q_k\right) &= 0 \,,
\end{aligned} \tag{1}$$

plus additional field equations for the stress deviator,

$$\frac{\partial \sigma_{ij}}{\partial t} + \frac{\partial \sigma_{ij} v_k}{\partial x_k} + \frac{4}{5}\frac{\partial q_{\langle i}}{\partial x_{j\rangle}} + 2p\frac{\partial v_{\langle i}}{\partial x_{j\rangle}} + 2\sigma_{k\langle i}\frac{\partial v_{j\rangle}}{\partial x_k} + \frac{\partial m_{ijk}}{\partial x_k} = -\frac{p}{\mu}\sigma_{ij}, \tag{2}$$

and the heat flux,

$$\frac{\partial q_i}{\partial t} + \frac{\partial q_i v_k}{\partial x_k} + \frac{5}{2} pR \frac{\partial T}{\partial x_i} + \frac{5}{2} \sigma_{ik} R \frac{\partial T}{\partial x_k} + RT \frac{\partial \sigma_{ik}}{\partial x_k} - \sigma_{ik} RT \frac{\partial \ln \varrho}{\partial x_k} - \frac{\sigma_{ij}}{\varrho} \frac{\partial \sigma_{jk}}{\partial x_k} + \frac{7}{5} q_k \frac{\partial v_i}{\partial x_k} + \frac{2}{5} q_k \frac{\partial v_k}{\partial x_i} + \frac{2}{5} q_i \frac{\partial v_k}{\partial x_k} + \frac{1}{2} \frac{\partial R_{ik}}{\partial x_k} + \frac{1}{6} \frac{\partial \Delta}{\partial x_i} + m_{ijk} \frac{\partial v_j}{\partial x_k} = -\frac{2}{3} \frac{p}{\mu} q_i. \tag{3}$$

Here, $p = \varrho RT$ is the pressure, and μ denotes the viscosity. Indices in angular brackets denote the symmetric trace-free parts of tensors. The above equations contain the additional quantities m_{ijk}, R_{ik} , Δ, and constitutive equations are required to close the equations. With the choice

$$m_{ijk} = R_{ik} = \Delta = 0 \tag{4}$$

the above set of equations is reduced to the well-known set of 13 moment equations of Grad [17,18]. The regularization of the Grad equations yields

$$\begin{aligned}
m_{ijk} &= -2\frac{\mu}{p}\left[RT\frac{\partial \sigma_{\langle ij}}{\partial x_{k\rangle}} - RT\sigma_{\langle ij}\frac{\partial \ln \varrho}{\partial x_{k\rangle}} + \frac{4}{5} q_{\langle i}\frac{\partial v_j}{\partial x_{k\rangle}} - \frac{\sigma_{\langle ij}}{\varrho}\frac{\partial \sigma_{k\rangle l}}{\partial x_l}\right], \\
R_{ij} &= -\frac{24}{5}\frac{\mu}{p}\left[RT\frac{\partial q_{\langle i}}{\partial x_{j\rangle}} + Rq_{\langle i}\frac{\partial T}{\partial x_{j\rangle}} - RTq_{\langle i}\frac{\partial \ln \varrho}{\partial x_{j\rangle}} - \frac{1}{\rho} q_{\langle i}\frac{\partial \sigma_{j\rangle k}}{\partial x_k}\right. \\
&\quad \left. + \frac{5}{7} RT\left(\sigma_{k\langle i}\frac{\partial v_{j\rangle}}{\partial x_k} + \sigma_{k\langle i}\frac{\partial v_k}{\partial x_{j\rangle}} - \frac{2}{3}\sigma_{ij}\frac{\partial v_k}{\partial x_k}\right) - \frac{5}{6}\frac{\sigma_{ij}}{\varrho}\frac{\partial q_k}{\partial x_k} - \frac{5}{6}\frac{\sigma_{ij}}{\varrho}\sigma_{kl}\frac{\partial v_k}{\partial x_l}\right], \\
\Delta &= -12\frac{\mu}{p}\left[RT\frac{\partial q_k}{\partial x_k} + \frac{5}{2} Rq_k\frac{\partial T}{\partial x_k} - RTq_k\frac{\partial \ln \varrho}{\partial x_k} + RT\sigma_{ij}\frac{\partial v_i}{\partial x_j} - \frac{1}{\varrho} q_j\frac{\partial \sigma_{jk}}{\partial x_k}\right].
\end{aligned} \tag{5}$$

In the resulting system (1)-(3) with (5), second order derivatives appear in the balance equations of the stress tensor and heat flux which lead to a regularization of the original 13 moment case of Grad.

The R13 equations were derived from the Boltzmann equations for the special case of Maxwell molecules, that is, particles that interact in a repulsive 5-th power potential. The corresponding viscosity is proportional to temperature as

$$\mu = \mu_0 \left(\frac{T}{T_0}\right)^s \tag{6}$$

with $s = 1$. It is well-known [4] that the viscosity is also of this form for other interaction potentials if one only adjusts the exponent s. In particular one computes $s = 1/2$ for hard spheres, and one measures $s \approx 0.8$ for argon. For the purpose of this paper we shall use $s = 1$ exclusively.

4 Chapman-Enskog expansions

The idea of the Chapman-Enskog expansion is to expand the distribution function in a series in the Knudsen number Kn as

$$f = f^{(0)} + \mathrm{Kn} f^{(1)} + \mathrm{Kn}^2 f^{(2)} + \mathrm{Kn}^3 f^{(3)} + \cdots ,$$

where the $f^{(\alpha)}$ are obtained from the Boltzmann equation [7,2]. In our case, we operate on the level of moments and moment equations, and thus we expand the pressure deviator and heat flux in a series as

$$\sigma_{ij} = \sigma_{ij}^{(0)} + \mathrm{Kn}\sigma_{ij}^{(1)} + \mathrm{Kn}^2\sigma_{ij}^{(2)} + \mathrm{Kn}^3\sigma_{ij}^{(3)} + \cdots,$$
$$q_i = q_i^{(0)} + \mathrm{Kn}q_i^{(1)} + \mathrm{Kn}^2 q_i^{(2)} + \mathrm{Kn}^3 q_i^{(3)} + \cdots.$$

In order to expand properly, one needs to consider the dimensionless forms of Eqs. (2) and (3), into which the above expressions are inserted. Then terms with equal powers in Kn are equated to find the $\sigma_{ij}^{(\alpha)}, q_i^{(\alpha)}$. Note that the dimensionless equations have $\mathrm{Kn}\mu$ instead of μ in Eqs. (2), (3), (5). The dimensions are restored after the expansion is performed.

In the Chapman-Enskog method it is customary to express the time derivatives of $\sigma_{ij}^{(\alpha)}, q_i^{(\alpha)}$ by time derivatives of the hydrodynamic variables ϱ, T, v_i. Some details on how this must be done successively can be found in [5] for the linear case, and for the non-linear case in [25].

From the R13 equations as given above we find the Euler equations at zeroth order,

$$\sigma_{ij}^{(0)} = q_i^{(0)} = 0, \tag{7}$$

and the first order corrections are the Navier-Stokes-Fourier equations

$$\sigma_{ij}^{(1)} = -2\mu\frac{\partial v_{\langle i}}{\partial x_{j\rangle}} \quad \text{and} \quad q_i^{(1)} = -\frac{15}{4}R\mu\frac{\partial T}{\partial x_i}. \tag{8}$$

The second order terms yield the Burnett equations for Maxwell molecules, that can be written as

$$\begin{aligned}\sigma_{ij}^{(2)} = \frac{\mu^2}{p}\Bigg[&R\frac{\partial^2 T}{\partial x_{\langle i}\partial x_{j\rangle}} - 2\frac{RT}{\varrho}\frac{\partial^2\varrho}{\partial x_{\langle i}\partial x_{j\rangle}} + 2\frac{RT}{\varrho^2}\frac{\partial\varrho}{\partial x_{\langle i}}\frac{\partial\varrho}{\partial x_{j\rangle}} - 2\frac{R}{\varrho}\frac{\partial T}{\partial x_{\langle i}}\frac{\partial\varrho}{\partial x_{j\rangle}}\\ &+3\frac{R}{T}\frac{\partial T}{\partial x_{\langle i}}\frac{\partial T}{\partial x_{j\rangle}} + \frac{10}{3}S_{ij}\frac{\partial v_k}{\partial x_k} - 4S_{k\langle i}\frac{\partial v_k}{\partial x_{j\rangle}} - 2\frac{\partial v_k}{\partial x_{\langle i}}\frac{\partial v_{j\rangle}}{\partial x_k} + 8S_{k\langle i}S_{j\rangle k}\Bigg]\end{aligned} \tag{9}$$

and

$$\begin{aligned}q_i^{(2)} = \frac{\mu^2}{p}\Bigg[&-\frac{13}{4}RT\frac{\partial^2 v_k}{\partial x_i\partial x_k} + \frac{3}{2}RT\frac{\partial^2 v_i}{\partial x_k\partial x_k} - 3\frac{RT}{\varrho}S_{ij}\frac{\partial\varrho}{\partial x_j}\\ &-\frac{25}{8}R\frac{\partial v_k}{\partial x_k}\frac{\partial T}{\partial x_i} + \frac{15}{8}R\frac{\partial v_k}{\partial x_i}\frac{\partial T}{\partial x_k} + \frac{105}{8}R\frac{\partial v_i}{\partial x_k}\frac{\partial T}{\partial x_k}\Bigg],\end{aligned} \tag{10}$$

where we have used the abbreviation

$$S_{ij} = \frac{\partial v_{\langle i}}{\partial x_{j\rangle}}.$$

It is not surprising that the Burnett equations arise from the second order expansion of the R13 equations, since it is an established fact that the Burnett equations can

already be obtained from Grad's 13 moment equations [31,32,25], i.e., with the Grad closure (4).

Indeed, a closer inspection of the closure relations (5) of the R13 equations shows that these contribute terms of super-Burnett order. The derivation of the super-Burnett equations is a cumbersome task, and they are difficult to find in the literature. Thus, we expanded the R13 equations for two special cases only: the three-dimensional linear equations and the one-dimensional non-linear equations. For the three-dimensional linear case one obtains [5]

$$\begin{aligned}
\sigma_{ij}^{(3)} &= \frac{\mu^3}{p^2}\left(\frac{5}{3}RT\frac{\partial^2}{\partial x_{\langle i}\partial x_{j\rangle}}\frac{\partial v_k}{\partial x_k} - \frac{4}{3}RT\frac{\partial^2}{\partial x_k\partial x_k}\frac{\partial v_{\langle i}}{\partial x_{j\rangle}}\right), \\
q_i^{(3)} &= \frac{\mu^3}{p^2}\left(-\frac{157}{16}RT\frac{\partial^3\theta}{\partial x_i\partial x_k\partial x_k} - \frac{5}{8}\frac{R^2T^2}{\rho}\frac{\partial^3\varrho}{\partial x_i\partial x_k\partial x_k}\right).
\end{aligned} \tag{11}$$

These are the same equations that Shavaliev found from the Boltzmann equation [9]. It can also be shown that the third order Chapman-Enskog expansion of the non-linear one-dimensional R13 equations agrees with the corresponding super-Burnett equations [6], but we refrain from printing these here.

From the above discussion it follows that the R13 equations agree up to the super-Burnett order with the Boltzmann equation. Note that Grad's classical 13 moment equations agree up to Burnett order, but not to super-Burnett order.

Moreover, the R13 equations have several advantages over the Burnett and super-Burnett equations. (a) They can be derived much easier, and faster, so that errors can be excluded with higher certainty. (b) The R13 equations contain only space derivatives of first and second order while the super-Burnett equations contain derivatives of up to fourth order. Thus, the R13 equations fit existing numerical methods more conveniently. Note that their mathematical structure is very similar to the NSF equations, so that methods for these can be used as well for solving the R13 equations. (c) Most important, however, is the fact that the R13 equations are linearly stable [5] as is shown below, while the Burnett and super-Burnett equations are linearly unstable [8,10].

5 Linear stability

We start our analysis of the R13 equations by considering the linear stability. For this, we consider small deviations from an equilibrium state given by $\varrho_0, T_0, v_{i,0} = 0$, and consider one-dimensional processes where $x_1 = x$, and $v_i = \{v(x,t), 0, 0\}$. Dimensionless variables $\hat{\varrho},\ \hat{T},\ \hat{v},\ \hat{\sigma},\ \hat{q}$ are introduced as

$$\begin{aligned}
&\varrho = \varrho_0\left(1+\hat{\varrho}\right)\ ,\quad T = T_0\left(1+\hat{T}\right)\ ,\quad p = \varrho_0 RT_0\left(1+\hat{\varrho}+\hat{T}\right)\ , \\
&v = \sqrt{RT_0}\hat{v}\ ,\quad \sigma_{11} = \varrho_0 RT_0\hat{\sigma}\ ,\quad q_1 = \varrho_0\sqrt{RT_0}^3\hat{q}\ .
\end{aligned}$$

Moreover, we identify a relevant length scale L of the process, and use it to non-dimensionalize the space and time variables according to

$$x = L\hat{x} \ , \quad t = \frac{L}{\sqrt{RT_0}}\hat{t} \ .$$

The corresponding dimensionless collision time is then given by the Knudsen number, which we define here as

$$\mathrm{Kn} = \frac{\tau\sqrt{RT_0}}{L} = \frac{\mu_0}{\varrho_0\sqrt{RT_0}L} \ .$$

Linearization in the deviations from equilibrium $\hat{\varrho}, \ \hat{T}, \ \hat{v}, \ \hat{\sigma}, \ \hat{q}$ yields the dimensionless linearized system in one dimension as

$$\begin{aligned}
&\frac{\partial\hat{\varrho}}{\partial\hat{t}} + \frac{\partial\hat{v}}{\partial\hat{x}} = 0 \ , \\
&\frac{\partial\hat{v}}{\partial\hat{t}} + \frac{\partial\hat{\varrho}}{\partial\hat{x}} + \frac{\partial\hat{T}}{\partial\hat{x}} + \frac{\partial\hat{\sigma}}{\partial\hat{x}} = 0 \ , \\
&\frac{3}{2}\frac{\partial\hat{T}}{\partial\hat{t}} + \frac{\partial\hat{q}}{\partial\hat{x}} + \frac{\partial\hat{v}}{\partial\hat{x}} = 0 \ , \\
&\frac{\partial\hat{\sigma}}{\partial\hat{t}} + \frac{8}{15}\frac{\partial\hat{q}}{\partial\hat{x}} + \frac{4}{3}\frac{\partial\hat{v}}{\partial\hat{x}} - \frac{6}{5}\mathrm{Kn}\frac{\partial^2\hat{\sigma}}{\partial\hat{x}^2} = -\frac{\hat{\sigma}}{\mathrm{Kn}} \ , \\
&\frac{\partial\hat{q}}{\partial\hat{t}} + \frac{5}{2}\frac{\partial\hat{T}}{\partial\hat{x}} + \frac{\partial\hat{\sigma}}{\partial\hat{x}} - \frac{18}{5}\mathrm{Kn}\frac{\partial^2\hat{q}}{\partial\hat{x}^2} = -\frac{2}{3}\frac{\hat{q}}{\mathrm{Kn}} \ .
\end{aligned} \tag{12}$$

This set of equations is equivalent to the equations proposed by Karlin et al. [33], who, however, did not give explicit numerical expressions for the factors that multiply the second derivatives of $\hat{\sigma}$ and $\hat{q}$, but presented them as integrals over the linearized collision operator which are not further evaluated.

For comparison, we shall also consider the Chapman-Enskog expansion to various orders (7-10), in which case we have to replace the last two equations with the relevant terms of

$$\begin{aligned}
\hat{\sigma}_{CE} &= -\mathrm{Kn}\frac{4}{3}\frac{\partial\hat{v}}{\partial\hat{x}} - \mathrm{Kn}^2\left[\frac{4}{3}\frac{\partial^2\hat{\varrho}}{\partial\hat{x}^2} - \frac{2}{3}\frac{\partial^2\hat{T}}{\partial\hat{x}^2}\right] + \mathrm{Kn}^3\frac{2}{9}\frac{\partial^3\hat{v}}{\partial\hat{x}^3} + \cdots \ , \\
\hat{q}_{CE} &= -\mathrm{Kn}\frac{15}{4}\frac{\partial\hat{T}}{\partial\hat{x}} - \mathrm{Kn}^2\frac{7}{4}\frac{\partial^2\hat{v}}{\partial\hat{x}^2} - \mathrm{Kn}^3\left[\frac{157}{16}\frac{\partial^3\hat{T}}{\partial\hat{x}^3} + \frac{5}{8}\frac{\partial^3\hat{\varrho}}{\partial\hat{x}^3}\right] + \cdots \ .
\end{aligned}$$

We assume plane wave solutions of the form

$$\phi = \tilde{\phi}\exp\left\{i\left(\omega\hat{t} - k\hat{x}\right)\right\} ,$$

where $\tilde{\phi}$ is the complex amplitude of the wave, ω is its frequency, and k is its wave number. The equations can be written as

$$\mathscr{A}_{AB}\left(\omega, k\right)\tilde{u}_B = 0 \quad \text{with} \quad \tilde{u}_B = \left\{\tilde{\varrho}, \tilde{T}, \tilde{v}, \tilde{\sigma}, \tilde{q}\right\}$$

and nontrivial solutions require

$$\det\left[\mathcal{A}_{AB}\left(\omega, k\right)\right] = 0\ ;$$

the resulting relation between ω and k is the dispersion relation.

If a disturbance in space is considered, then the wave number k is real, and the frequency is complex, $\omega = \omega_r\left(k\right) + i\omega_i\left(k\right)$. The phase velocity v_{ph} and damping α of the corresponding waves are given by

$$v_{ph} = \frac{\omega_r\left(k\right)}{k} \quad \text{and} \quad \alpha = \omega_i\left(k\right)$$

Stability requires damping, and thus $\omega_i\left(k\right) \geq 0$.

If a disturbance in time at a given location is considered, then the frequency ω is real, while the wave number is complex, $k = k_r\left(\omega\right) + ik_i\left(\omega\right)$. The phase velocity v_{ph} and damping α of the corresponding waves are given by

$$v_{ph} = \frac{\omega}{k_r\left(\omega\right)} \quad \text{and} \quad \alpha = -k_i\left(\omega\right)\ .$$

For a wave traveling in the positive x-direction ($k_r > 0$), the damping must be negative ($k_i < 0$), while for a wave traveling in the negative x-direction ($k_r < 0$), the damping must be positive ($k_i > 0$).

It is convenient to chose the mean free path as reference length, and the mean free time as reference time, so that $\mathrm{Kn} = 1$. Then the wave number is measured in units of the inverse mean free path, and the wave frequency in terms of the collision frequency $1/\tau$. This implies that the Knudsen number for an oscillation with dimensionless frequency ω is $\mathrm{Kn}_\omega = \omega$, and for a given wave number k the Knudsen number is $\mathrm{Kn}_k = k$.

We test the stability against local disturbances of frequency ω. As we have seen, stability requires different signs of real and imaginary part of $k\left(\omega\right)$. Thus, if $k\left(\omega\right)$ is plotted in the complex plane with ω as parameter, the curves should not touch the upper right nor the lower left quadrant.

Figure 1 shows the solutions for the different sets of equations considered in this paper; the dots mark the points where $\omega = 0$. Grad's 13 moment equations (Grad13), and Navier-Stokes-Fourier equations (NSF) give two different modes each, and none of the solutions violates the condition of stability (upper left in Fig. 1). This is different for the Burnett (3 modes, upper right) and super-Burnett (4 modes, lower left) equations: the Burnett equations have one unstable mode, and the super-Burnett have two unstable modes. The R13 equations, shown in the lower right, have 3 modes, all of them stable.

In a similar manner it can be shown that the R13 equations are stable with respect to a disturbance of given wave length, or wave number k, while the Burnett and super-Burnett equations are unstable [5,8].

6 Dispersion and damping

Next we compare phase speed and damping with experiments performed by Meyer and Sessler [39]. Figure 2 shows the inverse phase speed and the damping (as α/ω)

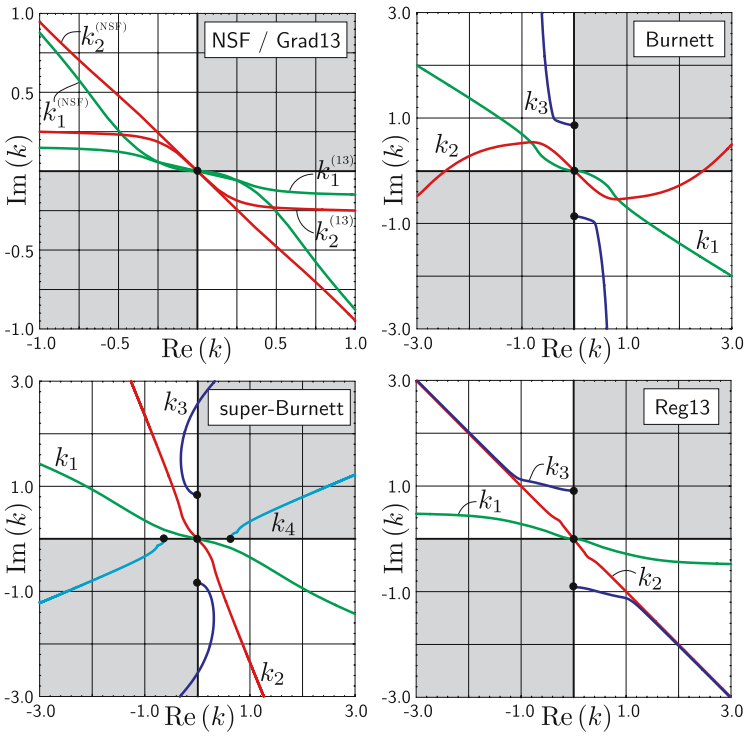


Fig. 1. The solutions $k\,(\omega)$ of the dispersion relation in the complex plane with ω as parameter for Navier-Stokes-Fourier, Grad's 13 moments, Burnett, super-Burnett, and Regularized 13 moment equations. The dots denote the points where $\omega = 0$

as functions of the dimensionless inverse frequency $1/\omega$, computed with the NSF, Grad 13, and R13 equations, and experimental data from [39]. Here we consider only those modes that yield the speed of sound as $\omega \to 0$.

As can be seen, the R13 equations reproduce the measured values of the damping coefficient α for all dimensionless frequencies less than unity, while the NSF and Grad13 equations already fail at $\omega = 0.25$ and $\omega = 0.5$, respectively. The agreement of the R13 prediction for the phase velocity is less striking, but the other theories also do not match well. One reason for this might be insufficient accuracy of the measurement. Altogether, the R13 equations give a remarkably good agreement with the measurements for values of $\omega < 1$.

Equations from expansions in the Knudsen number can be expected to be good only for $\mathrm{Kn} < 1$. We conclude that the R13 equations allow a proper description of processes quite close to the natural limit of their validity of $\mathrm{Kn}_\omega = 1$. It is not surprising that all theories show discrepancies to the experiments for larger frequencies. The reasonable agreement between the NSF phase speed and experiments must be seen as coincidence.

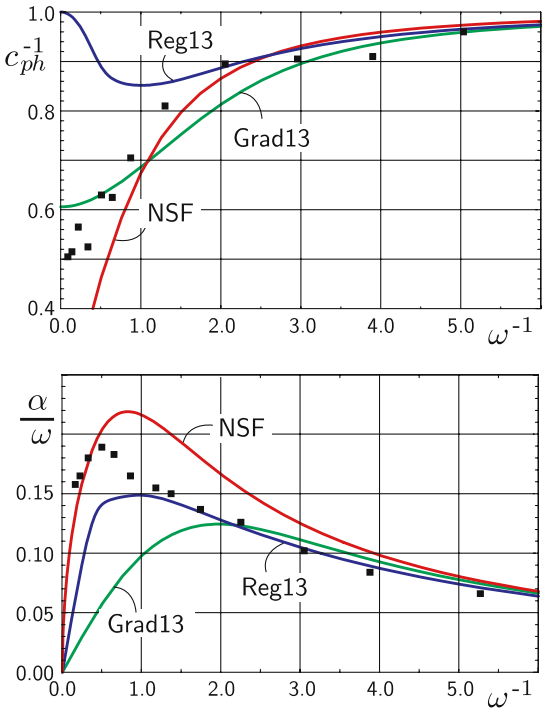


Fig. 2. Inverse phase velocity (above) and damping (below), theoretical results from Navier-Stokes-Fourier, Grad's 13 moments, and regularized 13 moments and measurements by Meyer and Sessler [39] (squares)

7 Knudsen boundary layers

In this section we briefly study boundary value problems for the linearized R13 equations. The goal is to show that the R13 equations lead to Knudsen boundary layers.

To this end we consider a simple steady state Couette flow problem: two infinite, parallel plates move in the $\{x_2, x_3\}$-plane with different speeds in the x_2 direction. The plate distance is $L = 1$ in dimensionless units, and the plates have different temperatures. In this setting, we expect that all variables will depend only on the coordinate $x_1 = x$. Since matter cannot pass the plates, we will have $v_1 = 0$. Moreover, for symmetry reasons, there will be no fluxes in the x_3 direction, so that

$$v_i = \{0, v(x), 0\} \text{ and } q_3 = \sigma_{13} = \sigma_{23} = 0 \,.$$

Under these assumptions, the linearized R13 equations can be split into the flow problem with the equations

$$\frac{\partial v}{\partial x} + \frac{2}{5}\frac{\partial q_2}{\partial x} = -\frac{\sigma_{12}}{\mathrm{Kn}} = \text{const} \quad , \quad \mathrm{q_2} = \frac{9}{5}\mathrm{Kn}^2\frac{\partial^2 \mathrm{q_2}}{\partial \mathrm{x}^2} \,,$$

and the heat transfer problem with the equations

$$\frac{5}{2}\frac{\partial T}{\partial x}+\frac{\partial \sigma_{11}}{\partial x}=-\frac{2}{3}\frac{q_1}{\mathrm{Kn}}=\mathrm{const}\ ,\quad \sigma_{11}=\frac{6}{5}\mathrm{Kn}^2\frac{\partial^2\sigma_{11}}{\partial \mathrm{x}^2}\ .$$

Two more non-trivial equations serve to compute ϱ, and σ_{22}, viz.,

$$\sigma_{22}=\mathrm{Kn}^2\left[\frac{2}{3}\frac{\partial^2\sigma_{22}}{\partial x^2}-\frac{4}{15}\frac{\partial^2\sigma_{11}}{\partial x^2}\right]\quad ,\quad \frac{\partial \varrho}{\partial x}+\frac{\partial T}{\partial x}+\frac{\partial \sigma_{11}}{\partial x}=0\ .$$

The linear equations are easy to integrate, and we obtain the solution of the flow problem as

$$v(x)=v_0-\sigma_{12}\frac{x}{\mathrm{Kn}}-\frac{2}{5}q_2(x) \tag{13}$$
$$\text{with}\ \ q_2(x)=A\sinh\left(\sqrt{\frac{5}{9}}\frac{x-\frac{1}{2}}{\mathrm{Kn}}\right)+B\cosh\left(\sqrt{\frac{5}{9}}\frac{x-\frac{1}{2}}{\mathrm{Kn}}\right),$$

where v_0, σ_{12}, A, B are constants of integration.

The solution of the heat transfer problem reads

$$T(x)=T_0-\frac{4}{15}q_1\frac{x}{\mathrm{Kn}}-\frac{2}{5}\sigma_{11}(x) \tag{14}$$
$$\text{with}\ \ \sigma_{11}(x)=C\sinh\left(\sqrt{\frac{5}{6}}\frac{x-\frac{1}{2}}{\mathrm{Kn}}\right)+D\cosh\left(\sqrt{\frac{5}{6}}\frac{x-\frac{1}{2}}{\mathrm{Kn}}\right),$$

where T_0, q_1, C, D are constants of integration.

Thus, in order to obtain the fields of temperature and velocity between the plates, we need 8 boundary conditions. The velocities and temperatures of the two plates give only four boundary conditions, and thus additional boundary conditions are required. As of now, the problem of how to prescribe meaningful boundary conditions for the R13 equations is unsolved, and we hope to be able to present proper boundary conditions (that, of course, allow for temperature jumps and velocity slips) in the future.

Nevertheless, it is worthwhile to study the general solutions (13, 14). In the linear Navier-Stokes-Fourier case, both, temperature and velocity, are straight lines according to

$$v_{NSF}(x)=v_0-\sigma_{12}\frac{x}{\mathrm{Kn}}\quad\text{and}\quad T_{NSF}(x)=T_0-\frac{4}{15}q_1\frac{x}{\mathrm{Kn}}\ ,$$

that is, for the NSF case one finds $q_2(x)=\sigma_{11}(x)=0$.

With the R13 equations, on the other hand, these functions are non-zero as given in (13, 14). From that, we identify $-\frac{2}{5}q_2(x)$ and $-\frac{2}{5}\sigma_{11}(x)$ as the Knudsen boundary layers for the velocity and temperature according to the R13 equations. Indeed, these functions have the typical shape of a boundary layer, their largest values are found at the walls, and the curves decrease to zero within several mean free paths away from the walls.

The curves are governed by the Knudsen number, so that, for small Knudsen numbers, $q_2(x)$ and $\sigma_{11}(x)$ are equal to zero almost everywhere between the plates. The boundary layers are confined to a small region adjacent to the wall, and contribute to temperature jump and velocity slip. In this case, the Navier-Stokes-Fourier theory can be used with proper jump and slip boundary conditions.

As Kn grows, the width of the boundary layers is growing as well. For Knudsen numbers above ∼0.05 one can no longer speak of boundary layers, since the functions $q_2(x)$, $\sigma_{11}(x)$ as given in (13, 14) are non-zero everywhere in the region between the plates. In this case boundary effects have an important influence on the flow pattern.

Since, at this point, we have no recipe for prescribing all boundary values required, we cannot say whether the boundary layers obtained from the R13 equations coincide well with those of the Boltzmann equation. Note that similar problems arise with the Burnett and super-Burnett equations which, however, lead to unphysical oscillations in space [10].

8 Shock structure computations

Now we turn to the non-linear equations. The shock profile connects the equilibrium states of density ρ_0, velocity v_0, and temperature T_0 before the shock at $x \to -\infty$ with the equilibrium ρ_1, v_1, T_1 behind the shock at $x \to \infty$. The process is modeled as one-dimensional flow. Hence, velocity, pressure deviator and heat flux have only one single non-trivial component in the direction normal to the shock wave. The field quantities are related to their values at $x \to -\infty$ by definition of the non-dimensional quantities

$$\hat{\rho} = \frac{\rho}{\rho_0},\ \hat{v} = \frac{v}{\sqrt{RT_0}},\ \hat{T} = \frac{T}{T_0},\ \hat{\sigma} = \frac{\sigma}{\rho_0 RT_0},\ \hat{q} = \frac{q}{\rho_0\sqrt{RT_0}^3},$$
$$\hat{\mu} = \frac{\mu}{\mu_0} = \hat{T}^s\,.$$

As in the linear case, $\hat{\sigma} = \sigma_{\langle 11\rangle}$ represents the non-trivial component of the pressure deviator, called stress in the following, and $\hat{q}$ denotes the normal heat flux.

A dimensionless space variable is introduced as

$$\hat{x} = \frac{x\rho_0\sqrt{RT_0}}{\mu_0},$$

where μ_0 is the viscosity of the state before the shock. From the viscosity follows the mean free path (see, e.g., [4] or [2]) calculated for $x \to -\infty$, viz.,

$$\bar{\lambda}_0 = \frac{4}{5}\frac{\mu_0}{\rho_0\sqrt{\frac{\pi}{8}RT_0}}. \tag{15}$$

Thus, the relation

$$\frac{x}{\bar{\lambda}_0} = \frac{5}{4}\sqrt{\frac{\pi}{8}}\,\hat{x} \approx 0.783\,\hat{x} \tag{16}$$

holds for our dimensionless space variable. In the plots we shall always use x/λ_0 as space variable. For the sake of simplicity we drop the "hats" of non-dimensional variables in the sequel.

The Mach number of the shock

$$M_0 = v_0/\sqrt{\frac{5}{3}}$$

acts as parameter for the computations. Shock structures are formally solutions of the one-dimensional R13 equations with the boundary conditions:

$$\text{at } (x \to -\infty): \ \varrho_0 = 1 \ , \ \ v_0 = \sqrt{\frac{5}{3}} M_0 \ , \ \ T_0 = 1 \ ,$$

$$\text{at } (x \to \infty): \ \varrho_1 = \frac{\varrho_0 v_0}{v_1} \ , \ \ v_1 = \sqrt{\frac{5}{3}} \frac{M_0^2 + 3}{4 M_0} \ ,$$

$$T_1 = \frac{(5M_0^2 - 1)(M_0^2 + 3)}{16 M_0^2} \ .$$

and $\sigma_0 = \sigma_1 = 0, \ q_0 = q_1 = 0$. The values behind the shock are given by the Rankine-Hugoniot relations. The density follows from the velocity by means of the mass balance as

$$\rho(v) = \sqrt{\frac{5}{3}} \frac{M_0}{v} \ , \tag{17}$$

and the relations for the stress σ and heat flux q as functions of velocity and temperature follow from the conservation laws for momentum and energy as

$$\sigma(v, T) = 1 + \frac{5}{3} M_0^2 - M_0 \sqrt{\frac{5}{3}} \left(\frac{T}{v} + v \right), \tag{18}$$

$$q(v, T) = \sqrt{\frac{5}{12}} M_0 \left(\frac{5}{3} M_0^2 + 5v^2 - 3T \right) - v \left(1 + \frac{5}{3} M_0^2 \right). \tag{19}$$

The R13 equations were solved numerically with a method outlined in [6]. We proceed to discuss the general behavior of shock structure solutions of the R13 equations.

8.1 Transition from Grad's 13 moment equations

Grad's 13 moment case was derived as an improvement on the NSF theory in the description of rarefied flows. Unfortunately, the equations fail to describe continuous shock structures, since they suffer from a subshock in front of the shock beyond the Mach number $M_0 = 1.65$; see [17] and [18]. This subshock grows with higher Mach numbers and at $M_0 \approx 3.5$ a second subshock appears in the middle of the shock. Both subshocks are artefacts from the hyperbolic nature of the 13-moment equations [40]. It turned out that any hyperbolic moment theory yields continuous shock structures only up to the Mach number corresponding to the highest characteristic

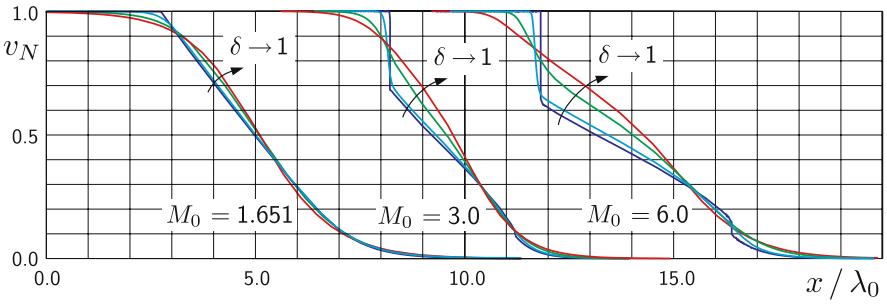


Fig. 3. Regularization process of Grad's 13 moment equation. Profiles for three different Mach numbers are shown with different values of $\delta = 0.0,\ 0.1,\ 0.5,\ 1.0$. The results of Grad's equation ($\delta = 0$) exhibit kinks as well as up to two subshocks of increasing strength. These singularities vanish in the R13 case where $\delta = 1$

velocity, see [41] and [26]. Further validation of results with measurements shows that moment theories succeed in describing shock thickness data accurately only for Mach numbers far below this critical value. In particular, Grad's 13 moment case describes the shock thickness accurately only up to $M_0 \approx 1.1$. Recent results from [22] required up to 900 moments to calculate a smooth shock structure for $M_0 = 1.8$ that fits to experimental data. For more information on shock structures in moment theories see the textbook [19].

One of the reasons for deriving the regularized 13 moment equations (R13) in [5] was to obtain field equations which lead to smooth and stable shock structures for any Mach number. Since the equations are based on Grad's 13 moment case, it must be emphasized that physicality of the R13 solutions is still restricted to small Mach numbers. However, the range of validity is extended by including higher order expansion terms into the R13 equations.

Figure 3 shows the transition to smooth shock structures for three different Mach numbers by means of the normalized velocity field v_N. The results are obtained with $s = 1$, i.e., Maxwell molecules. For this, we multiplied the right-hand sides of Eqs. (5) with a parameter δ that assumes values between zero and unity. The structures with $\delta = 0$ represent solutions of the classical 13 moment case, Eqs. (4). For these, at $M_0 = 1.651$ a kink at the beginning of the shock indicates that the highest characteristic velocity is reached before the shock. The kink develops into a pronounced subshock at $M_0 = 3$. In the case $M_0 = 6$ a second subshock is present towards the end of the structure.

The curves for $\delta = 0.1$ follow mainly the results of Grad's 13 moment case. The subshocks are still clearly visible, albeit smoothed out by increased dissipation.

At $\delta = 1$, however, the additional terms in the regularized 13 moment equations succeed in completely annihilating the subshocks and an overall smooth shock structure is obtained. At $M_0 = 6$ the R13 solution ($\delta = 1$) exhibits obvious asymmetries which start to appear in the structure with Mach numbers $M_0 > 3$. Since experiments

as in [42] or DSMC simulations predict almost perfect s-shaped profiles we conclude that the validity of R13 solutions may be lost beyond Mach numbers $M_0 \approx 3.0$.

8.2 Comparison with DSMC results

In this section we compare the shock structures obtained with the R13 equations to the results obtained with the direct simulation Monte-Carlo method (DSMC) of Bird [4]. For the DSMC results we used the shock structure code which is available from Bird's website. For the actual setup (interval length, upstream temperature, etc.) we adopted the values of Pham-Van-Diep et al. [43]. Note that the calculation of a single low Mach number shock structure by a standard DSMC program takes several hours which is several orders of magnitude slower than the calculation with a continuum model.

We compare results to DSMC solutions for Maxwell molecules, computed with Bird's code; see [4]. Since the DSMC code uses physical units we fixed the mean free path of the upstream region λ_0 as $\lambda_0 = 0.0014$m, which corresponds to our definition (15) and also reproduces the shock thickness results of [43].

In the next figures we compare the profiles of density and heat flux. The heat flux in a shock wave follows solely from the temperature and velocity via the relation (19). Hence, its profile gives a combined impression of the quality of the temperature and velocity profile. The soliton-like shape of the heat flux also helps to give a more significant judgement of the quality of the structure. Since it is a higher moment the heat flux is more difficult to match than the stress. We suppress the profiles of velocity, temperature and stress in the following. The density is normalized to give values between zero and unity for each Mach number. Similarly, the heat flux is normalized so that the DSMC result gives a maximal heat flux of 0.9.

Before we present the results of the regularized 13 moment equations we discuss briefly the failure of the classical theories and the standard Burnett models. Figure 4 shows the density and heat flux profile of an $M_0 = 2$ shock calculated with the NSF and Grad's 13 moment system as well as with the Burnett and super-Burnett equations. The NSF results simply mismatch the profile, while the Grad 13 solution shows a strong subshock. Burnett and super-Burnett solutions are spoiled by oscillations in the back of the shock.

In the Burnett case the oscillations arise if the length of a grid cell is below half of the mean free path. This is in correspondence to the result of the linear analysis which predicts spatial instabilities. It also explains the appearance of the oscillations in the downstream region, because the mean free path is smaller in that region. Since the oscillations stick to a wave length corresponding to the length of a grid cell, high resolution calculations are impossible. The super-Burnett result shows the same behavior; however the oscillation wave length is a multiple of the length of a grid cell. Still, the oscillations increase with grid refinement and convergence cannot be established.

The oscillations of both models, Burnett and super-Burnett, increase for shocks with higher Mach number and are also present for other values of the viscosity

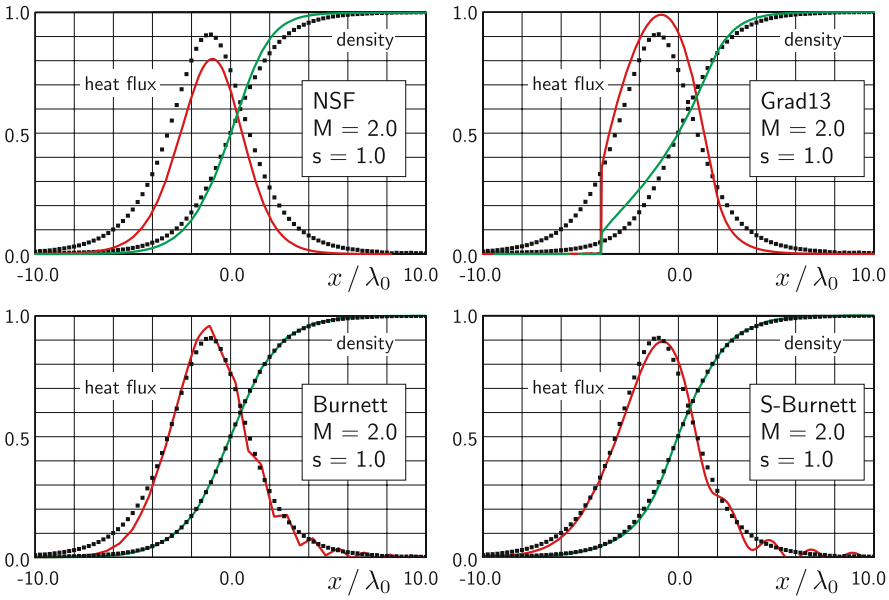


Fig. 4. Shock structure solutions of the system of Navier-Stokes-Fourier, the classical 13 moment case of Grad, and Burnett and super-Burnett equations for Maxwell molecules at Mach number $M_0 = 2$ (solid lines). Both Burnett results exhibit non-physical oscillations in the downstream region. The squares represent the DSMC solution

exponent. Hence, for the description of shock structures the Burnett equations and super-Burnett equations have to be rejected.

Figure 5 shows shock structures for the Mach numbers $M_0 = 1.5, 2, 3, 4$ calculated with the R13 equations, displayed together with the DSMC results. For smaller Mach numbers the shape of the heat flux is captured very well by the R13 equations, and the density profiles exhibit no visible differences to DSMC. The deviations from DSMC solutions become more pronounced for higher Mach numbers. The R13 results begin to deviate from the DSMC solution in the upstream part.

From the figures presented we may conclude that the results of the R13 system for Maxwell molecules agree well with DSMC results. For higher Mach numbers, however, the R13 equations deviate from DSMC data, and the applicability of the theory is no longer given, when quantitative features must be captured.

9 Conclusions

We conclude that the R13 equations are superior to all competing models, i.e., Burnett and super-Burnett equations, models derived from them, and Grad's 13 moment equations. They are unconditionally stable, and stand in good agreement with experiments for dispersion and damping, and shock structures. The equations discussed above are derived for the special case of Maxwell molecules. Other molecular in-

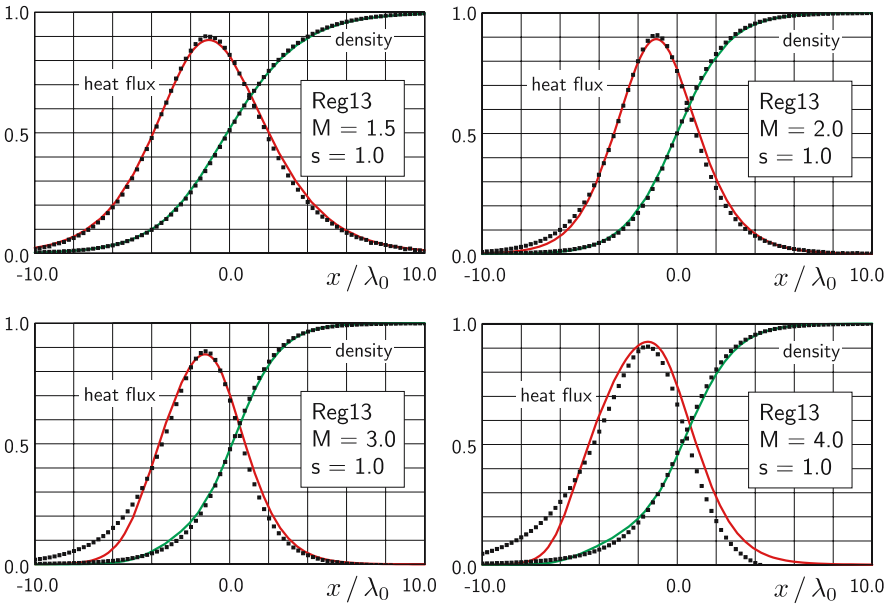


Fig. 5. Shock structures in a gas of Maxwell molecules with Mach numbers $M_0 = 1.5$, 2.0, 3.0, 4.0. Solid lines show the solution of the R13 equations, while squares correspond to the DSMC solution

teraction models can be incorporated ad hoc by adjusting the viscosity coefficient s in Eq. (6). However, the proper derivation is discussed in [36] where it becomes clear that more than 13 moments will be needed for a proper third order theory for non-Maxwellian molecules. The most pressing question at present is to find proper boundary conditions for the R13 equations, and we hope to be able to present these in the future.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council (NSERC).

References

[1] Cercignani, C. (1975): Theory and application of the Boltzmann equation. Scottish Academic Press, Edinburgh

[2] Chapman, S., Cowling, T.G. (1970): The mathematical theory of non-uniform gases. 3rd edition. Cambridge University Press, Cambridge

[3] Ohwada, T. (1996): Heat flow and temperature and density distributions in a rarefied gas between parallel plates with different temperatures. Finite-difference analysis of the nonlinear Boltzmann equation for hard-sphere molecules. Phys. Fluids **8**, 2153–2160

[4] Bird, G. (1994): Molecular gas dynamics and the direct simulation of gas flows. Clarendon Press, Oxford
[5] Struchtrup, H., Torrilhon, M. (2003): Regularization of Grad's 13-moment-equations. Derivation and linear analysis. Phys. Fluids **15**, 2668–2680
[6] Torrilhon, M., Struchtrup, H. (2002): Smooth shock structures for high Mach numbers with regularized 13-moment-equations. Submitted.
[7] Ferziger, J.H., Kaper, H.G. (1972): Mathematical theory of transport processes in gases. North-Holland, Amsterdam
[8] Bobylev, A.V. (1982): The Chapman-Enskog and Grad methods for solving the Boltzmann equation. (Russian). Dokl. Akad. Nauk SSSR **262**, 71–75; translation: Soviet Phys. Dokl. **27**, 29–31
[9] Shavaliev, M.S. (1993): Super-Burnett corrections to the stress tensor and the heat flux in a gas of Maxwellian molecules. (Russian). Prikl. Mat. Mekh. **57**, 168–171; translation: J. Appl. Math. Mech. **57**, 573–576
[10] Struchtrup, H. (2004): Failures of the Burnett and Super-Burnett equations in steady state processes. Contin. Mech. Thermodyn., to appear
[11] Karlin, I.V., Gorban, A.N. (2002): Hydrodynamics from Grad's equations: what can we learn from exact solutions? Ann. Phys. (8) **11**, 783–833
[12] Zheng, Y., Struchtrup, H. (2004): Burnett equations for the ellipsoidal statistical BGK model. Cont. Mech. Thermodyn. **16**, 97–108
[13] Zhong, X., MacCormack, R.W., Chapman, D.R. (1991): Stabilization of the Burnett equations and applications to high-altitude hypersonic flows. AIAA Paper 91-0770. American Institute of Aeronautics and Astronautics, Reston, VA
[14] Zhong, X., MacCormack, R.W., Chapman, D.R. (1993): Stabilization of the Burnett equations and applications to hypersonic flows, AIAA J. **31**, 1036–1043
[15] Jin, S., Slemrod, M. (2001): Regularization of the Burnett equations via relaxation. J. Statist. Phys. **103**, 1009–1033
[16] Jin, S., Pareschi, L., Slemrod, M. (2002): A relaxation scheme for solving the Boltzmann equation based on the Chapman-Enskog expansion. Acta Math. Appl. Sin. (Eng. Ser.) **18**, 37–62
[17] Grad, H. (1949): On the kinetic theory of rarefied gases. Comm. Pure Appl. Math. **2**, 331–407
[18] Grad, H. (1958): Principles of the kinetic theory of gases. In: Flügge, S. (ed.): Handbuch der Physik. Bd. 12. Thermodynamik der Gase. Springer, Berlin, pp. 205–294
[19] Müller, I., Ruggeri, T. (1998): Rational extended thermodynamics. 2nd edition. (Springer Tracts in Natural Philosophy, vol. **37**). Springer, New York
[20] Struchtrup, H. (2002): Heat transfer in the transition regime. Solution of boundary value problems for Grad's moment equations via kinetic schemes. Phys. Rev. E **65**, 041204
[21] Struchtrup, H. (1997): An extended moment method in radiative transfer: the matrices of mean absorption and scattering coefficients. Ann. Physics **257**, 111–135
[22] Au, J.D. (2003): Lösung nichtlinearer Probleme in der Erweiterten Thermodynamik. Dissertation. Technische Universität Berlin, Berlin
[23] Au, J.D., Torrilhon, M., Weiss, W. (2001): The shock tube study in extended thermodynamics. Phys. Fluids **13**, 2423–2432
[24] Struchtrup, H. (2000): Kinetic schemes and boundary conditions for moment equations. Z. Angew. Math. Phys. **51**, 346–365
[25] Struchtrup, H. (2004): Some remarks on the equations of Burnett and Grad. In: Ben Abdallah, N. et al. (eds.): Transport in transition regimes. (The IMA Volumes in Mathematics and its Applications, vol. **135**). Springer, New York, pp. 265–277

[26] Weiss, W. (1995): Continuous shock structure in extended thermodynamics. Phys. Rev. E **52**, R5760–R5763
[27] Reitebuch, D., Weiss, W. (1999): Application of high moment theory to the plane Couette flow. Contin. Mech. Thermodyn. **11**, 217–225
[28] Struchtrup, H. (2003): Grad's moment equations for microscale flows. In: Ketsdever, A.D., Muntz, E.P. (eds.): Rarefied Gas Dynamics. (AIP Conference Proceedings, vol. 663). American Institute of Physics, Melville, NY, pp. 792–799
[29] Ikenberry, E., Truesdell, C. (1956): On the pressures and the flux of energy in a gas according to Maxwell's kinetic theory. I. J. Rational Mech. Anal. **5**, 1–54
[30] Truesdell, C., Muncaster, R.G. (1980): Fundamentals of Maxwell's kinetic theory of a simple monatomic gas. Academic Press, New York
[31] Reinecke, S., Kremer, G.M. (1990): Method of moments of Grad. Phys. Rev. A **42**, 815–820
[32] Reinecke, S., Kremer, G.M. (1996): Burnett's equations from a (13+9N)-field theory. Cont Mech. Thermodyn. **8**, 121–130
[33] Karlin, I.V., Gorban, A.N., Durek, G., Nonnenmacher, T.F. (1998): Dynamic correction to moment approximations. Phys. Rev. E, **57**, 1668–1672
[34] Serre, D. (1999): Systems of conservation laws. I. Hyperbolicity, entropies, shock waves. Cambridge Univ. Press, Cambridge
[35] Jin, S., Xin, Z.P. (1995): The relaxation schemes for systems of conservation laws in arbitrary space dimensions. Comm. Pure Appl. Math. **48**, 235–276
[36] Struchtrup, H. (2004): Stable transport equations for rarefied gases at high orders in the Knudsen number. Phys. Fluids, to appear
[37] Müller, I., Reitebuch, D., Weiss, W. (2003): Extended thermodynamics – consistent in order of magnitude. Contin. Mech. Thermodyn. **15**, 113–146
[38] Bhatnagar, P.L., Gross, E.P., Krook, M. (1954): Model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems. Phys. Rev. (2) **94**, 511–525
[39] Meyer, E., Sessler, G. (1957): Schallausbreitung in Gasen bei hohen Frequenzen und sehr niedrigen Drucken. Z. Physik **149**, 15–39
[40] Torrilhon, M. (2000): Characteristic waves and dissipation in the 13-moment-case. Contin. Mech. Thermodyn. **12**, 289–301
[41] Ruggeri, T. (1993): Breakdown of shock-wave-structure solutions. Phys. Rev. E (3) **47**, 4135–4140
[42] Alsmeyer, H. (1976): Density profiles in argon and nitrogen shock waves measured by the absorbtion of an electron beam. J. Fluid Mech. **74**, 497–513
[43] Pham-Van-Diep, G. C., Erwin, D. A., Muntz, E.P. (1991): Testing continuum descriptions of low-Mach-number shock structures. J. Fluid Mech. **232**, 403–413

Hydrodynamic calculation for extended differential mobility in semiconductors

M. Trovato

Abstract. By using the maximum entropy principle (MEP) we present a general theory to obtain a closed set of balance hydrodynamic equations (HD) for hot carriers including the full-band effects with a total energy scheme. Furthermore, under spatially homogeneous conditions, a closed set of balance equations for the fluctuations of these variables is constructed. We analyze, in the linear case, the different coupling processes, as functions of the electric field, with a full set of scalar and vectorial moments. We prove that, for n-type Si, the coupling between the different moments can lead to a strongly non-exponential decay of the corresponding response functions. To check the validity of this theoretical approach numerical HD calculations are found to compare well with those obtained by an ensemble Monte Carlo (MC) simulator.

1 Introduction

In recent years significant attention has been given to the hydrodynamics models [1–4] based on the *moment balance equations* and to the possibility of their use to describe charge transport in submicrometer devices when extremely high electric fields and field gradients are present locally. Recently the formal derivation of HD moment equations from the microscopic dynamics of the system has been intensively studied using the extended thermodynamic and the maximum entropy principles [5–17]. The MEP approach offers a definite procedure for the construction of a macroequivalent distribution function [5,18,19] which determines the microstate corresponding to the given set of macroscopic variables. The use of this theoretical approach has also been proven to be a useful tool to describe the small-signal analysis in the homogeneous case [14,15]. The small-signal coefficients are of fundamental and applied importance for the description and characterization of the thermodynamic state of hot carriers in semiconductor materials and devices [20–28]. In particular, the study of the eigenvalues of the response matrix and the analysis of the decay in time of the response functions provides valuable information both on the coupling processes and on the relaxation processes of the relevant macroscopic variables of interest [14,15,23–27].

The aim of this paper is to develop and apply a theoretical study of MEP for the case of semiconductor materials under the influence of arbitrarily high electric fields. To this end, by generalizing the results given in earlier papers [14,15,21,22,25–27], we consider as relevant variables a full set of scalar and vectorial moments using a linear approximation of the MEP. With this approach it is possible to include the full-band effects with a total energy scheme and to obtain a closed set of coupled differential equations for the macroscopic variables of interest.

Furthermore, the moment equations for the charge carriers can be generalized, in the homogeneous case, to a set of balance equations describing the fluctuations around the stationary state of the macroscopic variables. This enables us to calculate both the generalized response matrix and the response functions of the relevant macroscopic variables in parabolic and non-parabolic approximations. Because of the electric field these equations are coupled and the time behavior of the response functions deviates from a simple exponential decay. The theory is applied to the case of n-type Si. In particular we report numerical results for the small-signal response in homogeneous bulk materials; a comparison with MC simulations is displayed.

2 General theory for a hydrodynamic approach

The microscopic description of hot carrier transport is governed on the kinetic BTE for the single particle distribution function $\mathcal{F}(\boldsymbol{k}, \boldsymbol{r}, t)$,

$$\frac{\partial\, \mathcal{F}}{\partial\, t} + u_i \frac{\partial\, \mathcal{F}}{\partial\, x_i} - \frac{e}{\hbar}\, E_i \frac{\partial\, \mathcal{F}}{\partial\, k_i} = Q(\mathcal{F}), \tag{1}$$

coupled with Poisson's equation for the self-consistent electric field E_i,

$$\boldsymbol{E} = -\nabla\phi, \quad \varepsilon\, \triangle\, \phi = -e\,(N_D - N_A - n), \tag{2}$$

where e is the unit charge, ε the dielectric constant, ϕ the electrical potential, N_D and N_A the donor and acceptor concentration respectively, u_i the carrier group velocity, k_i the wavevector, $\hbar$ the reduced Planck constant and

$$Q(\mathcal{F}) = \frac{V}{(2\pi)^3} \left\{ \int d\boldsymbol{k}'\, S(\boldsymbol{k}, \boldsymbol{k}')\mathcal{F}(\boldsymbol{k}', \boldsymbol{r}, t) - \int d\boldsymbol{k}'\, S(\boldsymbol{k}', \boldsymbol{k})\mathcal{F}(\boldsymbol{k}, \boldsymbol{r}, t) \right\} \tag{3}$$

the collision integral under non-degenerate conditions, where $S(\boldsymbol{k}, \boldsymbol{k}')$ is the total electron scattering rate for the transition $\boldsymbol{k}' \rightarrow \boldsymbol{k}$ and V the crystal volume.

In the framework of the moment theory, to pass from the kinetic level of the BTE to the extended HD level, within the general many-valley band model, we must consider the set of generalized kinetic fields

$$\psi_A(\boldsymbol{k}) = \{\varepsilon^m,\ \varepsilon^m u_{i_1},\ \ldots,\ \varepsilon^m u_{i_1} \cdots u_{i_s}\}, \tag{4}$$

where $\varepsilon(\boldsymbol{k})$ is a general single particle band energy dispersion of arbitrary form, $m = 0, 1, ..., N$ and $s = 1, 2, ..., M$. With this approach, we have the corresponding macroscopic quantities $F_A = \{F_{(m)},\ F_{(m)|i_1},\ \ldots,\ F_{(m)|i_1 \cdots i_s}\}$, where

$$F_A = \int \psi_A(\boldsymbol{k})\, \mathcal{F}(\boldsymbol{k}, \boldsymbol{r}, t)\, d\boldsymbol{k}, \tag{5}$$

and the set of moment equations [5–16]

$$\frac{\partial F_A}{\partial t} + \frac{\partial F_{Ak}}{\partial x_k} = -\frac{e}{\hbar} R_{Ai} E_i + P_A\,, \qquad A = 1, \ldots, \mathcal{N}, \tag{6}$$

where $\mathcal{N}$ is the fixed number of moments used, and F_{Ak}, R_{Ai}, P_A indicate, respectively, the fluxes, external field productions, and collisional productions defined as:

$$F_{Ak} = \int \psi_A(\boldsymbol{k})\, u_k \mathcal{F}(\boldsymbol{k}, \boldsymbol{r}, t)\, d\boldsymbol{k}\,, \tag{7}$$

$$R_{Ai} = \int \frac{\partial \psi_A(\boldsymbol{k})}{\partial k_i} \mathcal{F}(\boldsymbol{k}, \boldsymbol{r}, t)\, d\boldsymbol{k}\,, \tag{8}$$

$$P_A = \int \psi_A(\boldsymbol{k})\, Q(\mathcal{F})\, d\boldsymbol{k}\,. \tag{9}$$

In particular, for $N = M = 1$, we have the usual physical quantities, which have a direct physical interpretation, such as $F_{(0)} = n$ *(numerical density)*, $F_{(1)} = W$ *(total energy density)*, $F_{(0)|i} = nv_i$ *(velocity flux density)*, $F_{(1)|i} = S_i$ *(energy flux density)*, while, for $N, M > 1$, we obtain macroscopic additional field variables, which become the fluxes of the preceding equations. With this procedure, we obtain a system of balance equations of finite order in which there are unknown *constitutive functions* $H_A = \{F_{Ak},\ R_{Ai},\ P_A\}$ that must be determined in terms of the variables F_A.

Following information theory, one can determine systematically the unknown constitutive functions by introducing the MEP in terms of the generalized kinetic fields (4). The MEP is based on the assumption that the least biased distribution function assignment to a physical system is obtained from the solution of the variational problem of maximizing the entropy subject to the constraints imposed by the available information. For this reason, assuming that the information expressed by a fixed number $\mathcal{N}$ of moments describes the thermodynamical state of hot carriers satisfactorily, we look for the distribution that makes best use of this information [5,6,9,10,12,14,17–19]. In this approach the distribution function has the explicit form

$$\mathcal{F} = \mathcal{F}_M\ \exp(-\Pi), \qquad \Pi = \sum_{A=1}^{\mathcal{N}} \psi_A\, \widehat{\Lambda}_A\,, \tag{10}$$

where $\widehat{\Lambda}_A$ are the non-equilibrium part of the Lagrange multipliers [5,6,9,10,12,14,17] and $\mathcal{F}_M$ the local Maxwellian. Since, for a band of general form, only the total average electron energy is a well-defined quantity, the MEP must be applied with a total energy scheme [11,14–16]. Consistently with this choice, the local distribution function should be defined in terms of the total average energy of the single carrier, as $\mathcal{F}_M = \gamma \exp(-\beta\varepsilon(\boldsymbol{k}))$, where $\gamma = \gamma(n, W)$ and $\beta = \beta(W/n)$ are appropriate functions which can be determined by means of the local equilibrium conditions [14]

$$n(\boldsymbol{r}, t) = \int \mathcal{F}_M\, d\boldsymbol{k}\,, \quad W(\boldsymbol{r}, t) = \int \varepsilon(\boldsymbol{k})\, \mathcal{F}_M\, d\boldsymbol{k}\,. \tag{11}$$

By expanding the distribution function $(10)_1$ around the Maxwellian $\mathcal{F}_M$, up to the fixed order R, we obtain by means of the moments (5) a set of non-linear equations

in the non-equilibrium quantities $\widehat{\Lambda}_A$, namely,

$$F_A - F_A\,|_E = \int \psi_A\, \mathcal{F}_M \sum_{r=1}^{R} \frac{(-1)^r}{(r)!} \left(\sum_{B=1}^{\mathcal{N}} \psi_B\, \widehat{\Lambda}_B \right)^r d\boldsymbol{k}. \tag{12}$$

By expressing $\widehat{\Lambda}_B$ in Eqs. (12) in polynomial terms of the non-equilibrium variables F_B, the non-linear system can be inverted and the Lagrange multipliers obtained [5,6,9,10,12,17].

In this way, with an analytic expression for the $\widehat{\Lambda}_A$ determined, both the distribution function $\mathcal{F}$ and the constitutive functions H_A can be estimated, up to order R, as polynomials in the non-equilibrium variables whose coefficients depend on the local equilibrium quantities $\{n(\boldsymbol{r}, t),\ W(\boldsymbol{r}, t)\}$.

In particular, to evaluate the collisional production P_A, we consider in Eq. (3) the collision rate for acoustic intravalley transitions, within the elastic and equipartition approximations,

$$S_{ac}(\boldsymbol{k}, \boldsymbol{k}') = 2\frac{\pi\, E_l^2 K_B T_0}{\hbar V\, \rho\, U_l^2}\ \delta[\varepsilon(\boldsymbol{k}') - \varepsilon(\boldsymbol{k})]\ , \tag{13}$$

and, for intervalley transitions with no polar optical and acoustic phonons,

$$S_\eta(\boldsymbol{k}, \boldsymbol{k}') = \frac{\pi\, \Delta_\eta^2}{V\, \rho\, \omega_\eta} \left[N_\eta + \frac{1}{2} \pm \frac{1}{2} \right] \delta[\varepsilon(\boldsymbol{k}') - (\varepsilon(\boldsymbol{k}) \pm \hbar\, \omega_\eta)]\ , \tag{14}$$

where E_l is the acoustic deformation potential, ρ the crystal density, U_l the longitudinal sound velocity, Δ_η the intervalley deformation potential, ω_η the phonon angular frequency, and N_η the phonon occupation number, here taken as the equilibrium Planck distribution, with the $\pm$ signs referring to emission and absorption cases, respectively.

We stress that the previous closure scheme can be developed using two different levels of approximation that depends both on the number of moments used as constraints and on the order of the expansion of $\mathcal{F}$. A first level of approximation is closely related to the Grad moment method, and is obtained by considering a linear expansion of the distribution function but using an arbitrary number of moments [14]. A second level of approximation, that clearly differs from that in Grad's method, is obtained by considering the maximum entropy formalism in a strong non-linear context but using only the most important macroscopic variables of direct physical interpretation (the first 13 moments of the distribution function [6,9–11]). We note that, although the linear approach is capable of describing accurately the transport properties of hot carriers both in spatial homogeneous conditions and in the small gradient approximation, only a higher-order expansion of the distribution function can be fruitfully applied to describe transport phenomena in conditions far from thermodynamic equilibrium, as those present in submicron devices, with very high electric fields and field gradients (see, e.g., [9–11]).

3 General theory for small-signal analysis

Linear-response functions around the bias point are known to play a fundamental role in the investigation of hot-carrier transport in bulk semiconductors [21,22,25–27]. In the time domain they reflect both dynamic and relaxation processes inherent in the hot-carrier system. In the frequency domain they provide the a.c. coefficients of interest such as the usual differential mobility spectrum [20–22,25,26] and the noise temperature [27]. The aim of this section is to provide a general theoretical investigation for the linear response analysis in the framework of moment theory.

3.1 Hydrodynamic approach

Under spatially homogeneous conditions, the balance equations of the single particle moments $\widetilde{F}_A$ take the form

$$\frac{\partial \widetilde{F}_A}{\partial t} + \frac{e}{\hbar}\,\widetilde{R}_{Ai}E_i + \widetilde{P}_A = 0. \tag{15}$$

By assuming that at the initial time the system of carriers is perturbed by an electric field $\delta E\xi(t)$ along the direction of $\boldsymbol{E}$ (where $\xi(t)$ is an arbitrary function of time satisfying $|\xi(t)| \leq 1$), we calculate the deviations from the average values of the moments denoted by $\delta\widetilde{F}_A$. After linearizing Eqs. (15) around the stationary state, we obtain a system of equations which can be written as

$$\frac{d\ \delta\widetilde{F}_\alpha(t)}{d\ t} = \Gamma_{\alpha\beta}\delta\widetilde{F}_\beta(t) - e\,\delta E\,\xi(t)\,\Gamma_\alpha^{(E)}, \tag{16}$$

where the relaxation of the system to the stationary state is related to the response matrix $\Gamma_{\alpha\beta}$ which describes the time evolution of the moments after the perturbation of the electric field E and where the $-e\delta E\xi(t)\ \Gamma_\alpha^{(E)}$ are the fluctuating forces. Equation (16) has the formal solution [25]

$$\boldsymbol{\delta}\widetilde{\mathbf{F}}(t) = \exp(\boldsymbol{\Gamma}t)\ \boldsymbol{\delta}\widetilde{\mathbf{F}}(0) - e\,\delta E\int_0^t \mathbf{K}(s)\xi(t-s)\,ds, \tag{17}$$

where

$$\exp(\boldsymbol{\Gamma}t) = \boldsymbol{\Phi}\ \mathrm{diag}\{\exp(\lambda_1 \mathrm{t}), \cdots, \exp(\lambda_{N-1}\mathrm{t})\}\ \boldsymbol{\Phi}^{-1}, \tag{18}$$

$$\mathbf{K}(s) = \exp(\boldsymbol{\Gamma}s)\ \boldsymbol{\Gamma}^{(E)} \tag{19}$$

with λ_α the eigenvalues of $\Gamma_{\alpha\beta}$ and $\boldsymbol{\Phi}$ the matrix of its eigenvectors.

The eigenvalues λ_α can be real or complex and they correspond to the generalized relaxation rates $\nu_\alpha = -\lambda_\alpha$, even if an exact correspondence between these rates and the respective relaxation processes exists only in the relaxation time approximation for the collision integral (3) and in the absence of coupling between the variables $\widetilde{F}_\alpha$. The response function $\mathbf{K}(t)$ depends on the eigenvalues λ_α and determines the linear response of the moments $\widetilde{F}_A$ to an arbitrary perturbation of the electric field.

The initial values of the response functions can be calculated in an analytic way as functions of the electric field using Eq. (19) for $s = 0$. It is worth noting that in this case we have

$$K_A(0) = \Gamma_A^{(E)}. \tag{20}$$

Since at time $t = 0$ the moments are as yet unperturbed, we assume that $\boldsymbol{\delta}\widetilde{\mathbf{F}}(0) = 0$ in Eq. (17), and the small-signal analysis, in the time and frequency domains, is described by the explicit form of the function $\xi(t)$. In particular, the linear responses of hot carriers to a step-like switching of electric field and to a small harmonic field are of special interest.

In the first case $\xi(t) = 1$ for all $t > 0$ and the differential response $\delta\widetilde{F}_\alpha(t)/\delta E$ is the solution of the differential equation

$$K_\alpha(t) = -\frac{1}{e\,\delta E}\,\frac{d\;\delta\widetilde{F}_\alpha(t)}{d\;t}. \tag{21}$$

This means that, to a step-like variation of electric field, the linear response function $K_\alpha(t)$ is proportional to the time derivative of the corresponding perturbation $\delta\widetilde{F}_\alpha(t)$ and that $K_\alpha(\bar{t}) = 0$ corresponds to one extreme position of $\delta\widetilde{F}_\alpha$ at time $\bar{t}$ which evidently represents the same relaxation phenomenon. Analogously, by a further derivation of relation (21) we observe that

$$\frac{d\;K_\alpha(t)}{d\;t} = -\frac{1}{e\,\delta E}\,\frac{d^2\;\delta\widetilde{F}_\alpha(t)}{d\;t^2}, \tag{22}$$

to one extreme position of the response function $K_\alpha(t)$ at time t' is associated a flex point of the corresponding perturbation $\delta\widetilde{F}_\alpha(t')$.

As a second case we consider a small harmonic perturbation $\xi(t) = \exp(i\omega t)$ (along the direction of $\boldsymbol{E}$) applied to the electron system. For large values of time the upper limit in the integral of Eq. (17) can be replaced by infinity and we obtain a perturbation of the single-carrier moments which is also harmonic $\boldsymbol{\delta}\widetilde{\mathbf{F}}(t) = \boldsymbol{\delta}\widetilde{\mathbf{F}}(\omega)\exp(i\omega t)$, since

$$\delta\widetilde{F}_\alpha(\omega) = \mu'_\alpha(\omega)\;\delta E, \quad \text{with} \quad \mu'_\alpha(\omega) = -e\int_0^\infty K_\alpha(s)\exp(-i\omega s)\,ds. \tag{23}$$

If we consider real and imaginary parts separately, $\mu'_\alpha(\omega) = X_\alpha(\omega) + iY_\alpha(\omega)$, we observe that, in the low frequency limits, the real parts $\mathrm{Re}[\mu'_\alpha(\omega)] = X_\alpha(\omega)$ of the a.c. generalized differential mobility tend to the corresponding d.c. generalized differential mobility values $d\widetilde{F}_\alpha/dE$.

If we consider the integrals of the functions X_α and Y_α/ω over the entire range of frequencies, we find that

$$\int_0^\infty X_\alpha\,d\omega = -\frac{\pi}{2}\;e\;\Gamma_\alpha^{(E)} = -\frac{\pi}{2}\;e\;K_\alpha(0)\;, \tag{24}$$

$$\int_0^\infty \frac{1}{\omega}\,Y_\alpha\,d\omega = -\frac{\pi}{2}\;e\;\Gamma_{\alpha\beta}^{-1}\;\Gamma_\beta^{(E)} = -\frac{\pi}{2}\;X_\alpha(0)\,; \tag{25}$$

analogously, if we consider the integrals of the functions $[\omega Y_\alpha - e\Gamma_\alpha^{(E)}]$ and $[\omega^2 X_\alpha - e\,\Gamma_{\alpha\beta}\,\Gamma_\beta^{(E)}]$, we have

$$\int_0^\infty \left[\omega\, Y_\alpha - e\,\Gamma_\alpha^{(E)}\right] d\omega = \frac{\pi}{2}\, e\,\Gamma_{\alpha\beta}\,\Gamma_\beta^{(E)} = \frac{\pi}{2}\, e\, \left[\frac{dK_\alpha}{dt}\right]_{0^+}, \tag{26}$$

$$\int_0^\infty \left[\omega^2\, X_\alpha - e\,\Gamma_{\alpha\beta}\,\Gamma_\beta^{(E)}\right] d\omega = \frac{\pi}{2}\, e\,\Gamma_{\alpha\beta}^2\,\Gamma_\beta^{(E)} = \frac{\pi}{2}\, e\, \left[\frac{d^2K_\alpha}{dt^2}\right]_{0^+}. \tag{27}$$

It is worth noting that all the previous relations (17-27) are natural generalizations of results found in [14,21,25,26] for a.c. and d.c. generalized differential mobility in the framework of moment theory.

The advantages of the approach proposed here, based on the MEP with a total energy scheme, are that:

i) the formulation of a.c. and d.c. theory can be obtained, as at kinetic level, without the need of introducing external parameters and can carried out by using an energy dispersion of general form (full-band approach);

ii) if we explicitly know the response matrix $\Gamma_{\alpha\beta}$ and the vector $\Gamma_\alpha^{(E)}$, we can construct an analytic formulation of the theory.

In the following sections we consider an explicit application of the total energy scheme; in particular, the HD equations with numerical and analytic results are explained in detail with the purpose of validating the maximum entropy approach in a linear context.

4 Application of the total energy scheme

A simplified way to consider the total-energy scheme is to describe the full complexity of the band modeled in terms of a single particle with an effective mass which is a function of the average total energy $\widetilde{W}$ of the single carrier [11,14–16]. In this way the mass becomes a new constitutive function which should be independently determined by fitting experiments and/or from MC calculations of the bulk material [11]. The advantage of this simple approach of applying the total-energy scheme is that (i) all the constitutive relations are obtained in an analytic way, and (ii) the same set of balance equations describe the transport properties of hot carriers for both the parabolic (where m^* is constant) and the full-band cases (where $m^* = m^*(\widetilde{w})$).

4.1 Expansion with an arbitrary number of moments

A general formulation of the MEP and the construction of self-consistent closure relations with an arbitrary number of moments, was recently developed in [14]. With this approach we have as unique independent mean quantities the traceless parts $F_{(p)|<i_1\cdots i_s>}$ of the tensors $F_{(p)|i_1\cdots i_s}$. In particular, for problems with axial symmetry we assume that $E_i = \{E, 0, 0\}$, so that only the independent components

$$F_{(p)|\langle s\rangle} = F_{(p)|(\underbrace{1\cdots 1}_{\text{s times}})}$$

are of concern and, in homogeneous conditions, we obtain for the initial value of the response functions the analytic expression

$$K_{(p)|<s>}(0) = p\,\widetilde{F}_{(p-1)|<s+1>} + \frac{s^2}{2s-1}\left[\frac{2(p+s)+1}{2s+1}\right]\frac{1}{m^*}\,\widetilde{F}_{(p)|<s-1>}. \quad (28)$$

In this way for $s = 0$ we obtain the scalar moments $\widetilde{F}_{(p)}$, for $s = 1$ the vectorial moments of components $\widetilde{F}_{(p)|i} = \{\widetilde{F}_{(p)|1}, 0, 0\}$ and, in general, for $s > 1$ the traceless tensorial moment of rank s of which the unique independent component is $\widetilde{F}_{(p)|\langle s\rangle} = \widetilde{F}_{(p)|\langle 1\cdots 1\rangle}$. We remark that, for the quantities of direct physical interpretation $\{\widetilde{W} = \widetilde{F}_{(1)},\ v = \widetilde{F}_{(0)|1},\ \widetilde{S} = \widetilde{F}_{(1)|1}\}$, we have

$$K_{\widetilde{w}}(0) = v, \qquad K_v(0) = \frac{1}{m^*(\widetilde{w})}, \qquad K_{\widetilde{s}}(0) = \widetilde{F}_{(0)|<11>} + \frac{5}{3}\,\frac{\widetilde{W}}{m^*(\widetilde{w})}, \quad (29)$$

where, in particular, $\{v,\ 1/m^*\}$ are the well-known [21] response functions for moments $\{\widetilde{W}, v\}$ evaluated at $t = 0$.

4.2 Linear expansion with full set of scalar and vectorial moments

In this section we consider a linear expansion of the distribution function using only scalar and vectorial quantities and apply in general terms the theory for small-signal analysis to a full set of these moments. In this way, by considering only the variables $\{\widetilde{F}_{(p)},\ \widetilde{F}_{(p)|i}\}$, in a linear context, all the constitutive functions $F_{(p)|\langle ij\rangle}$ are zero and, in homogeneous conditions, the balance equations for $p = 1, \cdots, N$ and $q = 0, 1, \cdots, N$ read as

$$\frac{\partial\,\widetilde{F}_{(p)}}{\partial\,t} = -epF_{(p-1)|1}\,E - \widetilde{P}^{(0)}_p - \sum_{l=2}^{N}\alpha^{(0)}_{p\,l}\,\widetilde{\Delta}_{(l)}\,, \quad (30)$$

$$\frac{\partial\widetilde{F}_{(q)|1}}{\partial t} = -\frac{e}{m^*}\,\frac{2q+3}{3}\,\widetilde{F}_{(q)}E - \sum_{l=0}^{N}\alpha^{(1)}_{q\,l}\;\widetilde{F}_{(l)|1}\,, \quad (31)$$

where $\widetilde{\Delta}_{(p)} = \widetilde{F}_{(p)} - \widetilde{F}_{(p)|E}$ are the non-equilibrium parts of the scalar moments $\widetilde{F}_{(p)}$ (where $\widetilde{F}_{(p)|E} = (2p+1)!!/3^p\ \widetilde{W}^p$), and the closure relations for the quantities $\{\widetilde{P}^{(0)}_p, \alpha^{(0)}_{p\,l}, \alpha^{(1)}_{q\,l}\}$ are explicitly reported in [14]. In stationary conditions Eqs. (30-31) consist of a system of algebraic equations whose numerical solution [14] allows us to determine the moments as a function of the electric field E. By considering the time evolution of a small perturbation of the scalar and vectorial moments, around the stationary state, system (16) can be expressed in term of the $2N + 1$ quantities

$$\delta\widetilde{F}_\alpha(t) = \left\{\delta\widetilde{W},\ \delta\widetilde{F}_{(2)}, \cdots, \delta\widetilde{F}_{(N)}\,, \delta v,\ \delta\widetilde{S},\ \delta\widetilde{F}_{(2)|1}, \cdots, \delta\widetilde{F}_{(N)|1}\right\}\,,$$

with

$$\Gamma^{(E)}_\alpha = \left\{v, 2\,\widetilde{S}, \cdots, N\,\widetilde{F}_{(N-1)|1}, \frac{1}{m^*}, \frac{5}{3}\,\frac{\widetilde{W}}{m^*}, \frac{7}{3}\,\frac{\widetilde{F}_{(2)}}{m^*}, \cdots, \frac{(2N+3)}{3}\,\frac{\widetilde{F}_{(N)}}{m^*}\right\}$$

and the non-symmetric $(2N+1)\times(2N+1)$ response matrix $\Gamma_{\alpha\beta}$ given by

$$\begin{bmatrix}
\Gamma_{(1)w} & -\alpha_{12}^{(0)} & -\alpha_{13}^{(0)} & \cdots & -\alpha_{1N}^{(0)} & -eE & 0 & \cdots & 0 & 0 \\
\Gamma_{(2)w} & -\alpha_{22}^{(0)} & -\alpha_{23}^{(0)} & \cdots & -\alpha_{2N}^{(0)} & 0 & -2eE & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\Gamma_{(N)w} & -\alpha_{N2}^{(0)} & -\alpha_{N3}^{(0)} & \cdots & -\alpha_{pN}^{(0)} & 0 & 0 & \cdots & -NeE & 0 \\
\Gamma_{(0)|1w} & 0 & 0 & \cdots & 0 & -\alpha_{00}^{(1)} & -\alpha_{01}^{(1)} & \cdots & -\alpha_{0(N-1)}^{(1)} & -\alpha_{0N}^{(1)} \\
\Gamma_{(1)|1w} & 0 & 0 & \cdots & 0 & -\alpha_{10}^{(1)} & -\alpha_{11}^{(1)} & \cdots & -\alpha_{1(N-1)}^{(1)} & -\alpha_{1N}^{(1)} \\
\Gamma_{(2)|1w} & -\frac{7}{3}\frac{eE}{m^*} & 0 & \cdots & 0 & -\alpha_{20}^{(1)} & -\alpha_{21}^{(1)} & \cdots & -\alpha_{2(N-1)}^{(1)} & -\alpha_{2N}^{(1)} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\Gamma_{(N)|1w} & 0 & 0 & \cdots & -\frac{2N+3}{3}\frac{eE}{m^*} & -\alpha_{N0}^{(1)} & -\alpha_{N1}^{(1)} & \cdots & -\alpha_{N(N-1)}^{(1)} & -\alpha_{NN}^{(1)}
\end{bmatrix}$$

where in general the elements of the first column are complicated functions of the quantities $\{\mu_{(p)},\ \mu_{(q)|1},\ \mu'_{(p)},\ \mu'_{(q)|1}\}$, where $\{\mu_{(0)|1}=v/E,\ \mu'_{(0)|1}=dv/dE\}$ are the usual chord mobility and differential mobility respectively, and where the quantities $\{\mu_{(p)}=\widetilde{F}_{(p)}/E,\ \mu_{(q)|1}=\widetilde{F}_{(q)|1}/E,\ \mu'_{w}=d\widetilde{W}/dE,\ \mu'_{(p)}=d\widetilde{F}_{(p)}/dE,\ \mu'_{(q)|1}=d\widetilde{F}_{(q)|1}/dE\}$ are the generalized chord mobility and the generalized differential mobility of the remaining moments.

With this approach all the elements of the matrix $\Gamma_{\alpha\beta}$ can be explicitly evaluated to start from stationary values of the system. Analogously, the vectorial response function $\mathbf{K}(t)$ is expressed in terms of its $2N+1$ components for the fluctuations of scalar and vectorial moments

$$\mathbf{K}(t)=\{K_w,\ K_{(2)},\cdots,K_{(N)},\ K_v,\ K_s,\ K_{(2)|1},\cdots,K_{(N)|1}\}.$$

In closing we remark that, in accord with Eq. (20), we have, for $p=1,2,\cdots,N$ and $q=0,1,\cdots,N$,

$$K_{(p)}(0)=p\,\widetilde{F}_{(p-1)|1},\qquad K_{(q)|1}(0)=\frac{2q+3}{3}\,\frac{1}{m^*(\widetilde{w})}\,\widetilde{F}_{(q)}, \tag{32}$$

which, at the present level of approximation, coincides with the results (28).

5 Numerical results for n-Silicon

In this section we consider the application of the theoretical results reported in the preceding sections to the case of n-Si. By considering the electric field applied along the $\langle 111\rangle$ crystallographic axes we keep the axial symmetry; full-band effects have been described by introducing an effective mass as a function of the electron total energy [11]. For the collisional processes, scattering with phonons of f and g type are considered with six possible transitions (g_1, g_2, g_3, f_1, f_2, f_3). The HD calculations were carried out by using the physical scattering parameters used in [29] and the MC simulations were obtained by using a full-band model [30]. For the differential

mobility μ_v' as a function of electric field, we also report experimental data evaluated both in the low frequency limits [31] and at $f = 123.3$ GHz [20] for n-Si samples oriented in the $\langle 111 \rangle$ crystallographic direction.

5.1 Eigenvalue spectrum

Figure 1 reports the generalized relaxation rates $\nu_\alpha = -\lambda_\alpha$ obtained using, in a linear approximation, an increasing number of scalar and vectorial moments (i.e., $N = 2$, $N = 3$, $N = 5$), both in the parabolic (P) and non-parabolic (NP) band models, respectively.

As a general trend, all the vectorial rates increase with increasing field because of the increased efficiency of the scattering mechanisms while all the scalar rates decrease at the highest field because of the smaller efficiency of scattering to dissipate the excess energy gained by the field. The eigenvalue spectrum behavior is intricate and, in the case of complex values, the continuous lines and the dashed

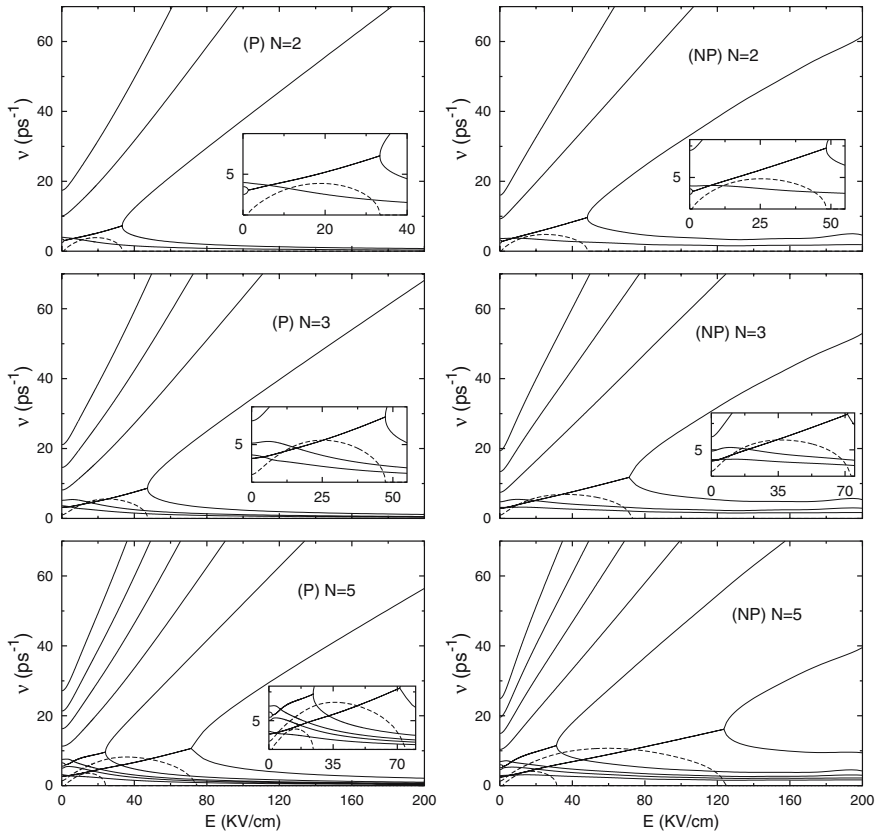


Fig. 1. Eigenvalues of the relaxation matrix as a function of the electric field

lines represent the real part $-\lambda_R$ and the imaginary part $-\lambda_I$ of the eigenvalues of $\Gamma_{\alpha\beta}$, respectively. The numerical results show different regions which correspond to the different characters of the eigenvalues. For small, intermediate and large values of the electric field there are couples of complex conjugate eigenvalues due to the strong coupling between scalar and vectorial moments. In particular the velocity and energy relaxation rates are coupled by the electric field for all values of N with an extension of the coupling region up to about 120 KV/cm for $N = 5$ in a non-parabolic approximation and with an imaginary part $-\lambda_I$ comparable with the real part $-\lambda_R$ on the right-hand side of this region.

From Fig. 1 we see that the width of the region with complex values and the number of coupled eigenvalues depends both on the increasing number of moments used and on the non-parabolicity. In fact, in both cases we find that, while the generalized vectorial rates are squeezed towards lower values, on the other hand, the generalized scalar rates are squeezed towards higher values with the consequent extension of the coupling regions.

In particular, for $N = 5$ the spectrum of $\Gamma_{\alpha\beta}$ shows another pair of complex conjugate eigenvalues with a smaller coupling region when compared with the analogous region for velocity and energy relaxation rate.

A complex eigenvalue indicates the presence of deterministic relaxation [23] in the system, in the present case, the joint action of electric field and emission of optical phonons. In the extreme case this is well-known as the condition of streaming motion [23,24]. The carrier is accelerated by the field up the energy of the optical phonon. From there, by emitting an optical phonon, it is scattered back to the bottom of the band and the cycle starts again. As a matter of fact, for the case of electrons in Si at $T_0 = 300$ K, carriers undergo many scattering events apart from optical phonon emissions, and therefore the streaming-motion regime is not fully achieved. By using many moments, the regions with complex values of the eigenvalues are deeply enlarged towards higher fields. At these fields, the other scattering mechanisms are still efficient and the processes of dissipation are now so strong that, to describe the ordering in the system, it is necessary to use many scalar and vectorial moments.

When the electric field is increased further, the eigenvalues again become real. At these very high fields energy thermalization of the carrier system becomes so efficient that any deterministic character is washed out and the transport takes on a full chaotic character. It should be noted that the dissipative processes associated to the streaming character of the transport have been observed in previous papers [23–27] by using only the usual HD equations for v and $\widetilde{w}$ with the relaxation time approximation. Although the results are similar to those obtained in this work, there are differences in the extension of the region where velocity and energy relaxations are strongly coupled.

This discrepancy is mainly attributed to the number of moments used to calculate the spectrum of $\Gamma_{\alpha\beta}$. The eigenvalue spectrum is rather sensitive both to the increasing number of moments and to the order of expansion with the direct consequence of a much more pronounced extension of the coupling regions and of the number of coupled complex eigenvalues. Probably, in a strong regime, which is far from equilibrium, the variables $\{v, \widetilde{w}\}$ no longer constitute a complete set of relevant variables.

In fact the analysis of the eigenvalue spectrum suggests that, for large values of the field, a detailed investigation of processes of dissipation involves higher moments of velocity and energy. On the basis of these results, it is reasonable to think that only by using many moments is it possible to describe the complete spectrum of dissipation processes for the whole range of values of electric field.

5.2 Response functions and differential response

Figure 2 shows the initial values ($t = 0$) of the response functions for the full set of scalar, and vectorial moments $\{\widetilde{W}, \widetilde{F}_{(2)}, \widetilde{F}_{(3)}, \widetilde{F}_{(4)}, \widetilde{F}_{(5)}\}$ and $\{v, \widetilde{S}, \widetilde{F}_{(2)|1}, \widetilde{F}_{(3)|1}, \widetilde{F}_{(4)|1}, \widetilde{F}_{(5)|1}\}$ respectively, as functions of the electric field strength, with parabolic (P) and non-parabolic (NP) band models at $T_0 = 300^o$K. As general trend, the hot-carrier effects are responsible for a systematic increase of almost all the initial values $\{K_{(p)}(0), K_{(p)|1}(0)\}$ which exhibit asymptotic behaviors steeper for higher moments. The net effect of non-parabolicity is to suppress systematically the increase of all moments with field (cf. Figs. 1–3 in [14]) and, consequently, in accordance with the analytical calculations (32), also of the corresponding initial values of the response functions. In particular K_v is practically constant with small changes due to increasing values of the mass m^* for the non-parabolic band, so that $K_v = 1/m^*(\tilde{w})$ slightly decreases. At increasing fields $K_{\tilde{w}}(0)$ increases with saturation effects that coincide automatically (see Eq. $(29)_1$) with the analogous region of saturation for the drift velocity.

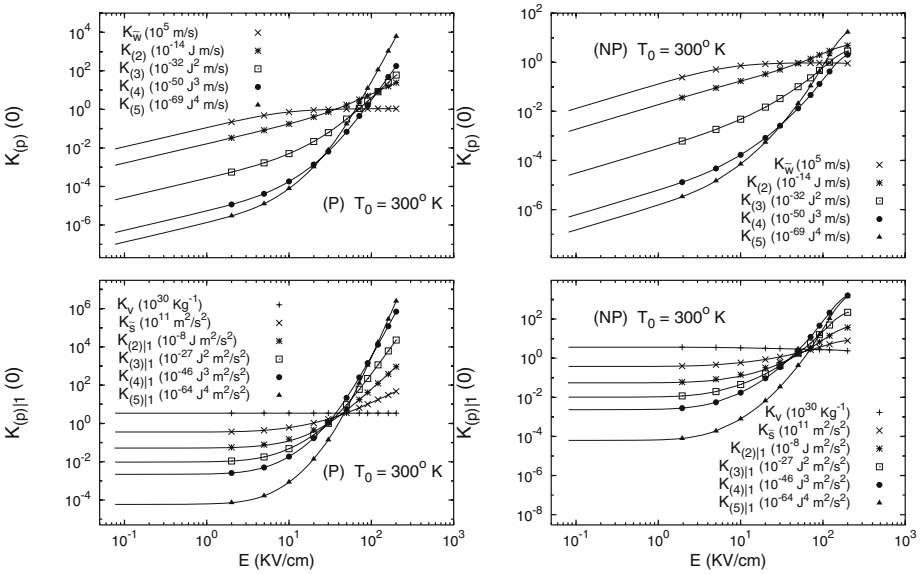


Fig. 2. Initial values (t = 0) of the response functions $\{K_{\tilde{w}}, K_{(2)}, K_{(3)}, K_{(4)}, K_{(5)}\}$ and $\{K_v, K_{\tilde{s}}, K_{(2)|1}, K_{(3)|1}, K_{(4)|1}, K_{(5)|1}\}$ as functions of the electric field

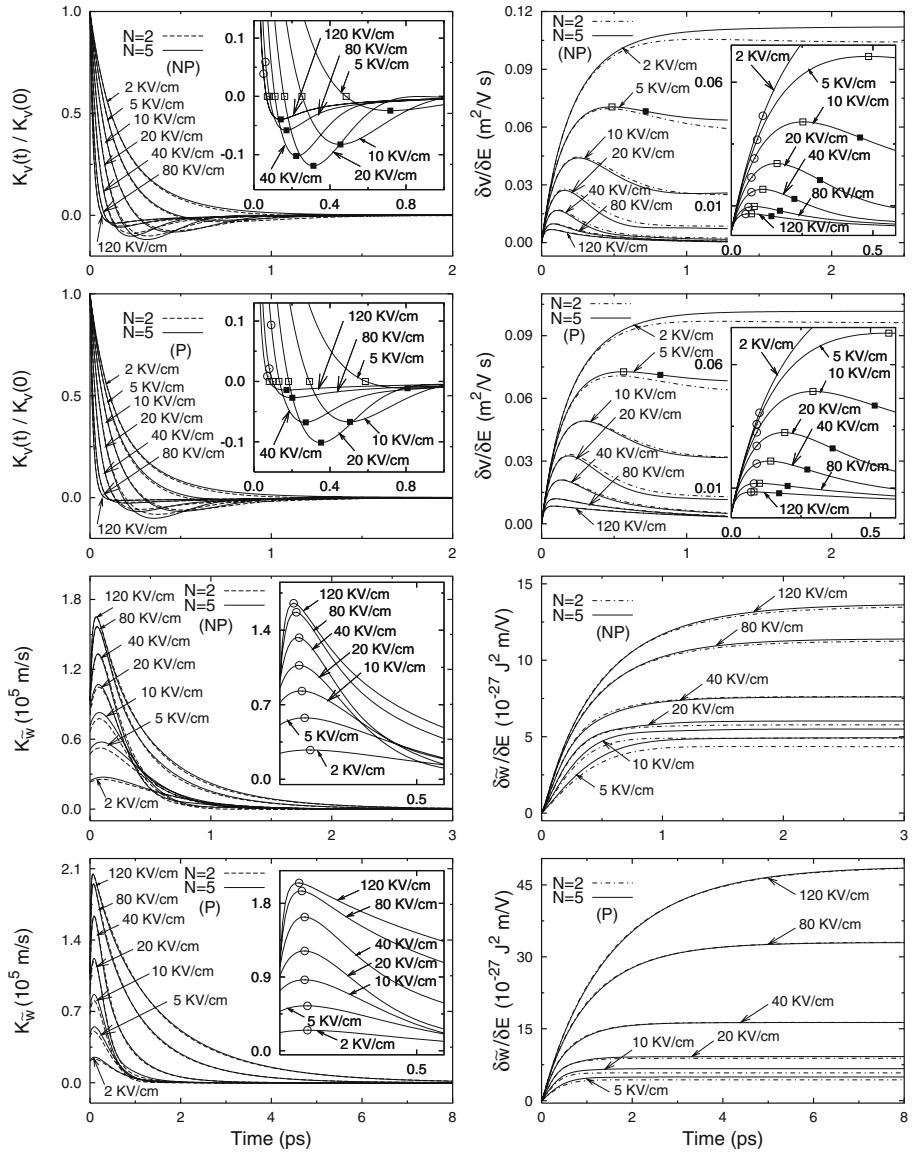


Fig. 3. Time dependencies of response functions $\{K_v,\, K_{\tilde{w}}\}$ and of differential responses $\{\delta v/\delta E,\, \delta\tilde{w}/\delta E\}$ to the step-like switch-on of electric field for $N = 2$ and $N = 5$

Figure 3 shows the velocity-response function K_v normalized to its initial value, the energy-response function $K_{\tilde{W}}$ and the corresponding differential responses $\{\delta v/\delta E,\, \delta\tilde{W}/\delta E\}$ to the step-like switch-on of electric field for parabolic and non-parabolic band models and increasing electric fields. The decay with time of the response functions is controlled essentially by the momentum and energy relaxation

rates. Figure 3 shows that, at low electric fields, the shape of the velocity-response function is practically exponential, with a characteristic time constant which corresponds to momentum relaxation. The presence of higher electric fields couples the two relaxations processes, thus giving a non-exponential shape to the decay. The shape of K_v becomes more complicated by exhibiting a negative part which correspond to the velocity overshoot of carriers. In fact the differential response $\delta v/\delta E$ quickly increases with t when $K_v(t) > 0$ reaches a maximum at time $\bar{t}$ corresponding to $K_v(\bar{t}) = 0$ and then falls with t when $K_v(t) < 0$. This is in agreement with the general relations (21-22), where (see the inserts of Fig. 3) to one extreme position of $\delta\widetilde{F}_\alpha$ (light square) corresponds a zero value of K_α, and, analogously, to one extreme position of the response function K_α (dark square) is associated a flex point of the corresponding perturbation $\delta\widetilde{F}_\alpha$.

The response function K_w clearly shows the coupling between velocity and energy relaxation through a non-monotonic behavior with a maximum which separates the velocity from the energy relaxation [28] while the corresponding differential response $\delta\widetilde{W}/\delta E$ increases in a monotone way as a function of the time for different values of the electric field. The response $K_{\tilde{w}}$ is always positive with a maximum which is reached at times t' shorter (see circles in the inserts) than the minimum of K_v. At increasing fields, because of the increased efficiency of the scattering mechanisms, the corresponding value $K_v(t')$ tends to approach the value $K_v(\bar{t}) = 0$ and analogously the corresponding value of $\delta v(t')/\delta E$ tends to approach its maximum value. Therefore if initially the carriers, gained by the field, obtain extra velocity (since their initial momentum relaxation time is somewhat longer than that in the new steady state), at a given time t' the energy relaxation starts to affect the momentum relaxation time which becomes shorter; at a later time, due to the scattering mechanisms, the corresponding fluctuation reaches its maximum, decreases and the extra velocity of the carriers is lost.

5.3 Differential mobility

The validity of this approach has been confirmed by the satisfactory agreement with the numerical results of full-band Monte Carlo (MC) simulations and available experimental data for the case of electrons in Si bulk. In Fig. 4 we report the differential mobility $\{\mu'_v, \mu'_{\widetilde{w}}\}$ for the moments $\{v, \widetilde{W}\}$, as a function of electric field for electrons in Si at $T_0 = 300\ K$. The lines refer to parabolic (P) and non-parabolic (NP) calculations obtained, for $N = 5$, from the real parts of the a.c. differential response coefficients $\mathrm{Re}[\mu'_v(f)]$ and $\mathrm{Re}[\mu'_{\widetilde{w}}(f)]$ in the low frequency limits ($f \approx 10^8 Hz$). The symbols refer to the d.c. differential mobility $\mu'_v = dv/dE$ and $\mu'_{\widetilde{w}} = d\widetilde{W}/dE$ obtained from the full-band MC simulations performed along the $\langle 111\rangle$ crystallographic direction [30]. For μ'_v we also report the derivative dv/dE of the experimental data for the drift velocity obtained with the microwave time-of-light technique along the $\langle 111\rangle$ crystallographic direction [31].

We note that in the non-parabolic case the HD results are in good agreement both with the full-band MC calculations and with the experimental data for the whole range of electric field $1KV \leq E \leq 200KV$.

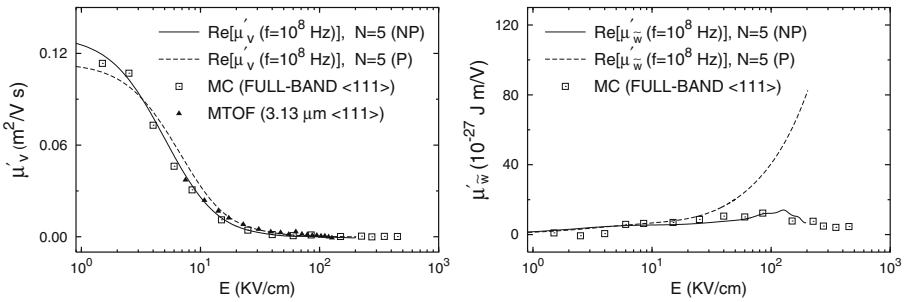


Fig. 4. Differential mobility $\{\mu_v', \mu_{\widetilde{w}}'\}$ vs electric field for electrons in Si at $T_0 = 300^o$K. The lines refer to HD calculations , with $N = 5$, for $\mathrm{Re}[\mu_v'(f)]$ and $\mathrm{Re}[\mu_{\widetilde{w}}'(f)]$ in the low frequency limits. The symbols refer to $\mu_v' = dv/dE$ and $\mu_{\widetilde{w}}' = d\widetilde{W}/dE$ obtained from the full-band MC simulations [30]. For μ_v' we also report the derivative dv/dE of the experiment [31]

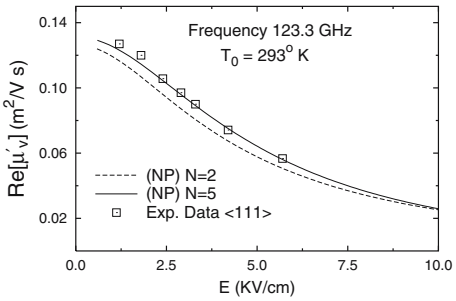


Fig. 5. Real a.c. differential mobility $\mathrm{Re}[\mu_v']$ of electrons in Si at $T_0 = 293^o$K evaluated at $f = 123.3$ GHz as a function of low electric fields. Symbols refer to the experimental data [20]. Lines refer to non-parabolic (NP) calculations with $N = 2$ and $N = 5$

As a test to validate the HD model in the high frequency range as well, we calculate the differential mobility for electrons in Si at $T_0 = 293^o$K with $f = 123.3$ GHz and we compare the numerical results with the experimental data. Thus, we report the real part of the mobility that we obtained, for the non-parabolic case (NP), from HD simulation (for $N = 2$ and $N = 5$) at 123.3 GHz in Fig. 5 as a function of low electric fields. We note that the HD calculations exhibit small variations (al most within 15%) from the number of moments used. In any case the numerical results converge for $N = 5$ in both the parabolic and non-parabolic case.

The non-parabolic HD results (for $N = 5$) agree well with the experimental data obtained at this frequency. The experimental data were deduced with the help of a measurement method of microwave transmission of n-Si samples in the $\langle 111 \rangle$ crystallographic axis with an experimental uncertainty that can be estimated at about 5% [20].

We believe that the present hydrodynamic method can be fruitfully applied to describe transport properties of hot carriers with the advantage of providing a closed

analytical approach and a reduced computational effort in comparison with other competitive numerical methods at kinetic level.

References

[1] Woolard, D.L., Tian, H., Trew, R.J., Littlejohn, M.A., Kim, K.W. (1991): Hydrodynamic electron-transport model: nonparabolic corrections to the streaming terms. Phys. Rev. B **44**, 11119–11132

[2] Thoma, R., Emunds, A., Meinerzhagen, B., Peifer, H.-J., Engl, W.L. (1991): Hydrodynamic equations for semiconductors with nonparabolic band structure. IEEE Trans. Electron Devices **38**, 1343–1353

[3] Rudan, M., Vecchi, M.C., Ventura, D. (1995): The hydrodynamic model in semiconductors – coefficient calculation for the conduction band of silicon. In: Marcati, P. et al. (eds.): Mathematical problems in semiconductor physics (Pitman Research Notes in Mathematics Series, vol. 340). Longman, Harlow

[4] Starikov, E., Shiktorov, P., Gruzinskis, V., Gonzalez, T., Martin, M.J., Pardo, D., Reggiani, L., Varani, L. (1996): Hydrodynamic and Monte Carlo simulation of steady-state transport and noise in submicrometre n^+nn^+ silicon structure. Semiconductor Sci. Tech. **11**, 865–872

[5] Müller, I., Ruggeri, T. (1998): Rational extended thermodynamics. (Springer Tracts in Natural Philosophy, vol. 37). Springer, New York

[6] Anile, A.M., Trovato, M. (1997): Nonlinear closures for hydrodynamical semiconductor transport models. Phys. Lett. A **230**, 387–395

[7] Falsaperla, P., Trovato, M. (1998): Ahydrodynamic model for transport in semiconductors without free parameters. VLSI Design **8**, 527–531

[8] Trovato, M., Falsaperla, P. (1998): Hydrodynamic model for hot carriers in silicon based on the maximum entropy formalism. In: De Meyer, K., Biesemans, S. (eds.): Simulation of semiconductor processes and devices, 1998. SISPAD 98. Springer, Vienna, pp. 320–323

[9] Trovato, M., Falsaperla, P. (1998): Full nonlinear closure for a hydrodynamic model of transport in silicon. Phys. Rev. B **57**, 4456–4471; erratum: Phys. Rev. B **57**, 12617

[10] Trovato, M., Reggiani, L. (1999): Maximum entropy principle for hydrodynamic transport in semiconductor devices. J. Appl. Phys. **85**, 4050–4065

[11] Trovato, M., Falsaperla, P., Reggiani, L. (1999): Maximum entropy principle for nonparabolic hydrodynamic transport in semiconductor devices. J. Appl. Phys. **86**, 5906–5908

[12] Struchtrup, H. (2000): Extended moment method for electrons in semiconductors. Physica A **275**, 229–255

[13] Liotta, S.F., Struchtrup, H. (2000): Moment equations for electrons in semiconductors: comparison of spherical harmonics and full moments. Solid State Electronics **44**, 95–103

[14] Trovato, M., Reggiani, L. (2000): Maximum entropy principle within a total energy scheme: application to hot-carrier transport in semiconductors. Phys. Rev. B **61**, 16667–16681

[15] Trovato, M., Reggiani, L. (2001): Maximum entropy principle within a total energy scheme for hot-carrier transport in semiconductor devices. VLSI Design **13**, 381–386

[16] Trovato, M. (2002): Hydrodynamic analysis for hot-carriers transport in semiconductors. In: Monaco, R. et al. (eds.): "WASCOM 2001" – 11th conference on waves and stability in continuous media. World Scientific, River Edge, NJ, pp. 585–590

[17] Mascali, G., Trovato, M. (2002): A non-linear determination of the distribution function of degenerate gases with an application to semiconductors. Physica A **310**, 121–138

[18] Zubarev, D.N. (1974): Nonequilibrium statistical mechanics. Consultants Bureau, London
[19] Drabold, D.A., Carlsson, A.E., Fedders, P.A. (1989): Applications of maximum entropy to condensed matter physics. In: Skilling, J. (ed.): Maximum entropy and Bayesian methods. Kluwer, Dordrecht, pp. 137 ff.
[20] Zimmermann, J., Leroy, Y., Constant, E. (1978): Monte Carlo calculation of microwave and far-infrared hot-carrier mobility in N-Si: efficiency of millimeter transit-time oscillators. J. Appl. Phys. **49**, 3378–3383
[21] Price, P.J. (1982): Dispersion relations for hot electrons. J. Appl. Phys. **53**, 8805–8808
[22] Price, P.J. (1983): On the calculation of differential mobility. J. Appl. Phys. **54**, 3616–3617
[23] Kuhn, T., Reggiani, L., Varani, L. (1990): Correlation functions and electronic noise in doped semiconductors. Phys. Rev. B **42**, 11133–11146
[24] Kuhn, T., Reggiani, L., Varani, L. (1992): Coupled-Langevin-equation analysis of hot-carrier transport in semiconductors. Phys. Rev. B **45**, 1903–1906
[25] Gruzhinskis, V., Starikov, E., Shiktorov, P., Reggiani, L., Saraniti, M., Varani, L. (1993): Hydrodynamic analysis of DC and AC hot-carrier transport in semiconductors. Semiconductor Sci. Tech. **8**, 1283–1290
[26] Varani, L., Vaissiere, J.C., Nougier, J.P., Houlet, P., Reggiani, L., Starikov, E., Shiktorov, P., Gruzhinskis, V., Hlou, L. (1995): A model hyperfrequency differential-mobility for nonlinear transport in semiconductors. J. Appl. Phys. **77**, 665–675
[27] Reggiani, L., Starikov, E., Shiktorov, P., Gruzinskis, V., Varani, L. (1997): Modelling of small-signal response and electronic noise in semiconductor high-field transport. Semiconductor Sci. Tech. **12**, 141–156
[28] Nedjalkov, M., Kosina, H., Selberherr, S. (1999): Monte Carlo method for direct computation of the small signal kinetic coefficients. In: 1999 International conference on simulation of semiconductor processes and devices. SISPAD'99. Business Center for Academic Societies Japan, Tokyo, pp. 155–158
[29] Canali, C., Jacoboni, C., Nava, F., Ottaviani, G., Alberigi-Quaranta, A. (1975): Electron drift velocity in silicon. Phys. Rev. B **12**, 2265–2284
[30] Fischetti, M. (1991): Monte Carlo simulation of transport in technologically significant semiconductors of the diamond and zinc-blende structures. I. Homogeneous transport. IEEE Trans. Electron Devices **38**, 634–649
[31] Smith, P.M., Inoue, M., Frey, J. (1980): Electron velocity in Si and GaAs at very high electric fields. Appl. Phys. Lett. **37**, 797–798

Small planar oscillations of an incompressible, heavy, almost homogeneous liquid filling a container

D. Vivona*

1 Introduction

The problem of small oscillations of an inviscid, incompressible, heavy, heterogeneous liquid in a container, was studied by a few authors: Rayleigh [1], Love [2] and Lamb [3].

In a paper published in 1993, Capodanno proved that the problem is not a classical problem of eigenvalues [4]. But he described the spectrum exactly only in a particular case, not for an arbitrary container.

This research was solicited from Capodanno by Philips and Transoft International, which create software for the transport of the liquids.

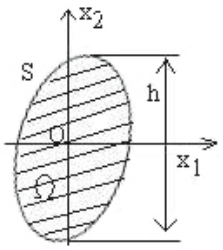
It is thus of interest to study the problem for an arbitrary container, at least for particular densities. We study the case where the density of the liquid in the equilibrium position can be approximated by a linear function of the height of the particle, which differs very little from a constant. In this case, the liquid is called *almost homogeneous*.

For an inviscid liquid, we prove the existence of the essential spectrum.

For a viscous liquid, we have a point spectrum, analogous to that of small oscillations of a liquid with free surface.

2 Inviscid liquid

Following Rayleigh and Love, we consider only planar motions. The liquid fills the domain Ω bounded by a regular curve S. The origin O belongs to Ω and the axis Ox_2 is vertical.



* This research was supported by GNFM of MURST (Italy).

2.1 Equations of motion

We denote by $\boldsymbol{u}(x_1, x_2, t)$ the displacement of a particle with respect to the equilibrium position, $\rho_0(x_2)$ the known density at the equilibrium, $p(x_1, x_2, t)$ the dynamical pressure. At the first order we have the equations:

$$\rho_0 \ddot{\boldsymbol{u}} = - \boldsymbol{grad}\ p + \rho_0^{'}\ g u_2 \boldsymbol{e_2} \quad \text{in } \Omega\,, \tag{1}$$

$$div\ \boldsymbol{u} = 0 \qquad \text{in } \Omega\,, \tag{2}$$

$$\boldsymbol{u} \cdot \boldsymbol{n} = 0 \qquad \text{on } S\,, \tag{3}$$

where $\rho_0^{'}\ g u_2 \boldsymbol{e_2}$ is obtained by the continuity equation. We assume $\rho_0^{'}(x_2) < 0$.

2.2 Variational formulation and spectral problem

It is easy to see that

$$\int_\Omega \rho_0\ \ddot{\boldsymbol{u}}\ \overline{\boldsymbol{w}}\ d\Omega = - \int_\Omega \rho_0^{'}(x_2)\ u_2\ \overline{w_2}\ d\Omega$$

for all $\boldsymbol{u}$ such that $div\ \boldsymbol{u} = 0$ in Ω and $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ on S.

We seek solutions of the form $\boldsymbol{u}(x_1, x_2, t) = e^{i\omega t}\boldsymbol{U}(x_1, x_2)$; putting

$$\boldsymbol{U} = (\frac{\partial \psi}{\partial x_2}, -\frac{\partial \psi}{\partial x_1}),$$

$$\boldsymbol{w} = (\frac{\partial \varphi}{\partial x_2}, -\frac{\partial \varphi}{\partial x_1}),$$

we introduce the space V_0 of the functions of H_0^1 equipped with the scalar product

$$(\psi, \varphi)_{V_0} = \int_\Omega \rho\ \boldsymbol{grad}\ \psi \cdot \boldsymbol{grad}\ \varphi\ d\Omega$$

and so we obtain the following eigenvalue problem.

Find $\psi(x_1, x_2) \in V_0$ and a real positive number ω^2 such that

$$\omega^2(\psi, \varphi)_{V_0} = a(\psi, \varphi) \quad \forall\ \varphi \in V_0\ ,$$

where $a(\psi, \varphi) = \int_\Omega -\rho_0^{'}(x_2)\ g\ \frac{\partial \psi}{\partial x_2}\ \frac{\partial \psi}{\partial x_1}\ d\Omega$.

As $a(\cdot, \cdot)$ is continuous on $V_0 \times V_0$, there exists a bounded linear operator A of V_0 in V_0 such that $a(\psi, \varphi) = (A\psi, \varphi)_{V_0}$.

We get the spectral problem

$$A\psi = \omega^2\psi, \quad \psi \in V_0\ .$$

It is easy to see that A is symmetric, and we can prove that A is positive definite. But A is not compact and its spectrum is not discrete.

One can describe the spectrum exactly if the container is a rectangle [5], in particular in Rayleigh's case $\rho_0 = ke^{-\beta x_2}(k, \beta \text{ constant} > 0)$. We can calculate the eigenvalues which form a set which is dense in $[0, \beta\ g]$, and the *essential spectrum* is $[0, \beta\ g]$ (formed by the eigenvalues and their points of accumulation).

2.3 Definition of an almost homogeneous liquid

Let h be the maximal vertical dimension or height of the container; thus, $|x_2| < h$ in Ω.

Now we assume that $\rho_0(x_2)$ has the form

$$\rho_0(x_2) = f(\beta x_2),$$

where $f(0) > 0$, $f'(0) < 0$ and β is a positive constant such that βh is small enough so that $(\beta h)^2$, $(\beta h)^3$, ... are negligible with respect to βh. Then, as $|\beta x_2| < \beta h$ in Ω, we get

$$\rho_0(x_2) = f(0) + \beta x_2 f'(0) + \dots \ .$$

In this case the liquid is called *almost homogeneous in* Ω.

Now, changing notation, we write

$$\rho_0(x_2) = \rho(1 - \beta x_2) + o(\beta h) \ .$$

(In particular, this is the Rayleigh's case under the preceding condition)

2.4 Equations of small oscillations and operator of the problem

By replacing ρ_0 with ρ and $\rho_0^{'}$ with $-\rho\beta$, we have the approximate equation of small oscillations, analogous to that of Boussinesq [6] in the theory of convective fluid motions:

$$\ddot{\boldsymbol{u}} = -\frac{1}{\rho}\, \boldsymbol{grad}\ p - \beta g u_2 \boldsymbol{e_2} \tag{4}$$

to which we must add Eqs. (2) and (3).

We can assume that $\boldsymbol{u} \in \mathcal{J}_0(\Omega)$, where

$$\mathcal{J}_0(\Omega) = \{\boldsymbol{u} \in \mathcal{L}^2(\Omega) = [L^2(\Omega)]^2 \ : \ div\, \boldsymbol{u} = 0,\ u_n = 0 \text{ in } H^{-1/2}(S)\} \tag{5}$$

and that $p \in H^1(\Omega)$, so that $\boldsymbol{grad}\ p$ and $-\frac{1}{\rho}\ \boldsymbol{grad}\ p$ belong to $\mathcal{G}(\Omega)$ which is the space of potential fields [6].

By using the orthogonal decomposition of Weyl [6, 7],

$$\mathcal{L}^2(\Omega) = \mathcal{J}_0(\Omega) \oplus \mathcal{G}(\Omega),$$

and by denoting the projector of $\mathcal{L}^2(\Omega)$ on $\mathcal{J}_0(\Omega)$ by $\mathcal{P}_0$, we have

$$\ddot{\boldsymbol{u}} = -\beta g \mathcal{P}_0(u_2 \boldsymbol{e_2}) \ .$$

By introducing the operator K of $\mathcal{J}_0$ to itself, defined by $K\boldsymbol{u} = \beta g \mathcal{P}_0(u_2 \boldsymbol{e_2})$, we get the equations of small oscillations

$$\ddot{\boldsymbol{u}} + K\boldsymbol{u} = 0 \quad \forall\, \boldsymbol{u} \in \mathcal{J}_0 \ . \tag{6}$$

It is easy to see that K is symmetric, bounded ($||K|| \leq \beta g$) and non-negative.

2.5 Study of the spectrum $\sigma(K)$ of the operator K

We want to prove that $\sigma(K) = [0, \beta g]$. By using Weyl's theorem [6], it is sufficient to prove that, for every μ, $0 < \mu < 1$, there exists a sequence $\{\boldsymbol{u}_l\} \in \mathcal{J}_0(\Omega)$ such that

$$\frac{||\frac{1}{\beta g}K\boldsymbol{u}_l - \mu\boldsymbol{u}_l||}{||\boldsymbol{u}_l||} \to 0 \text{ when } l \to +\infty .$$

In order to construct $\{\boldsymbol{u}_l\}$, we consider $q \in \mathcal{D}(\Omega)$ and $\boldsymbol{u} \in \mathcal{J}_0(\Omega))$ with

$$\boldsymbol{u} = (u_1 = \frac{\partial \Delta q}{\partial x_2}, u_2 = -\frac{\partial \Delta q}{\partial x_1}) .$$

It is possible to calculate $K\boldsymbol{u}$.

We take the sequence $\{u_{nm}\}$, with $q(x) = q_{nm}(x) = e^{i(nx_1+mx_2)}\psi(x)$, where $\psi(x) \in \mathcal{D}(\Omega)$ and is equal to 1 in the disk $C : |x - x_0| \le r$, $C \subset \Omega$. We prove that

$$\frac{1}{\beta g}K\boldsymbol{u}_{nm} - \frac{n^2}{n^2+m^2}\boldsymbol{u}_{nm} = O(n^2+m^2),$$

where $\frac{O(n^2+m^2)}{n^2+m^2}$ is uniformly bounded in Ω and

$$c_0(n^2+m^2)^3 \le ||\boldsymbol{u}_{nm}||^2 \le c_1(n^2+m^2)^3$$

with c_1, c_0 constant > 0.

Let $0 < \mu < 1$. For every $\varepsilon > 0$, we can find a rational number $\frac{\bar{m}}{\bar{n}}$ such that

$$\mu < \frac{\bar{n}^2}{\bar{n}^2+\bar{m}^2} < \mu + \varepsilon .$$

Choosing $m = l\bar{m}$, $n = l\bar{n}$, we find easily that the sequence $\{\boldsymbol{u}_{\,l\bar{n},\,l\bar{m}}\}$ satisfies Weyl's theorem.

Remark
If $\omega^2 \in \sigma(K)$, by Weyl's theorem there exists a sequence $\{\boldsymbol{v}_i\}$ such that

$$||\boldsymbol{v}_i|| = 1 \text{ and } (K - \omega^2 I)\boldsymbol{v}_i \to 0 \quad \text{in } \mathcal{J}_0(\Omega);$$

we can say that there is a kind of *resonance*.

3 Viscous liquid

3.1 Operator equation of the problem

Always in the almost homogeneous case, the approximate equation becomes

$$\ddot{\boldsymbol{u}} = -\frac{1}{\rho}\,\boldsymbol{grad}\; p + \nu\Delta\dot{\boldsymbol{u}} - \beta g u_2 \boldsymbol{e}_2 , \tag{7}$$

where the constant ν is an approximate value of the kinematic coefficient of viscosity and the condition on the wall is the adhesion condition

$$\dot{\boldsymbol{u}}_{|S} = 0 \, .$$

In order to obtain the variational equation of the problem, it is sufficient to recall that it expresses the principle of virtual work. Now we introduce, together with $\mathcal{J}_0(\Omega)$ given by (5), the space

$$\mathcal{J}_0{}^1(\Omega) = \{\boldsymbol{u} \in \mathcal{H}_0{}^1(\Omega) = [H_0^1(\Omega)]^2 : \ div\, \boldsymbol{u} = 0\} \, ,$$

equipped with the norm associated to the scalar product

$$E(\boldsymbol{u}, \boldsymbol{v}) = 2 \int_\Omega \varepsilon_{ij}(\boldsymbol{u}) \dot{\varepsilon_{ij}}(\bar{\boldsymbol{v}}) \, d\Omega,$$

which is equivalent in $\mathcal{J}_0{}^1(\Omega)$ to the classical norm of $\mathcal{H}_0{}^1(\Omega)$.

As $\rho\nu E(\dot{\boldsymbol{u}}, \boldsymbol{v})$ is the virtual work of the viscosity forces in the fields of displacements $\boldsymbol{v}$, we obtain the variational equation

$$\int_\Omega \ddot{\boldsymbol{u}} \cdot \bar{\boldsymbol{v}} \, d\Omega + \nu E(\dot{\boldsymbol{u}}, \boldsymbol{v}) + (\beta g u_2 \boldsymbol{x}_2, \boldsymbol{v})_{\mathcal{L}^2(\Omega)} = 0 \tag{8}$$

$\forall\, \boldsymbol{v} \in \mathcal{J}_0{}^1(\Omega)$.

The injection of $\mathcal{J}_0{}^1(\Omega)$ in $\mathcal{J}_0(\Omega)$ is dense, continuous and compact. Therefore, calling A_0 the unbounded operator associated to $E(\cdot, \cdot)$, we can easily deduce from the variational equation (8), the operator equation

$$\ddot{\boldsymbol{u}} + \nu A_0 \dot{\boldsymbol{u}} + K\boldsymbol{u} = 0 \quad \forall\, \boldsymbol{u} \in \mathcal{J}_0{}^1(\Omega) \, . \tag{9}$$

3.2 The spectrum of the problem

Seeking the normal oscillations, i.e., the solution of the form $\boldsymbol{u} = e^{-\lambda t}\boldsymbol{U}(x)$ and, then, putting $A_0^{1/2}\boldsymbol{U} = \boldsymbol{V} \in \mathcal{J}_0{}^1(\Omega)$, we obtain the equation

$$\boldsymbol{V} = \lambda\nu^{-1} A_0^{-1}\boldsymbol{V} + \frac{1}{\lambda}\nu^{-1} K_0 \boldsymbol{V} \, , \quad \boldsymbol{V} \in \mathcal{J}_0(\Omega) \, ,$$

where A_0^{-1} and $K_0 = A_0^{-1/2} K A_0^{-1/2}$ are operators of $\mathcal{J}_0(\Omega)$ to itself. The operator A_0^{-1} is classically self-adjoint, positive definite and compact. The operator K_0 is self-adjoint and non-negative; it is compact, as K is bounded and $A_0^{-1/2}$ is compact.

Therefore, we can apply the Askerov-Kreĭn-Laptev theorem [8] which states that there are countably many eigenvalues whose real part is positive, and that there are infinitely many aperiodic arbitrary strongly damped motions ($\lambda \to +\infty$) and infinitely many aperiodic arbitrary weakly damped motions ($\lambda \to 0$).

If $\nu^2 > 4||A_0^{-1}||\ ||K_0||$, i.e., the viscosity is large, all eigenvalues are real and there are no oscillatory damped motions.

If $\nu^2 \leq 4||A_0^{-1}||\ ||K_0||$, there are at most a finite number of non-real eigenvalues, which are in the annulus

$$\frac{\nu}{2||A_0^{-1}||} \leq |\lambda| \leq \frac{2||K_0||}{\nu}\ ,$$

corresponding to oscillatory damped motions.

Thus, *viscosity suppresses the essential spectrum*, which is replaced by a point spectrum analogous to the spectrum of oscillations of a viscous liquid with free surface.

References

[1] Rayleigh, Lord (1883): Investigation of the character of the equilibrium of an incompressible heavy liquid of variable density. Proc. London Math. Soc. **14**, 170–177

[2] Love, A.E.H. (1891): Wave motion in a heterogeneous heavy liquid. Proc. London Math. Soc. **22**, 307–316

[3] Lamb, H. (1932): Hydrodynamics. 6th ed. Cambridge University Press, London

[4] Capodanno, P. (1993): Un exemple simple de problème non standard de vibration: oscillations d'une liquide heterogene pesant dans un container. Mech. Res. Comm. **20**, 257–262

[5] Capodanno, P., Vivona, D. (2003): Mathematical study of the small oscillations of a heavy almost homogeneous liquid in a container. Rev. Roumaine Sci. Tech. Sér. Méc. Appl., submitted

[6] Kopachevsky, N.D., Krein, S.G. (2001): Operator approach to linear problems of hydrodynamics. Vol. 1. Self-adjoint problems for an ideal fluid. Birkhäuser, Basel

[7] Weyl, H. (1940): The method of orthogonal projection in potential theory. Duke Math. J. **7**, 411–444

[8] Askerov, N.G., Kreĭn, S.G., Laptev, G.I. (1964): On a class of non-self-adjoint boundary value problems. Dokl. Akad. Nauk SSSR **155**, 499–502

Thermodynamics of simple two-component thermo-poroelastic media

K. Wilmanski

Abstract. The paper is devoted to the thermodynamic construction of a two-component model of poroelastic media undergoing, in contrast to earlier works on this subject, nonisothermal processes. Under the constitutive dependence on partial mass densities, deformation gradient of skeleton, relative velocity, temperature, temperature gradient and porosity (simple poroelastic material) as well as the assumption of small deviations from the thermodynamic equilibrium we construct explicit relations for fluxes, prove the splitting of the free energy into partial contributions without mechanical couplings and construct a chemical potential for the fluid component important for the formulation of boundary conditions on permeable boundaries. We discuss as well a modification of the porosity balance equation in which we account for time changes of equilibrium porosity. This modification yields the behavior of the model characteristic for granular materials.

1 Introduction

Thermodynamic modeling of saturated poroelastic materials by means of a two-component continuum has been limited to isothermal processes. R. M. Bowen who initiated the work in this field [1,2] constructed a model for a multicomponent system with large deformations of the skeleton and internal variables (volume fractions) for which he postulated evolution equations. Such an approach may be appropriate for some biomaterials but it fails for granular media. For the latter, relaxation processes for internal variables are almost immaterial and the main mechanism driving changes of porosity are volume changes of components. Such phenomena are described within the model with the porosity balance equation. For a two-component system this was proposed in [4] and developed in [5]. Fundamental properties, its microscopic motivation and a transition to a linear model are presented in [6]. As a model belonging to the so-called extended thermodynamics it was extended to multicomponent systems in [7,9]. All these papers concern so-called simple poroelastic materials in which there is no constitutive dependence on higher gradients of fields. Linearization of such models does not lead to Biot's model which is successfully applied in various fields of geotechnics. It has been shown that additional couplings appearing in Biot's model require an extension of constitutive variables; the thermodynamics of such a model in which the gradient of porosity is the constitutive variable is the subject of [8].

In the present work we present a thermodynamic analysis of a model of simple poroelastic materials which differs from those mentioned above in two essential points:

- processes are not isothermal; the system is characterized by a single temperature field which may vary in space and time;
- the porosity balance equation is extended by changes of equilibrium porosity.

Changes in the porosity balance equation result from the analysis of earlier models for granular materials. Micro-macro transition as well as solutions of simple boundary problems show that the original model yields a very stiff behavior of the skeleton which is appropriate for rocks but not for granular materials (compare [10]). In addition, if we neglect the relaxation of porosity the modified equation of porosity yields in the linear case changes of porosity indicated by Gassmann relations.

We base our considerations on the assumption that deviations from thermodynamic equilibrium are small. As these deviations are described by three variables, temperature gradient $\mathbf{G} := \mathrm{Grad}\, T$, Lagrangian (relative) velocity $\dot{\mathbf{X}}^F$ and a deviation of porosity from its equilibrium value $\Delta_n := n - n_E$, this assumption means that we assume the dissipation to be a quadratic function of these variables.

Under this assumption we prove the following properties: 1) the Helmholtz free energy splits into two partial potentials which are not coupled by mechanical variables (*simple mixture*), 2) thermal parts of energy and entropy fluxes are connected by the classical Fourier relation, and 3) the flux of porosity contains only a linear contribution of the Lagrangian velocity with a coefficient proportional only to volume changes of the skeleton.

We complete the work with a presentation of a few simplified models. We show that the fully linear model does not coincide with the classical Biot's model due to the lack of coupling between partial stresses. This property was proven earlier for isothermal processes in simple poroelastic materials.

2 Balance equations

We use the Lagrangian description referring to the reference configuration $\mathcal{B}$ of the skeleton [3] in which its deformation gradient $\mathbf{F}^S = \mathbf{1}$. In the two-component medium considered in this work the partial balance equations for the skeleton are defined on a family of volume measurable sets $\left\{ \mathcal{P}^S \,\middle|\, \mathcal{P}^S \subset \mathcal{B} \right\}$ material with respect to the skeleton, i.e., independent of time. Simultaneously partial balance equations for the fluid are defined on a time dependent family of volume measurable sets $\left\{ \mathcal{P}^F \,\middle|\, \mathcal{P}^F \subset \mathcal{B} \right\}$ with the kinematics defined by the Lagrangian field of the relative velocity

$$\dot{\mathbf{X}}^F(\mathbf{X}, t) = \mathbf{F}^{S-1}\left(\dot{\mathbf{x}}^F - \dot{\mathbf{x}}^S\right), \quad \mathbf{X} \in \mathcal{B} \subset \mathcal{R}^3, \quad t \in \mathcal{T} \subset \mathcal{R}, \tag{1}$$

in which $\dot{\mathbf{x}}^F$, $\dot{\mathbf{x}}^S$ are the velocities of the fluid component and of the skeleton, respectively. Clearly, for the existence of a function of motion of the skeleton, we require the following conditions to be satisfied:

$$\dot{\mathbf{F}} := \frac{\partial \mathbf{F}^S}{\partial t} + \mathrm{Grad}\, \dot{\mathbf{x}}^S = 0, \quad \mathrm{Grad}\, \mathbf{F}^S = \left(\mathrm{Grad}\, \mathbf{F}^S\right)^{\overset{23}{T}}. \tag{2}$$

We say that a field φ^S, describing a property of the skeleton, whose flux is $\mathbf{\Psi}^S$ and supply is $\hat{\varphi}^S$, satisfies a balance equation if, for any set $\mathcal{P}^S$,

$$\frac{d}{dt}\int_{\mathcal{P}^S}\varphi^S dV = \oint_{\partial\mathcal{P}^S}\mathbf{\Psi}^S\cdot\mathbf{N}dS + \int_{\mathcal{P}^S}\hat{\varphi}^S dV, \tag{3}$$

where $\partial\mathcal{P}^S$ is the oriented boundary of $\mathcal{P}^S$, and $\mathbf{N}$ is the field of its unit outward normal vectors. A similar equation is assumed to hold for a field φ^F describing a property of the fluid.

Quantities appearing in the above equations are assumed to have at most finite singularities on a set of volume measure zero. For the purpose of this work it is sufficient to assume that this set forms an oriented surface $\mathcal{S}$ given by the equation

$$S(\mathbf{X},t)=0 \quad\Longrightarrow\quad \mathbf{N}=\frac{\operatorname{Grad}S}{|\operatorname{Grad}S|},\quad U=-\frac{\frac{\partial S}{\partial t}}{|\operatorname{Grad}S|},\quad \mathbf{X}\in\mathcal{B}, \tag{4}$$

where $U(\mathbf{X},t)$ is its normal speed of propagation through the reference configuration $\mathcal{B}$.

Under this assumption we can write the above equations in the following local form (e.g., [6]):

- at regular points a.e. in $\mathcal{B}$

$$\frac{\partial\varphi^S}{\partial t}=\operatorname{Div}\mathbf{\Psi}^S+\hat{\varphi}^S,$$
$$\frac{\partial\varphi^F}{\partial t}+\operatorname{Div}\left(\varphi^F\dot{\mathbf{X}}^F\right)=\operatorname{Div}\mathbf{\Psi}^F+\hat{\varphi}^F, \tag{5}$$

- at singular points on $\mathcal{S}$

$$-U\left[\left[\varphi^S\right]\right]=\left[\left[\mathbf{\Psi}^S\right]\right]\cdot\mathbf{N},\quad \left[\left[\varphi^F\left(\dot{\mathbf{X}}^F\cdot\mathbf{N}-U\right)\right]\right]=\left[\left[\mathbf{\Psi}^F\right]\right]\cdot\mathbf{N}, \tag{6}$$

where $[[\ldots]]=(\ldots)^+-(\ldots)^-$ denotes the difference of limits on both sides of the surface $\mathcal{S}$.

Furthermore, the fields appearing in the balance equations are partial mass densities in the reference configuration ρ^S, ρ^F, partial momentum densities $\rho^S\dot{\mathbf{x}}^S$, $\rho^F\dot{\mathbf{x}}^F$, partial energies $\rho^S\left(\varepsilon^S+\frac{1}{2}\dot{x}^{S2}\right)$, $\rho^F\left(\varepsilon^F+\frac{1}{2}\dot{x}^{F2}\right)$, porosity n, and partial entropies $\rho^S\eta^S$, $\rho^F\eta^F$. If we neglect a mass exchange between components then they have the form:

- partial mass conservation laws at regular points a.e. in $\mathcal{B}$ and singular points on $\mathcal{S}$

$$R^S:=\frac{\partial\rho^S}{\partial t}=0,\quad R^F:=\frac{\partial\rho^F}{\partial t}+\operatorname{Div}\left(\rho^F\dot{\mathbf{X}}^F\right)=0, \tag{7}$$
$$U\left[\left[\rho^S\right]\right]=0,\quad \left[\left[\rho^F\left(\dot{\mathbf{X}}\cdot\mathbf{N}-U\right)\right]\right]=0;$$

- partial momentum balance equations at regular points a.e. on $\mathcal{B}$

$$\mathbf{M}^S := \rho^S \frac{\partial \acute{\mathbf{x}}^S}{\partial t} - \operatorname{Div} \mathbf{P}^S - \hat{\mathbf{p}} = 0, \tag{8}$$
$$\mathbf{M}^F := \rho^F \left(\frac{\partial \acute{\mathbf{x}}^F}{\partial t} + \acute{\mathbf{X}}^F \cdot \operatorname{Grad} \acute{\mathbf{x}}^F \right) - \operatorname{Div} \mathbf{P}^F + \hat{\mathbf{p}} = 0,$$

where $\mathbf{P}^S$, $\mathbf{P}^F$ denote Piola-Kirchhoff partial stress tensors, and $\hat{\mathbf{p}}$ is the source of momentum, and at singular points on $\mathcal{S}$

$$-U \left[\left[\rho^S \acute{\mathbf{x}}^S \right] \right] = \left[\left[\mathbf{P}^S \right] \right] \mathbf{N}, \quad \left[\left[\rho^S \acute{\mathbf{x}}^S \left(\acute{\mathbf{X}}^F \cdot \mathbf{N} - U \right) \right] \right] = \left[\left[\mathbf{P}^S \right] \right] \mathbf{N}; \tag{9}$$

- partial energy balance equations at regular points a.e. on $\mathcal{B}$

$$\frac{\partial \left(\rho^S \varepsilon^S + \frac{1}{2} \rho^S \acute{x}^{S2} \right)}{\partial t} + \operatorname{Div} \left(\mathbf{Q}^S - \mathbf{P}^{ST} \acute{\mathbf{x}}^S \right) = 0, \tag{10}$$

$$\frac{\partial \left(\rho^F \varepsilon^F + \frac{1}{2} \rho^F \acute{x}^{F2} \right)}{\partial t} + \operatorname{Div} \left(\left(\rho^F \varepsilon^F + \frac{1}{2} \rho^F \acute{x}^{F2} \right) \acute{\mathbf{X}}^F + \mathbf{Q}^F - \mathbf{P}^{FT} \acute{\mathbf{x}}^F \right) = 0,$$

and at singular points on $\mathcal{S}$

$$-U \left[\left[\rho^S \left(\varepsilon^S + \tfrac{1}{2} \acute{x}^{S2} \right) \right] \right] + \left[\left[\mathbf{Q}^S - \mathbf{P}^{ST} \acute{\mathbf{x}}^S \right] \right] \cdot \mathbf{N} = 0,$$

$$\left[\left[\rho^F \left(\acute{\mathbf{X}}^F \cdot \mathbf{N} - U \right) \left(\varepsilon^F + \tfrac{1}{2} \acute{x}^{F2} \right) \right] \right] + \left[\left[\mathbf{Q}^F - \mathbf{P}^{FT} \acute{\mathbf{x}}^F \right] \right] \cdot \mathbf{N} = 0; \tag{11}$$

- porosity balance equation at regular points a.e. on $\mathcal{B}$

$$N := \frac{\partial \Delta_n}{\partial t} + \operatorname{Div} \mathbf{J} - \hat{n} = 0, \quad \Delta_n := n - n_E, \tag{12}$$

where $\mathbf{J}$ denotes the flux of porosity, $\hat{n}$ is its source, and n_E is the porosity in thermodynamic equilibrium, and at singular points on $\mathcal{S}$

$$-U \left[\left[\Delta_n \right] \right] + \left[\left[\mathbf{J} \right] \right] \cdot \mathbf{N} = 0; \tag{13}$$

- partial entropy balance equations at regular points a.e. on $\mathcal{B}$

$$\frac{\partial \left(\rho^S \eta^S \right)}{\partial t} + \operatorname{Div} \mathbf{H}^S = \hat{\eta}^S, \quad \frac{\partial \left(\rho^F \eta^F \right)}{\partial t} + \operatorname{Div} \left(\rho^F \eta^F \acute{\mathbf{X}}^F + \mathbf{H}^F \right) = \hat{\eta}^F, \tag{14}$$

and at singular points on $\mathcal{S}$

$$-U \left[\left[\rho^S \eta^S \right] \right] + \left[\left[\mathbf{H}^S \right] \right] = 0, \tag{15}$$
$$\left[\left[\rho^F \left(\acute{\mathbf{X}}^F \cdot \mathbf{N} - U \right) \eta^F \right] \right] + \left[\left[\mathbf{H}^F \right] \right] \cdot \mathbf{N} = 0.$$

The partial energy balance equations and partial entropy balance equations are used solely in the bulk form which we explain further in this work. This means that we add corresponding partial equations to each other. After easy calculations, the following balance equation for the internal energy follows:

$$E := \frac{\partial(\rho\varepsilon)}{\partial t} + \operatorname{Div}\mathbf{Q} - \mathbf{P}^S \cdot \operatorname{Grad}\acute{\mathbf{x}}^S - \mathbf{P}^F \cdot \operatorname{Grad}\acute{\mathbf{x}}^F - \left(\mathbf{F}^{ST}\hat{\mathbf{p}}\right)\cdot\acute{\mathbf{X}}^F = 0, \quad (16)$$

where

$$\begin{aligned} &\rho := \rho^S + \rho^F, \quad \rho\varepsilon := \rho^S\varepsilon^S + \rho^F\varepsilon^F, \\ &\mathbf{Q} := \mathbf{Q}^S + \mathbf{Q}^F + \rho^F\varepsilon^F\acute{\mathbf{X}}^F, \end{aligned} \quad (17)$$

i.e., ε, $\mathbf{Q}$ are the so-called intrinsic parts of the bulk internal energy and energy flux, respectively.

Simultaneously for the entropy we obtain

$$\hat{\eta}^S + \hat{\eta}^F = \frac{\partial(\rho\eta)}{\partial t} + \operatorname{Div}\mathbf{H}, \quad (18)$$

where

$$\rho\eta := \rho^S\eta^S + \rho^F\eta^F, \quad \mathbf{H} := \mathbf{H}^S + \mathbf{H}^F + \rho^F\eta^F\acute{\mathbf{X}}^F. \quad (19)$$

For the purpose of this work we use conditions on the singular surface only for the boundary of the skeleton on which $U \equiv 0$. We then have

$$\begin{aligned} &\left[\left[\rho^F\acute{\mathbf{X}}^F\cdot\mathbf{N}\right]\right] = 0, \quad \left[\left[\mathbf{P}^S\right]\right]\mathbf{N} = 0, \quad \left[\left[\rho^F\acute{\mathbf{X}}^F\cdot\mathbf{N}\acute{\mathbf{x}}^F\right]\right] = \left[\left[\mathbf{P}^F\right]\right]\mathbf{N}, \\ &\left[\left[\rho^F\acute{\mathbf{X}}^F\cdot\mathbf{N}\left(\tfrac{1}{2}\acute{x}^{F2}\right)\right]\right] + \left[\left[\mathbf{Q} - \mathbf{P}^{FT}\acute{\mathbf{x}}^F\right]\right]\cdot\mathbf{N} = 0, \\ &[[\mathbf{H}]]\cdot\mathbf{N} = 0. \end{aligned} \quad (20)$$

In addition, according to the compatibility condition $(2)_1$, on $\mathcal{S}$ we have

$$-U\left[\left[\mathbf{F}^S\right]\right] = \left[\left[\acute{\mathbf{x}}^S\right]\right]\otimes\mathbf{N} \quad \Longrightarrow \quad \left[\left[\acute{\mathbf{x}}^S\right]\right] = 0 \quad \text{for } U \equiv 0. \quad (21)$$

3 Fields and field equations

For two-component poroelastic materials we have the (macroscopic) fields

$$\mathcal{F} := \left\{\rho^S, \rho^F, \acute{\mathbf{x}}^S, \acute{\mathbf{x}}^F, \mathbf{F}^S, T, n\right\}, \quad (22)$$

where the first two fields are partial mass densities of the skeleton, and the fluid in the reference configuration, respectively, $\acute{\mathbf{x}}^S$, $\acute{\mathbf{x}}^F$ are macroscopic velocities of these two components, $\mathbf{F}^S$ is the deformation gradient of skeleton, T denotes the common temperature of components, and n is the porosity.

The balance equations of the previous section form field equations for the seven fields $\mathcal{F}$ (22) provided the constitutive quantities

$$\mathcal{C} := \left\{ \mathbf{P}^S, \mathbf{P}^F, \hat{\mathbf{p}}, \varepsilon, \mathbf{Q}, n_E, \mathbf{J}, \hat{n} \right\}, \tag{23}$$

are given as sufficiently smooth functions of the constitutive variables

$$\mathcal{V} := \left\{ \rho^S, \rho^F, \mathbf{F}^S, \hat{\mathbf{X}}^F, \Lambda_n, T, \mathbf{G} \right\}, \quad \mathbf{G} := \operatorname{Grad} T. \tag{24}$$

This set of constitutive variables defines the *simple two-component thermo-poroelastic medium.*

Substitution of the functions $\mathcal{C}(\mathcal{V})$ in the balance equations yields field equations whose solutions are called *thermodynamic processes*. These processes are thermodynamically *admissible* if the entropy production $\hat{\eta}^S + \hat{\eta}^F$ is nonnegative, i.e., the entropy inequality

$$\frac{\partial(\rho\eta)}{\partial t} + \operatorname{Div} \mathbf{H} \geq 0, \quad \eta = \eta(\mathcal{V}), \quad \mathbf{H} = \mathbf{H}(\mathcal{V}), \tag{25}$$

where η is the entropy density, and $\mathbf{H}$ – its flux, is identically satisfied. This is the *second law of thermodynamics* for thermo-poroelastic media.

All constitutive quantities depend as well on an initial constant porosity n_0. This dependence is not limited by the second law because the initial porosity does not evolve in time. It shall not be indicated in subsequent relations in this work.

In the next section we exploit the second law of thermodynamics.

4 Conditions following from the second law of thermodynamics

In the exploitation of the second law we use the standard procedure of Lagrange multipliers. According to Liu's theorem the following inequality should hold for arbitrary fields:

$$\begin{aligned} &\frac{\partial(\rho\eta)}{\partial t} + \operatorname{Div} \mathbf{H} - \Lambda^{\rho^S} R^S - \Lambda^{\rho^F} R^F - \\ &- \boldsymbol{\Lambda}^S \cdot \mathbf{M}^S - \boldsymbol{\Lambda}^F \cdot \mathbf{M}^F - \Lambda^{\varepsilon} E - \boldsymbol{\Lambda}^F \cdot \mathbf{F} - \Lambda^n N \geq 0, \end{aligned} \tag{26}$$

where the multipliers Λ^{ρ^S}, Λ^{ρ^F}, $\boldsymbol{\Lambda}^S$, $\boldsymbol{\Lambda}^F$, Λ^{ε}, $\boldsymbol{\Lambda}^F$, Λ^n are functions of variables $\mathcal{V}$.

After application of the chain rule of differentiation we see that the above inequality is linear with respect to the time derivatives

$$\left\{ \frac{\partial \rho^S}{\partial t}, \frac{\partial \rho^F}{\partial t}, \frac{\partial \dot{\mathbf{x}}^S}{\partial t}, \frac{\partial \dot{\mathbf{x}}^F}{\partial t}, \frac{\partial T}{\partial t}, \frac{\partial \Lambda_n}{\partial t}, \frac{\partial \mathbf{G}}{\partial t}, \frac{\partial \mathbf{F}^S}{\partial t} \right\}. \tag{27}$$

This yields the relations

$$\Lambda^{\rho^S} = \frac{\partial \rho\eta}{\partial \rho^S} - \Lambda^{\varepsilon} \frac{\partial \rho\varepsilon}{\partial \rho^S}, \quad \Lambda^{\rho^F} = \frac{\partial \rho\eta}{\partial \rho^F} - \Lambda^{\varepsilon} \frac{\partial \rho\varepsilon}{\partial \rho^F}, \tag{28}$$

$$\mathbf{\Lambda}^S = \mathbf{\Lambda}^F = 0, \tag{29}$$

$$\Lambda^n = \frac{\partial \rho\eta}{\partial \Delta_n} - \Lambda^\varepsilon \frac{\partial \rho\varepsilon}{\partial \Delta_n}, \quad \mathbf{\Lambda}^F = \frac{\partial \rho\eta}{\partial \mathbf{F}^S} - \Lambda^\varepsilon \frac{\partial \rho\varepsilon}{\partial \mathbf{F}^S}, \tag{30}$$

$$\frac{\partial \rho\eta}{\partial T} - \Lambda^\varepsilon \frac{\partial \rho\varepsilon}{\partial T} = 0, \quad \frac{\partial \rho\eta}{\partial \mathbf{G}} - \Lambda^\varepsilon \frac{\partial \rho\varepsilon}{\partial \mathbf{G}} = 0. \tag{31}$$

The linearity with respect to the spatial derivatives

$$\left\{ \operatorname{Grad} \rho^S, \operatorname{Grad} \rho^F, \operatorname{Grad} \dot{\mathbf{x}}^S, \operatorname{Grad} \dot{\mathbf{x}}^F, \operatorname{Grad} \mathbf{F}^S, \operatorname{Grad} \mathbf{G}, \operatorname{Grad} \Delta_n \right\}, \tag{32}$$

is investigated under two simplifying assumptions.

First of all we assume that a dependence on vectorial variables $\dot{\mathbf{X}}^F, \mathbf{G}$ is linear. This is justified later. Consequently, for *isotropic* materials, constitutive vector functions must have the representation

$$\begin{aligned} &\mathbf{Q} = Q_V \dot{\mathbf{X}}^F - K\mathbf{G}, \quad \mathbf{H} = H_V \dot{\mathbf{X}}^F + H_T \mathbf{G}, \\ &\mathbf{J} = \Phi \dot{\mathbf{X}}^F + J_T \mathbf{G}, \quad \mathbf{F}^{ST}\hat{\mathbf{p}} = \Pi_V \dot{\mathbf{X}}^F + \Pi_T \mathbf{G}, \end{aligned} \tag{33}$$

where all scalar coefficients are independent of $\dot{\mathbf{X}}^F$ and $\mathbf{G}$.

Secondly we assume that the dissipation $\mathcal{D}$ is quadratic in variables describing a deviation from the thermodynamic equilibrium. The dissipation $\mathcal{D}$ is determined by the *residual inequality* which follows after the elimination of the linear part containing the derivatives (27) and (32). Under assumption (33) it has the form

$$\begin{aligned} \mathcal{D} := &\frac{\partial H_V}{\partial T} \dot{\mathbf{X}}^F \cdot \mathbf{G} + \frac{\partial H_T}{\partial T} G^2 - \Lambda^\varepsilon \left(\frac{\partial Q_V}{\partial T} \dot{\mathbf{X}}^F \cdot \mathbf{G} - \frac{\partial K}{\partial T} G^2 \right) - \\ &- \Lambda^n \left(\frac{\partial \Phi}{\partial T} \dot{\mathbf{X}}^F \cdot \mathbf{G} + \frac{\partial J_T}{\partial T} G^2 \right) + \Lambda^\varepsilon \left(\Pi_V \dot{\mathbf{X}}^F \cdot \dot{\mathbf{X}}^F + \Pi_T \dot{\mathbf{X}}^F \cdot \mathbf{G} \right) + \Lambda^n \hat{n} \geq 0, \\ G^2 := &\, \mathbf{G} \cdot \mathbf{G}. \end{aligned} \tag{34}$$

As the quantity Δ_n describes the deviation of porosity from its equilibrium value the above assumption yields

$$\hat{n} = -\frac{\Delta_n}{\tau}, \tag{35}$$

where τ is independent of vector variables and of Δ_n, and, simultaneously, the multiplier Λ^n must be a homogeneous linear function of Δ_n. Consequently,

$$\frac{\partial \Phi}{\partial T} = 0, \quad \frac{\partial J_T}{\partial T} = 0. \tag{36}$$

In addition Λ^ε must be independent of Δ_n. Then, according to $(30)_1$, both ε and η are quadratic even functions of Δ_n.

We return to the conditions following from the linearity with respect to the derivatives (32). We have:

$$\frac{\partial H_V}{\partial \rho^S}\acute{\mathbf{X}}^F + \frac{\partial H_T}{\partial \rho^S}\mathbf{G} - \Lambda^\varepsilon\left(\frac{\partial Q_V}{\partial \rho^S}\acute{\mathbf{X}}^F - \frac{\partial K}{\partial \rho^S}\mathbf{G}\right) -$$

$$-\Lambda^n\left(\frac{\partial \Phi}{\partial \rho^S}\acute{\mathbf{X}}^F + \frac{\partial J_T}{\partial \rho^S}\mathbf{G}\right) = 0; \tag{37}$$

$$\frac{\partial H_V}{\partial \rho^F}\acute{\mathbf{X}}^F + \frac{\partial H_T}{\partial \rho^F}\mathbf{G} - \mathbf{\Lambda}^{\rho^F} - \Lambda^\varepsilon\left(\frac{\partial Q_V}{\partial \rho^S}\acute{\mathbf{X}}^F - \frac{\partial K}{\partial \rho^S}\mathbf{G}\right)$$

$$-\Lambda^n\left(\frac{\partial \Phi}{\partial \rho^S}\acute{\mathbf{X}}^F + \frac{\partial J_T}{\partial \rho^S}\mathbf{G}\right) = 0; \tag{38}$$

$$-\left[H_V - \rho^F\Lambda^{\rho^F} - \Lambda^\varepsilon Q_V - \Lambda^n\Phi\right]\mathbf{F}^{S-T} + \Lambda^\varepsilon\mathbf{P}^S + \mathbf{\Lambda}^F = 0; \tag{39}$$

$$\left[H_V - \rho^F\Lambda^{\rho^F} - \Lambda^\varepsilon Q_V - \Lambda^n\Phi\right]\mathbf{F}^{S-T} + \Lambda^\varepsilon\mathbf{P}^F = 0; \tag{40}$$

$$\underset{23}{\mathrm{sym}}\left\{-\left[H_V - \rho^F\Lambda^{\rho^F} - \Lambda^\varepsilon Q_V - \Lambda^n\Phi\right]\left(\mathbf{F}^{S-T}\otimes\acute{\mathbf{X}}^F\right) + \right.$$
$$+\left(\frac{\partial H_V}{\partial \mathbf{F}^S}\otimes\acute{\mathbf{X}}^F + \frac{\partial H_T}{\partial \mathbf{F}^S}\otimes\mathbf{G}\right) -$$
$$\left.-\Lambda^\varepsilon\left(\frac{\partial Q_V}{\partial \mathbf{F}^S}\otimes\acute{\mathbf{X}}^F - \frac{\partial K}{\partial \mathbf{F}^S}\otimes\mathbf{G}\right) - \Lambda^n\left(\frac{\partial \Phi}{\partial \mathbf{F}^S}\otimes\acute{\mathbf{X}}^F + \frac{\partial J_T}{\partial \mathbf{F}^S}\otimes\mathbf{G}\right)\right\} = 0; \tag{41}$$

$$H_T + \Lambda^\varepsilon K - \Lambda^n J_T = 0; \tag{42}$$

$$\frac{\partial H_V}{\partial \Delta_n}\acute{\mathbf{X}}^F + \frac{\partial H_T}{\partial \Delta_n}\mathbf{G} - \Lambda^\varepsilon\left(\frac{\partial Q_V}{\partial \Delta_n}\acute{\mathbf{X}}^F - \frac{\partial K}{\partial \Delta_n}\mathbf{G}\right) - \Lambda^n\left(\frac{\partial \Phi}{\partial \Delta_n}\acute{\mathbf{X}}^F + \frac{\partial J_T}{\partial \Delta_n}\mathbf{G}\right) = 0. \tag{43}$$

These conditions must hold for arbitrary $\acute{\mathbf{X}}^F$, $\mathbf{G}$, Δ_n. Hence we obtain a series of identities which we proceed to investigate.

Condition (42) immediately yields

$$J_T = 0, \quad H_T + \Lambda^\varepsilon K = 0. \tag{44}$$

According to the conditions following from (37), (41) for coefficients of $\mathbf{G}$, we obtain

$$\frac{\partial H_T}{\partial \rho^S} + \Lambda^\varepsilon\frac{\partial K}{\partial \rho^S} = 0, \quad \frac{\partial H_T}{\partial \rho^F} + \Lambda^\varepsilon\frac{\partial K}{\partial \rho^F} = 0, \quad \frac{\partial H_T}{\partial \mathbf{F}^S} + \Lambda^\varepsilon\frac{\partial K}{\partial \mathbf{F}^S} = 0.$$

Consequently, bearing (44) in mind, we see that

$$\Lambda^\varepsilon = \Lambda^\varepsilon(T).$$

We now turn our attention to the coefficients of $\acute{\mathbf{X}}^F$. We have

$$\frac{\partial H_V}{\partial \rho^S} - \Lambda^\varepsilon \frac{\partial Q_V}{\partial \rho^S} - \Lambda^n \frac{\partial \Phi}{\partial \rho^S} = 0, \quad \frac{\partial H_V}{\partial \rho^F} - \Lambda^\varepsilon \frac{\partial Q_V}{\partial \rho^F} - \Lambda^n \frac{\partial \Phi}{\partial \rho^F} = \Lambda^{\rho^F},$$

$$-\left[H_V - \rho^F \Lambda^{\rho^F} - \Lambda^\varepsilon Q_V - \Lambda^n \Phi\right] \mathbf{F}^{S-T} + \left(\frac{\partial H_V}{\partial \mathbf{F}^S} - \Lambda^\varepsilon \frac{\partial Q_V}{\partial \mathbf{F}^S} - \Lambda^n \frac{\partial \Phi}{\partial \mathbf{F}^S}\right) = 0.$$

Consequently,

$$\frac{\partial \Phi}{\partial \rho^S} = 0, \quad \frac{\partial \Phi}{\partial \rho^F} = 0, \quad \Phi \mathbf{F}^{S-T} = \frac{\partial \Phi}{\partial \mathbf{F}^S} \quad \Longrightarrow \quad \Phi = \Phi_0 J^S, \quad \Phi_0 = \text{const}. \tag{45}$$

where (36) was used.

There remain the identities

$$\frac{\partial}{\partial \rho^S}\left(H_V - \Lambda^\varepsilon Q_V\right) = 0, \quad \frac{\partial}{\partial \rho^F}\left(H_V - \Lambda^\varepsilon Q_V\right) = \Lambda^{\rho^F},$$

$$\left[H_V - \rho^F \Lambda^{\rho^F} - \Lambda^\varepsilon Q_V\right] \mathbf{F}^{S-T} = \frac{\partial}{\partial \mathbf{F}^S}\left(H_V - \Lambda^\varepsilon Q_V\right). \tag{46}$$

The integrability condition of the first two conditions immediately yields

$$\frac{\partial \Lambda^{\rho^F}}{\partial \rho^S} = 0. \tag{47}$$

On the other hand, substitution of $(46)_2$ in $(46)_3$ leads to the equation

$$\rho^F \frac{\partial \Lambda^{\rho^F}}{\partial \rho^F} \mathbf{F}^{S-T} + \frac{\partial \Lambda^{\rho^F}}{\partial \mathbf{F}^S} = 0.$$

This equation can easily be integrated[1] and we obtain

$$\Lambda^{\rho^F} = \Lambda^{\rho^F}\left(T, \rho_t^F\right), \quad \rho_t^F := \rho^F J^{S-1}. \tag{48}$$

[1] For isotropic materials considered later in this section the dependence of Λ^{ρ^F} on $\mathbf{F}^S$ reduces to a dependence on the three invariants I, II, III of the tensor $\mathbf{C}^S$. Then

$$\frac{\partial \Lambda^{\rho^F}}{\partial \mathbf{F}^S} = \frac{\partial \Lambda^{\rho^F}}{\partial I}\frac{\partial I}{\partial \mathbf{F}^S} + \frac{\partial \Lambda^{\rho^F}}{\partial II}\frac{\partial II}{\partial \mathbf{F}^S} + \frac{\partial \Lambda^{\rho^F}}{\partial III}\frac{\partial III}{\partial \mathbf{F}^S} =$$
$$= 2\frac{\partial \Lambda^{\rho^F}}{\partial I}\mathbf{F}^S + 2\frac{\partial \Lambda^{\rho^F}}{\partial II}\mathbf{F}^S(I\mathbf{1} - \mathbf{C}^S) + \frac{\partial \Lambda^{\rho^F}}{\partial J^S} J^S \mathbf{F}^{S-T}.$$

Hence we obtain the equation

$$\left(\rho^F \frac{\partial \Lambda^{\rho^F}}{\partial \rho^F} + J^S \frac{\partial \Lambda^{\rho^F}}{\partial J^S}\right)\mathbf{1} + 2\left(\frac{\partial \Lambda^{\rho^F}}{\partial I} + I\frac{\partial \Lambda^{\rho^F}}{\partial II}\right)\mathbf{C}^S - 2\frac{\partial \Lambda^{\rho^F}}{\partial II}\mathbf{C}^{S2} = 0.$$

According to the Cayley-Hamilton theorem tensors $\left\{\mathbf{1}, \mathbf{C}^S, \mathbf{C}^{S2}\right\}$ span the space of tensor functions of $\mathbf{C}^S$. Consequently, the coefficients in this equation should vanish separately,

It is convenient to introduce the notation

$$\psi := \varepsilon - \Lambda^{\varepsilon-1}\eta. \tag{49}$$

Obviously ψ corresponds to the classical Helmholtz free energy function.

Before we proceed with the exploitation of the above results we summarize the results for multipliers which follow from the above considerations. We have

$$\begin{aligned} &\Lambda^{\rho^S} = -\Lambda^{\varepsilon}\frac{\partial\rho\psi}{\partial\rho^S}, \quad \Lambda^{\rho^F} = -\Lambda^{\varepsilon}\frac{\partial\rho\psi}{\partial\rho^F}, \\ &\Lambda^{n} = -\Lambda^{\varepsilon}\frac{\partial\rho\psi}{\partial\Delta_n}, \quad \mathbf{\Lambda}^{F} = -\Lambda^{\varepsilon}\frac{\partial\rho\psi}{\partial\mathbf{F}^S}, \quad \frac{\partial\rho\psi}{\partial\mathbf{G}} = 0. \end{aligned} \tag{50}$$

Consequently, the integration of (48) yields the splitting of the free energy into two constitutive parts

$$\rho\psi = \rho^S\psi^S + \rho^F\psi^F, \quad \psi^S = \psi^S\left(T, \rho^S, \mathbf{F}^S, \Delta_n\right), \quad \psi^F = \psi^F\left(T, \rho_t^F, \Delta_n\right). \tag{51}$$

This separation justifies the name *simple porous materials*. As with simple mixtures of fluids, partial free energies of components depend only on their own measures of deformation: the skeleton on the deformation gradient $\mathbf{F}^S$, and the fluid on the current mass density of the fluid ρ_t^F. There is no energy of interaction between components.

We are now in a position to integrate the relations between H_V and Q_V. After intergration of $(46)_2$ we obtain

$$H_V - \Lambda^{\varepsilon}Q_V = -\Lambda^{\varepsilon}\rho^F\psi^F. \tag{52}$$

Hence the fluxes can be written in the following final form:

$$\begin{aligned} &\mathbf{H} = \Lambda^{\varepsilon}\left(\mathbf{Q} - \rho^F\psi^F\dot{\mathbf{X}}^F\right), \quad \text{i.e.} \quad \mathbf{H}^S + \mathbf{H}^F = \Lambda^{\varepsilon}\left(\mathbf{Q}^S + \mathbf{Q}^F\right), \\ &\mathbf{J} = \Phi_0 J^S\dot{\mathbf{X}}^F, \\ &H_T = -\Lambda^{\varepsilon}K, \quad H_V = \Lambda^{\varepsilon}Q_V - \Lambda^{\varepsilon}\rho^F\psi^F\dot{\mathbf{X}}^F, \end{aligned} \tag{53}$$

where the relations (17) and (19) have been used.

We now consider an impermeable boundary between a saturated porous material and a fluid which is physically identical with the fluid filling pores of the skeleton. If the temperature is continuous on this boundary and simultaneously jump conditions

i.e.,

$$\frac{\partial\Lambda^{\rho^F}}{\partial I} = 0, \quad \frac{\partial\Lambda^{\rho^F}}{\partial II} = 0,$$

$$\rho^F\frac{\partial\Lambda^{\rho^F}}{\partial\rho^F} + J^S\frac{\partial\Lambda^{\rho^F}}{\partial J^S} = 0.$$

Integration of this equation yields the relation (48). It can easily be shown that this is also a solution in the general case without the assumption of isotropy.

(20) are satisfied, we call it an *ideal wall* for the fluid component. This surface is material simultaneously with respect to the skeleton and the fluid, i.e.,

$$\begin{aligned}
&\dot{\mathbf{X}}^F = 0 \quad \Longrightarrow \quad \left[\left[\dot{\mathbf{x}}^F\right]\right] = 0 \quad \Longrightarrow \\
\Longrightarrow \quad &[[\mathbf{Q}]] \cdot \mathbf{N} = 0 \;\&\; [[\mathbf{H}]] \cdot \mathbf{N} = 0 \quad \Longrightarrow \\
\Longrightarrow \quad &[[\Lambda^\varepsilon]] = 0.
\end{aligned}$$

As the multiplier in the fluid outside the porous material is the reciprocal of the temperature the above result yields

$$\Lambda^\varepsilon = \frac{1}{T}, \tag{54}$$

in the porous material as well.

Finally inspection of relations (39), (40) for partial stresses shows that they satisfy the relations

$$\begin{aligned}
\mathbf{P}^F &= -\rho_t^{F2} \frac{\partial \psi^F}{\partial \rho_t^F} J^S \mathbf{F}^{S-T} - \Phi_0 \frac{\partial \rho\psi}{\partial \Delta_n} J^S \mathbf{F}^{S-T}, \\
\mathbf{P}^S &= \frac{\partial \rho^S \psi^S}{\partial \mathbf{F}^S} + \Phi_0 \frac{\partial \rho\psi}{\partial \Delta_n} J^S \mathbf{F}^{S-T}.
\end{aligned} \tag{55}$$

Transformation to Cauchy stresses yields

$$\mathbf{T}^F := J^{S-1} \mathbf{P}^F \mathbf{F}^{ST} = -p^F \mathbf{1}, \quad p^F := \rho_t^{F2} \frac{\partial \psi^F}{\partial \rho_t^F} + \beta \Delta_n, \tag{56}$$

$$\begin{aligned}
&\mathbf{T}^S := J^{S-1} \mathbf{P}^S \mathbf{F}^{ST} = \rho_t^S \frac{\partial \psi^S}{\partial \mathbf{F}^S} \mathbf{F}^{ST} + \beta \Delta_n \mathbf{1}, \quad \rho_t^S := \rho^S J^{S-1}, \\
&\beta := \Phi_0 \frac{\partial^2 \rho\psi}{\partial \Delta_n^2},
\end{aligned} \tag{57}$$

where we use the property that the free energy is a quadratic even function of Δ_n.

Bearing the above results in mind we analyze a jump condition on the permeable boundary of the skeleton. According to relations (20) and (21) we have

$$\begin{aligned}
&\left[\left[\rho^F \dot{\mathbf{X}}^F \cdot \mathbf{N} \tfrac{1}{2} \dot{x}^{F2} + \mathbf{Q} \cdot \mathbf{N}\right]\right] = \left[\left[\dot{\mathbf{x}}^F \cdot \mathbf{P}^F \mathbf{N}\right]\right] = \\
&= \left[\left[\rho^F \dot{\mathbf{X}}^F \cdot \mathbf{N} \left(\frac{p^F}{\rho^F} J^S - \dot{\mathbf{x}}^F \cdot \dot{\mathbf{x}}^S\right)\right]\right].
\end{aligned}$$

Hence the relations $(20)_5$ and $(53)_1$ immediately yield

$$\left[\left[\mu^F\right]\right] = 0, \quad \mu^F := \psi^F + \frac{p^F}{\rho_t^F} + \tfrac{1}{2} \left(\dot{\mathbf{x}}^F - \dot{\mathbf{x}}^S\right) \cdot \left(\dot{\mathbf{x}}^F - \dot{\mathbf{x}}^S\right). \tag{58}$$

The quantity μ^F is the *chemical potential* of the fluid component. Its continuity on the permeable boundary replaces the mechanical condition on continuity of partial pressures on impermeable boundaries.

It remains to rewrite the residual inequality (34) in which we account for the above results. We obtain

$$\mathcal{D} = \frac{1}{T} K G^2 - \left[\frac{Q_V}{T} + T \frac{\partial}{\partial T} \left(\frac{\rho^F \psi^F}{T} \right) - \Pi_T \right] \dot{\mathbf{X}}^F \cdot \mathbf{G} + \Pi_V \dot{\mathbf{X}}^F \cdot \dot{\mathbf{X}}^F + \frac{\beta}{\Phi_0 \tau} \Delta_n^2 \geq 0. \tag{59}$$

It is now obvious that the simplifying assumptions which we have made in this section amount indeed to a quadratic form of the dissipation $\mathcal{D}$, i.e., to small deviations of processes from the thermodynamic equilibrium in which $\dot{\mathbf{X}}^F \Big|_E = 0$, $\mathbf{G}|_E = 0$, $\Delta_n|_E = 0$. Consequently, material parameters should satisfy the conditions

$$K > 0, \quad \Pi_V > 0, \quad K \Pi_V + \left[\frac{Q_V}{T} + T \frac{\partial}{\partial T} \left(\frac{\rho^F \psi^F}{T} \right) - \Pi_T \right]^2 > 0, \tag{60}$$
$$\frac{\beta}{\Phi_0 \tau} > 0.$$

Further restrictions on material parameters follow from the stability analysis of thermodynamic equilibrium. We do not consider this problem in the present work.

5 Particular cases

In this section we demonstrate a few examples of simplified models of a thermo-poroelastic material which have an important bearing in applications. We begin with the exploitation of the isotropy assumption with respect to deformations of the skeleton. Obviously, contributions of the fluid are already isotropic.

We use the right and left Cauchy-Green deformation tensors

$$\begin{aligned} &\mathbf{C}^S := \mathbf{F}^{ST} \mathbf{F}^S, \quad \mathbf{B}^S := \mathbf{F}^S \mathbf{F}^{ST}, \\ &I = \operatorname{tr} \mathbf{C}^S \equiv \operatorname{tr} \mathbf{B}^S, \quad II = \tfrac{1}{2} \left(I^2 - \operatorname{tr} \mathbf{C}^{S2} \right) \equiv \tfrac{1}{2} \left(I^2 - \operatorname{tr} \mathbf{B}^{S2} \right), \\ &III \equiv J^{S2} = \det \mathbf{C}^S \equiv \det \mathbf{B}^S, \end{aligned} \tag{61}$$

where I, II, III are the main invariants common to both deformation tensors.

According to the polar decomposition theorem we have

$$\mathbf{F}^S = \mathbf{R}^S \sqrt{\mathbf{C}^S}, \quad \mathbf{R}^{ST} = \mathbf{R}^{S-1}. \tag{62}$$

Under the assumption of material objectivity the free energy function ψ^S is independent of rotations $\mathbf{R}^S$. Consequently, if we drop a trivial dependence on ρ^S, we have

$$\psi^S = \psi^S \left(T, \mathbf{C}^S, \Delta_n \right) = \psi^S \left(T, I, II, III, \Delta_n \right), \tag{63}$$

where the second part of the relation follows from the assumption on isotropy.

Bearing this relation in mind we obtain from (56) the following relation for the partial Cauchy stress in the skeleton:

$$\mathbf{T}^S = \beth_{-1}\mathbf{B}^{S-1} + \beth_0\mathbf{1} + \beth_1\mathbf{B}^S + \beta\Delta_n\mathbf{1}, \tag{64}$$

where

$$\beth_{-1} := -2\rho_t^S III\left.\frac{\partial\psi^S}{\partial II}\right|_{\Delta_n=0}\mathbf{B}^{S-1}, \quad \beth_0 := 2\rho_t^S\left.\left(II\frac{\partial\psi^S}{\partial II} + III\frac{\partial\psi^S}{\partial III}\right)\right|_{\Delta_n=0},$$

$$\beth_1 := 2\rho_t^S\left.\frac{\partial\psi^S}{\partial I}\right|_{\Delta_n=0}, \tag{65}$$

and the Cayley-Hamilton theorem,

$$\mathbf{B}^{S3} - I\mathbf{B}^{S2} + II\mathbf{B}^S - III\mathbf{1} = 0, \tag{66}$$

has been used.

Problems with the practical determination of the elasticities $\beth_{-1}, \beth_0, \beth_1$ for poroelastic materials yield the necessity of a further simplification. In classical elasticity theory a quadratic isotropic model was proposed by Signorini. The constitutive relation for this model follows from the above model by the truncation on the second term in the expansion around the point $(T = T_0, \mathbf{F}^S = \mathbf{1})$ (i.e., $I = 3, II = 3, III = 1$) and it has the form

$$\mathbf{T}^S = \mathbf{T}_0^S + \left[\lambda^S I_e + c^S II_e + \tfrac{1}{2}\left(\lambda^S + \mu^S - \tfrac{1}{2}c^S\right)I_e^2\right]\mathbf{1} +$$

$$+2\left[\mu^S - \left(\lambda^S + \mu^S + \tfrac{1}{2}c^S\right)I_e\right]\mathbf{e}^S + 2c^S\mathbf{e}^{S2} - \alpha_T^S\frac{T - T_0}{T_0}\mathbf{1} + \beta\Delta_n\mathbf{1}, \tag{67}$$

where

$$\mathbf{e}^S := \frac{1}{2}\left(\mathbf{1} - \mathbf{B}^{S-1}\right), \quad I_e := \operatorname{tr}\mathbf{e}^S, \quad II_e := \tfrac{1}{2}\left(I_e^2 - \operatorname{tr}\mathbf{e}^{S2}\right), \tag{68}$$

$\mathbf{e}^S$ is the Almansi-Hamel deformation tensor, α_T^S is the thermal expansion coefficient of the skeleton which may be linearly dependent on I_e, while material parameters λ^S, μ^S, c^S depend only on the reference temperature T_0.

Finally the classical linear model for small deformations,

$$\left\|\mathbf{e}^S\right\| \ll 1, \quad \left\|\mathbf{e}^S\right\| := \min\left(\{\lambda_1, \lambda_2, \lambda_3\}\right),$$

$$|\epsilon| \ll 1, \quad \epsilon := \frac{\rho_0^F - \rho_t^F}{\rho_0^F},$$

$$\det\left(\mathbf{e}^S - \lambda_i\mathbf{1}\right) = 0, \quad i = 1, 2, 3,$$

follows from (67) in the form

$$\mathbf{T}^S = \mathbf{T}_0^S + \lambda^S I\mathbf{1} + 2\mu^S\mathbf{e}^S - \alpha_T^S\frac{T - T_0}{T_0}\mathbf{1} + \beta\Delta_n\mathbf{1}, \tag{69}$$

$$\mathbf{T}^F = \mathbf{T}_0^F - \kappa\rho_0^F\epsilon\mathbf{1} - \alpha_T^F\frac{T - T_0}{T_0}\mathbf{1} - \beta\Delta_n\mathbf{1},$$

where κ is the compressibility coefficient of the fluid, and α_T^S, α_T^F denote constant thermal expansion coefficients of the skeleton and fluid, respectively.

The last model corresponds to the classical Biot's model but it does not contain Biot's coupling term.

6 Concluding remarks

We have shown that the assumption on a quadratic form of the dissipation yields a quite explicit form of constitutive relations for thermo-poroelastic materials. Their mechanical part does not differ from that derived earlier for isothermal processes. Relations for energy and entropy fluxes justify the assumption made in the earlier papers on the proportionality of their intrinsic parts (see (53)) which was basic for the formulation of the second law of thermodynamics for isothermal processes. In addition we have shown that, as with classical miscible mixtures, a chemical potential for the fluid component is continuous on the permeable boundary of the porous body. This property is fundamental for the formulation of boundary conditions on such a boundary.

References

[1] Bowen, R.M. (1980): Incompressible porous models by use of the theory of mixtures. Internat J. Engrg. Sci. **18**, 1129–1148
[2] Bowen, R.M. (1982): Compressible porous media models by use of the theory of mixtures. Internat J. Engrg. Sci. **20**, 697–763
[3] Wilmanski, K. (1995): Lagrangean model of two-phase porous material. J. Non-Equilibrium Thermodyn. **20**, 50–77
[4] Wilmanski, K. (1996): Porous media at finite strains. The new model with the balance equation of porosity. Arch. Mech. **48**, 591–628
[5] Wilmanski, K. (1998): A thermodynamic model of compressible porous materials with the balance equation of porosity. Transp. Porous Media **32**, 21–47
[6] Wilmanski, K. (1998): Thermomechanics of continua. Springer, Berlin
[7] Wilmanski, K. (2000): Toward extended thermodynamics of porous and granular materials. In: Iooss, G. et al. (eds.): Trends in applications of mathematics to mechanics. Chapman&Hall/CRC, Boca Raton, FL, pp. 147–160
[8] Wilmanski, K. (2002): Thermodynamical admissibility of Biot's model of poroelastic saturated materials. Arch. Mech. **54**, 709–736
[9] Wilmanski, K. (2003): On thermodynamics of nonlinear poroelastic materials. J. Elasticity **71**, 247–261
[10] Kirchner, N.P. (2001): Thermodynamics of structured granular materials. Dissertation. (Reihe Thermodynamik). Shaker, Aachen